УДК 330.1
doi:10.18720/SPBPU/2/id24-496

Guo Xin [1],
Master's Student;
Galina F. Malykhnia [2],
Professor, Doctor of Technical Sciences

PYRAMID CONVOLUTIONAL ATTENTION FUSION NETWORK IMPROVES U-NET FOR MEDICAL IMAGE SEGMENTATION

[1,2] Peter the Great St.Petersburg Polytechnic University, St. Petersburg, Russia,
[1] go7.s@edu.spbstu.ru, [2] g_f_malychina@mail.ru

Abstract. Pneumonia is one of the leading causes of death in humans due to infection. The construction of Pyramidal Convolutional Attention Fusion Network (PCAF-Net) based on Convolutional Networks for Biomedical Image Segmentation (U-Net) enhances the important features in the original feature maps, reduces the loss of information in the process of feature transfer, and improves the segmentation performance of the network.

Keywords: X-Ray image, medical image processing, pyramid convolution, attention mechanism, spatial attention, channel attention, feature extraction, medical image segmentation.

Го Синь [1],
магистрант;
Малыхина Галина Федоровна [2],
профессор, д-р техн. наук, профессор

ПИРАМИДАЛЬНАЯ КОНВОЛЮЦИОННАЯ СЕТЬ СЛИЯНИЯ ВНИМАНИЯ УЛУЧШАЕТ U-NET ДЛЯ СЕГМЕНТАЦИИ МЕДИЦИНСКИХ ИЗОБРАЖЕНИЙ

[1,2] Россия, Санкт-Петербург,
Санкт-Петербургский политехнический университет Петра Великого,
[1] go7.s@edu.spbstu.ru, [2]g_f_malychina@mail.ru

Аннотация. Пневмония – одна из основных причин смерти людей в результате инфекции. Построение пирамидальной конволюционной сети слияния внимания (PCAF-Net) на основе конволюционных сетей для сегментации биомедицинских изображений (U-Net) позволяет выделить важные признаки в исходных картах признаков, уменьшить потерю информации в процессе передачи признаков и улучшить сегментационные характеристики сети.

Ключевые слова: рентгеновское изображение, обработка медицинских изображений, свертка пирамиды, механизм внимания, пространственное внимание, канальное внимание, извлечение признаков, сегментация медицинских изображений.

Introduction

Pneumonia, a condition in which the lungs become inflamed, mainly affecting the alveoli. Globally, about 450 million people develop pneumonia each year, and about 4 million people die each year as a result [1].

In recent years, machine learning based image segmentation algorithms have been continuously proposed by researchers. Phillips et al [2] used unsupervised FCM algorithm to segment MRI dataset into organizational classifications based on three different weighting parameters. Ray et al [3] proposed an improved K-means algorithm based on an intra-distance measure and an inter-cluster distance measure, allowing the number of classification clusters to be determined automatically. Zheng et al [4] proposed an adaptive K-mean image segmentation method that avoids the interactive input of K-values. Khilkhal et al [5] combined an algorithm based on K-means, thresholding, and morphological operations to solve the problem of tumor segmentation in brain MRI images.

Influenced by FCN, Ranneberger et al [6] proposed U-Net, which is built based on an encoder-decoder structure, where the encoder and decoder are connected and feature fusion is achieved through jump connections, and a small amount of data can be used to learn a model that is robust to edge extraction.

Although the above image segmentation methods have achieved good results, the following two problems still exist:

1) information is damaged when transferring features between the encoder and the decoder;

2) insufficient feature information is extracted when passing the feature map forward in the codec.

To solve the above two problems, we constructed the Pyramid Convolutional Attention Fusion Network (PCAF-Net) based on the U-Net.

1. Dataset

X-ray images are very important in target segmentation, especially in the diagnosis of pneumonia diseases. Pneumonia (Chest X-ray images) Dataset were selected from a retrospective cohort of children aged 1 to 5 years from the Guangzhou Women's and Children's Medical Centre in China. All chest X-ray images were performed as part of the patient's routine clinical care.

Table 1

Pneumonia (Chest X-Ray Images) Dataset meta-information

Dimension	Modal	Anatomical structure	Data volume	File Formats
2D	X-ray	Lung	5,863	JPEG

X-rays are a non-invasive, painless, and side-effect free procedure that can provide information about a patient's internal structure and function. In patients with pneumonia, X-rays can determine the location, size, and extent of the inflammation and can be complementary to treatment.

Fig. 1. Three types of Pneumonia (Chest X-Ray Images)

Viral pneumonia (right image) shows a more diffuse 'interstitial' pattern in both lungs. Bacterial pneumonia (middle image) usually shows focal lobar solidity, in this case the right upper lobe (white arrow). Viral pneumonia (right image) shows a more diffuse interstitial pattern in both lungs.

We divide the dataset at 70 % as training set, 15 % as validation set and remaining 15 % as test set. This division ratio ensures that the model has enough data to learn during training, while model tuning is performed on the validation set to avoid overfitting and the performance of the final model is evaluated on the test set.

2. Mathematical and algorithmic justification of ANN

2.1. Pyramid Convolutional Attention Fusion Network

The overall flow of the Pyramid Convolutional Attention Fusion Network (PCAF-Net) proposed in this study is shown in Fig. 2 and the network structure is shown in Fig. 3 PCAF-Net is built on U-Net as the basic model and mainly consists of four parts: encoder, decoder, skip connection and bottleneck layer.

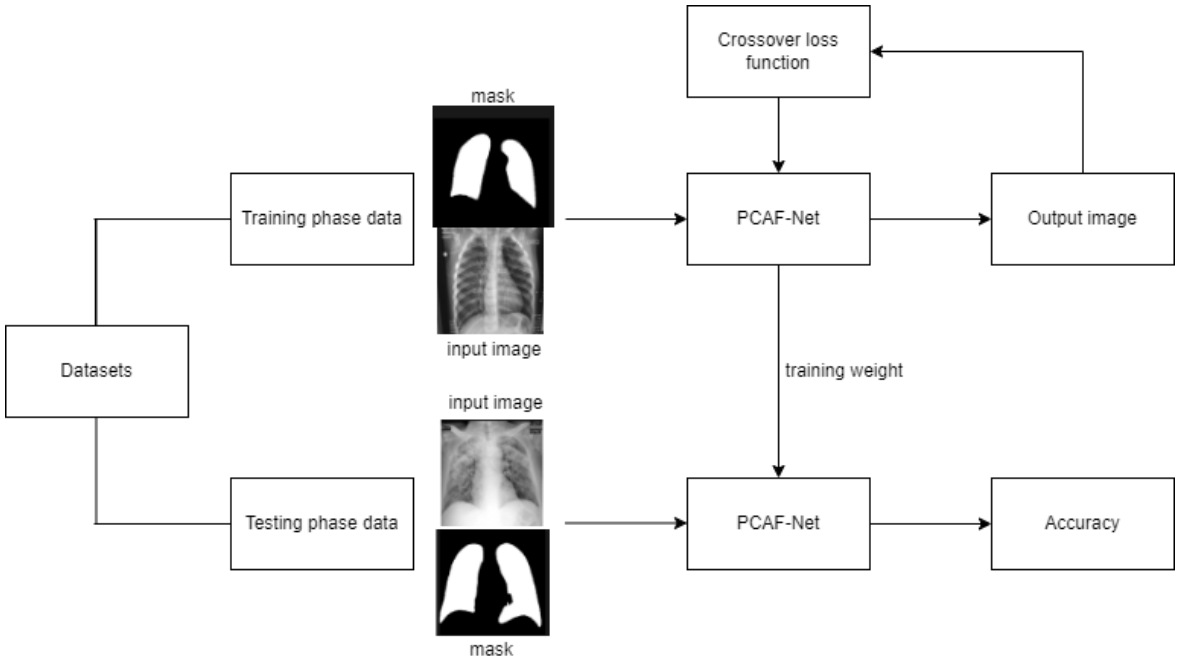

Fig. 2. PCAF-Net Overall Flowchart

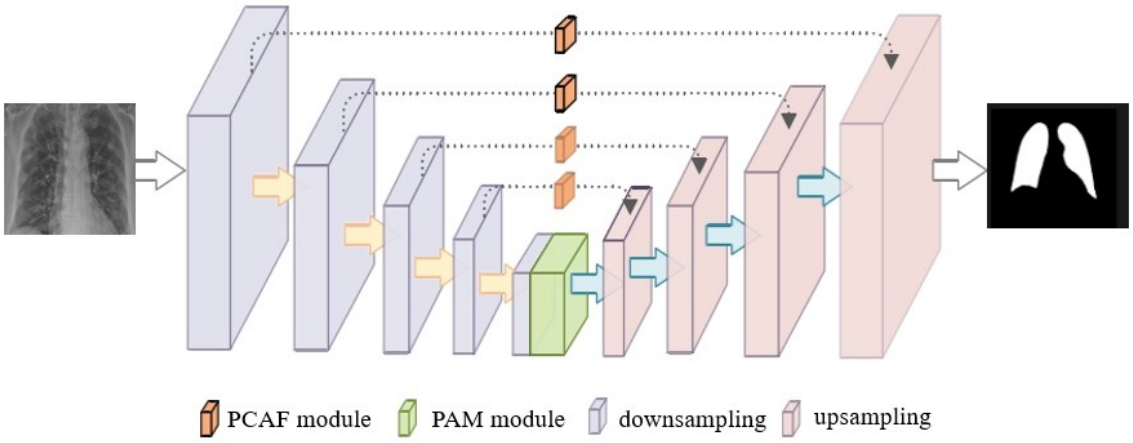

Fig. 3. PCAF-Net overall network structure

The pyramid convolution attention fusion module in this network introduces the idea of multi-scale processing and adds a channel attention mechanism to recalibrate the channel dimension of the feature map in each processing path, so that the feature map channels in different processing paths Having different weights makes up for the shortcoming of the pyramid convolution structure that cannot filter features when processing feature information at different scales, enhances important features in the original feature map, and reduces information loss during feature transfer.

Semantic expression ability alleviates the semantic difference in feature transmission between the encoder and the decoder and introduces the positional attention PAM module [7] in the bottleneck layer part of the network to achieve the function of aggregating global semantic information and using the spatial attention mechanism to Advanced feature maps are processed to improve the performance of the network.

2.2. Pyramid Convolutional Attention Fusion Module

To enhance the information transfer capability of the hopping connection and to reduce the semantic feature differences between the encoder and decoder, we have designed the PCAF module.

The input feature maps in the pyramid convolutional layer are processed in three paths, with convolution kernel sizes of 5, 3, and 1 from the top to the bottom convolutional layer.

As the kernel size increases, the number of channels decreases, forming a positive pyramid structure at the level of the number of channels and an inverted pyramid structure at the level of the kernel size.

The structure of the PCAF module is shown in Fig. 4 and consists of two parts:

1) the pyramidal convolutional layer;

2) the attention layer.

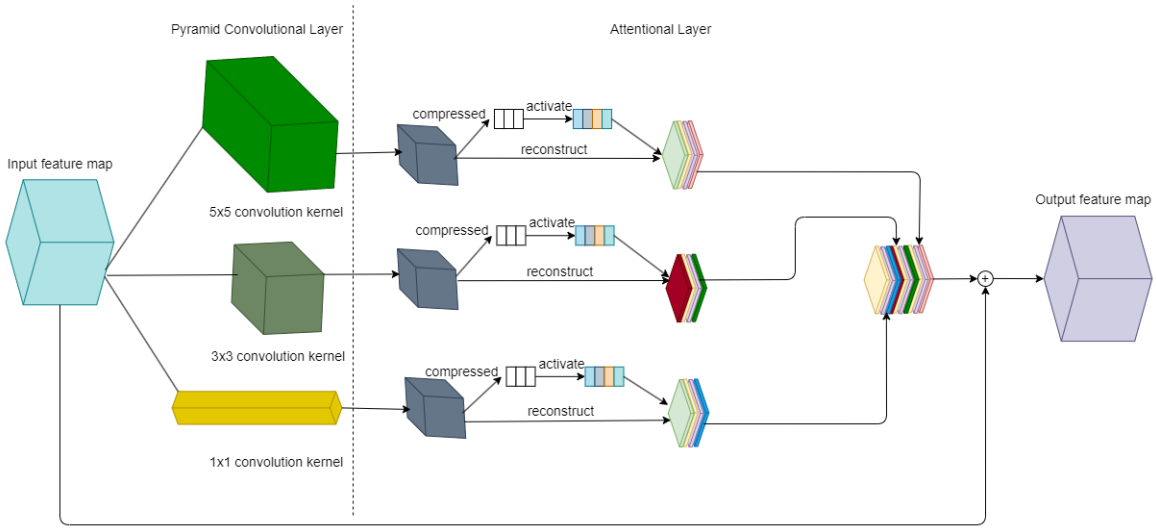

Fig. 4. Pyramid Convolutional Attention Fusion Module

The pyramidal convolution is divided into three classes, with convolution kernel sizes 5, 3, and 1 in descending order, and we replaced the convolution kernel of size 5 with two convolution kernels of size 3 during our experiments, increasing the depth of the network and strengthening the nonlinearity of the network while keeping the sense field unchanged.

The structure of the pyramid convolutional layer is shown in Fig. 5.

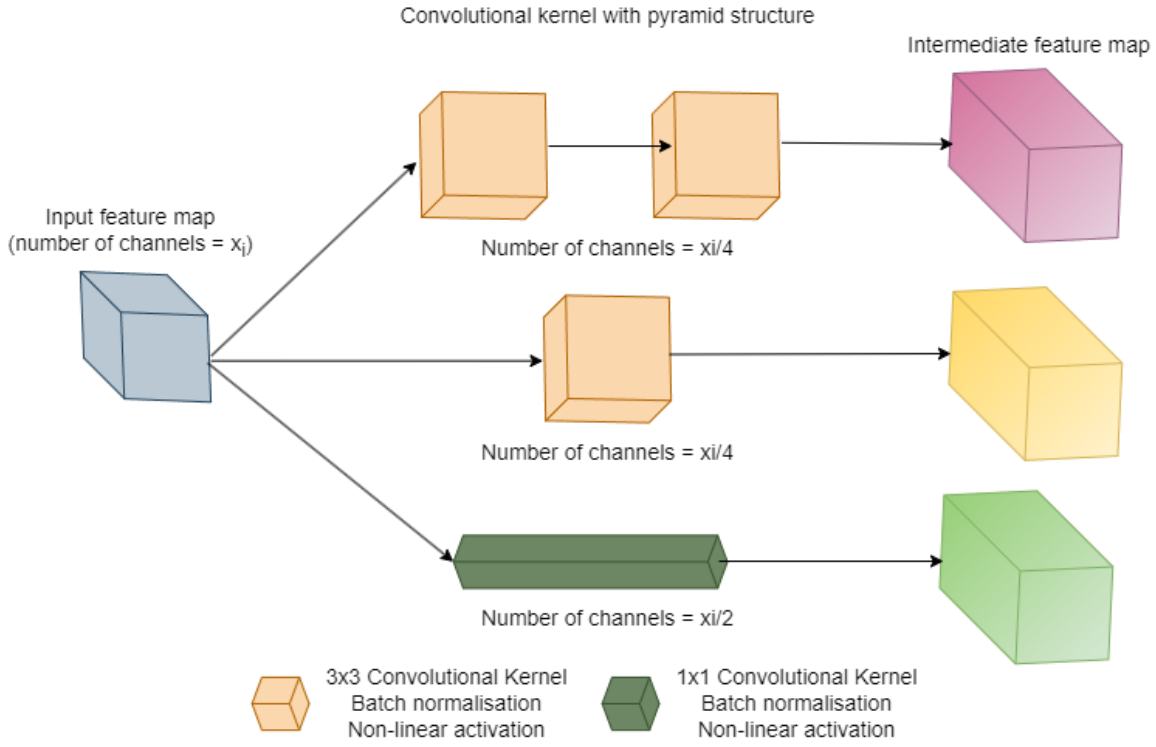

Fig. 5. Convolutional kernel with pyramid structure

The corresponding channel numbers are Xi/4, Xi/4 and Xi/2 in that order (Xi is the number of input channels), and we can use the following formula for parameter calculation:

$$P = \sum_i^n K_i^2 \times I_i \times O_i, \quad (1)$$

where i = 1, 2, 3 is the number of layers in the pyramid, K = 1, 3 is the size of the convolution kernel, i is the number of input channels to the feature map and O is the number of output channels from the feature map.

This structure does not increase the parameters of the model too much under the condition of introducing multiple scales, and the number of parameters introduced is compared to the number of parameters of a normal 3 × 3 convolution kernel as shown in table 2.

Table 2

Pyramid Convolution vs. Traditional Convolution Parametric Quantities

Input channel	Output channel	Pyramidal convolution	Traditional convolution
64	64	29696	36864
128	128	118784	147456
256	256	475136	589824
512	512	1900544	2359296

When the number of input and output channels are 64, 128, 256, and 512 respectively, the pyramid convolution reduces 7168, 28672, 114688, and 458752 compared to the number of traditional convolution parameters, respectively, and it can be seen from Table 2 that the advantages of introducing pyramid convolution become more and more obvious with the deepening of the network and the increase in the number of channels.

The feature maps processed by the pyramidal convolutional layer enter the attention layer and undergo channel recalibration through the attention mechanism, which enhances features useful for the segmentation task while suppressing irrelevant feature expressions.

Among them, the attention mechanism is divided into two operations, compression and motivation, and its specific structure is shown in Fig. 6.

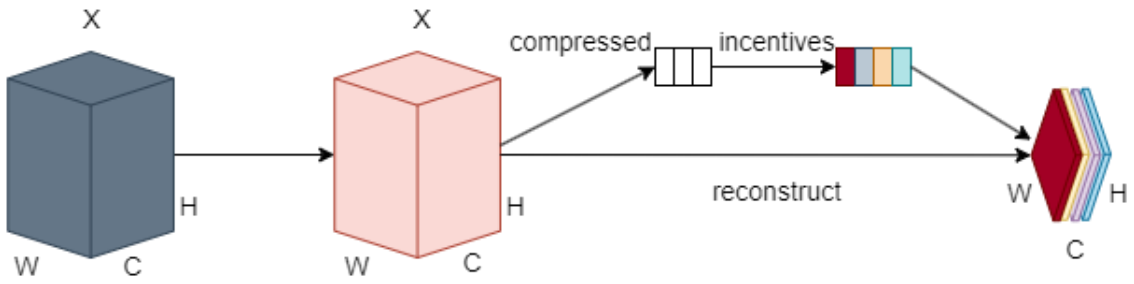

Fig. 6. Self-Attentional layer

For an input feature map $X \in R^{HxWxC}$, it is first changed to $U \in R^{HxWxC}$ by a convolution operation with a convolution kernel $V = \{V_1, V_2, V_3, \dots, V_c\}$, where v_c denotes the c convolution kernel, then the output $U = \{U_1, U_2, \dots, U_c\}$.

$$U_n = V_n X \ (n = 1, \dots, c), \quad (2)$$

where X represents the input feature map, V_n represents the convolution kernel, and U_n represents the output feature map. The fact that the convolution operation is performed in a localised space makes it difficult for U to obtain enough information to extract the relationships between the channels, and this effect is even more pronounced in shallow networks because of the small receptive field.

The compression operation uses global average pooling to compress the spatial information on a channel into a global feature:

$$Z_c = F_{sq}(U_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c \ (i, j), \tag{3}$$

where F_{sq} represents the feature compression operation, the essence of equation 3 is to sum and average the sentence ownership represented by the feature map.

After the compression operation to obtain 1 × 1 × C feature vector, based on which it sequentially Full connected (FC) operation, rectified linear unit (ReLU) operation, FC, and sigmoid activation operation, and ultimately to achieve the purpose of controlling the amount of information flow, the specific process as shown in formula 4 is shown:

$$S = F_{ex}(Z, W) = \sigma\big(W_2 \delta(W_1 Z)\big), \tag{4}$$

where δ is the ReLU activation function, σ is the Sigmoid activation function, W_1 and W_2 are fully connected operations, the first fully connected layer is used to reduce the model complexity and computation, and in the second fully connected layer it is used to recover the original dimensions of the feature vectors, and S denotes the attention vector of the obtained channel dimensions.

Finally, S is multiplied with U to get the final output feature map:

$$\tilde{X} = F_{scale}(U, S), \tag{5}$$

where F_{scale} represents the reconstruction operation and $\tilde{X}$ represents the output feature map obtained after channel attention computation.

2.3. Position Attention Module

Position Attention Module (PAM), which establishes rich contextual links in local features through the spatial attention mechanism, encodes broader contextual information into local features. Aggregate global semantic information, thus enhancing the feature expression and feature transfer ability of the network and improving the segmentation performance of the network.

The structure of the Position Attention Module is shown in Fig. 7.

The three paths are processed separately and then spliced in the channel dimension, and then fused with the input feature map to get the final output feature map.

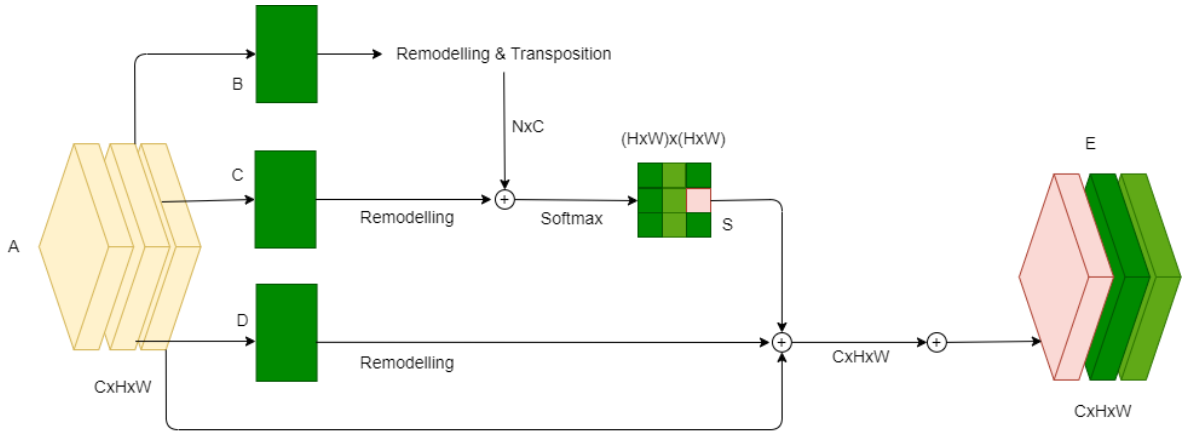

Fig. 7. Position Attention Module

The specific process as shown in formula 6 is shown:

$$S_{ij} = \frac{\exp\left(B_i' \; o \; C_j'\right)}{\sum_{i,j=1}^{N} \exp\left(B_i' \; o \; C_j'\right)}, \tag{6}$$

where o represents the matrix multiplication and S_{ij} denotes the effect of the i position on the j position, the more similar the features of the two positions are the greater the effect on S_{ij}.

The input feature map $A \in R^{CxHxW}$ is subjected to a convolution operation with a Batch Normalisation (BN) layer and a ReLU layer to obtain two new feature maps B, C, where $\{B, C\} \in R^{CxHxW}$, and B, C are reshaped into $\{B', C'\} \in R^{CxN}$, where N = H x W, after which the multiplication of B' transposed with the applied matrices to C' is applied, and the location-attention feature map $S \in R^{NxN}$ is computed after the Softmax activation layer.

The feature map A was fed into a convolutional layer with BN and ReLU layers to produce the feature map $D \in R^{CxHxW}$, which was similarly reshaped into $D' \in R^{CxN}$, and the output was reshaped into R^{CxHxW} after applying matrix multiplication between D' and S. The final output $E \in R^{CxHxW}$ was obtained by multiplying with the scaling factor α and then summing the principal elements with A.

The specific process as shown in formula 7 is shown:

$$E_j = \alpha \sum_{i=1}^{N} (S_{ji} D_i + A_j), \tag{7}$$

where the scaling factor α is initialised to 0, after which it is gradually learnt during the training process, and from equation 7 it can be found that each position of E is a weighted sum of the features of all the positions and the original features, and thus it aggregates the global semantic information.

3. Results

We compare the accuracy of PCAF-Net and U-net trained on our dataset while keeping everything else constant.

When we used U-net for the segmentation task on the Pneumonia (Chest X-Ray Images) dataset, the accuracy was 88 %.

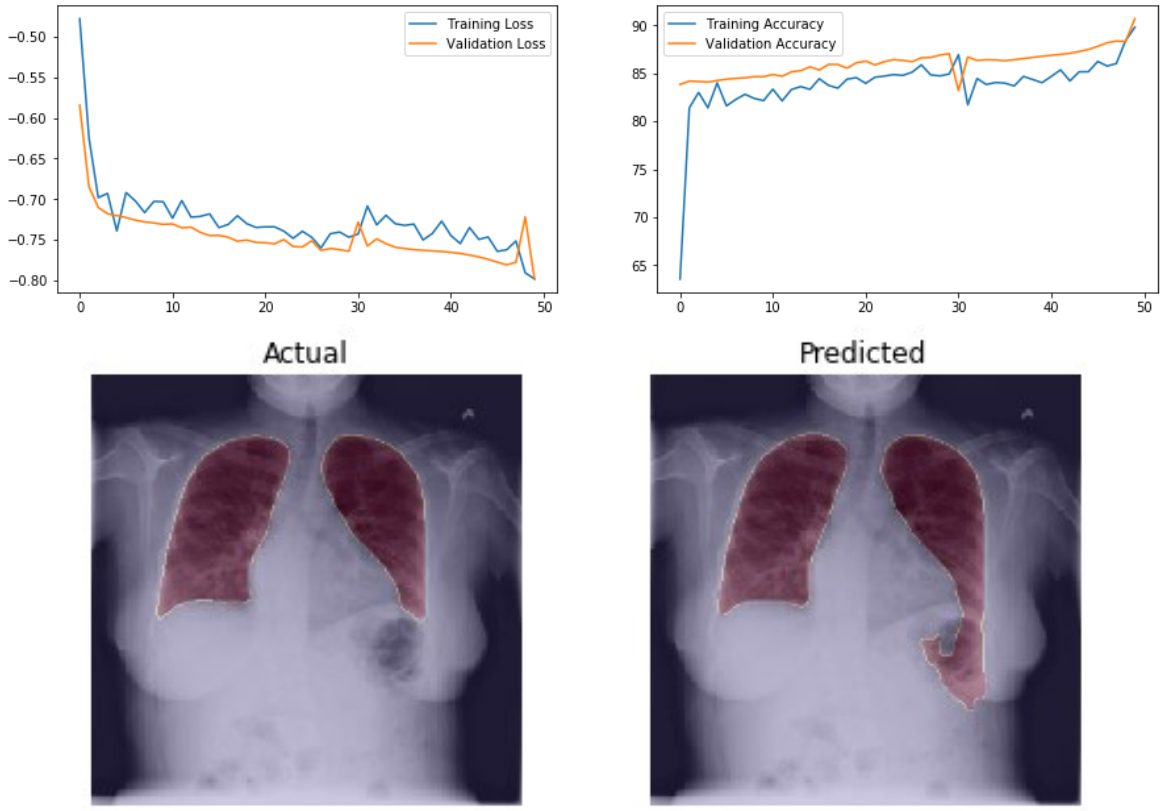

Fig. 8. U-net Experimental Groups

When we use PCFA-net to perform the segmentation task on the Pneumonia (Chest X-Ray Images) dataset, the accuracy is 96 %.

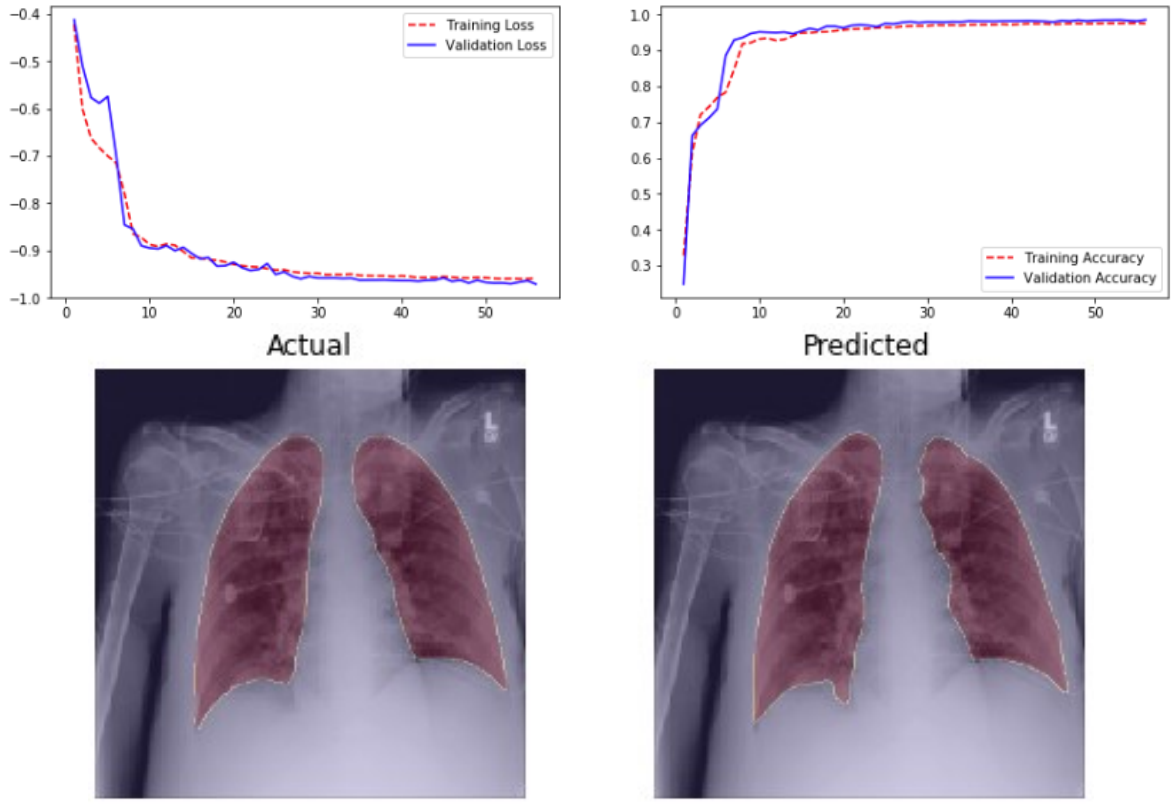

Fig. 9. PCFA-net Experimental Groups

Conclusion

We propose a novel intelligent segmentation network, PCAF-Net, which incorporates a PCAF module in the hopping connection of U-Net to reduce the variability of feature maps in the coding and decoding paths before feature fusion.

And the positional attention module is added to the transition layer of the U-Net network to enhance the ability of the network to extract features. The effectiveness of the PCAF and PAM modules is verified by extracting the feature maps in the intermediate process of the network.

Compared to U-net, our proposed PCAF-net improves the accuracy by 8 % for lung image segmentation task.

References

1. Lodha R., Kabra S. K., Pandey R. M. Antibiotics for community-acquired pneumonia in children // Cochrane database of systematic reviews. – 2013. – Vol. 6 (June 4). – CD004874. – PMID 23733365. – DOI:10.1002/14651858.CD004874.pub4.

2. Phillips II W. E., Velthuizen R. P., Phuphanich S. [et al.] Application of fuzzy c-means segmentation technique for tissue differentiation in MR images of a hemorrhagic glioblastoma multiforme // Magnetic resonance imaging. – 1995. – Vol. 13(2). – Pp. 277–290.

3. Ray S., Turi R. H. Determination of number of clusters in k-means clustering and application in colour image segmentation // Proceedings of the 4th international conference on advances in pattern recognition and digital techniques. – 1999. – Vol. 137. – P. 143.

4. Zheng X., Lei Q., Yao R. [et al.] Image segmentation based on adaptive K-means algorithm // EURASIP Journal on Image and Video Processing. – 2018. – Vol. 1. – Pp. 1–10.

5. Khilkhal R., Ismael M. Brain tumor segmentation utilizing thresholding and K-means clustering // Proc. 2022 Muthanna International Conference on Engineering Science and Technology (MICEST). – IEEE, 2022. – Pp. 43–48.

6. Ronneberger O., Fischer P., Brox T. U-net: convolutional networks for biomedical image segmentation // Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. – Springer International Publishing, 2015. – Pp. 234–241.

7. Fu J., Liu J., Tian H. [et al.] Dual attention network for scene segmentation // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2019. – Pp. 3146–3154.