IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 June 2024, accepted 4 July 2024, date of publication 11 July 2024, date of current version 13 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3426518

RESEARCH ARTICLE

An Enhanced U-Network by Combining PPM and CBAM for Medical Image Segmentation

ZHONGMING FU[1], HEJIAN CHEN[1], (Graduate Student Member, IEEE), MENGSI HE[1,2], AND LI LIU[1]

[1]College of Computer Science and Technology, University of South China, Hengyang, Hunan 421001, China
[2]College of Information Science and Engineering, Hunan University, Changsha, Hunan 410082, China

Corresponding author: Zhongming Fu (fuzhongming@hnu.edu.cn)

This work was supported in part by Hunan Natural Science Foundation Project of China under Grant 2023JJ40555, and in part by Hunan Provincial Department of Education Scientific Research Project of China under Grant 22B0451.

ABSTRACT U-network is a comprehensive convolutional neural network that is widely utilized in medical image segmentation domain. However, it is not accurate enough in detail segmentation and resulting in unsatisfactory segmentation results. To solve this problem, this paper proposes an enhanced U-network that combines an improved Pyramid Pooling Module (PPM) and a modified Convolutional Block Attention Module (CBAM). Its whole network is U-Net architecture, where the PPM is improved by reducing the number of bin species and increasing the pooling connection multiples. It is used in the downsampling part of the network, which can extract input image features of various dimensions. And the CBAM is modified by using 1×1 convolutional layers instead of the original fully connected layers. It is used in the upsampling part of the network, which can combine convolution and attention mechanism. This pays attention to the image from two aspects of space and channel. Besides, the network is trained with novel RGB training to further improve the segmentation ability of the network. Experimental results show that our network outperforms traditional U-shaped segmentation networks by 30% to 40% in metrics Dice, IoU, MAE, and BFscore respectively. What's more, it is better than U-Net ++, U^2-Net, ResU-Net, ResU-Net++, and UNeXt in terms of segmentation effect and training time.

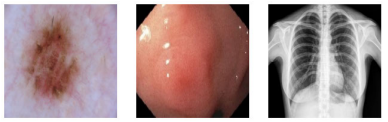
INDEX TERMS U-Net, pyramid pooling module, convolutional block attention module, RGB train.

I. INTRODUCTION

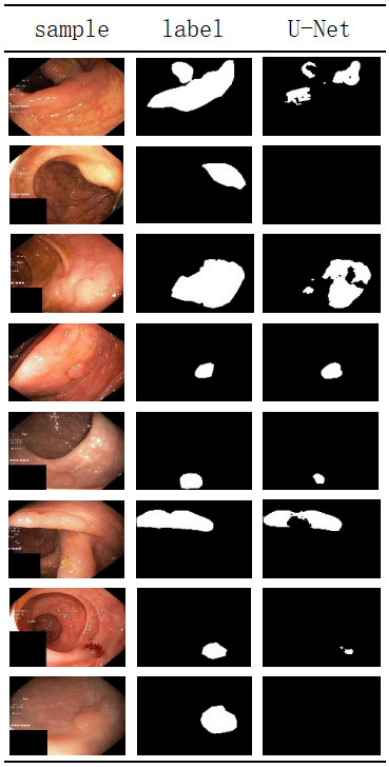
With the development of medical imaging technology, medical image segmentation has become an important research direction in the field of medical image processing. Fig. 1 shows medical images of three different modalities, namely Skin mirror images, Endoscopic images, and X-ray images. The purpose of medical image segmentation is to segment different organ tissues or lesion areas in the image, so that doctors can diagnose and treat patients' diseases. Traditional medical image segmentation methods rely mainly on image processing techniques, such as thresholding [1], [2], [3], region growing [4], [5], region merging and splitting [4], clustering [6], and edge detection [7], [8]. Nevertheless, these methods have obvious shortcomings: First, they are difficult to deal with complex background and noise, so the segmentation effect is not good. Second, their algorithms are difficult to deal with irregularly shaped objects, so the segmentation accuracy is limited. Third, the segmentation results require more manual participation, which is time-consuming and labor-intensive. Fourth, they are difficult to process massive image data due to slow computation.

In recent years, many researchers were turned their attention to deep learning. It has shown great potential in advancing research on medical image segmentation, such as AlexNet [9], VGGNet [10], GoogleNet [11], Faster-RCNN [12], and extensions on top of these basic network models. In terms of network structure, FCNs [13] are better than the most advanced segmentation methods by using a fully convolutional network structure that enables end-to-end, pixel-to-pixel training. In terms of tricks, resnet [14] is a

The associate editor coordinating the review of this manuscript and approving it for publication was Zhan-Li Sun.



(a) Medical images.



(b) Segmentation results of U-Net.

FIGURE 1. **Medical images and segmentation results.**

residual network that solves the problem of decreased effect when the network depth is deepened in deep learning.

As time demands increase, the general trend leans towards a network structure that is both simple and efficient. U-Net [15] network structure is a type of convolutional neural network that is widely used in medical image segmentation applications, as well as a variety of other image processing tasks. This type of network is characterized by its simplicity, as it consists of two main parts: the encoder and decoder. The encoder part of the network is to extracting feature maps from the input image and compressing the spatial size of the feature maps, with each layer of max pooling being applied to improve resolution. This involves dividing the image into smaller blocks, and extracting features of different sizes in order to capture the image's fine details and textures. The decoder part of the network is responsible for recovering the spatial resolution of the feature maps that are compressed by the encoder, thereby allowing for a seamless and high-quality reconstruction of the original image. More importantly, the decoder can use the feature maps in the contract path to grab the context information of the required resolution, which is crucial for accurately segmenting the object's boundaries and fine details. The efficiency and efficacy of U-Net in object segmentation has been validated by its ability to effectively represent complex structural and boundary details of objects. However, it has certain limitations as its structure is relatively simple and only a small number of layers are utilized to extract multi-level features, which leads to the segmentation results often being slightly fuzzy. Therefore, it is difficult to accurately restore the boundary contour of the object. This makes it challenging to accurately segment intricate objects with fine structures, as shown in Fig. 1b, which are the segmentation results of U-Net on the Kvasir-SEG dataset [16].

To solve these problems, some studies have been proposed for improving U-Net. In particular, U-Net++ [17] utilizes full scale skip connections and deep supervision. U^2Net [18] embeds a U-Net structure in each sub module of the U-Net structure. ResUNet [19] combines the strengths of residual learning and U-Net. ResUNet++ [20] is an improved ResUNet architecture that combines the characteristics of U-Net++and ResUNet. UneXt [21] designs a tokenized MLP block to effectively label and project convolutional features. However, because of complex structures, these networks require long training time. This is difficult to accept for those non-cell level medical segmentation tasks. Accordingly, a simple structure and the ability of detail segmentation is required for medical segmentation networks.

This paper proposes an enhanced U-network by combining PPM (pyramid pooling module [22]) and CBAM(convolutional block attention module [23]) for medical image segmentation. For feature extraction, the basic U-Net uses a simple convolutional pooling structure to lift features, which may result in the feature extraction of samples being incomplete. We enhance the PPM model to extract more comprehensive image features from various perspectives. For feature synthesis, the basic U-Net employs a straightforward convolution and pooling structure to reconstruct features, which may result in loss of crucial information. We modify the CBAM to provide better reconstruction of extracted features from both spatial and channel dimensions.

The primary contributions can be summarized as follows.

- We improve PPM by reducing the number of bin species and increasing the pooling connection multiples, and use it in the upsampling part of U-Net to improve network feature extraction ability, which is more suitable for medical image segmentation tasks.
- We modify CBAM by using 1×1 convolutional layers instead of the original fully connected layers, and use it in the downsampling part of U-Net, which focuses on both channel and spatial dimensions to improve the prediction accuracy.
- We use the RGB training that trains the network separately with information from the three color channels of the sample red, green, and blue. This training method can increase the model's understanding and utilization

of the feature information of different color channels of samples.

The rest of this article is structured as follows. Section II presents a review of the relevant research. Section III delves into the intricate design of the proposed network. Section IV provides a detailed breakdown of the data used and the resulting experimental results. Lastly, Section V provides a comprehensive summary of the conclusions.

II. RELATED WORK

Traditional image segmentation techniques can be broadly classified into five categories:

A. THRESHOLDING

Thresholding approaches are a simple and effective method for image segmentation [1], [2], [3]. Its basic idea is to divide the image into different regions by setting a threshold to classify the pixels.

B. REGION GROWING

It is an algorithm based on neighborhood pixel comparison. It segments object regions in the image by selecting seed points and expanding pixel domains with similar properties based on connectivity gradients [5].

C. REGION SPLIT AND MERGE APPROACH

Region segmentation first divides the image into multiple non overlapping regions. Then, region merging is based on the measurement of pixel similarity between regions, iteratively merging adjacent regions that meet certain conditions to obtain the segmentation results of the image target region [24].

D. CLUSTERING

Clustering is the process of grouping data that share similar characteristics together, based on certain criteria that compare and analyze the data. The FCM algorithm is a clustering method based on fuzzy theory. It allows a sample point to belong to multiple classes and provides metrics for the classes it belongs to. FCM achieves fuzzy clustering of data samples to various centers by iteratively optimizing the membership of class centers and the categories to which each sample point belongs [6].

E. EDGE DETECTION

Edge detection, a traditional methodology, is utilized to identify irregularities in an image. The Canny detector, an effective edge enhancement technique, uses a gradient extent threshold to identify potential edges [8]. It stifles them through the system of non-maximal suppression and hysteresis thresholding [25].

Compared with traditional segmentation methods that are often based on pixel-by-pixel processing, model based methods offer a more comprehensive network result that can be used to segment targets more accurately and efficiently [26], [27], [28], [29]. In the field of medical image segmentation, Dong et al. [30] proposed a new mesh network (MNet) for anisotropic medical image segmentation. The authors highlight the limitations of vanilla 2D/3D convolutional neural networks (CNNS) in representing sparse inter-slice information and dense intra-slice information in a balanced manner. The latent fusion of representation processes in MNet allows flexible selection of representation processes to balance inter-slice and intra-slice information.

Liu et al. [31] introduced a hybrid architecture named PHTrans, specifically designed for medical image segmentation. PHTrans efficiently combines the capabilities of CNNs and transformers to produce hierarchical representations that are rich in both global and local features. The architecture follows a U-shaped encoder-decoder design and introduces parallel hybrid modules in deep stages. These modules consist of convolution blocks and modified 3D Swin transformers that learn local features and global dependencies separately. The outputs of these modules are then aggregated using a sequence-to-volume operation.

Cheng et al. [32] proposed a Learnable Oriented-Derivative Network (LOD-Net) for polyp segmentation. This network utilizes the representation capacity of eight oriented derivatives at each pixel to determine a candidate border region for a polyp. This step involves the selection of pixels with large derivative values. The network then refines the border prediction by fusing features from the border region with high-level semantic features. Overall, LOD-Net has demonstrated superior performance in polyp segmentation by utilizing the representation ability of oriented derivatives for border region searching. The above networks utilize complex network structures to improve the segmentation ability.

Solanki et al. [33], [34], [35] presented an extensive survey on brain tumor classification and segmentation methods based on MRI images. They construct a model to detect brain tumors from 2D magnetic resonance images of the brain using hybrid deep learning techniques. Then, this method is combined with traditional classification techniques and deep learning methods. The application of this concept in clinical settings is the ultimate goal.

Patel and Kashyap [36] proposed a two-dimensional Flexible analytical wavelet transform (FAWT) based on a novel technique. This method is decomposed pre-processed images into sub-bands. Then statistical-based relevant features are extracted, and principal component analysis (PCA) is used to identify robust features. After that, robust features are ranked with the help of the Student's t-value algorithm. Finally, features are applied to Least Square-SVM (RBF) for classification.

Saxena et al. [37] proposed a chaotic algorithm based on Marine Predator Algorithm (MPA) named as Marine Predator Chaotic Algorithm (MPCA). A normalized fusion of chaotic function-is first proposed. Based on this function, position update mechanism is developed for improving the performance of the original MPA. The COVID-19 dataset

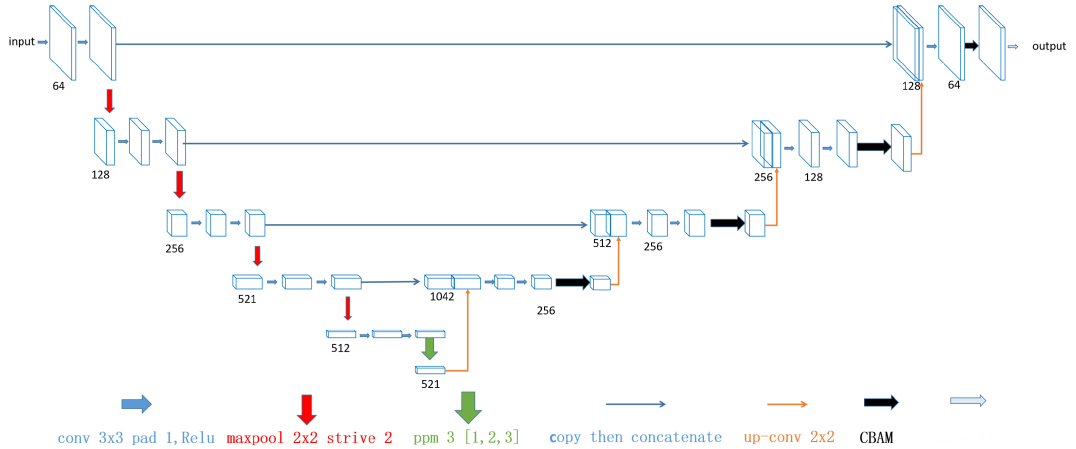


FIGURE 2. Our network architecture (PPM is improved and used in the last downsampling layer of the network and CBAM is modified and used in each upsampling layer of the network).

has been employed for judging the efficacy of the proposed algorithms.

There has also been a great development of deep learning-based models in other image segmentation fields. Liu et al. [38] proposed a network called SimpleNet for image anomaly detection and localization. SimpleNet consists of four components: a pre-trained feature extractor, a shallow feature adapter, a simple anomaly feature generator, and a binary anomaly discriminator. Network structure relies on the intuition that altering pre-trained features to be target-oriented can alleviate domain bias. Additionally, generating synthetic anomalies within the feature space is more efficient. It has been found that a straightforward discriminator is more practical and efficient. SimpleNet outperforms previous methods in terms of accuracy and efficiency on the MVTec AD benchmark. It also shows improvements in performance on the One-Class Novelty Detection task.

Compared to existing work, our network is model based and resolves the problem of insufficient segmentation ability of U-Net. It is designed for medical image segmentation that combines improved PPM and CBAM modules with new training methods.

III. PROPOSED NETWORK

In this section, we provide a comprehensive explanation of the proposed network (i.e., PCU-Net). Table 1 lists the abbreviations used in this paper. As depicted in Fig. 2, this network is composed of an encoder and a decoder. All the constituent modules are thoroughly explained in order.

A. ENCODER

The encoder structure is a popular model structure within the field of deep learning. It is designed to extract complex feature representations from input data. It is frequently employed in image processing, audio, text processing, and other areas. In this paper, we incorporate a convolutional neural network-based encoder to capture the local features of an input image. As previously referenced in [15], the structure of our encoder consists of four primary modules, each of which includes two repetitions of a 3×3 convolutional layer. The convolutional operation is applied with a step size of 1 and the output feature map size is maintained through a padding of 1, facilitating a seamless connection with the decoder section.

TABLE 1. The abbreviations used in this paper.

abbreviation	full name
PPM	Pyramid Pooling Module
CBAM	Convolutional Block Attention Module
MRI	Magnetic resonance imaging
Dice	Dice Coefficient
IoU	Intersection over Union
MAE	Mean Absolute Error
BFscore	Boundary F1 Score

Before performing the convolution operation, we first need to adjust the distribution of the input data by performing batch normalization. This ensures that the data distribution remains relatively consistent before and after convolution, which is beneficial for the training process and helps prevent the data from overfitting. Specifically, batch normalization adjusts the mean and variance of the input data to make the distribution more uniform, thereby improving the effectiveness of the subsequent convolution operation. Additionally, batch normalization can effectively address the problem of low efficiency gradient descent due to gradient disappearance or dense gradient change. This ensures that the algorithm can optimize and learn effectively from the data. Our encoder structure is based on the typical convolutional neural network architecture. The repeated use of convolutional and pooling operations allows us to layer by layer extract the local features of the input image, which are then fed to the decoder part to perform higher level feature reduction and reconstruction. This process generates high-level image features and improve the accuracy of the generated image.

In addition, each submodule employs the 2×2 Max pooling operation for downsampling to reduce the spatial size of

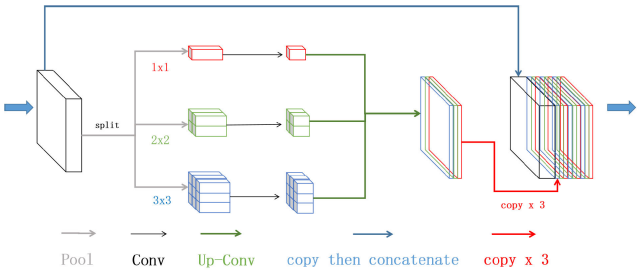


FIGURE 3. Improved PPM in the proposed network.

the feature map and change the number of feature channels, thereby extracting richer feature information. It helps to improve the overall efficiency and accuracy of the network. This method aims to extract specific features that are most relevant and useful to the task at hand, while reducing computational complexity. It is particularly effective in ensuring the stability of the network and its ability to more accurately process and interpret complex, diverse, and constantly changing data. Specifically, the number of feature channels in each sub-module is doubled (except the last one) to accommodate higher level feature expression requirements. By following this strategy, the network is able to progressively learn and represent increasingly complex features at each layer, resulting in enhanced performance across a wide range of tasks.

In the final step of our encoder architecture, a module named PPM is utilized to further extract global feature information from the input image. The detailed structure of this module can be seen in Fig. 3. This module employs a pyramid-shaped multi-scale pooling strategy, which is a proven technique to enhance the global context of the input image, thereby improving the segmentation effect. The algorithm utilized in this module is based on the well-known Atrous Spatial Pyramid Pooling (ASPP) method [39], which has been shown to preserve the global context better than simple pooling strategies. The general concept behind this module is to extract global feature information from different scales of the input image while maintaining local features that are crucial for object detection and segmentation. This is achieved by employing a pyramid-shaped multi-scale pooling strategy, which involves three different pooling operations, each with a different window size. The first pooling operation involves a 1×1 window size, the second pooling operation has a 2×2 window size, and the third pooling operation utilizes a 3×3 window size. Each pooling operation is applied to a distinct layer of the input image, thereby capturing global feature information from different scales. The output of the pyramid pooling module is then passed to the decoder network for subsequent processing steps. In summary, the pyramid pooling module is an integral part of our encoder architecture that enhances the global context of the input image and ultimately improves the overall segmentation performance.

To effectively and simultaneously utilize the different features from different scales of an image, we create a three-pool connection method. This method is designed to connect the pooled results from three image scales. Specifically, we pool the results from the maximum, medium, and minimum scales to capture the diversity of features within an image. With this advanced pooling method, the model's feature extraction ability can be significantly improved, and its segmentation performance is enhanced correspondingly. Therefore, the pyramid pool module plays a crucial role in the model's encoder structure as it is responsible for extracting and emphasizing the global feature information in the input image, ultimately boosting the segmentation ability of our model. The output size of PPM is calculated using the following formula:

$$C_{out} = C_{in} + (P \times N), \tag{1}$$

where C_{out} is the output channel, C_{in} is the input channel, P is multiple of pooled concatenate, and N is the number of pyramid levels.

For example, the pooling cascade of the PPM used in our network has a multiple of 3, with 3 pyramid layers of size 1×1, 2×2, and 3×3 respectively. The encoder part also has a submodule, which increases the original number of channels from 512 to 521. The purpose of this is to ensure that the number of channels in the feature map equals the number of channels in the output feature map of the encoder part PPM. This helps to obtain improved segmentation results.

Different pyramid levels and the size of each level can be employed for various feature maps. For instance, for our general medical segmentation dataset, compared to the feature fusion process of [22] at four pyramid scales, our three different pyramid scales exhibit superior efficiency in feature fusion and better segmentation ability.

B. DECODER

The decoder architecture of our model comprises four different submodules. First, we perform an upsampling operation to the final feature map layer, which has the lowest resolution. Each subsequent layer in the decoder architecture then applies an upsampling operation to the previous layer, progressively increasing the resolution. The process continues until the final feature layer reaches the same resolution as the original image. An innovative technique, skip concatenate, has also been employed, allowing the model to merge feature maps of the same resolution. This technique is executed in both the encoder and decoder components of the model, enhancing the model's ability to capture contextual information from the input image. In our architecture, we adopt a typical convolutional neural network architecture, each submodule including an upsampling operation and two repeated 3×3 convolutional layers. These components empower our decoder architecture to accurately decode and upsample the input image, thereby enhancing segmentation ability.

In the upsampling operation, a 2×2 transposed convolutional layer is utilized to double the size of the input feature map while halving the number of feature channels. Prior to

the convolution operation, the batch normalization operation is applied to the input data to ensure that the data distribution before and after convolution remains similar, promoting the training process. Additionally, the skip concatenate operation is utilized to fuse feature maps with the same resolution in the encoder section with the upsampling results in the decoder section, thereby enhancing the diversity and richness of feature expression.

Our network incorporates the feature mapping of the encoder and decoder to create a new input feature mapping, which is then employed to enhance the spatial and context information of the input image. This new feature mapping incorporates comprehensive information for subsequent processing and segmentation. We merge the feature maps from the encoding and decoding parts along the channel dimension to obtain a more comprehensive context of the input image, thereby improving the segmentation accuracy. Furthermore, the encoding part employs a 3×3 convolution layer with a step of 1 and a filling operation of 1, ensuring the input image remains unchanged, which is particularly significant for jump cascading. This guarantees that the network can transition smoothly between the encoding and decoding parts. Additionally, we use a 2×2 upsampling operation to expand the input feature mapping channel twice, which is essential for higher-level feature representation and further improves the segmentation accuracy. Ultimately, the decoder structure uses a typical convolutional neural network structure, enabling the network to effectively restore the segmentation results layer by layer through upsampling and jump connection operations. This method enables us to achieve stronger medical image segmentation ability.

After each upsampling, CBAM is improved and used. Given an intermediate feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ as input, CBAM sequentially infers a 1D channel attention map $\mathbf{M_c} \in \mathbb{R}^{C\times 1\times 1}$ and a 2D spatial attention map $\mathbf{M_s} \in \mathbb{R}^{1\times H\times W}$. The overall attention process can be summarized as:

$$\begin{aligned} \mathbf{F}' &= \mathbf{M_c}(\mathbf{F}) \otimes \mathbf{F}, \\ \mathbf{F}'' &= \mathbf{M_s}(\mathbf{F}') \otimes \mathbf{F}', \end{aligned} \tag{2}$$

where $\otimes$ denotes element-wise multiplication. During multiplication, the attention values are broadcast (copied) accordingly: channel attention values are broadcasted along the spatial dimension, and vice versa. $\mathbf{F}''$ is the final refined output. CBAM combines the channel attention mechanism [40] and the spatial attention mechanism [41], so that the model can adaptively select the method of feature fusion at different scales. This improves the accuracy and robustness of segmentation.

In the last layer of the decoder, 1×1 convolutions are used to map each 64 component feature vector to the desired number of categories.

C. RGB TRAINING

In the semantic segmentation task, we employ a unique training mode similar to [42], the RGB training, as opposed to traditional methods. This training mode exploits the information within the red, green, and blue channels of the training sample to train the network parameters separately, as shown in Fig.4. In particular, we calculate the respective loss values of the three channels separately during the training process. This approach helps our algorithm to accurately discern differences between channels, providing more accurate segmentations. We subsequently take the average of the three channel loss values to obtain the overall loss value. The overall loss value denotes the quality of the segmentation output and a lower value signifies higher accuracy. The network with the minimum overall loss value is then selected as the final training results.

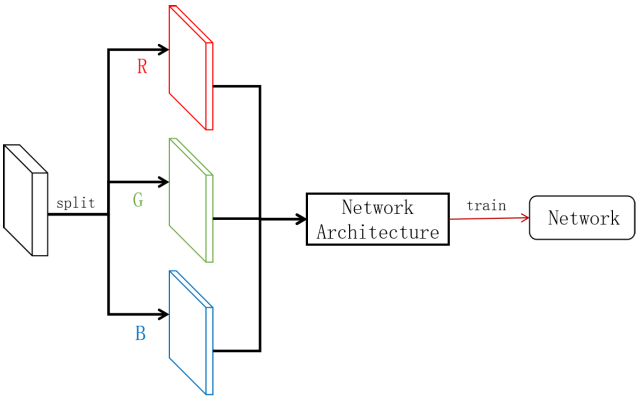


FIGURE 4. RGB training.

Compared with traditional training methods, the RGB training method can notably enhance the performance of semantic segmentation. This conclusion can be drawn by scrutinizing the specific benefits of the RGB training. First, the aberrant region of RGB color image typically possesses a certain color channel or several color channel information that is substantially disparate from the normal region. Through the RGB training, we can successfully capture and leverage this aberrant region information, which is instrumental in augmenting the accuracy of segmentation. Second, the RGB training can extract the features of red, green, and blue channels respectively, and can utilize the information of these three color channels to more precisely delineate the shape and edge of the segmented object. Third, the conventional training methods tend to amalgamate the information of three colors into the channel training, which fails to fully utilize the comprehensive information proffered by RGB three channel images. Therefore, by employing RGB three channels to dissect images, we can procure richer visual features, thereby enhancing the segmentation ability. This method can aid in segmenting the object and background more precisely, which is conducive to the segmentation procedure. Generally, the utilization of RGB three channel images can provide richer visual data. The RGB training method can learn the features of the red, green, and blue channels separately, and then amalgamate them. This can simulate the ability of the human eye to recognize the shape

of objects based on color, thereby enhancing the ability of semantic segmentation.

Algorithm 1 PCU-Net Training Pseudo-Code, Pytorch-Like

Input: F (This is input figure)
L(This is input label)
Output: N (This is save network)

```
1: bestLoss = inf
2: r, g, b = split(F)
3: R, G, B = split(L)
4: for r, g, b, R, G, B in trainLoader do
5:     lossR = lossFuction(r, R)
6:     lossG = lossFuction(g, G)
7:     lossB = lossFuction(b, B)
8:     loss = (lossR + lossG + lossB)/3
9:     if loss < bestLoss then
10:        bestLoss = loss
11:        save N
12:    end if
13: end for
```

Algorithm 1 shows the pseudo-code for our RGB training. Line 1 defines a variable that holds the minimum loss value. Lines 2-3 split the sample and the marked red, green, and blue channels, respectively. Line 4 starts to calculate the loss value in a cycle, and lines 5-8 calculate the loss value of the three channels and calculate the average value. Lines 9-10 hold the network model with the smallest loss. The end of the loop completes the training. The advantage of this training method is that it can make better use of the color information in the color image, improve the robustness and generalization ability of the model, and reduce the risk of overfitting.

When the network is in prediction mode, the RGB channels of the input predicted image are segmented and input into the network for prediction. Then, the three predicted images are merged into the final segmentation effect image, as shown in Fig.5. Algorithm 2 shows the pseudo-code for our RGB predicting. line 1 splits the red, green and blue channels of the sample, line 3-5 predicts the results of the three channels respectively, line 7-12 converts the prediction results to the corresponding pixel values, and line 13 merges the results of the three channels to obtain the segmentation map.

As shown in Fig. 6, the results of RGB three channel segmentation for some samples are not the same. For different samples, the focus of each channel is different. Separating the three channels and then fusing them can grasp the details of all aspects and get better segmentation results.

IV. EXPERIMENTS

Our experiments are conducted on an NVIDIA Tesla V100 32G, with 100 epochs of training, batch size of 1, and learning rate of 0.00001. The weight decay index of the RMSprop algorithm used is 1×10^{-8}, and the momentum is 0.9.

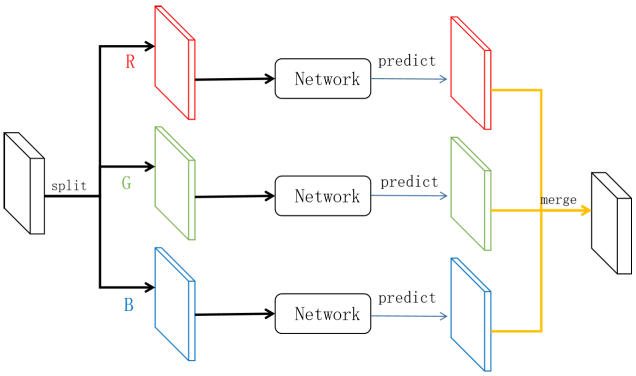


FIGURE 5. RGB predict.

Algorithm 2 PCU-Net Predicting Pseudo-Code, Pytorch-Like

Input: F (This is input figure)
Output: S (This is output segmentation result)

```
1: r, g, b = split(F)
2: //Predict
3: pred-r = N(r)
4: pred-g = N(g)
5: pred-b = N(b)
6: //Converting to pixels
7: pred-r[pred-r >= 0.5] = 255
8: pred-r[pred-r < 0.5] = 0
9: pred-g[pred-g >= 0.5] = 255
10: pred-g[pred-g < 0.5] = 0
11: pred-b[pred-b >= 0.5] = 255
12: pred-b[pred-b < 0.5] = 0
13: S = merge(pred-r,pred-g,pred-b)
```

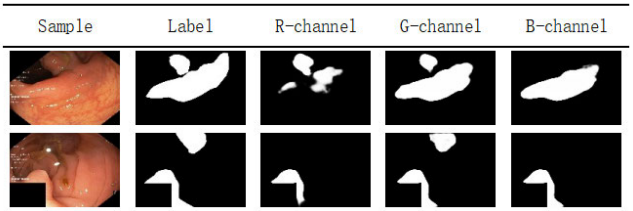


FIGURE 6. Segmentation results of RGB three channels.

A. DATASETS

The dataset mainly used in this paper is Kvasir-SEG. Besides, ISIC 2018 Task 1 dataset [43], [44] and COVID-19 Radiography Database [45], [46] are used to test the segmentation effect of different lesion areas and tissue organs.

Kvasir-SEG is a dataset for colorectal polyp segmentation, developed jointly by the Norwegian University and the Norwegian Institute for Health Research. The dataset contains images, annotations, and case information to help medical professionals and researchers in the field of computer vision to study and apply polyp segmentation. The dataset contains 1000 images of colonoscopic examinations, which are used as the training set. The resolution size of the images in the

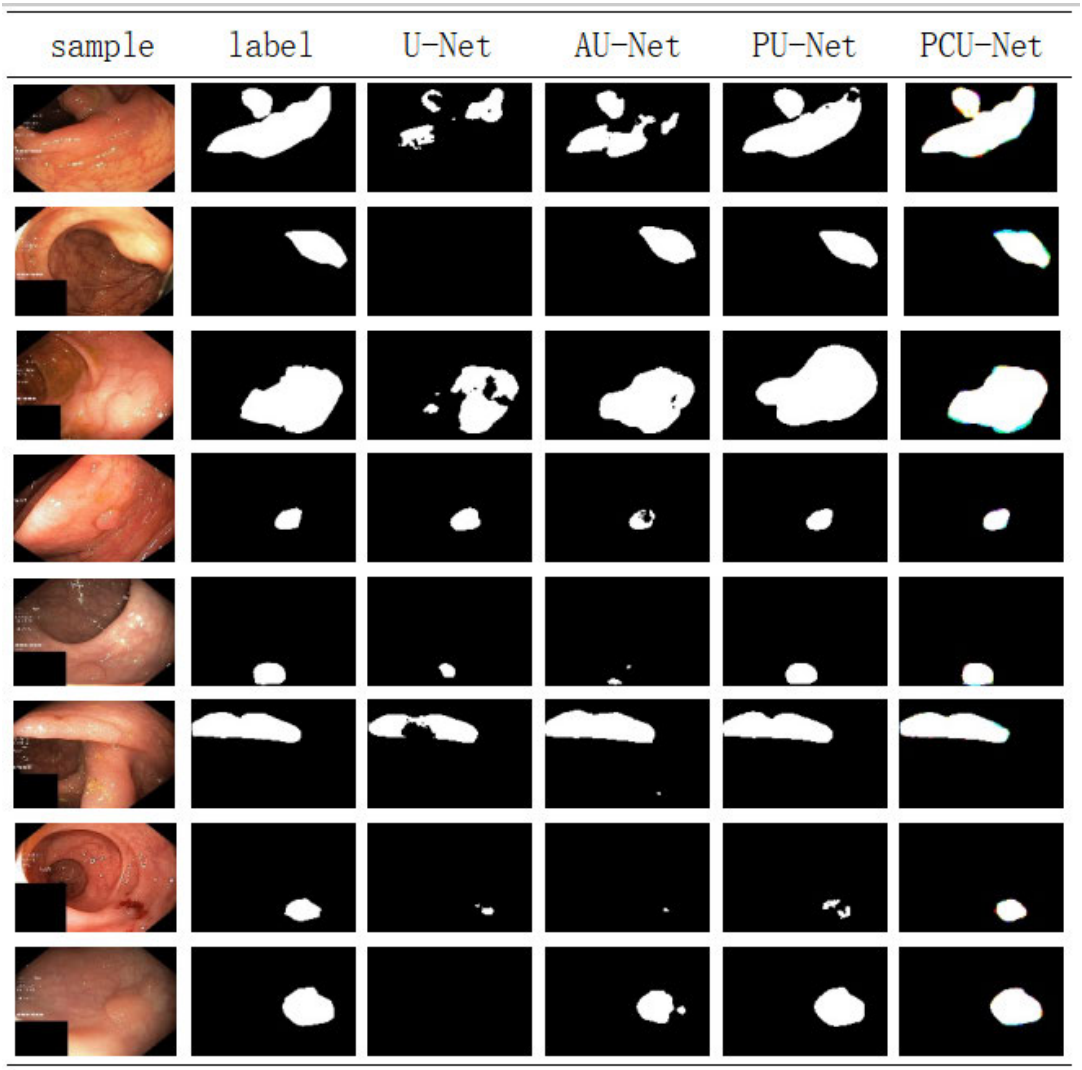


FIGURE 7. Segmentation results of our network and other networks on Kvasir-SEG (AU-Net is U-Net with CBAM and PU-Net is U-Net with PPM).

dataset is different, and the resolution size is 150 × 100 after unified processing for the convenience of the experiment. Each image has a corresponding binary mask, which is used to annotate whether each pixel belongs to the polyp region or not. The main application of this dataset is the automatic detection and segmentation of colorectal polyps to help medical professionals diagnose and treat patients more accurately. In addition, the dataset can also be used for research in the field of computer vision, such as image segmentation and convolutional neural networks.

The ISIC 2018 dataset was published by the International Skin Imaging Collaboration (ISIC) as a large-scale dataset of dermoscopy images. It contains a total of 2594 images as the training set, and 100 and 1000 images as the validation set and test set, respectively. This dataset is used for research on automatic detection of skin diseases based on medical images. The COVID-19 Radiograph Database was created by a team of researchers from Qatar University in Doha, Qatar and Dhaka University in Bangladesh, along with collaborators and doctors from Pakistan and Malaysia. This database contains the chest X-ray image database of COVID-19 positive cases, as well as normal and viral pneumonia images.

B. EVALUATION METRICS

In this paper, there are four metrics used to evaluate the segmentation performance: Dice coefficient, intersection over Union (IoU) [47], mean absolute error (MAE), and mean Boundary F1 Score (BFscore):

1) DICE COEFFICIENT

It is a measure of the accuracy of binary image segmentation, which calculates the proportion of overlap between the predicted segmentation results and the true label.

2) IOU COEFFICIENT

It is another widely used image segmentation performance metric, which calculates the ratio between the intersection

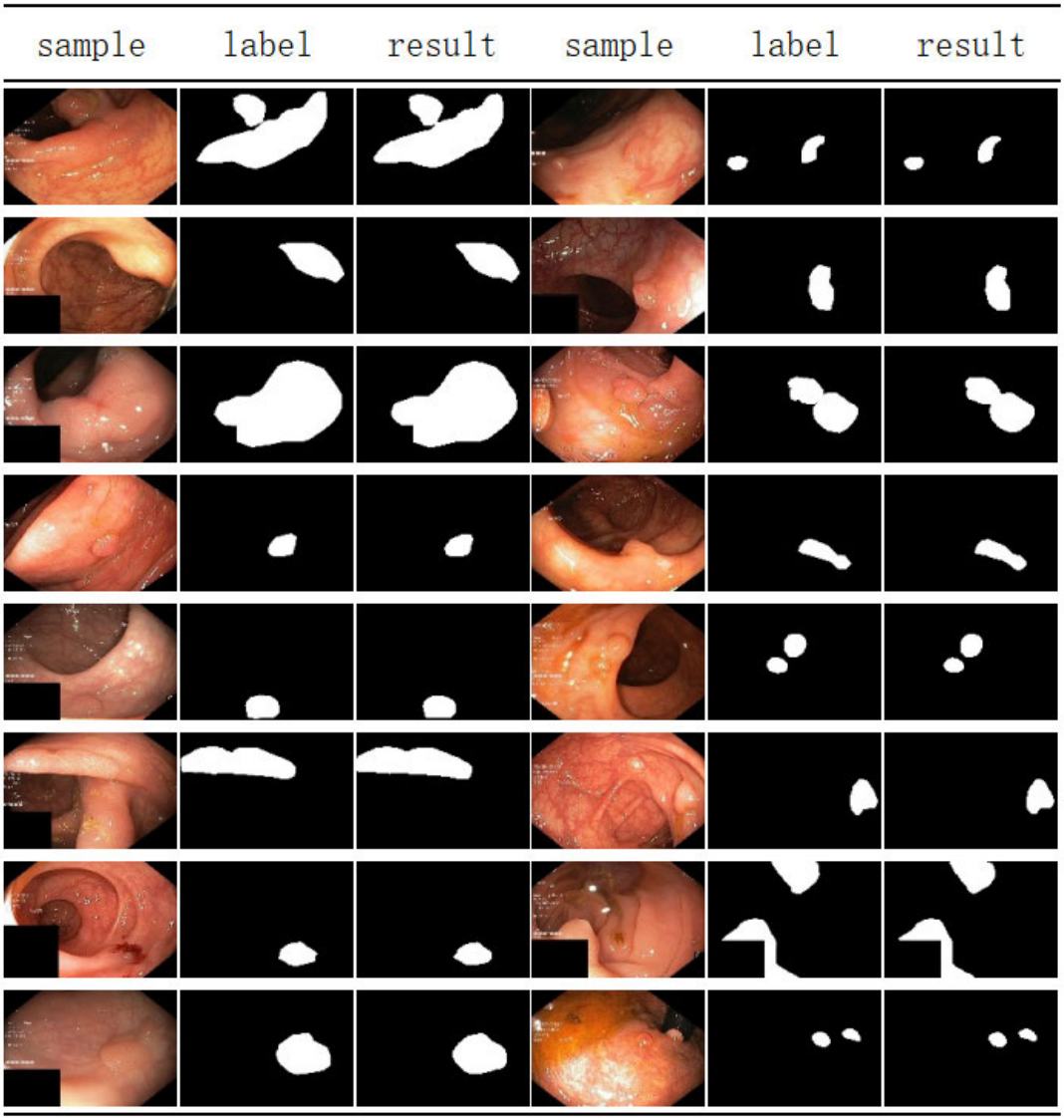


FIGURE 8. Segmentation results on Kvasir-SEG.

and union of the predicted segmentation results and the true label.

3) MAE COEFFICIENT

It is a measure of image segmentation error, which calculates the average of the absolute value of the error at each pixel between the predicted segmentation results and the true label.

4) BFSCORE COEFFICIENT

It is the harmonic mean of precision and recall, striking a balance between the two. When both accuracy and recall are high, the BFscore value will also be higher.

C. LOSS FUNCTION AND OPTIMIZER

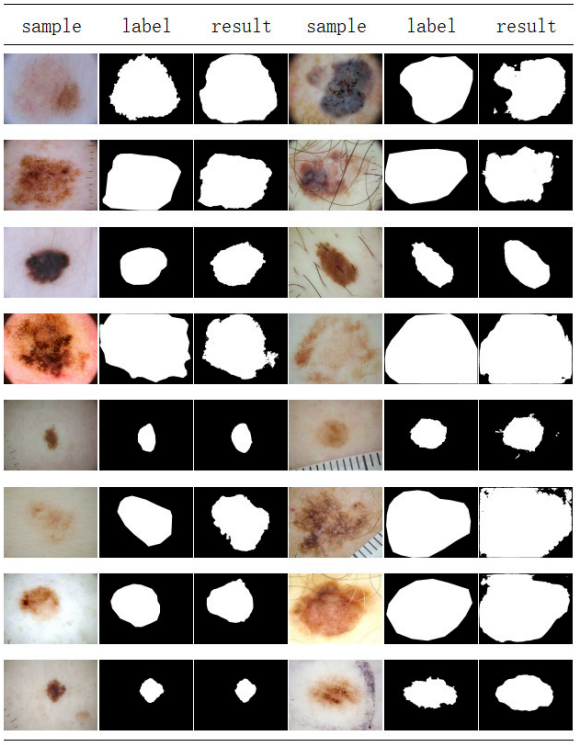
In this paper, the binary cross-entropy loss function is used [48]. It is one of the commonly used loss functions in binary classification problems. However, PyTorch provides a more convenient loss function called BCEWithLogitsLoss because it combines sigmoid with BCELoss and needs to transform the output of the model before calculating the cross-entropy loss. The BCEWithLogitsLoss formula decomposes as follows: Suppose there are N batches and each batch predicts n labels, then:

$$\begin{aligned} Loss = \{l_1, \ldots, l_N\}, l_n = -[y_n \cdot \log(\sigma(x_n)) \\ + (1 - y_n) \cdot \log(1 - \sigma(x_n))], \end{aligned} \tag{3}$$

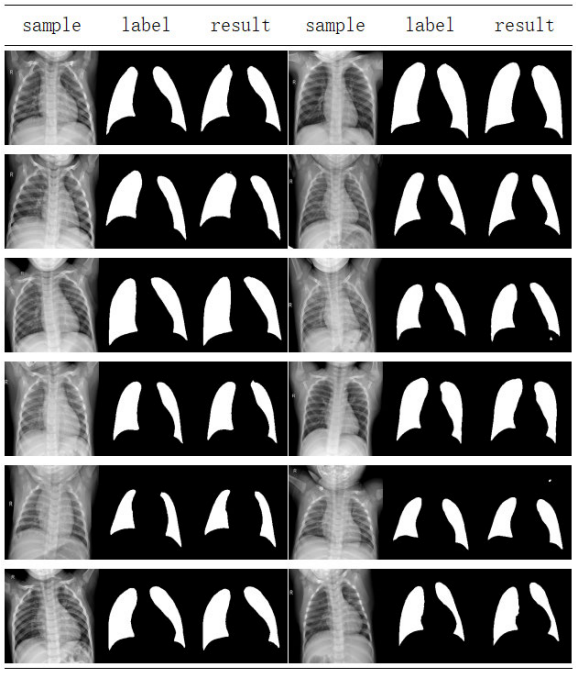
where $\sigma(x_n)$ is the Sigmoid function that maps x to the interval (0, 1):

$$\sigma(x_n) = \frac{1}{1 + \exp(-x)}. \tag{4}$$

The optimizer and loss function are two essential concepts in machine learning. They serve distinct roles, yet are interdependent. The role of the optimizer is to optimize the



(a) Segmentation results of our network on ISIC 2018 Task 1.



(b) Segmentation results of our network on COVID-19 Radiography Database.

FIGURE 9. Segmentation results on ISIC 2018 Task 1 and COVID-19 Radiography Database.

model's parameters based on the gradients calculated by the loss function, with the objective of achieving a better fit of the training data. The function of the loss function, on the other hand, is to calculate the difference between the model output and the label, thereby serving as an input to the optimizer to guide the model's training. In other words, the optimizer and loss function act on different aspects of the model but are interconnected, making them essential to the successful optimization and training of a machine learning model.

The optimizer used in this paper is RMSprop [49] optimizer. The main role of RMSprop is to update network parameters according to the gradient. Rmsprop is an adaptive learning rate method. It can adaptively adjust the learning rate according to the size of the gradient in different parameter update steps. It is mainly to solve the problem that the learning rate decays too fast in the Adagrad algorithm. Compared with Adagrad, a moving average is used when calculating the squared gradient. It makes the decay of the learning rate more gentle, smooth the change of the gradient, reduce the shock of the gradient, and further improve the stability of the model.

D. SEGMENTATION RESULTS

The segmentation results on Kvasir-SEG are shown in Fig.7. We can find that U-Net has a good segmentation effect on some simple samples, but it has a poor effect on complex segmentation tasks, and even fails to segment some non-obvious samples. The AU-Net has a certain improvement, but the effect is still not good because of some wrong segmentation phenomena. Our network segmentation results are significantly better than other networks. Especially for samples that cannot be segmented by other networks, we can not only complete the segmentation task well, but also achieve good results. Fig. 8 shows more segmentation results of our network. In the case of partial occlusion of the sample, our network still performs well for the details that are difficult to segment.

Table 2 shows that the Dice coefficient of our network is 0.85, the IoU of our network is 0.74, the MAE of our network is 0.0367%, and the BFscore of our network is 0.87511. Notably, U-Net achieves Dice coefficient of 0.66, IoU of 0.54, MAE of 0.06%, and BFscore of 0.79499. Our evaluation metrics exhibit enhancements of 28%, 38%, 39%, and 10%, correspondingly. Furthermore, AU-Net demonstrates Dice coefficient of 0.79, IoU of 0.67, MAE of 0.042%, and BFscore of 0.82972. Our evaluation metrics exhibit improvements of 8%, 11%, 13%, and 5% respectively.

The segmentation results on ISIC 2018 Task 1 are shown in Fig. 9a. Our network also has good results in lesion segmentation on the skin surface. However, because the abnormal boundaries of skin lesions are fuzzy, the segmentation results are not detailed enough in the boundaries of abnormal regions. Even so, it still has a good effect on assisting skin lesion segmentation.

The segmentation results on COVID-19 Radiography Database are shown in Fig. 9b. For the experiments in tissue

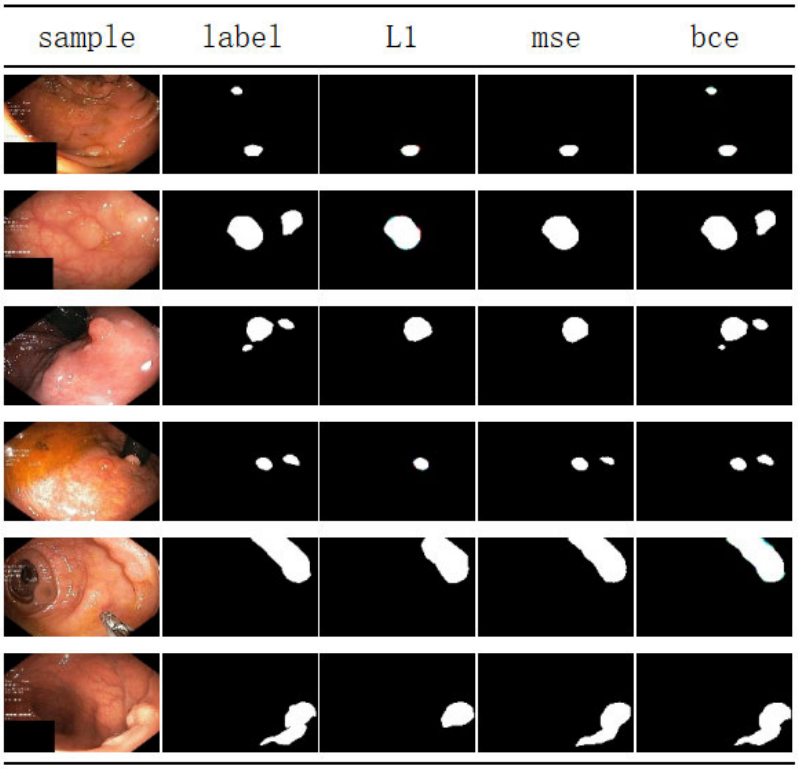


FIGURE 10. Comparison of segmentation results for different loss function on Kvasir-SEG.

TABLE 2. Metrics comparison of segmentation results between our network and other networks on Kvasir-SEG.

Network	Dice	IoU	MAE	BFscore
U-Net	0.66289	0.53715	0.0600%	0.79499
AU-Net	0.78699	0.66664	0.0421%	0.82972
PU-Net	0.84341	0.73319	0.0371%	0.86553
U-Net++	0.82545	0.70639	0.0391%	0.85751
U^2-Net	0.83046	0.71493	0.0368%	0.86488
ResU-Net	0.82217	0.70308	0.0392%	0.85845
ResU-Net++	0.82843	0.71068	0.0382%	0.86448
U-NeXt	0.81290	0.69138	0.0404%	0.85391
PCU-Net	**0.84977**	**0.74208**	**0.0367%**	**0.87511**

and organ segmentation, we divide the normal classification data under the dataset into a training set and a testing set in a 4:1 ratio. Fig. 9b shows that the proposed network also performs well in organ segmentation.

We have the following explanation on how our improvement can enhance the segmentation ability of the network. First, the improved PPM includes pooling layers of different specifications, allowing the network to not only better segment overall anomalies in the samples, but also to segment edge details more clearly. Second, the modified CBAM enhances the network's segmentation ability in both channel and spatial aspects, allowing the network to achieve good overall and detailed results [50].

E. ABLATION STUDY

1) PU-NET AND AU-NET

The network proposed in this paper has two structural improvements compared with U-Net. In this part, the two improvements are separated and experimented separately for ablation. We refer to the U-Net with PPM added in the downsampling part as PU-Net, and to the U-Net with CBAM added in the upsampling part as AU-Net. First, Table 2 shows the comparison between our proposed network and these two networks in terms of metrics. Compared with AU-Net, PCU-Net can improve Dice, IoU, MAE, and BFscore by 8%, 11%, 13%, and 5% respectively. Second, from Fig. 7, compared with PU-Net, although PCU-Net does not show significant improvement in metrics, it has better segmentation results than PU-Net that has obvious shortcomings in edges.

2) LOSS FUNCTION

In comparison with the widely used MSE loss and a simple truncated L1 loss function, the proposed loss function presented in Section IV-C is compared and assessed in Table 3. Our chosen cross-entropy loss function significantly outperforms both L1 loss and MSE loss. Compared with L1 loss, cross-entropy loss can improve Dice, IoU, MAE, and BFscore by 1%, 1%, 5%, and 0.1% respectively. Compared with MSE loss, cross-entropy loss can improve Dice, IoU, MAE, and BFscore by 5%, 7%, 33%, and 6% respectively. While Fig. 11 reveals that the convergence time of the three loss functions is not markedly different, Fig. 10 illustrates that L1 loss and MSE loss frequently suffer from instances of false segmentation and missing segmentation. For example, in samples 3 and 5, L1 loss and MSE loss fail to accurately segment the fine details. However, the proposed loss function precisely segments the results, demonstrating that it is a more

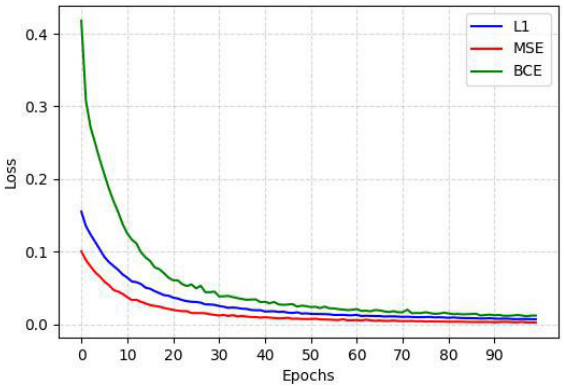


FIGURE 11. Training process of PCU-Net with different loss function on Kvasir-SEG.

TABLE 3. Metrics comparison with the segmentation results of the loss function on Kvasir-SEG.

Loss Function	Dice	IoU	MAE	BFscore
L1	0.84094	0.73041	0.0384%	0.87386
mse	0.80736	0.69091	0.0555%	0.82857
bce (ours)	**0.84977**	**0.74208**	**0.0367%**	**0.87511**

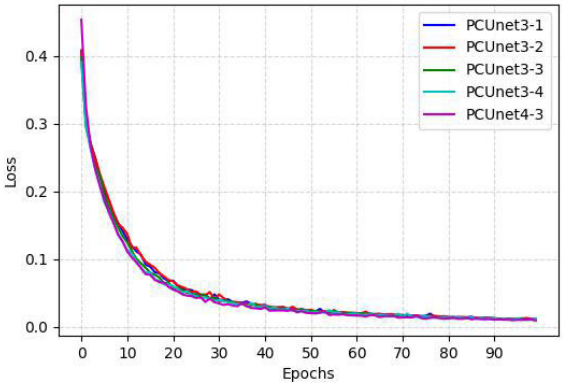


FIGURE 12. Training process of PCU-Net with different configurations on Kvasir-SEG.

effective choice for image segmentation. After extensive experimental comparison, we select the cross-entropy loss function as our primary loss function.

3) CONFIGURATION OF PPM

We experimentally compare several different configurations for PPM, including the number of pyramid levels, the number of pooling connection multiples, and the size of each level. As shown in Table 4, despite minimal differences among each configuration, even the segmentation results may not be distinguishable by the naked eye, it is evident from the metrics that the three bin sizes of pooling concatenation (1×1, 2×2, 3×3), particularly those selected by us, demonstrate superior efficiency and effectiveness. Although a study by Zhao et al [22] found that a 4 bin size (1×1, 2×2, 3×3, 6×6) had better performance in terms of metrics, this advantage was outweighed by the additional complexity of the training procedure. To evaluate these experiments, we utilize the three bin sizes for PPM. Fig. 12 illustrates the convergence process of PCU-Net training for each configuration. It can be seen that

TABLE 4. Experimental comparison of PPM configurations and mode of training(-3-1 is 3 bin sizes and 1 time concatenate) on Kvasir-SEG.

Network	Dice	IoU	MAE	BFscore
PCU-Net-3-1	0.84443	0.73408	0.0369%	0.87188
PCU-Net-3-2	0.84858	0.74019	0.0365%	0.85961
PCU-Net-3-4	0.84924	0.74105	0.0372%	0.87032
PCU-Net-4-3	**0.85004**	**0.74258**	**0.0363%**	**0.88128**
PCU-Net-3-3-noRGB	0.83405	0.71892	0.0373%	0.86014
PCU-Net-3-3 (ours)	**0.84977**	**0.74208**	**0.0367%**	**0.87511**

TABLE 5. Comparison of training time of the network on Kvasir-SEG.

Network	Time(seconds)
U^2-Net	8510
U-Net++	4249
ResU-Net	5315
ResU-Net++	4609
Ours	**4199**

although PCU-Net-4-3 has the fastest convergence due to its complexity. From a comprehensive perspective of time cost and performance, the properly configuration is PCU-Net-3-3. In conclusion, we select the configuration of PCU-Net-3-3 with three bin sizes and three layers of pooling connection.

4) TRAINING MODE

In order to comprehensively evaluate the performance and advantages of our RGB training method, we compare it with the traditional segmentation methods that are currently in use. These results are summarized and presented in Table 4. Based on the experimental data, our RGB training method has demonstrated significant improvements in segmentation accuracy and overall segmentation effectiveness. This can be attributed to the unique ability of our method to extract features from the three different channels, leading to improved segmentation object delineation and definition.

5) TRAINING TIME

We compare the training time and segmentation results of U-Net ++, U^2-Net, ResU-Net, and ResU-Net++ on the Kvasir-SEG dataset. The training time of our network and others networks is shown in Table 5. In the same training epochs of the 100 cases, the segmentation results of our network and others are shown in Fig.13. Table 2 shows that our network is also better in terms of metrics. In the case of the same number of training steps, U-Net++ has some incorrect segmentation. Due to its complex structure, U^2-Net increases training time and suffers from incomplete training of some detail segmentation abilities. ResU-Net and ResU-Net++ also have similar issues.

V. DISCUSSION

A. APPLICATION OF THE PROPOSED NETWORK

The above experiments demonstrate that the proposed network has better results than some traditional U-shaped networks on RGB images. However, because the RGB training is only applicable to RGB images, it cannot be applied to non RGB medical images such as CT, MRI and

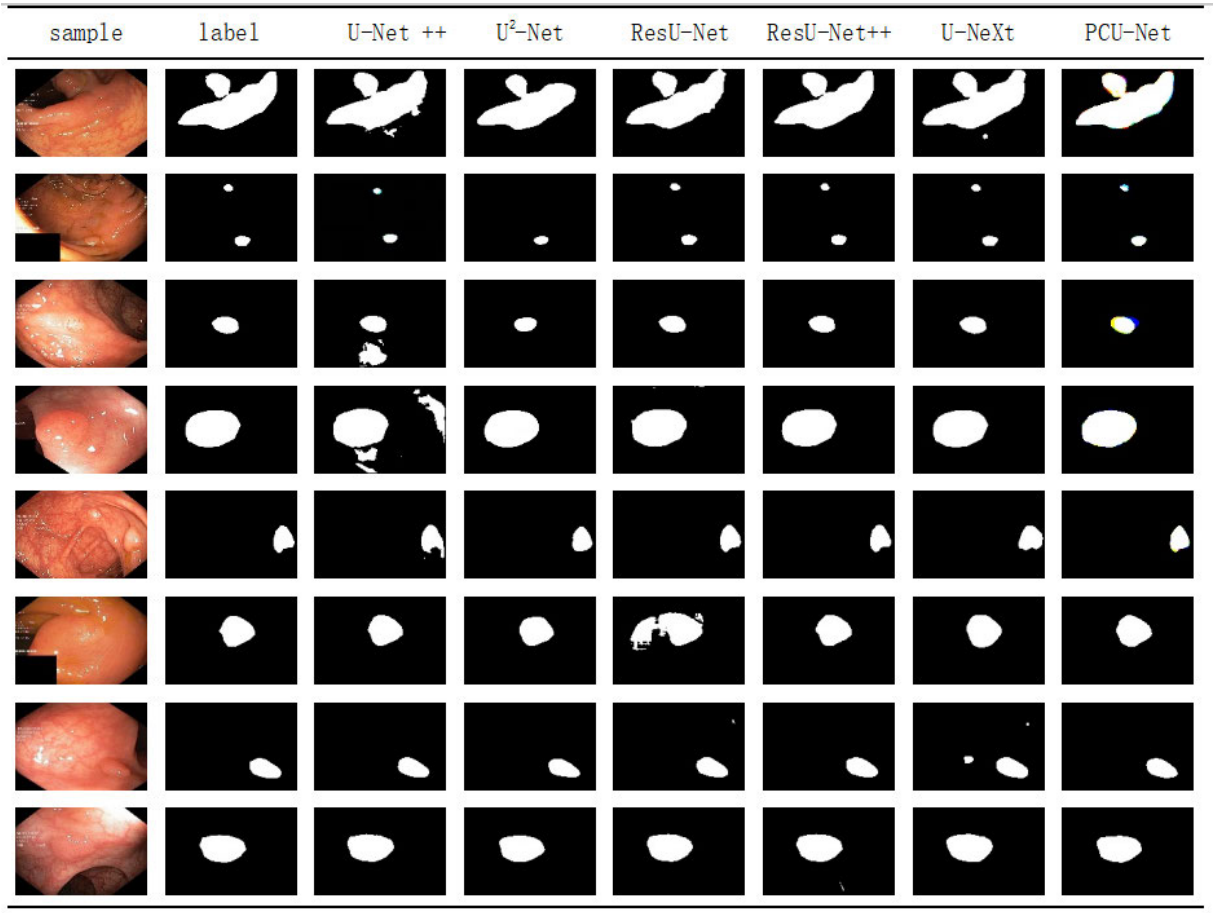


FIGURE 13. The contrast of network segmentation results on Kvasir-SEG.

X-ray. Then using the proposed network without the RGB training, by comparing Table 2 and Table 4, we can see that the PCU-Net-3-3-noRGB network is better than other networks in Dice, IoU, MAE, and BFscore. Moreover, Fig. 9b shows that the proposed network also performs well in segmenting X-ray images. Therefore, we believe that even if the RGB training is not used in the PCU-Net, it also has a good segmentation effect in the abnormal segmentation of the medical images of other modalities.

B. POSSIBLE DIRECTIONS FOR IMPROVEMENT

Compared with transformer and other lightweight CNNs, U-Net has many extensions and applications in the field of medical image segmentation due to its simple structure. We also consider using model pruning technique to make the network more efficient. However, U-Net itself is relatively simple, and model pruning may not significantly improve the network. Table 6 shows the parameter performance of our proposed network during experiments on the Kvasir-SEG dataset. However, after undergoing two pruning algorithms: L1Unstructured and RandomUnstructured, the parameter situation does not change. Nevertheless, it is still worth studying other pruning techniques to achieve the effect of simplifying the model.

TABLE 6. Parameter situation of PCU-Net.

Total params:	13,631,658
Trainable params:	13,631,658
Non-trainable params:	0
Input size (MB):	0.06
Forward/backward pass size (MB):	215.73
Params size (MB):	52.00
Estimated Total Size (MB):	267.79

VI. CONCLUSION

This paper proposes an enhanced U-network by combining PPM and CBAM for medical image segmentation. The proposed network compensates for the insufficient segmentation ability of the basic U-Net architecture. PPM is improved and used in the downsampling part of the encoder to extract multi-scale context information from the input feature map. It extracts features from different pyramid levels, which provides a multi view representation of input. And CBAM is modified and used in the upsampling part of the decoder. It helps the network focus on the important spatial regions and channel features of the input image.

We perform experiments on the Kvasir-SEG dataset. Experimental results show that compared with U-Net and AU-Net, our network improves Dice by 28% and 8%, IoU by 38% and 11%, MAE by 39% and 11%, BFscore by

10% and 5% respectively. Moreover, the performance of our network on the experimental data set is also better than those of improved versions of U-Net: U-Net++, U^2-Net, ResU-Net, ResU-Net++, and U-NeXt, with shorter training time and stronger segmentation ability under the same training rounds. In addition, the RGB training we use can improve the segmentation ability of the network compared with the ordinary training method.

REFERENCES

[1] M. A. Elaziz, A. A. Ewees, and D. Oliva, "Hyper-heuristic method for multilevel thresholding image segmentation," *Expert Syst. Appl.*, vol. 146, May 2020, Art. no. 113201.

[2] L. Chen, J. Gao, A. M. Lopes, Z. Zhang, Z. Chu, and R. Wu, "Adaptive fractional-order genetic-particle swarm optimization Otsu algorithm for image segmentation," *Int. J. Speech Technol.*, vol. 53, no. 22, pp. 26949–26966, Nov. 2023.

[3] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[4] A. Thakur and R. S. Anand, "A local statistics based region growing segmentation method for ultrasound medical images," *Int. J. Med. Health Sci.*, vol. 1, no. 10, pp. 564–569, 2007.

[5] J. Wu, S. Poehlman, M. D. Noseworthy, and M. V. Kamath, "Texture feature based automated seeded region growing in abdominal MRI segmentation," in *Proc. Int. Conf. Biomed. Eng. Informat.*, May 2008, pp. 263–267.

[6] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, and T. Moriarty, "A modified fuzzy *c*-means algorithm for bias field estimation and segmentation of MRI data," *IEEE Trans. Med. Imag.*, vol. 21, no. 3, pp. 193–199, Mar. 2002.

[7] G. Bertasius, J. Shi, and L. Torresani, "DeepEdge: A multi-scale bifurcated deep network for top-down contour detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4380–4389.

[8] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 26th Annu. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Dec. 2012, pp. 1106–1114.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, Y. Bengio and Y. LeCun, Eds., May 2015, pp. 1–14.

[11] 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)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, Munich, Germany, Oct. 2015, pp. 234–241.

[16] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *Proc. Int. Conf. Multimedia Modelling*, Daejeon, South Korea, Jan. 2020, pp. 451–462. [Online]. Available: https://datasets.simula.no/kvasir-seg/

[17] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, in Lecture Notes in Computer Science, vol. 11045, Granada, Spain, Sep. 2018, pp. 3–11.

[18] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.

[19] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[20] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, and H. D. Johansen, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, San Diego, CA, USA, Dec. 2019, pp. 2250–2255.

[21] J. Valanarasu and V. M. Patel, "UNeXt: MLP-based rapid medical image segmentation network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, in Lecture Notes in Computer Science, vol. 13435, Singapore, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., Sep. 2022, pp. 23–33.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[23] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[24] K. Ramesh, G. K. Kumar, K. Swapna, D. Datta, and S. S. Rajest, "A review of medical image segmentation algorithms," *EAI Endorsed Trans. Pervasive Health Technol.*, vol. 7, no. 27, p. e6, 2021.

[25] N. Mahmood, A. Shah, A. Waqas, A. Abubakar, S. Kamran, and S. B. Zaidi, "Image segmentation methods and edge detection: An application to knee joint articular cartilage edge detection," *J. Theor. Appl. Inf. Technol.*, vol. 71, no. 1, pp. 87–96, 2015.

[26] X.-F. Du, J.-S. Wang, and W.-Z. Sun, "UNet retinal blood vessel segmentation algorithm based on improved pyramid pooling method and attention mechanism," *Phys. Med. Biol.*, vol. 66, no. 17, Sep. 2021, Art. no. 175013.

[27] G. Kim, B.-S. Jeoun, S. Yang, J. Kim, S.-J. Lee, and W.-J. Yi, "CAPPU-Net: A convolutional attention network with pyramid pooling for segmentation of middle and inner ear structures in CT images," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2021, p. 5336.

[28] D. Tian, H. Yao, Z. Song, G. Zhan, and Z. Wang, "Improved parts drawing segmentation method based on U-Net," *J. Image Process. Theory Appl.*, vol. 5, no. 1, pp. 52–58, 2022.

[29] A. Beji, A. G. Blaiech, M. Said, A. B. Abdallah, and M. H. Bedoui, "An innovative medical image synthesis based on dual GAN deep neural networks for improved segmentation quality," *Int. J. Speech Technol.*, vol. 53, no. 3, pp. 3381–3397, Feb. 2023.

[30] Z. Dong, Y. He, X. Qi, Y. Chen, H. Shu, J.-L. Coatrieux, G. Yang, and S. Li, "MNet: Rethinking 2D/3D networks for anisotropic medical image segmentation," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Vienna, Austria, Jul. 2022, pp. 870–876.

[31] W. Liu, T. Tian, W. Xu, H. Yang, X. Pan, S. Yan, and L. Wang, "PHTrans: Parallelly aggregating global and local representations for medical image segmentation," in *Proc. 25th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, in Lecture Notes in Computer Science, vol. 13435, Singapore, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., Sep. 2022, pp. 235–244.

[32] M. Cheng, Z. Kong, G. Song, Y. Tian, Y. Liang, and J. Chen, "Learnable oriented-derivative network for polyp segmentation," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Strasbourg, France, Sep. 2021, pp. 720–730.

[33] S. Solanki, U. P. Singh, S. S. Chouhan, and S. Jain, "A systematic analysis of magnetic resonance images and deep learning methods used for diagnosis of brain tumor," *Multimedia Tools Appl.*, vol. 83, no. 8, pp. 23929–23966, Aug. 2023.

[34] S. Solanki, U. P. Singh, S. S. Chouhan, and S. Jain, "Brain tumour detection and classification by using deep learning classifier," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 2s, pp. 279–292, 2023.

[35] S. Solanki, U. P. Singh, and S. S. Chouhan, "Brain tumor classification using ML and DL approaches," in *Proc. IEEE 5th Int. Conf. Cybern., Cognition Mach. Learn. Appl. (ICCCMLA)*, Oct. 2023, pp. 204–208.

[36] R. K. Patel and M. Kashyap, "Automated diagnosis of COVID stages from lung CT images using statistical features in 2-dimensional flexible analytic wavelet transform," *Biocybernetics Biomed. Eng.*, vol. 42, no. 3, pp. 829–841, Jul. 2022.

[37] A. Saxena, S. S. Chouhan, R. M. Aziz, and V. Agarwal, "A comprehensive evaluation of marine predator chaotic algorithm for feature selection of COVID-19," *Evolving Syst.*, vol. 15, no. 4, pp. 1235–1248, Aug. 2024.
[38] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "SimpleNet: A simple network for image anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20402–20411.
[39] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[40] 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.
[41] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 2017–2025.
[42] R. K. Patel and M. Kashyap, "Automated screening of glaucoma stages from retinal fundus images using BPS and LBP based GLCM features," *Int. J. Imag. Syst. Technol.*, vol. 33, no. 1, pp. 246–261, Jan. 2023.
[43] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.
[44] 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, pp. 1–9, Aug. 2018.
[45] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, and M. T. Islam, "Can AI help in screening viral and COVID-19 pneumonia?" *IEEE Access*, vol. 8, pp. 132665–132676, 2020. [Online]. Available: https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database
[46] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier, M. S. Khan, and M. E. H. Chowdhury, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Comput. Biol. Med.*, vol. 132, May 2021, Art. no. 104319.
[47] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[48] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.
[49] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.
[50] Y. Yu and M. Yao, "When convolutional neural networks meet laser-induced breakdown spectroscopy: End-to-end quantitative analysis modeling of ChemCam spectral data for major elements based on ensemble convolutional neural networks," *Remote Sens.*, vol. 15, no. 13, p. 3422, Jul. 2023.

ZHONGMING FU received the Ph.D. degree from the College of Computer Science and Electronic Engineering, Hunan University, China, in 2020. He is currently a Lecturer with the School of Computer Science, University of South China. His research interests include machine learning, deep learning, and parallel distributed processing.

HEJIAN CHEN (Graduate Student Member, IEEE) is currently pursuing the master's degree with the College of Computer Science and Technology, University of South China. His research interests include deep learning and medical image segmentation.

MENGSI HE received the master's degree from the College of Computer Science and Electronic Engineering, Hunan University, China, in 2020. She is currently a Technician with the School of Computer Science, University of South China. Her current research interests include machine learning, deep learning, and parallel distributed processing.

LI LIU received the Ph.D. degree from the Huazhong University of Science and Technology, in 2008. He is currently a Professor with the School of Computer Science, University of South China. He has presided over or mainly participated in seven scientific research projects of the National Natural Science Foundation and Hunan Provincial Department of Education. He has published more than 20 academic papers in important domestic and foreign journals or international conferences, seven of which have been included in SCI\EI\ISTP. His current research interests include digital image processing and embedded systems.