IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 October 2023, accepted 23 December 2023, date of publication 25 December 2023,
date of current version 5 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347424

RESEARCH ARTICLE

SPCB-Net: A Multi-Scale Skin Cancer Image Identification Network Using Self-Interactive Attention Pyramid and Cross-Layer Bilinear-Trilinear Pooling

XIN QIAN[1], TENGFEI WENG[2], QI HAN[1], CHEN WU[1], HONGXIANG XU[1], MINGYANG HOU[1], ZICHENG QIU[1], BAOPING ZHOU[3], AND XIANQIANG GAO[3]

[1]School of Intelligent Technology and Engineering, Chongqing University of Science and Technology, Chongqing 401331, China
[2]School of Electrical Engineering, Chongqing University of Science and Technology, Chongqing 401331, China
[3]School of Information Engineering, Tarim University, Alar 843300, China

Corresponding author: Tengfei Weng (wengtf_cq@163.com)

This work was supported in part by the West Light Foundation of the Chinese Academy of Science, in part by the Research Foundation of the Natural Foundation of Chongqing City under Grant cstc2021jcyj-msxmX0146, in part by the Scientific and Technological Research Program of Chongqing Municipal Education Commission under Grant KJQN202301517 and Grant HZ2021015, in part by the Chongqing Science and Technology Military-Civilian Integration Innovation Project (2022), in part by the Bingtuan Science and Technology Program in China under Grant 2021AB026, in part by the Sichuan Science and Technology Program under Grant 2023JDRC0033, and in part by the Luzhou Science and Technology Program under Grant 2021-JYJ-92.

ABSTRACT Deep convolutional neural networks have made some progress in skin lesion classification and cancer diagnosis, but there are still some problems to be solved, such as the challenge of small inter-class feature differences and large intra-class feature differences, which might limit the classification performance of the model as high-level and low-level features are not properly utilized. This paper proposes a multi-scale skin cancer image identification network using self-interactive attention pyramid and cross-layer bilinear-trilinear pooling(SPCB-Net), which mainly consists of three proposed sub-modules that are the self-interacting attention pyramid (SAP), the across-layer bilinear-trilinear pooling operation and the global average algorithm(GAA). The SPCB-Net is applied to two representative datasets of medical images in dermatology and histopathology (HAM10000 and NCT-CRC-HE-100K) to demonstrate the effectiveness of in the skin lesion classification. SPCB-Net(ResNet101) achieves 97.10% and 99.87% accuracy on HAM10000 and NCT-CRC-HE-100K respectively, which are both achieved performance improvements of 0.4% compared to the state-of-the-art models. In addition, a large number of experiments on HAM10000 show that the interactive attention pyramid(SPA) proposed in this paper is superior to the common attention module, and the method with a cross-layer bilinear-trilinear pooling is superior to the cross-layer trilinear pooling method. SPCB-Net is configured on Vgg19 and ResNet101 to evaluate the effectiveness of our proposed module. The experimental results show that SPCB-Net has shown state-of-the-art performance in the two field of dermatology and histopathology. Therefore, it is not only well qualified for the task of identifying skin cancer image but also has the potential to identify skin cancer by identifying pathological tissue.

INDEX TERMS Deep learning, convolutional neural network, medical image classification, multi-scale fusion, attention mechanism, skin lesions.

The associate editor coordinating the review of this manuscript and approving it for publication was R. K. Tripathy.

I. INTRODUCTION

According to the Skin Cancer Foundation [1], one in five Americans will suffer from skin cancer in their life. In the

United States, the number of patients diagnosed as skin cancer exceeds the sum of all other cancers. Among all types of skin cancer [2], melanoma is the most lethal and dangerous cancer [3]. Timely identification and diagnosis of skin cancer can cure nearly 95% of cases [4]. It is a very effective means to identify cancer by skin lesion image in computer medical assistance. In addition, it can also be diagnosed by identifying various tissue cells, and the identification of different tissue types is the first step of histopathological image analysis.

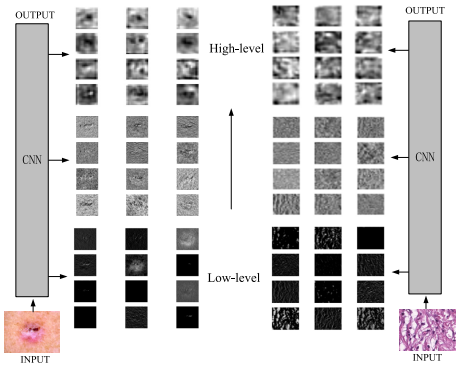


FIGURE 1. Low-level information and high-level information are shown through two images from HAM10000 and NCT-CRC-HE-100K datasets.

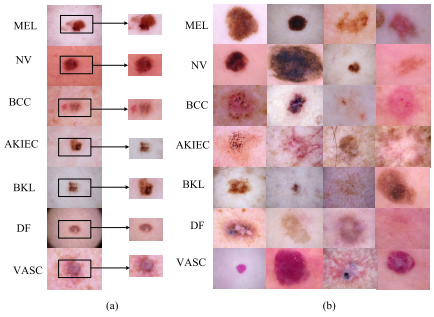


FIGURE 2. (a) The low-level features in different diseases are similar in some images. (b) The high-level features information in the same class of disease have dissimilarity.

Therefore, computer aided diagnosis is particularly important for the early diagnosis of some diseases. Traditionally, doctors use visual viewing. However, for professional dermatologists, early detection of skin cancer is also a challenge due to the pressure of repeated reading. Therefore, it is necessary to develop a fast, scalable and efficient auxiliary diagnostic method to help doctors detect diseases early.

With the development of artificial intelligence in the medical field, in the past few years, deep learning (DL) has been widely used in the detection, segmentation and classification of medical images [5], [6], [7], [8]. In image classification, prior knowledge and complex image preprocessing are very necessary. The hair contained in the dermoscopy image may hinder important information such as texture, shape, and boundary of skin damage. Therefore, data preprocessing is also very important for classification. Bansal et al. [9], [10], [11] processed the existing hair based on the black cap hair recognition and removal method (BHM) and the first expansion and then corrosion method (closed opration). The processed images can clearly show important features, which is convenient for us to classify.

It is proved that medical images can be classified by deep learning methods. Cheng et al. [12] proposed a attention block with modular group, which can capture the feature correlation in medical images in two independent dimensions (channel and space). Wang et al. [13] proposed a high-order interaction (HOI) method in the field of fine-grained visual classification, which introduced an an cross-layer three-linear pool to calculate the third-order interaction among three layers. HOI can produce more discriminative representations, combining attention mechanisms with HOI to obtain improvements. Zhang et al. [14] proposed a dermoscopic image classification method based on non-traditional convolutional neural network, which applied high efficiency and high performance large kernel in the underlying convolutional layer to obtain a larger receptive field than the original capsule network. Although the above networks can identify the corresponding categories to some extent, as the CNN structure becomes deeper and deeper, high-level neurons have a strong response to the whole image and rich semantics, but it is inevitable to lose detailed information from small-resolution regions, which is a big challenge for similar categories of disease images, and the improvement of the model is limited. Convolutional neural network(CNN) plays an important role in disease diagnosis and clinical treatmen [4]. More and more researchers are working on designing effective classification networks for computer aided diagnosis system, but many of the network used for classifying medical images exhibit poor generalization, leading to poor accuracy in predicting results [5], [6], [7].

In most CNN, the feature map with low-level information derived from the shallow network has abundant local information. The feature map at low-level has higher resolution and smaller receptive field of a single pixel, which can capture more information of small targets. High-level information comes from the deep network. With the increase of the number of deep network layers, the global information of the feature map is richer. The resolution of the feature map at high-level is relatively low, and the sensitivity field of a single pixel is relatively large, so more information of target can be captured. In the Figure 1, semantic information of high level and low level is shown from two medical images. In some medical images, low-level features are very similar between different classes, while high-level features are not similar between the same class, which happens in the dermatological data, as shown in the Figure 2 (a), where the low features of different categories are very similar in images of skin disease. In Figure 2 (b), images in the same class do not have similar high-level features. In this paper, the backbone will be focused to further improve the classification performance and

generalization ability from the similarity of low-level features within classes and the dissimilarity of high-level features between classes.

In this paper, inspired by the pyramid structure [19] and the attention module [20], [21], a module called the self-interactive attention pyramid(SAP) is proposed. SAP module can not only fully discover the multi-scale features with low semantics to high semantics extracted from the CNN backbone, but also effectively improve the reliability of classification through learning both high-level and low-level feature representations. The high-level features and low-level features of different disease images have similar information which makes image classification difficult. If the interaction between different level features is utilized, a convolutional neural network can learn more information about image features. Therefore, we propose a high order interaction (HOI) that combines a cross-layer bilinear and trilinear pooling to associate spatial features of different layers. Our method can solve the problem about intra-class similarity and inter-class dissimilarity which causes classification difficulties. At the same time, classification performance is improved by use of the interaction of high-level and low-level features.

The main contributions of this study are as follows:

(1) A multi-scale skin cancer image identification network using self-interactive attention pyramid and cross-layer bilinear-trilinear pooling(SPCB-Net) is proposed. The model can accurately locate the local regions, realize the interaction between low-level features and high-level features, reduce background noise and improve the classification effect.

(2) SK feature pyramid(SK-FP) is proposed, which includes feature pyramid [11] and SKNet [18]. SKNet is used to assign different receptive fields to the feature maps of different scales in FPN, so as to avoid the loss of some important information.

(3) A attention structure called self-interactive attention pyramid(SAP) is proposed, which combines the spatial attention pyramid [22], a channel attention pyramid [23] and a skip connection from the bottom layer to the top layer. SAP can accurately locate the important regions and improve network reliability.

(4) A cross-layer bilinear-trilinear pooling operation is proposed. This method can obtain the long-range correlation of cross-layer and learn more about different expressions with fewer parameters.

II. RELATED WORK

A. FEATURE FUSION NETWORK

One of the best ways to deal with classification problem is to utilize multiscale information, which has been extensively studied and researched in this area. In 2017, feature pyramid network(FPN) [18] was proposed to enhance the representational power of feature maps by fusing those of different scales. This was achieved through the development of a top-down architecture with lateral connections, which constructed a feature pyramid using the inherent multi-scale pyramid hierarchy of deep convolutional networks at minimal additional cost. PANet [24] was proposed next year, featuring bi-directional fusion from bottom-up. As opposed to the original FPN's uni-directional fusion from deep to shallow, it encompasses bi-directional fusion from both deep to shallow and shallow to deep. Then there was a big boost, for example in Recursive-FPN [25], the output from the fusion of Traditional FPN is fed back into the backbone network for a second iteration, leading to noticeable improvements in various tasks. Although different scales of information are utilized effectively in previous developments, it is worth noting that the use of uniform convolution kernels across different scales in each layer leads to the loss of important information. To overcome this problem, a dynamic convolution kernel method [18] is introduced in this paper, which employs varying receptive fields to capture information from different scales. This proposed approach builds on the original FPN model and can help retain important details more efficiently.

B. ATTENTION MECHANISM

In 2019, Hu et al. [21] proposed a new architectural unit, called the ''Squeeze and excitation'' (SE) block, which adaptively recalibrates the channel characteristic response by explicitly modeling the interdependencies between channels. In 2020, Wang et al. [26] by analyzing the channel attention module in SENet, proposed an efficient channel attention (ECA) module, which only involves a few parameters and brings obvious performance gain. Woo et al. [20] proposed a convolutional block attention module (CBAM), which is a lightweight convolution module. CBAM contains two sub-modules which are channel attention module (CAM) and spartial attention module (SAM). CBAM not only saves parameters and computing power, but also can be integrated into the existing network architecture as plug and play modules. In 2021, Li et al. [19] proposed a convolutional neural network with attention pyramid (AP-CNN) again, and proposed a new dual-path architecture, which combines top-down feature pathway and bottom-up attention pathway to enhance high-level semantics and low-level details representation for fine-grained classification learning. The previous work has made significant progress, but one crucial aspect that has not been fully utilized is the joint exploitation of channel information and spatial information across different levels, which is particularly crucial for addressing multi-scale problems. Inspired by the above work, this paper proposes a self-interactive attention pyramid(SAP). In addition to utilizing channel attention and spatial attention at each layer, SAP also involves multiple skip connections to capture information between different levels in terms of spatial and channel.

C. BILINEAR POOLING AND TRILINEAR POOLING METHODS

At present, most classification works based on bilinear pooling use second-order interactions, which have good performance in processing image features [13]. Lin et al.

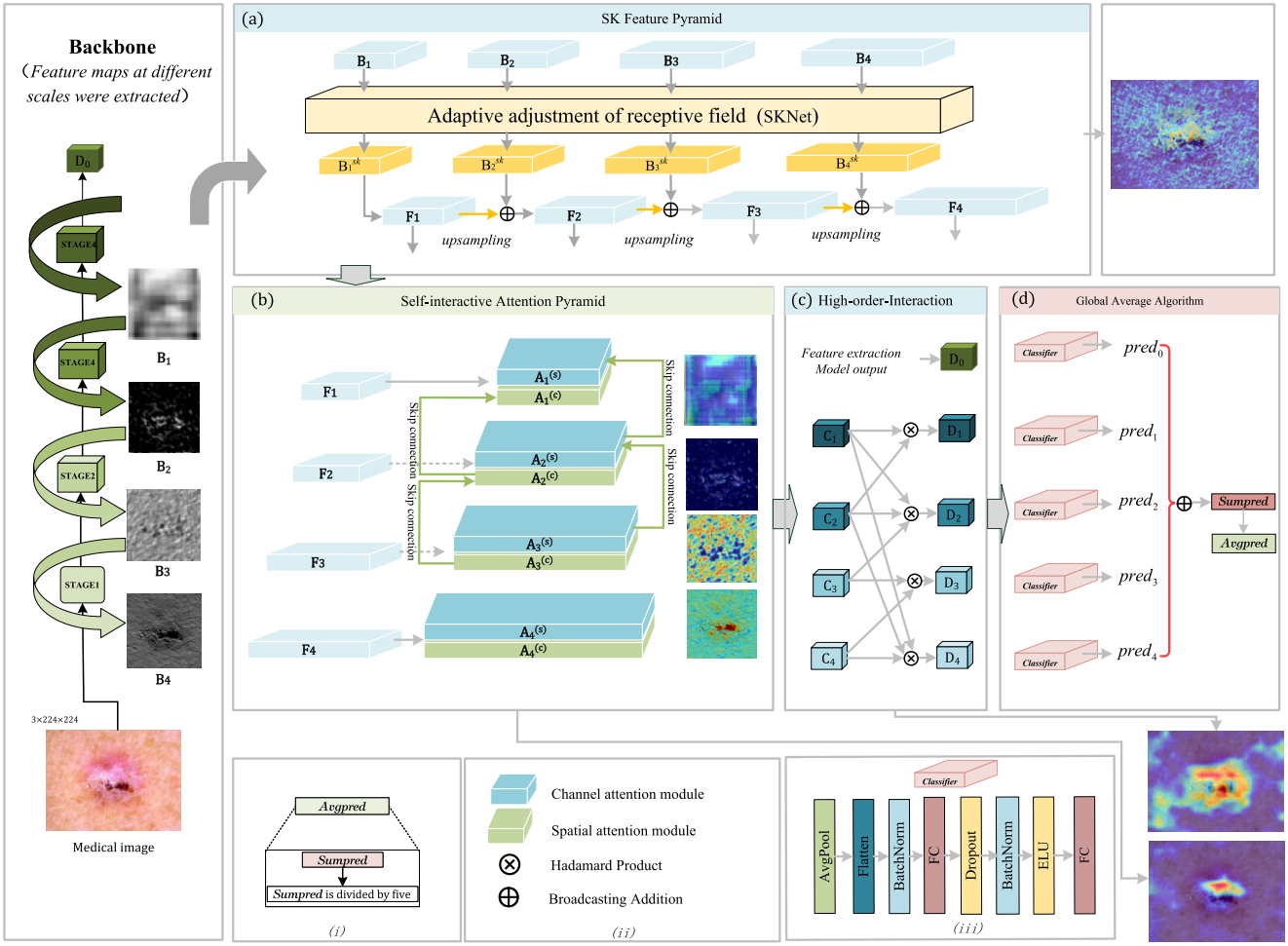


FIGURE 3. Illustration of SPCB-Net. SPCB-Net is divided into five parts. The first part is the backbone. The second part (a) is an extraction module of four-layer feature pyramid network [17] and selective kernel networks [18]. The third part (b) is a self-interactive attention pyramid structure module(SAP), which is a composite structure composed of channel attention pyramid and space attention pyramid. The fourth part (c) is the higher-order interaction module base on cross-layer bilinear-trilinear pooling operation, which can calculate the result of the higher-order interaction between D_1, D_2, D_3, and D_4. Finally, the fifth part (d) is the global average algorithm(GAA) module.

[27] proposed a bilinear model, which is an identification architecture composed of two feature extractor, where an output of the model is is multiplied by external product at each position of the image. The bilinear form simplifies the gradient calculation, and is much simpler and easier to train. Yu et al. [28] proposed a novel hierarchical bilinear pooling framework, which integrated multiple cross-layer bilinear features to enhance representation ability. Kong et al. [13] also proposed a low-order bilinear pooling, where the covariance feature was expressed as a matrix, and the low-rank bilinear classifier was applied. In addition to the above bilinear order interaction, some researchers also proposed higher order interaction methods. Jun et al. [13] paid attention to the high interaction among multiple layers, and introduced an effective cross-layer trilinear pooling to calculate the third-order interaction among three different layers. Bilinear pooling and trilinear pooling have different advantages. While bilinear pooling has lower performance than trilinear pooling, it requires fewer parameters. On the other hand, trilinear pooling has a higher number of parameters than bilinear pooling, it exhibits better performance. This paper one of aims is to combine the strengths of both bilinear pooling methods and trilinear pooling.

III. METHOD

In this section, in order to classify skin lesions, a main structure of the multi-scale deep learning network using self-interactive attention pyramid and cross-layer bilinear-trilinear pooling (SPCB-Net) is presented and described in detail, where SPCB-Net is shown in Figure 3(a+b+c+d). SPCB-Net is mainly divided into five parts, which are backbone, feature extraction module in Figure 3(a), interactive attention pyramid in Figure 3(b), high-order-Interaction module in Figure 3(c), and output result processing module in Figure 3(d).

Feature maps of different scales can be derived from backbone.

For the large intra-class feature differences of skin lesion in the paper, high and low level features are extracted, where feature pyramid network(FPN) [17] and SKNet [18] is used

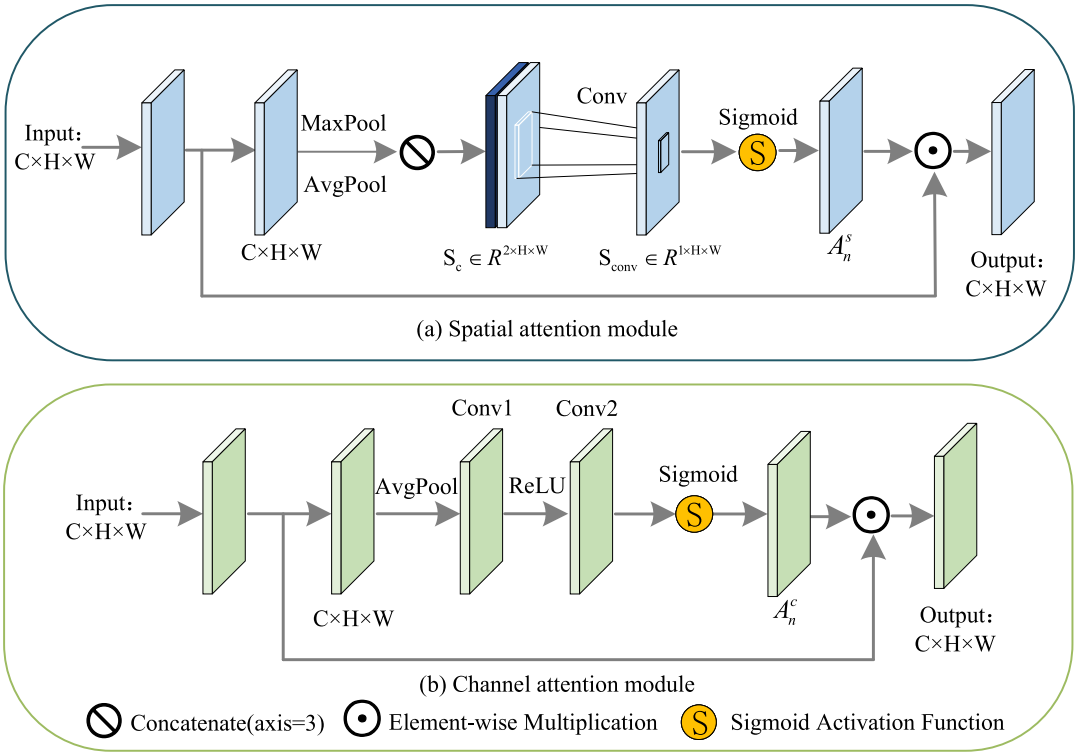


FIGURE 4. The spatial attention module and the channel attention module in the self-interacting pyramid.

as the basic structure to extract features at different scales from top to bottom, and four features with different scale are selected for processing. In order to improve the generalization performance of the network, it is necessary for the network to locate the important information and suppress the useless information to reduce the background noise. For this purpose, a module called the self-interactive Attention Pyramid (SAP) is proposed, which combines the spatial attention [20], a channel attention and a Skip connection from the bottom layer to the top layer. For the small inter-class feature differences of skin lesion, the correlation is used to carry out accurate learning of convolutional neural network between different feature levels. Therefore, a high-order interaction method based on cross-layer bilinear and trilinear pooling is proposed. Finally, an effective output result processing algorithm is proposed to given the output results.

A. SK FEATURE PYRAMID CONSTRUCTION(SK-FP)

For extracting high-level and low-level feature information, feature pyramid network (FPN) [17] and SKNet [18] is introduced into medical image processing to solve insufficient semantic information and low resolution of the underlying feature map. FPN can integrate feature maps with strong and low resolution semantic information, and has rich spatial information with less computational effort. SKNet [18] introduces a selection module to calculate the contribution of each channel to features at different scales and dynamically adjusts the size and shape of corresponding convolution kernels based on these contributions. As shown in Figure 3(a), SK feature pyramid is constructed by four layer feature maps of different scales, and a SKNet module that adjusts the receptive field. It is worth noting that the sizes of four inputs $\boldsymbol{B}_i(i = 1, 2, 3, 4)$ of the SK feature pyramid can be adjusted according to extracted feature maps of different scales. In the Figure 3(*a*), it is mainly divided into two steps. In the first step, $\boldsymbol{B}_i(i = 1, 2, 3, 4)$ is given different receptive fields through SKNet to get $\boldsymbol{B}_i^{sk}(i = 1, 2, 3, 4)$,choose B_1^{sk}=$\boldsymbol{F}_1$.

$$\boldsymbol{B}_i^{sk} = SK\left(\boldsymbol{B}_i\right)(i = 1, 2, 3, 4) \tag{1}$$

where SK is the SKNet [18] operation.

Each feature map in set $\{\boldsymbol{B}_2^{sk}, \boldsymbol{B}_3^{sk}, \boldsymbol{B}_4^{sk}\}$ will be processed with a 1×1 convolution layer to get the same channel size as in set $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3\}$.

$$\boldsymbol{B}_i^{sk} = \mathrm{Cov}_{1\times1}\left(\boldsymbol{B}_i^{sk}\right)(i = 2, 3, 4) \tag{2}$$

where $Cov_{1\times1}$ represents the 1×1 convolution operation to adjust channel.

In the second step, in order to ensure channel fusion of equal size, double up-sampling is performed on the high-level feature map $\boldsymbol{F}_i(i = 1, 2, 3)$ to get the same size with $\boldsymbol{B}_{i+1}^{sk}(i = 2, 3, 4)$. After the two steps, the size of the two feature map $\boldsymbol{F}_i$ and $\boldsymbol{B}_{i+1}^{sk}(i = 2, 3, 4)$ is exactly the same, and $\boldsymbol{F}_{i+1}(i = 2, 3, 4)$ is obtained after the broadcasting addition

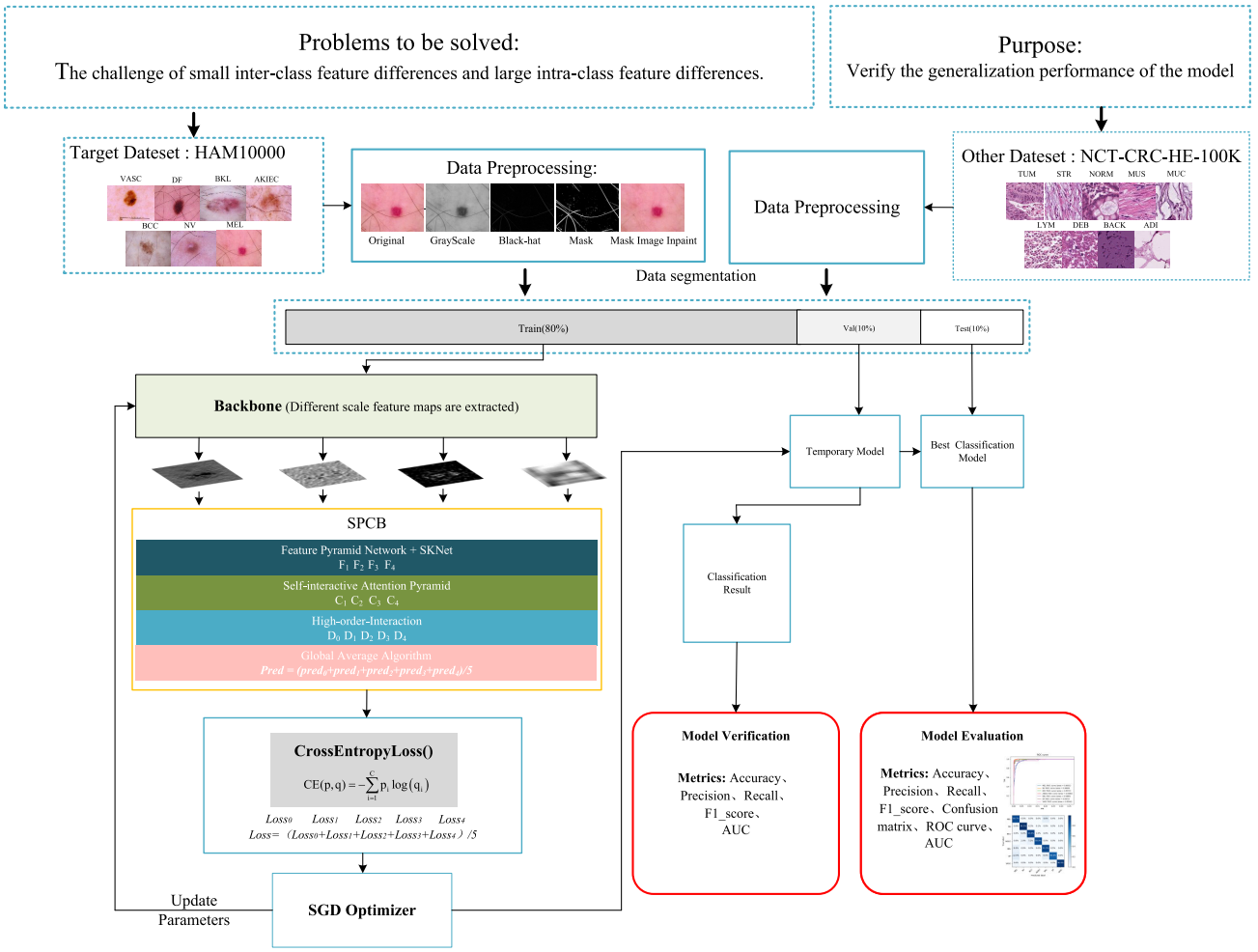


FIGURE 5. The flowchart of the experiment.

operation of $\boldsymbol{F}_i$ and $\boldsymbol{B}_{i+1}^{sk}(i = 1, 2, 3)$.

$$F_i = \begin{cases} \boldsymbol{B}_i^{sk} & i = 1 \\ \boldsymbol{B}_i^{sk} + \boldsymbol{upsampling}(\boldsymbol{F}_{i-1}) & i = 2, 3, 4 \end{cases} \tag{3}$$

where ***upsampling*** represents the double upsampling operation to adjust size.

B. SELF-INTERACTIVE ATTENTION PYRAMID(SAP)

A self-interactive attention pyramid (SAP) is proposed in Figure 3(b), where SAP combines a spatial attention pyramid [20], a channel attention pyramid and an attention skip connection pathway, which can locate the important area and suppress the useless information to reduce the background noise.

1) SPATIAL ATTENTION PYRAMID

Spatial attention is essentially to enhance the feature representation of importance regions. In essence, spatial attention transforms the spatial information in the original image into another space through the spatial conversion module and retain the key information, and generate a weight mask for each location weighted output, thereby enhancing the specific target region of interest while weakening the irrelevant background region [29].

In the paper, spatial attention module in CBAM [20] is used. The difference between spatial attention module in the paper and that in [18] is that the 7×7 kernel of the last convolutional layer is replaced with 3×3 kernels to extract more useful features and the amount of computation is reduced.

As shown in Figure 4(a), two feature maps can be obtained by the maximum pooling and average pooling in the channel dimension respectively. For each feature maps $\boldsymbol{F}_n(n = 1, 2, 3, 4)$ from Figure 3(a), the $MaxPool(\boldsymbol{F}_n) \in R^{1\times H\times W}$ and $AvgPool(\boldsymbol{F}_n) \in R^{1\times H\times W}$ are connected to obtain $\boldsymbol{S}_c \in R^{2\times H\times W}$. By a 3×3 convolution kernel, the feature information $\boldsymbol{S}_{conv} \in R^{1\times H\times W}$ is obtain in the spatial dimension. The activation function is Sigmoid, and the spatial mask $\boldsymbol{A}_n^{(s)}$ can be expressed as:

$$\boldsymbol{S}_c = MaxPool(\boldsymbol{F}_n) + AvgPool(\boldsymbol{F}_n) \tag{4}$$

$$\boldsymbol{A}_n^{(s)} = \sigma(f^{3\times3}(\boldsymbol{S}_c)) \tag{5}$$

where + indicates a feature map connection, σ represents sigmoid activation function, $f^{3\times3}$ represents a 3×3 convolution operation, *AvgPool* and *MaxPool* represent max pooling and average pooling, respectively.

2) CHANNEL ATTENTION PYRAMID

Channel attention structure to identify the importance of each feature channel. Channel attention automatically obtains the importance of each feature channel through network learning, and finally assigns different weight coefficients to each channel, so as to strengthen the suppression of non-important features by important features [29].

Inspired by SE-Net [21] and ECA-Net [26], the channel attention is shown in Figure 4(b), which is composed of an average pooling layer and two 2D convolution layers. From the feature maps $\boldsymbol{F}_n(n = 1, 2, 3, 4)$, the channel attention mask $\boldsymbol{A}_n^{(c)}(n = 1, 2, 3, 4)$ can be expressed as:

$$\boldsymbol{A}_n^{(c)} = \sigma(\boldsymbol{W}_2 \cdot ReLU(\boldsymbol{W}_1 \cdot AvgPool(\boldsymbol{F}_n))) \tag{6}$$

where σ and *ReLU* represent sigmoid and ReLU activation functions respectively, $\cdot$ represent element multiplication, and $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the weight matrixs of two 2D convolution layers.

3) SKIP CONNECTION

As shown in Figure 3(b), the skip connection represents the fusion between different channel attention and spatial attention, where the interaction is from low level to high level. The specific skip connection formula is as follows:

$$\boldsymbol{A}_n^{(c)} = (\boldsymbol{A}_{n+1}^{(c)} \oplus \boldsymbol{A}_n^{(c)})/2 \tag{7}$$

$$\boldsymbol{A}_n^{(s)} = (\boldsymbol{A}_{n+1}^{(s)} \oplus \boldsymbol{A}_n^{(s)})/2 \tag{8}$$

where $\oplus$ represents brodcadting addition. In this paper, choose $n = 1, 2$.

Actually, spatial attention treats the features in each channel equally and ignores the information interaction between channels. Channel attention is to directly process the information in a channel globally, and it is easy to ignore the information interaction in space [20]. So spatial attention and channel attention are added together to get comprehensive features. The addition operation will be performed by the feature map of spatial attention weight $\boldsymbol{A}_n^{(s)}(n = 1, 2, 3, 4)$ and the feature map of channel attention $\boldsymbol{A}_n^{(c)}(n = 1, 2, 3, 4)$, and then the result of addition operation multiply with the feature map $\boldsymbol{F}_n$ to obtain a scaled feature $\boldsymbol{F}_n^{(a)}(n = 1, 2, 3, 4)$. $\boldsymbol{F}_n$ can be expressed as:

$$\boldsymbol{F}_n^{(a)} = \boldsymbol{F}_n \otimes (\boldsymbol{A}_n^{(c)} \oplus \boldsymbol{A}_n^{(s)}) \tag{9}$$

where $\otimes$ represents element-wise multiplication and $\oplus$ represents brodcadting addition.

C. CROSS-LAYER BILINEAR-TRILINEAR POOLING OPERATION

In Figure 3(c), a cross-layer bilinear-trilinear pooling operation is proposed, which includes trilinear pooling operation and bilinear pooling operation. Most of the visual classification is mainly based on second-order interaction [13], [27], [28]. Wang et al. [13] already tried to use an efficient cross-layer trilinear pooling operation for computing the high order interaction. The trilinear pooling operation can be expressed by the following formula:

$$\boldsymbol{D}_n = AvgPool(\boldsymbol{C}_i \otimes \boldsymbol{C}_j \otimes \boldsymbol{C}_k) \tag{10}$$

where $\otimes$ represents the hadamard product, *AvgPool* is the average pooling operation, $i, j, k \in \{1, 2, 3, 4\}$ are not equal to each other and $n \in \{1, 2, 3, 4\}$. However, cross-layer trilinear pooling operation [13] for all layers not only requires extra computational cost, but also may affect the classification effect. Therefore, on the basis of Ref. [13], part of the cross-layer trilinear pooling operation is simplified to cross-layer bilinear pooling operation. A combination of cross-layer trilinear and bilinear pooling operation is proposed, which is called cross-layer bilinear-trilinear pooling operation. Cross-layer bilinear-trilinear pooling combines the advantages of cross-layer trilinear pooling and cross-layer bilinear pooling with fewer parameters. The formula of the cross-layer bilinear-Trilinear pooling operation is as follows:

$$\boldsymbol{D}_p = AvgPool(\boldsymbol{C}_a \otimes \boldsymbol{C}_b) \tag{11}$$

$$\boldsymbol{D}_q = AvgPool(\boldsymbol{C}_c \otimes \boldsymbol{C}_d \otimes \boldsymbol{C}_e) \tag{12}$$

where $\otimes$ represents the hadamard product, *AvgPool* is the average pooling operation, and the numbers in the $a, b, \in \{1, 2, 3, 4\}$, $c, d, e \in \{1, 2, 3, 4\}$ are not equal to each other, $p \in \{1, 2\}$, $q \in \{3, 4\}$.

As shown in Figure 3(c), the cross-layer trilinear pooling needs three inputs and the cross-layer bilinear pooling only needs two inputs. Through this bilinear-trilinear pooling operation, $\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3$ and $\boldsymbol{D}_4$ can be obtained after high-order interaction.

D. GLOBAL AVERAGE ALGORITHM (GAA)

A global average algorithm is proposed in this section as shown from, Figure 3(c). Feature maps D_1, D_2, D_3, D_4 and the feature map D_0 of a backbone are input five classifiers in Figure 3(*iii*), respectively, where classifier contains a global pooling (GAP) layer and two fully connected (FC) layers. Then five prediction results are obtained, which are $pred_1, pred_2, pred_3, pred_4$ and $pred_5$. *Sumpred* is divided by five according to Figure 3(*i*), and *Avgpred* is obtained.The procedure of the GAA is shown in algorithm 1. The *Avgpred* is the classification result of a network (SPCB-Net), where *Avgpred* is also the largest probability value of a class.

IV. EXPERIMENTS AND RESULTS

In this study, all experiments are implemented using PyTorch 1.8, python3.7, and run on server with Intel I7 CPU and RTX 3090 GPU. Experiments are conducted on two datasets, HAM10000 [30] and NCT-CRC-HE-100K [31]. Figure 5 is the flowchart of this paper experiment. The target dataset of this paper is HAM10000, and the SPCB-Net design aims

Algorithm 1 The Procedure of the GAA

Input: Feature maps of different level D_1, D_2, D_3, D_4 and D_0;

Output: Final classification result vector: *Avgpred*;

```
D0 ←backbone\\ Output feature matrix of backbone
for all Dk ∈ {D0, D1, D2, D3, D4, } do
    predk ←Dk by GAP and two FC layers;
end for
Sumpred ← pred1 + pred2 + pred3 + pred4 + pred5;
Avgpred ← Sumpred/5;
return Avgpred;
```

to better address the challenge of small inter-class feature differences and large intra-class feature differences. After data preprocessing, we trained the backbone+SPCB-Net directly. The loss function utilized is cross-entropy loss, which uses it to constrain the five outputs of the GAA algorithm. The model was updated using the SGD optimizer to adjust its parameters, and then evaluated with a validation set to select the best-performing model during training. Furthermore, to assess the model's ability to generalize, experiments were conducted in the field of pathology.

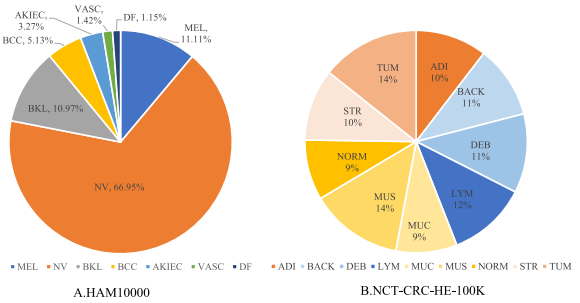


FIGURE 6. (a) Distribution of the seven types of skin types in HAM10000, (b) Distribution of the nine types of colorectal cancer images in NCT-CRC-HE-100K.

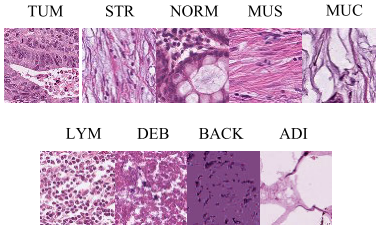


FIGURE 7. Nine different tissue images are shown in NCT-CRC-HE-100K, which are normal colonic mucosa (NORM), tumor-associated stroma (STR), colonic rectum adenocarcinoma epithelium (TUM),adipose (ADI), background (BACK), fragment (DEB),Lymphocyte (LYM), Mucus (MUC) and Smooth muscle (MUS).

A. DATASETS

A.HAM10000[1] The images in HAM10000 were collected from diverse populations and acquired and stored

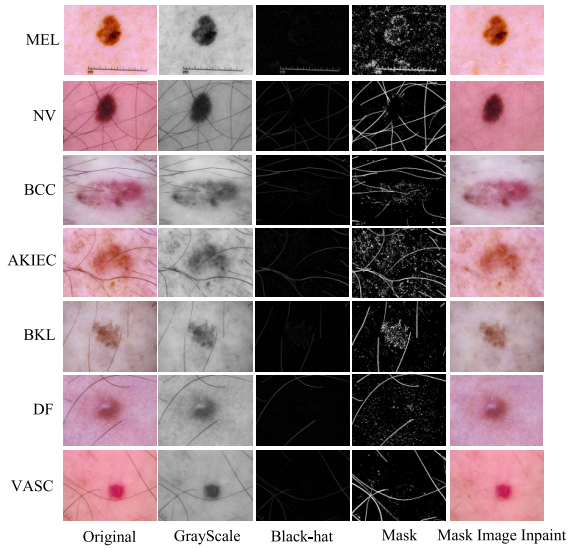


FIGURE 8. The removal method for skin image with hair and other impurities based on black hat operation.

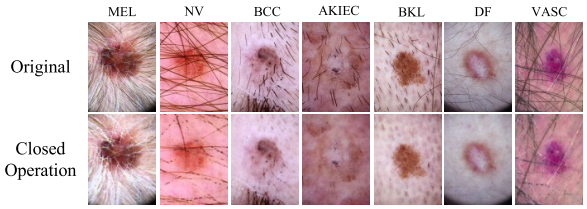


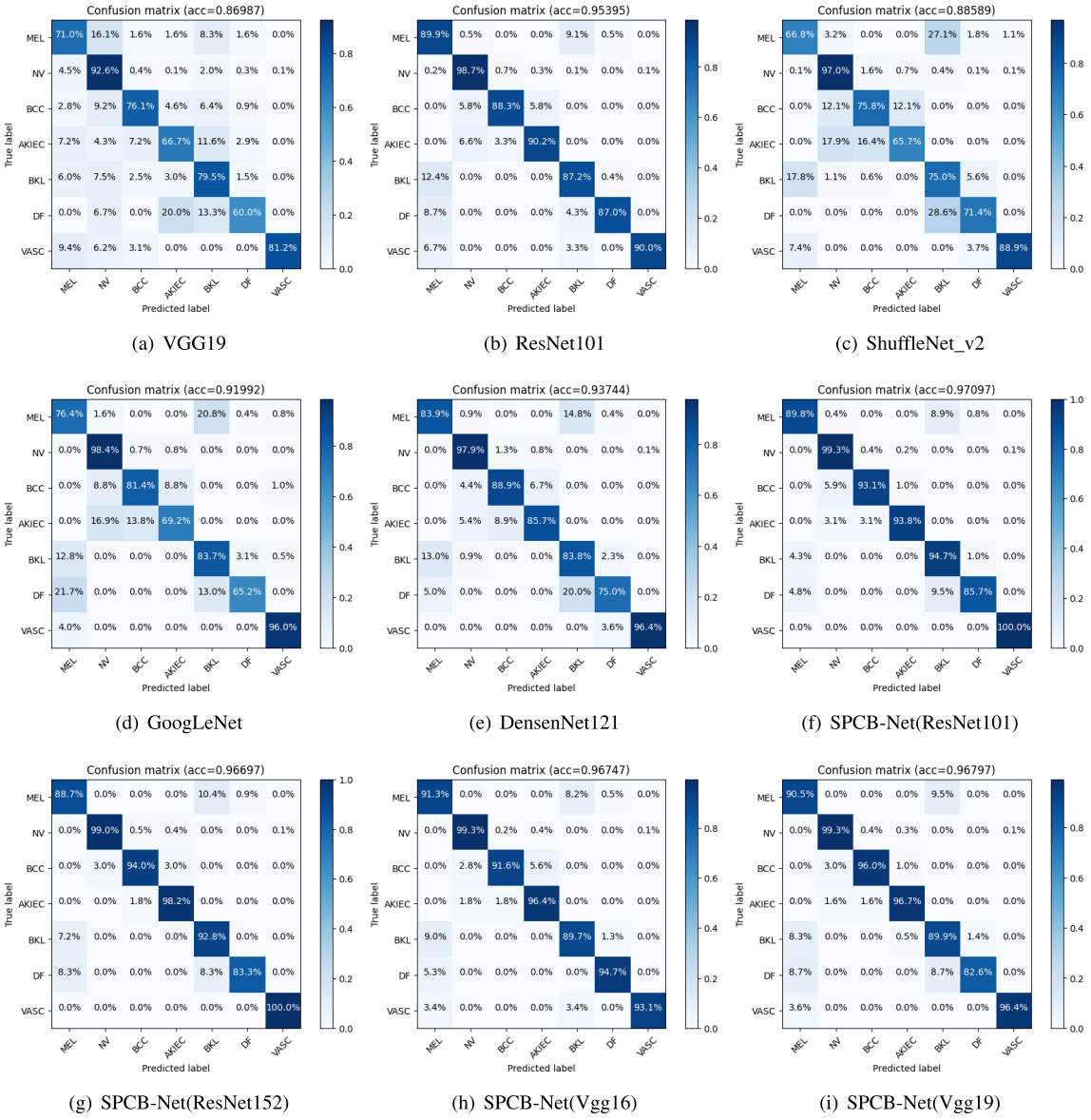
FIGURE 9. The closed operation method for removing luxuriant hairs from skin images.

through different modalities. The dataset contains a representative set of all important diagnostic categories in the field of pigmented lesions. More than half of the lesions in the dataset are confirmed via histopathology, while for the rest, basic facts are either confirmed through follow-up examination, expert consensus, or in vivo confocal microscopy. The data distribution map of HAM10000 is shown in Figure 6(A), where HAM10000 dataset contains 10015 dermoscopy images with a size of 450 × 600. There are seven types of skin lesions(cancer), which are melanoma (MEL), melanocytic nevi (NV), basal cell carcinoma (BCC), actinic keratosis/intraepithelial carcinoma (AKIEC), benign keratosis (BKL), dermatofibroma (DF), and vascular lesions (VASC). The HAM1000 dataset provides ground truth labels for 10015 images. All images in HAM1000 are divided into 80 % for training data (8017 images), 10 % valiging data(999 images) and 10 % test data (999 images).

B.NCT-CRC-HE-100K[2] The NCT-CRC-HE-100K dataset comprises a collection of 100,000 non-overlapping image patches extracted from 86 HE-stained human cancer tissue slides and normal tissue sourced from the NCT biobank (National Center for Tumor Diseases) and the UMM pathology archive (University Medical Center Mannheim). The data distribution map of NCT-CRC-HE-100K is shown in Figure 6(B), where NCT-CRC-HE-100K dataset contains

[1]https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000

[2]https://www.kaggle.com/datasets/imrankhan77/nct-crc-he-100k



(a) VGG19 (b) ResNet101 (c) ShuffleNet_v2

(d) GoogLeNet (e) DensenNet121 (f) SPCB-Net(ResNet101)

(g) SPCB-Net(ResNet152) (h) SPCB-Net(Vgg16) (i) SPCB-Net(Vgg19)

FIGURE 10. **Confusion matrix of nine different networks.**

100,000 non-overlapping patches of colorectal cancer images with a size of 224× 224. As shown in Figure 7, all colorectal cancer images in NCT-CRC-HE-100K are divided into nine categories, which are normal colonic mucosa (NORM), tumor-associated stroma (STR), colonic rectum adenocarcinoma epithelium (TUM),adipose (ADI), background (BACK), fragment (DEB),Lymphocyte (LYM), Mucus (MUC) and Smooth muscle (MUS). All images in dataset are divided into 80 % training data(80003 images), 10 % valiging data(10002 images) and 10 % test data (9995 images).

B. DATA PRE-PROCESSING

In HAM10000 dataset, hair in dermoscopy images may interfere with important information, such as texture, shape, and boundary of skin lesions. Due to the existence of hair, it is difficult to classify the dermoscopy images, so hair removal is a key step in the classification process. For recognizing and removing hair on the skin, two methods is used, which are black hat[3] and closed operation.[4] The closed operation method is to expand the image first and then corrode it. The two methods preprocess the images with hair, and the processed images can be more suitable for our model classfication. The processing effect of two methods is shown in the Figure 8 and Figure 9.

4All images in HAM10000 and NCT-CRC-HE-100K are normalized and their sizes are adjusted. Each channel pixel values of RGB image is normalized to [−1, 1] for having

[3]https://github.com/MangoWAY/medicalImageScriptDemo
[4]https://github.com/QIANXIN22/Corrosion-and-expansion

TABLE 1. Performance of backbone and SPCB-Net model in HAM10000 test set.

Model	Accuray(%)	Precision(%)	Recall(%)	Specificity(%)	F1(%)
VGG19 [31]	86.99	75.26	73.37	96.57	73.91
ResNet101 [32]	95.40	89.61	91.11	99.11	89.89
ShuffleNet_v2 [33]	88.59	77.17	69.57	97.64	71.31
GoogLeNet [34]	91.99	81.49	79.07	98.36	80.89
DensenNet121 [35]	93.74	87.37	83.64	98.60	85.36
SPCB-Net(Vgg16)	96.75	93.56	91.70	99.39	92.49
SPCB-Net(Vgg19)	96.80	93.06	92.49	99.36	92.76
SPCB-Net(ResNet50)	96.60	93.07	92.20	99.36	92.61
SPCB-Net(ResNet152)	96.70	93.71	92.27	99.29	92.90
SPCB-Net(ResNet101)	**97.10**	**93.66**	**93.34**	**99.40**	**92.25**

the same input range, which can speed up the convergence of model. In addition, the input data is adjusted to 224×224.

C. DATA AUGMENTATION

The automatic diagnosis of skin lesions still has some problems which are limited scale, insufficient existing database of dermoscopy images, and limited reliable annotation of basic facts. In order to solve these problems, the enhancement operation is performed on the training set to increase the number of training images and avoid the overfitting problem that may occur in the training process when using a small amount of training data. For HAM10000 datasets, each training dermoscopy image is randomly flipped at a given probability of 0.5. The brightness, contrast, saturation and hue of the image are randomly adjusted with a probability of 0.4. Finally, the transformation list with a given probability is randomly applied to generate the enhanced image for each original input image. For NCT-CRC-HE-100K datasets, there are sufficient and balanced data, so the data do not need to be processed.

D. EVALUATION METRICS

To quantitatively evaluate the capability of the proposed SPCB-Net on skin lesion classification, we select all the possibilities associated with classification, such as true positive (TP) classification, false positive (FP) classification, false negative (FN) classification and true negative (TN) classification, and obtain accuracy,precision, recall,specificity and F1 score. TP, FP, FN and TN are used to calculate performance indicators. In addition, the AUC is used, which is called area under curve and is used to measure the accuracy. A higher AUC value means a larger area under the curve, which means a better prediction accuracy. The closer the curve is to the top left, the higher the prediction accuracy.

E. RESULTS

1) THE STUDY OF BACKBONE COMPARE TO SPCB-NET

To show the effectiveness of our proposed module, some common classification networks are used to experiment on the HAM10000 dataset. The experimental results can be seen from Table1 and their confusion matrix diagrams are shown in Figure 10. In addition, experiments are carried out on different depths of backbone. From the Table 1, the conclusion is that SPCB-Net(ResNet101) and APBT(Vgg19) have better performance with some existing backbone in a comprehensive comparison.

2) THE STUDY OF SPCB-NET SUB-MODULES

In order to evaluate SPCB-Net and its sub-modules, we compared submodules of SPCB-Net with other attention models, differernt selection of skip connection, different combinations of bilinear and trilinear pooling operation. In the Table 2, Table3 and Table4, bc-BT-de indicates that bilinear pooling operation is used in the D_b and D_c, and trilinear pooling operation is used in the D_d and D_e, where $b, c, d, e \in \{1, 2, 3, 4\}$. PyramidAttention-efg indicates that the skip connection is used in the F_e, F_f and F_g, where $e, f, g, \in \{1, 2, 3\}$.

By reviewing a large number of literatures, two multi-output result processing algorithms and GAA algorithm are selected for experiments to obtain the best classification results, which are respectively processing algorithm 1 (PA1), processing algorithm 2 (PA2) and processing algorithm 3 (PA3).They are defined as follows:

(*i*)PA1: $(pred_1, pred_2, pred_3, pred_4)/4$;

(*ii*)PA2: Global average algorithm(GAA);

(*iii*)PA3: $\{\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3, \boldsymbol{D}_4\}$ are concatenated together by channels, then sent to the classifier.

a: SELF-INTERACTIVE ATTENTION PYRAMID(SAP) AND VISUALIZATION

By use of ResNet101 or Vgg19 as the backbone, cross-layer bilinear-trilinear pooling(12-HOI-34) and GAA algorithm(PA2), the classification performance is compared in three cases, where 1)SAP's self-interactive path is applied at different levels, 2) there is no self-interactive path, 3) the existing attention module is applied to backbone each level. As shown in Table 2, the performance of different attention sub-modules is shown on HAM10000 test set. From Table 2, we know that the best performance is achieved when the skip connection path is applied to $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$. Even compared to some traditional attention models, our SPA

TABLE 2. The performance of different attention sub-modules on HAM10000 test set.

Backbone	Attention	Accuray(%)	Precision(%)	Recall(%)	Specificity(%)	F1(%)
ResNet101	CBAM [18]	96.80	93.61	90.96	99.33	92.24
	SE [19]	96.75	**94.10**	93.30	99.39	93.10
	SK [36]	96.90	**94.10**	92.43	99.37	93.20
	ECA [24]	96.50	91.76	90.81	99.27	91.27
	PyramidAttention	96.75	93.33	**93.61**	99.31	92.23
	PyramidAttention123	96.45	92.81	92.11	99.24	92.34
	PyramidAttention23	96.65	93.27	93.37	99.33	**93.27**
	PyramidAttention13	96.74	92.79	93.71	99.33	92.50
	PyramidAttention12(**proposed**)	**97.10**	93.66	93.34	**99.40**	93.25
Vgg19	CBAM [18]	96.15	91.46	89.76	99.23	90.56
	SE [19]	96.55	92.09	90.89	99.30	91.47
	SK [36]	96.60	93.02	90.33	99.04	91.84
	ECA [24]	96.40	91.56	90.89	99.26	91.14
	PyramidAttention	96.75	92.29	92.54	99.36	92.36
	PyramidAttention123	96.70	91.87	93.21	99.35	92.47
	PyramidAttention23	96.75	91.64	**92.91**	**99.40**	92.17
	PyramidAttention13	96.65	92.41	93.20	99.34	92.74
	PyramidAttention12(**proposed**)	**96.80**	**93.06**	92.49	99.36	**92.76**

TABLE 3. The performance of different High-Order-Interaction methods on HAM10000 test set.

Backbone	Pooling Method	Accuray(%)	Precision(%)	Recall(%)	Specificity(%)	F1(%)
ResNet101	Bilinear Pooling	96.64	92.60	91.46	99.33	92.06
	Trilinear Pooling	96.95	93.89	93.31	99.37	93.53
	1-BT-134	96.80	94.16	92.21	99.36	93.10
	23-BT-14	96.90	93.36	93.06	99.49	93.19
	13-BT-24	96.55	92.84	92.41	99.33	92.60
	14-BT-23	96.70	93.10	92.56	99.36	92.80
	24-BT-13	96.80	**94.21**	93.30	99.36	**93.73**
	34-BT-12	96.45	92.37	91.73	99.31	92.01
	12-BT-34(**proposed**)	**97.10**	93.66	**93.34**	99.40	93.25
Vgg19	Bilinear Pooling	96.64	92.67	92.53	99.31	92.56
	Trilinear Pooling	96.70	90.76	92.40	99.33	91.53
	1-BT-134	96.45	91.83	92.40	99.31	92.06
	23-BT-14	96.60	91.79	92.31	**99.36**	92.04
	13-BT-24	96.50	92.89	90.73	99.31	91.69
	14-BT-23	96.65	92.06	92.97	99.30	92.36
	24-BT-13	96.70	92.10	**93.06**	99.34	92.47
	34-BT-12	96.75	92.28	92.54	**99.36**	92.13
	12-BT-34(**proposed**)	**96.80**	**93.06**	92.49	**99.36**	**92.76**

module shows efficient performance. In order to show the advantages of SPA more directly. The heat map can be seen in Figure13, which showed SAP can capture more details and general information well, showing better performance than SE attention and CBAM attention.

b: CROSS-LAYER BILINEAR-TRILINEAR POOLING OPERATION

As show in Table 3, By use of ResNet101 or Vgg19 as the backbone, the best classification performance is realized by applying cross-layer bilinear pooling at the layers of $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ and cross-layer trilinear pooling at the layers of $\boldsymbol{D}_3$ and $\boldsymbol{D}_4$. The experimental results prove the excellent performance of the proposed bilinear-trilinear pooling operation

c: RESULTS PROCESSING ALGORITHM

The classification performance is compared in four cases: 1) by use of processing algorithm 1(PA1), the four results ($pred_1, pred_2, pred_3, pred_4$) of SPCB-Net output are respectively sent to the classifier then averaged, 2) by use of processing algorithm 2 (PA2) [19], the five output results ($pred_1, pred_2, pred_3, pred_4, pred_5$) of SPCB-Net and backbone are averaged respectively, 3) by use of processing algorithm 3 (PA3) [13], The four feature output matrices (D1, D2, D3, D4) of SPCB-Net are concatenated together by channels, and finally sent to the classifier and 4) perform a hybrid operation of algorithm 1, algorithm 2 and algorithm 3

As show in Table 4, the best classification performance is the network using processing algorithm 2, and it is far

TABLE 4. The performance of differen multi-output processing algorithm on test set HAM10000.

CNN model	Different combinations of sub-modules	Accuray(%)	Precision(%)	Recall(%)	Specificity(%)	Mean F1(%)
ResNet101	PA1	96.75	93.61	90.71	99.37	92.04
	PA3	96.10	91.03	92.44	99.24	91.00
	PA1+PA3	96.30	91.14	92.49	99.27	91.67
	PA2+PA3	96.45	93.00	92.64	99.29	92.81
	PA1+PA2+PA3	96.70	92.81	93.16	**99.40**	92.97
	PA2(**Proposed**)	**97.10**	**93.66**	**93.34**	**99.40**	**93.25**
Vgg19	PA1	96.50	92.53	90.37	99.33	91.36
	PA3	96.60	92.91	93.23	99.27	92.03
	PA1+PA)	96.44	92.97	91.77	99.29	**93.63**
	PA2+PA3	96.65	91.89	92.37	**99.36**	92.07
	PA1+PA2+PA3	96.60	92.33	91.31	99.33	91.79
	PA2(**proposed**)	**96.80**	**93.06**	**92.49**	**99.36**	92.76

TABLE 5. The performance of each sub-modules of the SPCB-Net on a test set HAM10000.

Model	Accuray(%)	Precision(%)	Recall(%)	Specificity(%)	F1(%)
ResNet101	95.40	90.06	91.44	99.10	91.44
ResNet101+SK-FP	96.45	92.33	91.80	99.38	93.20
ResNet101+SK-FP+SAP	96.75	93.23	92.80	99.36	93.00
ResNet1SK+SK-FP+HOI	96.55	92.83	93.21	99.33	92.96
ResNet101+SPCB-Net(SK-FP+SPA+HOI)	**97.10**	**93.66**	**93.34**	**99.40**	**93.25**

TABLE 6. Comparison of the overall performance of the proposed method with state-of-the-art on HAM10000.

Model	Methods	Accuray(%)	AUC
DenseNet-II [37]	Base on DenseNet	96.27	0.96
DPE-BOTNET [38]	DensTeNet201 + DPESA	95.80	0.94
FixCaps [14]	Capsule + CBAM	96.49	0.98
Cubic SVM [39]	MESbS + SVM	96.70	0.98
SPCB-Net(ResNet101)(**proposed**)	Attention + HOI	**97.10**	**0.99**

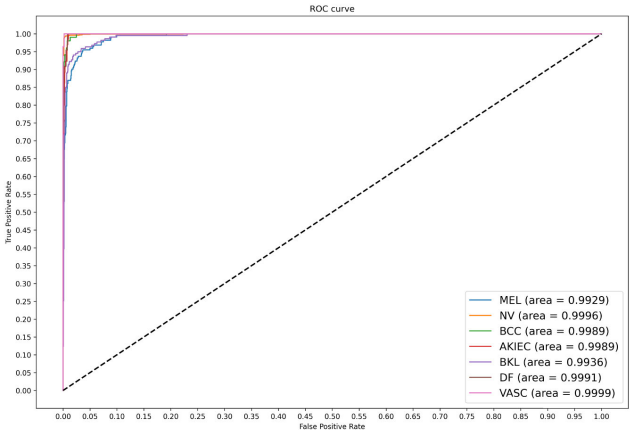


FIGURE 11. ROC curve of SPCB-Net(ResNet101) on the dataset HAM10000.

ahead in all performance. The experimental results prove the excellent performance of the proposed GAA algorithm.

3) ABLATION EXPERIMENTS OF SPCB-NET

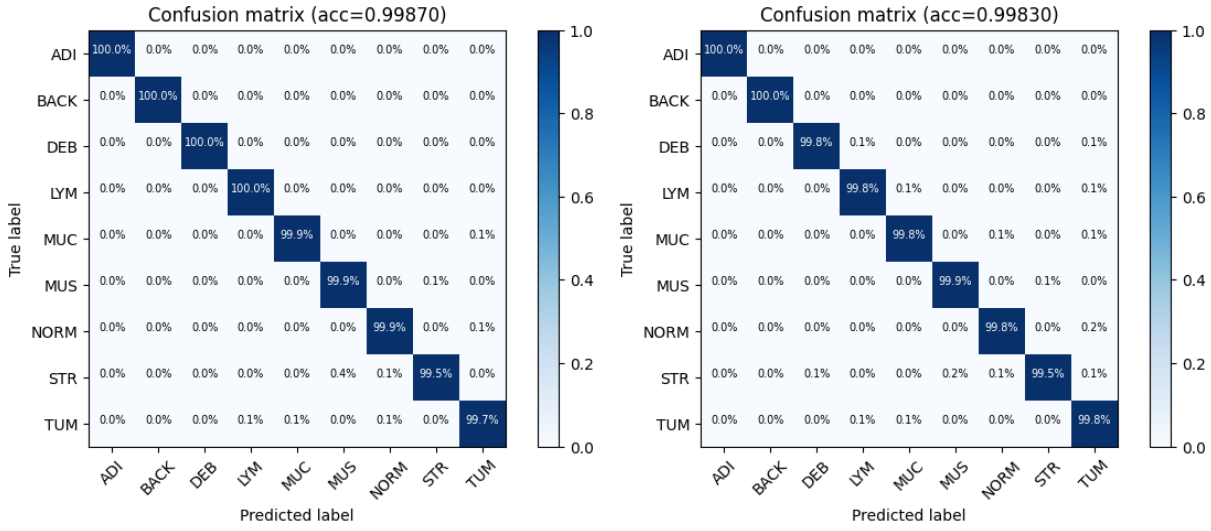
Ablation experiments are performed to analyze the contribution of each module of SPCB-Net. The ablation experiments are carried out on the HAM10000 dataset, using ResNet101 as the backbone.

a: EFFECTS OF ATTENTION PYRAMIDS

In Table 5, by comparing the ResNet101 with ResNet101 added attention mechanism, we can find that the attention mechanism can significantly improve the scores of indicators compared with the original backbone. The attention mechanism enhances feature representation and discrimination of location area to improve classification performance.

b: EFFECTS OF HIGH-ORDER-INTERACTION

As can be seen from Table 5, after joining HOI, scores of various indicators have been improved to varying degrees compared with ResNet101. But comparing with ResNet101+SPA, the scores of Accuray, Precision,



(a) SPCB-Net(ResNet101) (b) SPCB-Net(Vgg19)

FIGURE 12. Confusion matrixs of SPCB-Net(Vgg19) and APBT(ResNe101) on NCT-CRC-HE-100K respectively.

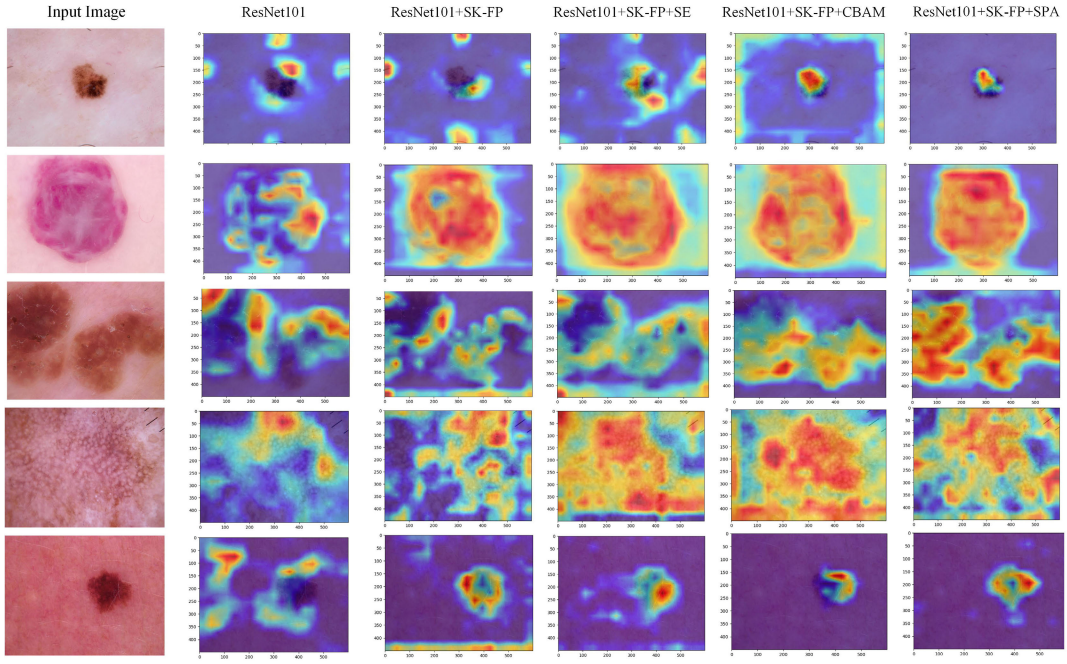


FIGURE 13. Visualized results of ablation study with the attention module. In the figure, a higher weight reflects a higher thermal value. The first column is the input images, ResNet101 is used in the second column, ResNet101 + SK-FP is used in the third column, ResNet101+ SK-FP + SE is used the fourth column, ResNet101+ SK-FP + CBAM is used in the fifth column, and ResNet101+ SK-FP + SAP is used in the sixth column.

Specificity and F1 have a small drop. The main reason for the decline of scores is that the residual structure of ResNet has a similar effect with our HOI. After adding HOI, we can see in Table 5 that in addition to the recall rate and F1 scores, there are different degrees of improvement.

Through the above two ablation experiments, it can be seen that each part of SPCB-Net plays an important role and has good performance. Figure 11 is the ROC curve of SPCB-Net prediction for different categories. The AUC of each category is close to 1 and near the upper left corner, which proves that SPCB-Net is an excellent classifier.

4) GENERALIZATION PERFORMANCE OF SPCB-NET

Experiments are performed on dataset NCT-CRC-HE-100K to evaluate the generalization performance of SPCB-Net. As shown in the Figure12 is the confusion matrix. Figure12 and Table 7 indicates that SPCB-Net module can be not only applied in dermatology, but also can be applied

TABLE 7. Comparison of the accuracy of the proposed method with state-of-the-art on NCT-CRC-HE-100K.

Model	Methods	Accuray(%)	AUC
Ensemble DNN [40](2021)	Ensemble Learning	99.13	**0.99**
DARC [41] (2022)	Clustering	99.76	0.97
IL-MCAM [42](2022)	Attention +Interactive	99.78	0.98
CRCCN-Net [43](2023)	CNN Models	99.26	0.98
SPCB-Net(VGG19)(**proposed**)	Attention + HOI	99.83	**0.99**
SPCB-Net(ResNet101)(**proposed**)	Attention + HOI	**99.87**	**0.99**

in histopathology with state-of-the-art performance, which proves good generalization of SPCB-Net. Therefore, it is not only well qualified for the task of identifying skin cancer,but has the potential to be applied to other medical image fields.

5) COMPARISON WITH STATE-OF-THE-ART METHODS

In this section, ResNet101+HIAP and Vgg+HIAP are used for comparing with state-of-the-art models on dataset HAM10000 and NCT-CRC-HE-100K.

The evaluation indexes are precision, recall specific, F1 scores and AUC, which are all the best performance on the model. In Table 6,7, our model is compared with the state-of-the-art model in dataset HAM10000 and NCT-CRC-HE100K, and SPCB-Net(ResNet) achieve the best performance, which are both achieved performance improvements of 0.4% compared to the state-of-the-art models. In summary, SPCB-Net(ResNet101) achieves the best performance on both datasets HAM10000 and NCT-CRC-HE-100K.

V. CONCLUSION

In this paper, a multi-scale skin cancer image identification network using self-interactive attention pyramid and cross-layer bilinear-trilinear pooling (SPCB-Net) to solve a problem that low-level features are very similar between different classes and high-level features have different to a certain extent between the same class. SPCB-Net is mainly composed of four sub-modules which are SK-FPN, self-interactive attention pyramid (SAP), high-order interaction based on bilinear-trilinear pooling operation and output result processing. SPCB-Net can accurately locate the local regions, realize the interaction between low-level features and high-level features, reduce background noise and improve the classification effect, and has good generalization ability.

Experimental results demonstrate the effectiveness of the SPCB-Net on HAM10000 and NCT-CRC-HE-100K. Compared with the state-of-the-art models, SPCB-Net(ResNet101) model exceeds them in most evaluation indicators. In particular, the predicted classification accuracy reaches 97.10 % on HAM10000 and 99.87% on NCT-CRC-HE-100K, which are both achieved performance improvements of 0.4% compared to the state-of-the-art models. Therefore, SPCB-Net can be very good for recognition of skin cancer image task, used in computer-aided medical diagnosis, improve the discovery rate of skin cancer in the crowd, has the potential to identify skin cancer by identifying pathological tissue. Good generalization ability shows that SPCB-Net can promote the improvement and popularization of medical cancer screening technology. However, it still has some shortcomings. For example, this paper does not fully discuss the computational complexity of SPCB-Net and its effectiveness for other classification tasks. One direction of our future work is to apply SPCB-Net to other domain tasks and make a theoretical analysis of their computational complexity.

REFERENCES

[1] Skin Cancer Foundation. (2022). *Skin Cancer Facts & Statistics,What You Need to Know*. [Online]. Available: https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/

[2] W. E. Damsky and M. Bosenberg, "Melanocytic nevi and melanoma: Unraveling a complex relationship," *Oncogene*, vol. 36, no. 42, pp. 5771–5792, Oct. 2017.

[3] R. Marks, "An overview of skin cancers," *Cancer*, vol. 75, no. S2, pp. 607–612, Jan. 1995.

[4] M. Thorn, F. Ponte, R. Bergstrom, P. Sparen, and H.-O. Adami, "Clinical and histopathologic predictors of survival in patients with malignant melanoma: A population-based study in Sweden," *JNCI J. Nat. Cancer Inst.*, vol. 86, no. 10, pp. 761–769, May 1994.

[5] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[6] G. J. Chowdary, G. V. S. N. D. Yathisha, G. Suganya, and M. Premalatha, "Automated skin lesion segmentation using multi-scale feature extraction scheme and dual-attention mechanism," in *Proc. 3rd Int. Conf. Adv. Comput., Commun. Control Netw. (ICAC3N)*, Dec. 2021, pp. 1763–1771, doi: 10.1109/ICAC3N53548.2021.9725739.

[7] Y. Dong, L. Wang, S. Cheng, and Y. Li, "FAC-Net: FeedBack attention network based on context encoder network for skin lesion segmentation," *Sensors*, vol. 21, no. 15, p. 5172, Jul. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/15/5172

[8] K. M. Hosny and M. A. Kassem, "Refined residual deep convolutional network for skin lesion classification," *J. Digit. Imag.*, vol. 35, no. 2, pp. 258–280, Apr. 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:245856976

[9] P. Bansal, R. Garg, and P. Soni, "Detection of melanoma in dermoscopic images by integrating features extracted using handcrafted and deep learning models," *Comput. Ind. Eng.*, vol. 168, Jun. 2022, Art. no. 108060.

[10] J. Zhao, G. Ji, X. Han, Y. Qiang, and X. Liao, "An automated pulmonary parenchyma segmentation method based on an improved region growing algorithmin PET-CT imaging," *Frontiers Comput. Sci.*, vol. 10, no. 1, pp. 189–200, Feb. 2016.

[11] H. D. Li and L. Y. Zhao, "The mathematical morphology image edge detection based on FPGA," *Appl. Mech. Mater.*, vol. 467, pp. 599–603, Dec. 2013.

[12] J. Cheng, S. Tian, L. Yu, C. Gao, X. Kang, X. Ma, W. Wu, S. Liu, and H. Lu, "ResGANet: Residual group attention network for medical image classification and segmentation," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102313. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841521003583

[13] J. Wang, N. Li, Z. Luo, Z. Zhong, and S. Li, "High-order-interaction for weakly supervised fine-grained visual categorization," *Neurocomputing*, vol. 464, pp. 27–36, Nov. 2021.

[14] Z. Lan, S. Cai, X. He, and X. Wen, "FixCaps: An improved capsules network for diagnosis of skin cancer," *IEEE Access*, vol. 10, pp. 76261–76267, 2022.

[15] M. Gridach, R. Yasrab, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "D2ANET: Densely attentional-aware network for first trimester ultrasound CRL and NT segmentation," in *Proc. IEEE 20th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2023, pp. 1–4.

[16] A. He, T. Li, N. Li, K. Wang, and H. Fu, "CABNet: Category attention block for imbalanced diabetic retinopathy grading," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 143–153, Jan. 2021.

[17] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," 2016, *arXiv:1612.03144*.

[18] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[19] Y. Ding, Z. Ma, S. Wen, J. Xie, D. Chang, Z. Si, M. Wu, and H. Ling, "AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2826–2836, 2021.

[20] J. Woo, S. Park, J.-Y. Lee, E. V. Kweon, M. Hebert, C. Sminchisescu, and Y. Weiss, "CBAM: Convolutional block attention module," in *Computer Vision—ECCV 2018*. Cham, Switzerland: Springer, 2018, pp. 3–19.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[22] H. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, and L. Uhlmann, "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, 2018.

[23] Y. Li and L. Shen, "Skin lesion analysis towards melanoma detection using deep learning network," 2017, *arXiv:1703.00577*.

[24] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[25] S. Qiao, L.-C. Chen, and A. Yuille, "DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10208–10219.

[26] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 13–19.

[27] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.

[28] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," 2018, *arXiv:1807.09915*.

[29] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 763–772.

[30] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, Aug. 2018, Art. no. 180161.

[31] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, L. Jansen, C. C. Reyes-Aldasoro, I. Zörnig, D. Jäger, H. Brenner, J. Chang-Claude, M. Hoffmeister, and N. Halama, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLOS Med.*, vol. 16, no. 1, Jan. 2019, Art. no. e1002730.

[32] H. Li and M. Wang, "Very deep convolutional network for large-scale image recognition," *Jisuanji Xitong Yingyong= Comput. Syst. Appl.*, vol. 60, no. 9, p. 330, 2021.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[34] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[36] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.

[37] N. Girdhar, A. Sinha, and S. Gupta, "DenseNet-II: An improved deep convolutional neural network for melanoma cancer detection," *Soft Comput.*, vol. 27, no. 18, pp. 13285–13304, Aug. 2022, doi: 10.1007/s00500-022-07406-z.

[38] K. Nakai and X.-H. Han, "DPE-BoTNeT: Dual position encoding bottleneck transformer network for skin lesion classification," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5.

[39] S. Maqsood and R. Damaševičius, "Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare," *Neural Netw.*, vol. 160, pp. 238–258, Mar. 2023.

[40] S. Ghosh, A. Bandyopadhyay, S. Sahay, R. Ghosh, I. Kundu, and K. C. Santosh, "Colorectal histology tumor detection using ensemble deep neural network," *Eng. Appl. Artif. Intell.*, vol. 100, Apr. 2021, Art. no. 104202.

[41] J. Li, J. Liu, H. Yue, J. Cheng, H. Kuang, H. Bai, Y. Wang, and J. Wang, "DARC: Deep adaptive regularized clustering for histopathological image classification," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102521.

[42] H. Chen, C. Li, X. Li, M. M. Rahaman, W. Hu, Y. Li, W. Liu, C. Sun, H. Sun, X. Huang, and M. Grzegorzek, "IL-MCAM: An interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach," *Comput. Biol. Med.*, vol. 143, Apr. 2022, Art. no. 105265.

[43] A. Kumar, A. Vishwakarma, and V. Bajaj, "CRCCN-Net: Automated framework for classification of colorectal tissue using histopathological images," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104172.

XIN QIAN received the bachelor's degree in computer science from Jinggangshan University, China, in 2021. He is currently pursuing the master's degree with the Chongqing University of Science and Technology. His research interests include machine learning and medical image analysis.

TENGFEI WENG received the B.S. and M.S. degrees in chemistry and chemical engineering from Chongqing University, China, in 2007 and 2010, respectively. Currently, she is with the Chongqing University of Science and Technology. Her current research interests include artificial intelligence and neural networks.

QI HAN received the B.S. degree in computer science and technology from Shandong University, China, in 2005, and the M.S. and Ph.D. degrees from Chongqing University, China, in 2009 and 2012, respectively. Currently, he is an Associate Professor with the Chongqing University of Science and Technology. His current research interests include artificial intelligence, system optimization, neural networks, and chaos control.

CHEN WU received the B.S. degree in computer science and technology from the Shanxi Institute of Technology, Shanxi, in 2019. He is currently pursuing the M.S. degree with the Chongqing University of Science and Technology. His research interests include fine-grained visual categorization and computer vision.

HONGXIANG XU received the B.S. degree from the School of Sanjiang University, Jiangsu, China, in 2020. He is currently pursuing the master's degree with the Chongqing University of Science and Technology. His research interests include deep learning, brain–computer interface, and medical image analysis.

MINGYANG HOU received the B.S. degree from the School of Information Engineering, Institute of Disaster Prevention, Hebei, China, in 2018, and the M.S. degree from the School of Intelligent Engineering, Chongqing University of Science and Technology, Chongqing, China, in 2023, where he is currently pursuing the Ph.D. degree. He has published several articles in reputable journals, including *Computers in Biology and Medicine*. His research interests include deep learning, image super-resolution, and medical image processing.

ZICHENG QIU was born in Wuhan, Hubei, China, in 1983. He received the B.S. degree in optoelectronic engineering from the Huazhong University of Science and Technology, Wuhan, in 2005, and the Ph.D. degree in optical engineering (advanced lithography technologies) from the Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 2010. From 2010 to 2011, he was a Software Engineer with Synopsys. From 2011 to 2015, he was an Assistant Professor with the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Science, Chongqing, China. From 2015 to 2020, he was an Associate Professor with the College of Information Engineering, Tarim University, Alar, Xinjiang, China. He is currently an Associate Professor with the College of Intelligent Technology of Engineering, Chongqing University of Science and Technology. He is the author of more than ten articles and three inventions. His research interests include AI, intelligent speech technologies, and security of AI.

BAOPING ZHOU received the B.S. degree in mathematics from the Kashgar Teachers College, in 1995, and the M.S. degree in agricultural electrification and automation from Northeast Agricultural University, in 2004. Since 2010, he has been a Professor with the College of Information Engineering, Tarim University. He is the author of seven books and more than 50 articles. One of his books was awarded as 12th Five-Year National Planning Textbook. He has been the person in charge of more than ten national and Xinjiang Production and Construction Corps (XPCC) projects, which includes the National Key Research and Development Program, the National Natural Science Foundation of China, and the National Natural Science Foundation of Xinjiang Joint Key Project. His research interests include applied mathematics, natural speech and language intelligent technologies, agricultural information processing, computer intelligent management, and agricultural information technologies. He is the Director of the Xinjiang Uyghur Autonomous Region Mathematical Society and the Physics Society. He is a member of the Xinjiang Organizing Committee of China Undergraduate Mathematical Contest in Modeling; and the "Smart Ecology" Professional Technical Group, Internet of Things Committee, Chinese Communications Society. He was a recipient of the Excellent Talent of XPCC, in 2013; and the Bronze XPCC Talent, in 2014, for his research work in southern Xinjiang, China.

XIANQIANG GAO received the bachelor's degree from Xidian University, in 2004, and the master's degree from the Xi'an University of Science and Technology, in 2009. He is currently a Professor with the School of Information Engineering, Tarim University.