THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A
Volume 24, Number 2/2023, pp. 191–201

DOI: 10.59277/PRA-SER.A.24.2.11

AN ATTENTION MECHANISM AND MULTI-FEATURE FUSION NETWORK FOR MEDICAL IMAGE SEGMENTATION

Xianxiang REN, Hu LIANG, Shengrong ZHAO

Qilu University of Technology (Shandong Academy of Sciences),
Department of Computer Science and Technology, Jinan, 250353, China
Corresponding author: Shengrong ZHAO, E-mail: zhaosr2006@126.com

Abstract. Recently, deep learning has been applied to medical image segmentation. However, existing methods based on deep learning still suffer from several disadvantages, such as blurred edge segmentation of image lesion regions and weak context information extraction. To tackle these problems, this paper proposes an attention mechanism and multi-feature fusion network with the encoder-decoder structure for medical image segmentation. In the proposed network, the convolutional group encoder module and the self-attention module are applied to divide images. The convolutional group encoder uses multiple convolution and dilated convolution to enhance the multi-scale information capturing capability of the model. The extracted image features will be useful for precise segmentation. Moreover, the self-attention module is introduced into the network for mining and complementing the edge details of segmented images. In the proposed model, convolutional group encoders and self-attention are applied repeatedly to capture changes in contextual relationships and continuously refine boundary information. Several experiments have been conducted on the BUSI and ISIC datasets to verify the effectiveness of the proposed method. Compared with other methods, the proposed method can achieve better segmentation results.

Key words: medical image segmentation, multi-feature fusion, encoder-decoder, self-attention.

1. INTRODUCTION

Image segmentation is receiving more and more attention as it's an essential component of medical-aided diagnosis. The boundary information of the focus area in the medical image is very critical for doctors to judge the patient's condition. Smooth boundaries usually represent positive lesions, while rough boundaries represent malignant lesions. Fig.1 shows medical skin cancer segmentation results. It indicates that image segmentation can help physicians focus on suspicious areas of lesions.

(a) Image (b) Label (c) AM-Net

Fig. 1 – An example of medical image segmentation: a) a sample image of skin cancer dataset; b) an image labels made by pathologists; c) segmentation result obtained by our proposed method.

Traditional image segmentation strategies have achieved valuable results. For example, threshold segmentation based on region image segmentation is a popular segmentation technique due to its computational simplicity and consistent performance. The edge detection method is an image segmentation method that searches for a target region's boundary. And the region extraction method starts from a single pixel and merges gradually to create a segmented region. The disadvantage of these methods is that they rely on manual manipulation [1,2].

Deep learning techniques have enabled breakthroughs in medical image segmentation. In 2014, Long et al. proposed a full convolutional network (FCN) [3] model based on a convolutional neural network (CNN), which solves the problem of image segmentation through pixel level classification. In 2015, Ronneberger et al. proposed U-Net [4], which is built on top of FCN. U-Net's U-shaped structure avoids the drawback of FCN, i.e., that it cannot consider global context information. Among U-Net's context information is obtained from neighborhood data, target annotations, and the spatial location of the target. In recent years, researchers have used the U-Net network or an improved network structure for medical image segmentation. Such as, Swin-Unet [5], Transformer-Unet [6] and UTNet [7] combined transform with U-Net for medical image segmentation to achieve optimal results. The above three approaches have modified the structure of U-Net and improved its model performance. However, despite their continued progress, these methods still encounter difficulties when faced tasks with high accuracy, especially complex spatial features and edge details.

In fact, more detailed image segmentation results are often driven by context information and boundary features, and many work has focused on this approach. For example, CO-Net [8] and T-Net [9] obtained regionally significant information by fusing cross-layer features. K-Net [10] used a learnable kernel for consistent segmentation. OCR-Net [11] enhanced the description of pixel features by learning the relationship between pixels and object region features. However, as pointed out in [12], ignoring semantic gaps often affects feature mining and fusion. References [13, 14] are closely related to feature extraction and medical field, and they achieve excellent results in the medical image segmentation. In addition, several researches [15–17] have investigated the direction of weakly supervised semantic segmentation, which aims to learn semantic segmentation with only image-level object category supervision. Therefore, it remains a challenge to design a network driven by context information and boundary features to obtain fine-grained results.

To solve the deficiencies of blurred edge segmentation and weak context information extraction in image lesion areas, we proposed an attention mechanism and multi-feature fusion network (named AM-Net for short). As shown in Fig. 2, AM-Net includes Convolutional Group Encoder (CGE) and Self-Attention (SA) modules. In our work, CGE is designed to extract multi-scale image features. In addition, the SA module is used to mine the edge details of the segmented image, so as to obtain a more accurate segmented image. In summary, the contributions of this paper include:

- A novel multi-scale feature fusion neural network model (called AM-Net for short) is proposed for medical image segmentation. This method can address the issues of poor context extraction and imprecise edge detection effectively.

- In the proposed AM-Net, a multi-scale feature fusion module, i.e., CGE, is designed to obtain fine segmentation details. The CGE structure could control the information flow to reduce redundancy, and the features of each layer gain access to richer context with the help of multi-feature.

- Following the CGE module, the self-attention module is applied, which will pay more attention to the edge information and solve the problem of unsmooth edge segmentation.

- The model is evaluated on a new ISIC and BUSI dataset and shows excellent performance. The experimental results also verify the superiority of AM-Net.

The remainder of the paper is organized as follows. Section 2 introduces the proposed AM-Net model, Section 3 explores experimental data, and Section 4 summarizes the study conclusions.

2. THE PROPOSED MODEL

We propose an attention mechanism and multi-feature fusion network (named AM-Net for short) for medical images segmentation. In this section we introduce the AM-Net structure.

2.1. AM-Net

We design an AM-Net with an encoder-decoder structure, which can segment medical images accurately. The structure of AM-Net is shown in Fig. 2. The encoder includes convolution, SA, CGE, down-sampling and spatial pyramid pooling [18]. The encoder is used to extract feature information, while the SA module is used to gradually improve edge structure information. The CGE module is used to enhance the contextual feature extraction capability of the network. A spatial pyramid pooling module is used to extract multi-scale information further. The decoder includes deconvolution and up-sampling. The decoder is also used to gradually recover the pixel coordinates of an image, where the Deconvolution Group Decoder (DGD) is used to retain shape information regarding the input image. Up-sampling is used to recover image size. In the segmentation

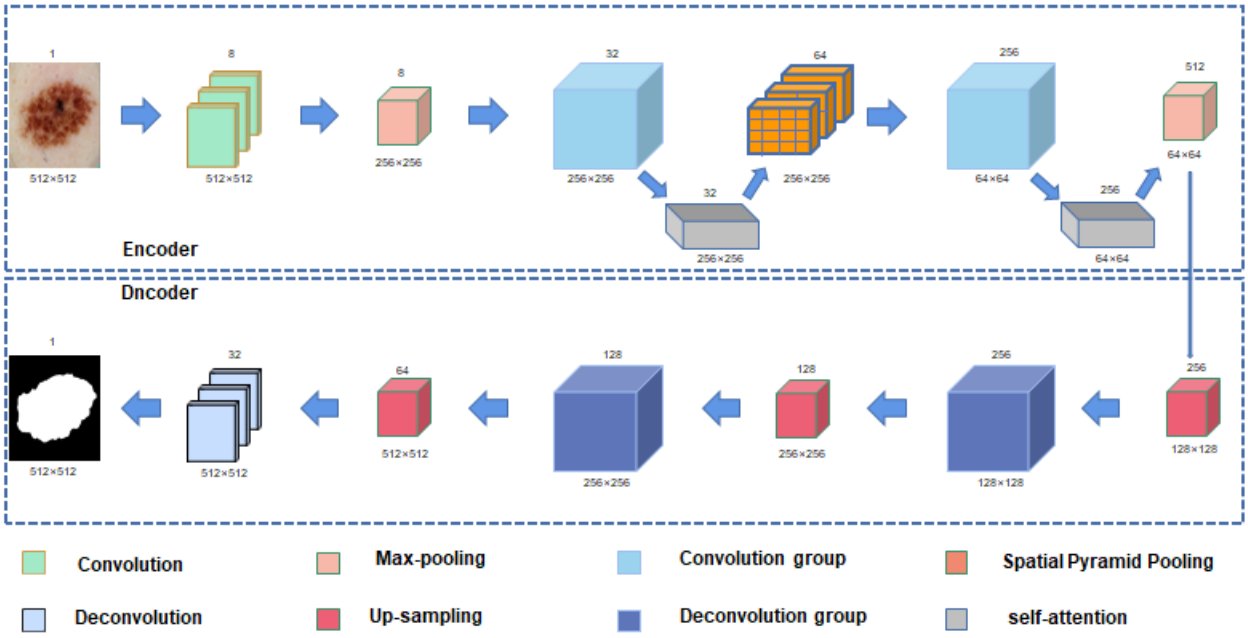


Fig. 2 – Overall architecture of the AM-Net.

process, a pre-processed image is first fed into the feature encoder to extract primary features. Then, the CGE is introduced into the encoding module to capture and dynamically fuse multi-scale contextual semantic information. At the same time, the SA module is used to capture more accurate edge structure information. Next, the features are recovered by the decoder. The encoder outputs a feature tensor, which passes through the decoder to take the output feature tensor as input. The features are recovered with deconvolution after down-sampling. After deconvolution, summation, and up-sampling, a prediction map of the same size as the input image is finally obtained.

2.2. Encoder

An encoder is used to extract the feature information of processed images. It continuously reduces a feature map to a lower dimension to obtain the features of an image. Herein, we propose the CGE, an encoder that employs convolution and null convolution to improve multi-scale information capturing capability. Figure 3 depicts the structure of the CGE. The CGE module is composed of dilated convolution and convolution.

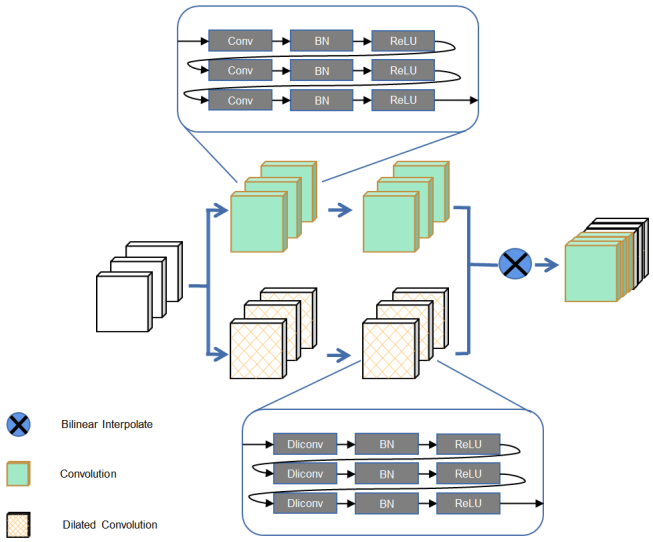


Fig. 3 – Convolutional Group Encoder (CGE).

Assume that the input of CGE is represented by Q. Then the specific calculation process is given as follows:

$$X_{CGE} = BI(Conv_{n\times n}(Q), Dilconv_m(Q)) \tag{1}$$

where BI is the splicing operation, Q is the characteristic image received by CGE, $Conv_{n\times n}$ contains three layers of convolution, three layers of batch normalization and three layers of activation function, $Dilconv_m$ contains three layers of hole convolution, and X_{CGE} is the output of CGE. n is the convolution kernel and m is the dilution rate. Assuming that the number of channels of Q is M, the number of channels of X_{CGE} is $4\times M$.

First we use a 3×3 convolution kernel to convolve the image. Then, the image features are processed using two CGEs and SA. The first CGE uses a 3×3 convolution kernel of convolution and dilated convolution, where the dilation rate of the dilated convolution is 3. The second CGE uses a 5×5 convolution kernel of convolution and dilated convolution, where the dilation rate is 5. We use different scale convolution and dilated convolution to acquire the features of the image and fuse them. Thus, the feature maps of different scales of the image can be obtained to the maximum extent possible, and the image can be segmented more effectively. Two pooling operations are used in our proposed framework, i.e., max pooling and Spatial Pyramid Pooling. We use two 2×2 max pooling layers for halving the resolution of the feature map to enforce spatial invariance, which helps to aggregate features from different spatial regions. Spatial pyramid pooling consisting of 5×5, 9×9 and 13×13 average pooling is used to divide image blocks into multiple regions to identify the local information of the images.

SA module is introduced into the network to strengthen the image features obtained by CGE and supplement the edge details of segmented images. The SA module can effectively improve the blurred edge of image segmentation. Figure 4 shows the structure of self-attention. In Fig. 4, the "·" represents dot multiplication. Matrices Q(query), K(key value) and V(value) are used in the calculation. The input of the SA module is the output of the CGE module. Q, K, and V are derived from the linear transformation of the input matrix. The calculation of the entire SA module can is represented in Algorithm 1. Q, K and V are calculated using the following formulars.

$$Q = W_{\mathrm{q}}X \tag{2}$$

$$K = W_{\mathrm{k}}X \tag{3}$$

$$V = W_{\mathrm{v}}X \tag{4}$$

where W_q, W_k, W_v are the parameter matrices of the outer linear mapping.

Algorithm 1 Algorithm flow of SA method

Input: Characteristic drawing X_{CGE}
Output: Enhanced feature map (H)

1: The characteristic graph X_{CGE} is linearly mapped to obtain Q, K, V(Calculate according to formula 2,3,4.);
2: Calculate the attention score of X_{CGE} : $Score$=QK^T;
3: Normalization : NL = Softmax($Score$);
4: $H = V*NL$

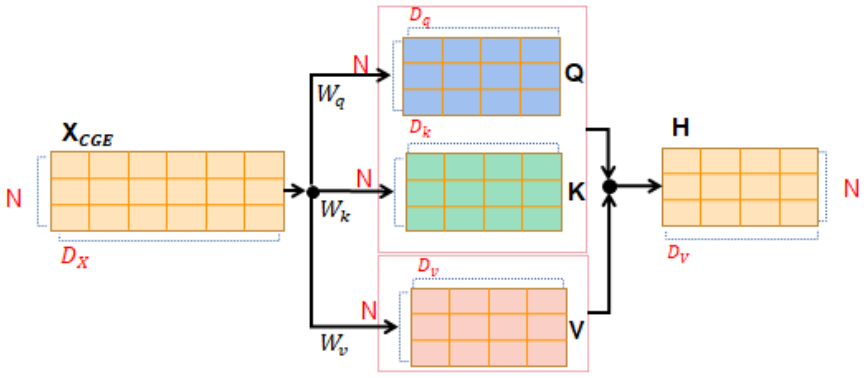


Fig. 4 – The structure of self-attention.

2.3. Decoder

After the up-sampling layer, the decoder module employs a deconvolution layer. To recover features from deconvolution layers, a feature stitching approach is used to connect to the deconvolution layer, which captures more information about edge structure. The DED is mainly responsible for reconstructing the shape of the input image. To preserve the shape information of the input image, a hierarchy of deconvolution layers is designed. In addition, after three down-samplings, the encoder output feature tensor is convolved, summed, and upsampled with operations to finally obtain a 512×512 prediction map with the same size as the input image. Assuming that P denotes the input, the process of DGD image restoration is given as follows:

$$X_{DGD} = F(Up(P)) \tag{5}$$

where F contains a layer of deconvolution, a layer of BN, and Up is the upsampling operation. Assuming that the number of channels of P is M, the number of channels of X_{DGD} is $\frac{M}{2}$.

To obtain high confidence segmentation results, we use the Binary Cross Entropy (named BCE for short) function, the most frequently used loss function in medical image segmentation. The BCE function is formulated as follows:

$$Loss_{\mathrm{BCE}} = -\sum_{i=1}^{C} p_i \log(q_i) \tag{6}$$

where C represents the number of categories, p_i is the true value, and q_i is the predicted value. The BCE loss function operates at the pixel-level.

3. EXPERIMENTS

To demonstrate the effectiveness and generalization of the proposed method, we conducted experiments on the ISIC and BUSI datasets.

3.1. Experiment setup

The experiment adopts the ISIC and BUSI datasets. The ISIC data set has 1 252 pictures, and BUSI data set has 630 pictures. We uniformly set the ISIC and BUSI images to a fixed size of 512×512. Then the image is divided into training set, verification set and test set according to the ratio of 6:2:2, and the division method is random. All experiments are carried out using the Python deep learning framework and run on the GeForce RTX 3090 GPU card. The batch size for training is set to 4. The experiment uses the RMSprop optimizer to adjust the parameters, and the initial learning rate is 0.0001. During the training of the AM-Net, we set a maximum training period of 120 epochs, and the actual training is manually stopped when the network shows signs of overfitting using the early stop method. Finally, the model is saved for subsequent validation when the detection accuracy reaches the maximum.

The proposed model uses Pixel Accuracy (PA), Mean Intersection over Union (MIoU), Precision (Pre), Recall (Rec) and F-Score to evaluate the measurement equations as shown below.

$$PA = \frac{TP+TN}{TP+TN+FP+FN} \cdot 100\% \tag{7}$$

$$MIoU = \frac{1}{K+1} \sum_{i=0}^{K} \frac{TP}{TP+FP+FN} \cdot 100\% \tag{8}$$

$$Pre = \frac{TP}{TP+FP} \cdot 100\% \tag{9}$$

$$Rec = \frac{TP}{TP+FN} \cdot 100\% \tag{10}$$

$$F\text{-}Score = 2 \cdot \frac{Pre \cdot Rec}{Pre+Rec} \cdot 100\% \tag{11}$$

where K is the number of categories; Pixels that are correctly predicted as areas of interest (true positive(TP)); Pixels that are correctly predicted as non-interesting (true negative(TN)); Non-interesting pixels incorrectly predicted as such(false positive(FP)); Interesting pixels incorrectly predicted as such (false negative(FN).

3.2. Analysis of the image preprocessing

In this sub-section, we first evaluate the effect of image preprocessing, because some images contain hair that corrupts the obtained images. Such types of noise have an effect on the image segmentation performed by the model. Thus, it is necessary to remove hair from the images. Noise such as hair on the skin is removed using morphological filters [19]. Figure 5 shows an example which presents the original image and processed image obtained by morphological filtering.

(a) Original Image (b) Processed Image

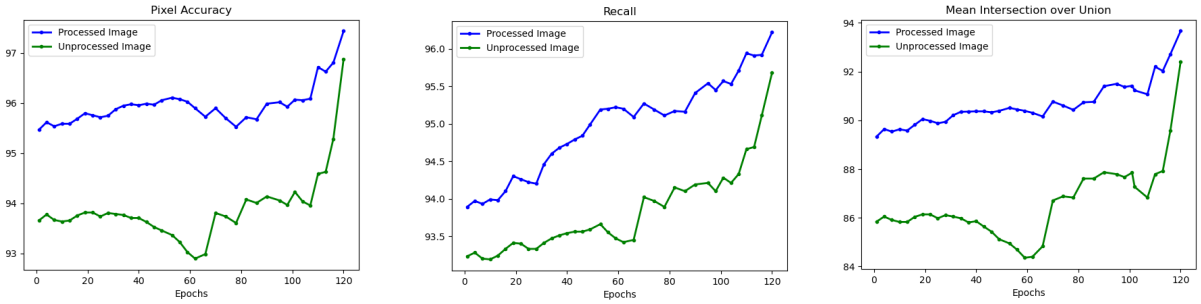
Fig. 5 – A sample image from the skin cancer dataset.

The ISIC dataset sets are trained with this paper's model AM-Net before and after pre-processing, respectively.The Pre and F-Score scores obtained by training the images before and after pre-processing are shown in Table 1. Figure 6 shows a comparison of Pa, Rec and Miou for the images before and after pre-processing. Clearly, the training effect of the pre-processed images is better and more stable than that of the original images. This indicates that the pre-processing image method we chose is suitable for the task.

Table 1

Comparison of Pre and F-Score of preprocessed images and original images

Type	Precision	F-Score
Original Image	91.44%	92.12%
Pre-Processed Image	**95.62%**	**95.05%**



(a) *PA* (b) *Rec* (c) *MIou*

Fig. 6 – Comparison of Pixel Accuracy, Recall and Mean Intersection over Union between preprocessed and original images.

3.3. Ablation experiment

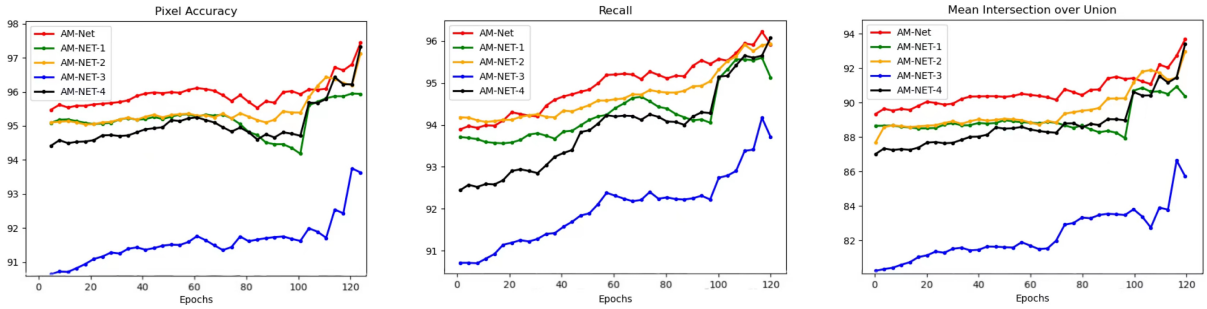
To determine the effectiveness of the various modules in AM-Net, ablation studies were conducted in the same experimental environment, and the performance of the network was compared after adding the modules. We compare AM-Net with its four variants on the ISIC dataset in Table 2. We use four variants of AM-Net, namely, AM-NET-1, AM-NET-2, AM-NET-3 and AM-NET-4, where:

- AM-NET-1 is built by removing the first SA module;
- AM-NET-2 is built by removing the second SA module;
- AM-NET-3 is constructed by transforming the CGE module into a normal convolutional module;
- AM-NET-4 is built using only the CGE module.

Table 2

Comparison of Precision and F-Score of ablation experiments

Modules	Precision	F-Score
AM-Net	**95.62%**	**95.05%**
AM-NET-1	93.68%	94.15%
AM-NET-2	94.20%	93.85%
AM-NET-3	87.26%	82.07%
AM-NET-4	95.35%	94.52%



(a) PA (b) Rec (c) MIou

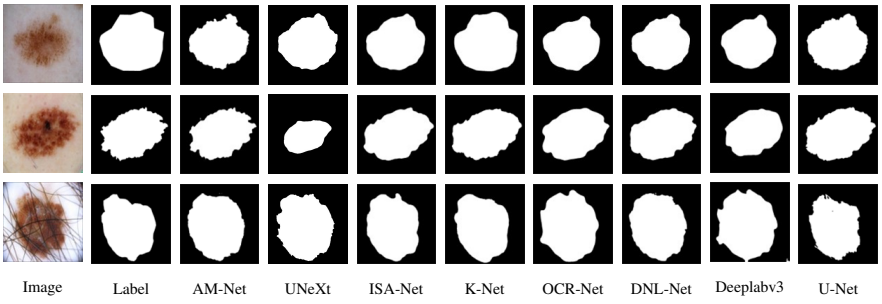
Fig. 7 – Comparison of Pixel Accuracy, Recall, and Mean Intersection over Union between ablation experiments.

Using the pre-processed dataset and comparing the AM-NET-1, AM-NET-2, AM-NET-3 and AM-NET-4 experiments, we obtained the following results. By comparing the results of AM-NET-1 and AM-NET-2, it can be seen that using the SA module alone will lead to poor performance. However, AM-Net's Pre and F-Score show that using two SA modules is effective. AM-Net and models AM-NET-1, AM-NET-2 and AM-NET-4 all achieve high segmentation accuracy, indicating the effectiveness of using SA to obtain region and contour information. The segmentation effect is unsatisfactory when the CGE module is converted to a normal convolutional module. Model AM-NET-3 has the lowest Pre and F-Score of 87.26% and 82.07%, respectively. This is significantly lower than those of models AM-NET-1, AM-NET-2, and AM-NET-3. This shows that the CGE can extract features at different image scales, thus enabling more effective segmentation of the image. Figure 7 shows a comparison of Pa, Rec and Miou for AM-Net and four variants. Clearly, the training effect of AM-Net is better and more stable than that of its four variants. By comparing the module results, it can be seen that each individual module significantly improves segmentation accuracy.

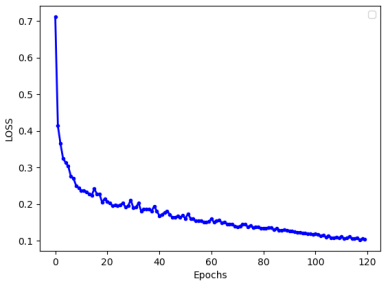
3.4. Contrast experiment

We compare the performance of AM-Net and seven other models, UNeXt [21], ISA-Net [22], K-Net [10], OCR-Net [11], DNL-Net [23], Deeplabv3 [24] and U-Net [4], on the skin cancer dataset. AM-Net scores are calculated according to the formulas described in Eqs.(7)–(11). The split result maps are annotated manually by a pathologist. In addition, qualitative and quantitative validation measures are conducted to measure the efficiency and efficacy of the proposed AM-Net framework.

Qualitative results. The qualitative results of AM-Net on the ISIC and BUSI datasets are shown in Fig. 8 and Fig. 9. In this case, the black area in the segmentation map is detected as the background and the white area is the lesion area.

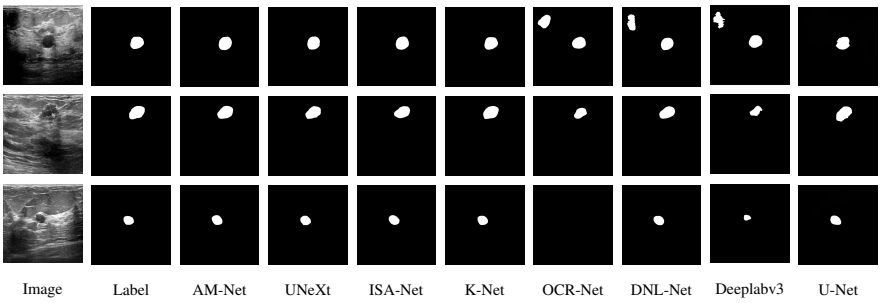


(a) ISIC dataset segmentation results of different methods.

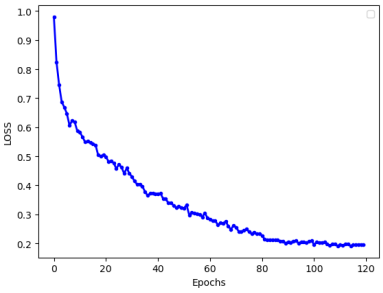


(b) loss curve on ISIC dataset.

Fig. 8 – Segmentation results obtained by different methods on ISIC dataset and loss curves corresponding to AM-Net.



(a) BUSI dataset segmentation results of different methods.



(b) loss curve on BUSI dataset.

Fig. 9 – Segmentation results obtained by different methods on BUSI dataset and loss curves corresponding to AM-Net.

Quantitative Results. Quantitative results are also essential for the statistical evaluation of the proposed framework. The quantification results shown in Table 3 and Table 4 show the performance measurements of the AM-Net model along with the experimental results of the seven existing models. Table 3 and Table 4 show that AM-Net is effective as it outperforms the seven segmentation models across all metrics of the skin cancer dataset.

Analysis. In Fig. 8 and Fig. 9, it shows the segmentation results obtained by different methods on the ISIC and BUSI datasets. Moreover, the loss curve conesponding to AM-Net is also presented in these two figures.

For the ISIC dataset, we conclude that using AM-Net to classify images at the pixel level leads to good segmentation performance. In contrast, U-Net bridges this gap, showing that using the encoder-decoder format is effective. UNeXt adds a convolutional multilayer perceptron on the basis of U-Net, which reduces the number of parameters and computational complexity, and improves the segmentation accuracy. K-Net performs better in detailed boundary processing, which shows the effectiveness of boundary information. OCR-Net is unable to extract rich contextual information with its encoder, resulting in a significant decrease in accuracy.

Table 3

Comparison of different methods on ISIC dataset

Models	Year	Params(M)	PA	MIoU	Recall	Precision	F-Score
UNeXt [21]	MICCAI-2022	1.47	95.86%	88.10%	92.39%	93.41%	93.66%
ISANet [22]	IJCV-2021	-	93.82%	86.14%	93.40%	91.65%	92.46%
K-Net [10]	NeurIPS-2021	37.26	95.81%	90.10%	94.45%	95.00%	94.77%
OCR-Net [11]	ECCV-2020	-	90.08%	78.99%	89.37%	86.99%	88.04%
DNL-Net [23]	ECCV-2020	71.48	94.42%	87.10%	92.81%	93.22%	93.01%
DeepLabv3 [24]	ArXiv-2017	-	84.91%	69.68%	82.40%	80.98%	81.63%
U-Net [4]	MICCAI-2016	13.39	93.78%	91.05%	86.32%	93.12%	89.72%
AM-Net	-	2.89	**96.16%**	**90.72%**	**94.48%**	**95.62%**	**95.05%**

Similarly, the networks are tested on the BUSI dataset, and the results of the experiment are shown in Table 4 and Fig. 9. AM-Net and ISA-Net incorporating the attention structure mostly achieve significant performance and effectively improve the accuracy of the edge segmentation compared with U-Net. In addition, compared with ISA-Net, which only uses attention as the basis for network construction, AM-Net uses the combination of attention and CGE to achieve better performance and further improve segmentation accuracy, which proves the correctness of AM-Net's idea of combining attention and CGE. In addition, compared with other segmentation networks, the multi-scale feature information fusion interaction capability of AM-Net again improves the accuracy of the model, and the segmentation of lesion edges is more accurate, which proves the importance and effectiveness of the CGE module for the final segmentation performance improvement.

4. CONCLUSION

Our AM-Net network for segmenting medical images tackles the present challenges of blurred edges in the segmentation of lesion regions in medical images and poor contextual information extraction. The segmentation results are affected by network acquired knowledge about the lesion area and its boundaries. The SA module of AM-Net captures the edge feature information and enhances the accuracy of the edge segmentation. For acquisition purposes, the CGE module of AM-Net fuses image information uses multiple scales of convolution and dilated convolution. As a result, the feature maps of the multiple scales of an image are collected to the maximum extent possible, and the image is segmented more efficiently.

In addition, our findings suggest that our image preparation strategy is well-suited to the ISIC dataset and the BUSI dataset, where the AM-Net network achieves 96.16% and 95.12% PA (Pixel Accuracy), 90.72% and 76.62% MIoU (Mean Intersection over Union), 94.48% and 82.84% Rec (Recall), 95.62% and 88.95% Pre (Precision), and 95.09% and 85.55% F-Scores. A large number of experiments on the skin cancer segmentation dataset also demonstrate the network's effectiveness.

Table 4

Comparison of different methods on BUSI dataset

Models	Year	Params(M)	PA	MIoU	Recall	Precision	F-Score
UNeXt	MICCAI-2022	1.47	94.86%	75.37%	80.19%	85.49%	85.24%
ISA-Net	IJCV-2021	-	94.91%	75.58%	81.67%	88.76%	84.75%
K-Net	NeurIPS-2021	37.26	94.15%	72.36%	78.82%	86.71%	82.14%
OCR-Net	ECCV-2020	-	91.60%	61.57%	68.08%	79.12%	71.89%
DNL-Net	ECCV-2020	71.48	94.39%	73.70%	80.46%	86.87%	83.27%
DeepLabv3	ArXiv-2017	-	93.09%	66.25%	71.61%	86.19%	76.56%
U-Net	MICCAI-2016	13,39	83.68%	71.92%	81.25%	81.47%	76.12%
AM-Net	-	2.89	**95.12%**	**76.62%**	**82.84%**	**88.95%**	**85.55%**

ACKNOWLEDGEMENTS

This work was supported by Natural Science Foundation of Shandong Province (No.ZR2020MF039 and No.ZR2022LZH008), The 20 Planned Projects in Jinan (No.2021GXRC046), Qilu University of Technology (Shan dong Academy of Sciences) Young doctor Cooperation Fund Project (No. 2019BSHZ009), and Basic Research enhancement ment Program of Qilu University of Technology (Shandong Academy of Sciences) (No.2021JC02015).

REFERENCES

1. S. MAYALA, J.B. HAUGSEN, *Threshold estimation based on local minima for nucleus and cytoplasm segmentation*, Medical Imaging, **22**, *1*, pp. 77–89, 2022.
2. M.H. SIDDIQI, I. ALRASHDI, *Edge detection-based feature extraction for the systems of activity recognition*, Computational Intelligence and Neuroscience, art. 8222388, 2022.
3. J. LONG, E. SHELHAMER, T. DARRELL, *Fully convolutional networks for semantic segmentation*, Proceedings of the IEEE Conference on Computer Vision and Pattern reRcognition (CVPR), 2015, pp. 3431–3440.
4. O. RONNEBERGER, P. FISCHER, T. BROX, *U-net: Convolutional networks for biomedical image segmentation*, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241.
5. H. CAO, Y. WANG, J. CHEN, D. JIANG, X. ZHANG, Q. TIAN, W. MANNING, *Swin-Unet: Unet-like pure transformer for medical image segmentation*, European Conference on Computer Vision (ECCV), 2022, pp. 205–218.
6. Y. SHA, Y. ZHANG X. JI, *Transformer-Unet: raw image processing with Unet*, arXiv preprint arXiv: 2109.08417, 2021.
7. Y. GAO, M. ZHOU, D.N. METAXAS, *UTNet: A hybrid transformer architecture for medical image segmentation*, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2021, pp. 61–71.

8. A. LIU, X. HUANG, T. LI, P. MA, *Co-Net: A collaborative region-contour-driven network for fine-to-finer medical image segmentation*, Conference on Applications of Computer Vision (WACV), 2022, pp. 1046–1055.
9. T.M. KHAN, A. ROBLES-KELLY, S.S. NAQVI, *T-Net: A resource-constrained tiny convolutional neural network for medical image segmentation*, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 644–653.
10. W. ZHANG, J. PANG, K. CHEN, C. LOY, *K-Net: Towards unified image segmentation*, CVPR, 2021, pp. 10326–10338.
11. Y. YUAN, X. CHEN, J. WANG, *Object-contextual representations for semantic segmentation*, European Conference on Computer Vision (ECCV), 2020, pp. 173–190.
12. Z. ZHANG, X. ZHANG, C. PENG, X. XUE, J. SUN, *Exfuse: Enhancing feature fusion for semantic segmentation*, Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 269-–284.
13. E. ARICAN, T. AYDIN, *An RGB-D descriptor for object classification*, Romanian Journal of Information Science and Technology (ROMJIST), **25**, *3–4*, pp. 338–349, 2022.
14. S. ÖĞÜTCÜ, M. İNAL, C. ÇELIKHASI, U. YILDIZ, N.Ö. DOĞAN, M. PEKDEMIR, *Early detection of mortality in COVID-19 patients through laboratory findings with factor analysis and artificial neural networks*, Romanian Journal of Information Science And Technology (ROMJIST), **25**, *3–4*, pp. 290–302, 2022.
15. Y.T. CHANG, Q. WANG, W.C. HUNG, R. PIRAMUTHU, Y.H. TSAI, M.H. YANG, *Weakly-supervised semantic segmentation via subcategory exploration*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8991–9000.
16. J. FAN, Z. ZHANG, C. SONG, T. TAN, *Learning integral objects with intra-class discrim-inator for weakly-supervised semantic segmentation*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4283–4292.
17. Y. LI, Z. KUANG, L. LIU, Y. CHEN, W. ZHANG *Pseudo-mask matters in weakly-supervised semantic segmentation*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6964–6973.
18. K. HE, X. ZHANG, S. REN, J. SUN, *Spatial pyramid pooling in deep convolutional networks for visual recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **37**, *9*, pp. 1904–1916, 2015.
19. R. MONDAL, M. DEY, B. CHANDA, *Image restoration by learning morphological opening-closing network*, Mathematical Morphology – Theory and Applications, **4**, pp. 87–107, 2020.
20. A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A.N. GOMEZ, L. KAISER, I. POLOSUKHIN, *Attention is all you need*, Advances in Neural Information Processing Systems (NIPS), 2017, vol. 30, pp. 5998–6008.
21. J.M.J. VALANARASU, V.M. PATEL, *UNeXt: MLP-based rapid medical image segmentation network*, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 202, pp. 23–33.
22. L. HUANG, Y. YUAN, J. GUO, C. ZHANG, X. CHEN, J. WANG, *Interlaced sparse self-attention for semantic segmentation*, IJCV, 2021, pp. 1-11.
23. M. YIN, Z. YAO, Y. CAO, X. LI, Z. ZHANG, S. LIN, H. HU, *Disentangled non-local neural networks*, European Conference on Computer Vision (ECCV), 2021, pp. 191–207.
24. L.C. CHEN, G. PAPANDREOU, F. SCHROFF, H. ADAM, *Rethinking atrous convolution for semantic image segmentation*, arXiv preprint arXiv:1706.05587, 2017.

Received October 24, 2022