2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 5-8, 2025. Vienna, Austria

2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 979-8-3315-3358-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/SMC58881.2025.11343391

MAR-Net: Multi-scale Attention Refinement Network for Enhanced Medical Image Segmentation

Hongyao Ma[1], Jing Bi[1], Ziqi Wang[2], Ning Li[1], Haitao Yuan[3], Yibo Li[1] and Jia Zhang[4]

[1]College of Computer Science, Beijing University of Technology, Beijing 100124, China
[2]School of Software Technology, Zhejiang University, Ningbo 315100, China
[3]School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China
[4]Dept. of Computer Science in Lyle School of Engineering, Southern Methodist University, Dallas, TX, USA

***Abstract*—Medical image segmentation challenges stem from complex lesion morphologies and real-time clinical needs. Here we present MAR-Net, a novel framework that integrates adaptive attention mechanisms, hierarchical contextual modeling, and efficient training strategies to address these challenges. The architecture employs a CBAM-based dual-attention module to dynamically enhance discriminative features while suppressing redundant information, improving lesion boundary localization. Cascaded dilated convolutions expand the receptive field for global context capture, complemented by a multi-scale decoder that integrates deep semantic and shallow spatial features. A multi-scale training strategy with hierarchical loss supervision optimizes model adaptability without compromising inference efficiency. Experimental validation on ISIC and CholecSeg8K datasets demonstrates MAR-Net's superiority: it outperforms mainstream methods across segmentation accuracy, recall rate, and other metrics, achieving notable improvements for complex lesions of varying sizes. Notably, MAR-Net maintains high performance on both dermoscopic and laparoscopic images, showcasing its broad applicability. These results confirm MAR-Net as a robust medical segmentation solution, balancing precision and efficiency for clinical use.**

***Index Terms*—Gabor filters, fine-grained features, convolutional neural networks, resNet101, image classification.**

I. Introduction

Medical image segmentation is critical for clinical diagnosis, surgical planning, and disease monitoring. However, the lesion regions in medical images often exhibit complex shapes, varying sizes, and blurred boundaries, making automatic segmentation a challenging task. Traditional medical segmentation methods [1]primarily use hand-crafted feature extraction or image processing techniques (e.g., thresholding, region growth, edge detection). However, these methods struggle to handle complex structures and fine details. Recently, deep learning, represented by convolutional neural networks (CNNs) and Transformers, has offered novel solutions for medical image segmentation and has become the dominant approach.

In the field of medical image segmentation, traditional segmentation networks such as U-Net [2], DeepLabV3 [3],FCN [4] and LR-ASPP [5] have made significant progress. U-Net is a classic medical segmentation model for its symmetric encoder-decoder design, effectively combining multi-scale features and performing exceptionally well in tasks like cell and organ segmentation. DeepLabV3 enhances lesion area recognition by introducing dilated convolutions to expand the receptive field. FCN, another important segmentation network, performs pixel-wise segmentation, making it widely applicable to medical image analysis. Despite their strong performance in many tasks, these models struggle with medical images featuring complex lesion shapes and blurred boundaries. As a result, Transformer-based models, such as SwinUnet [6] and TransUNet [7], have emerged as promising alternatives.

Despite the success of these segmentation networks in various applications, they still have some limitations. Many models use simple addition or concatenation operations in the feature fusion stage, which can introduce redundant information, thus weakening the representation of key features. Further, Transformer-based models tackle CNNs' limited receptive field issue, they still need further exploration in enhancing the ability to extract features from complexly shaped lesions in the encoder part.

To tackle these challenges, this study proposes a medical segmentation method combining a ResNet34 encoder, dilated convolution bridge, CBAM [8] feature enhancer, and multi-scale fusion decoder. The CBAM module improves feature representation by applying channel and spatial attention before and after feature fusion. The multi-scale feature fusion decoder effectively handles lesion regions of varying sizes and shapes, improving segmentation precision and robustness. Also, the dilated convolution bridge expands the receptive field, thereby improving context capture capability. Finally, a multi-scale training and single-scale inference strategy ensures training effectiveness and optimizes inference efficiency for real-time clinical use.

The main contributions of this research are as follows:

This work was supported by the Beijing Natural Science Foundation under Grants L233005 and 4232049, the National Natural Science Foundation of China under Grants 62173013 and 62473014, and in part by Beihang World TOP University Cooperation Program (Corresponding author: Jing Bi)

1) A CBAM-based feature fusion strategy is proposed, enhancing segmentation precision through channel and spatial attention mechanisms.
2) A multi-scale feature fusion decoder is designed to effectively address lesions with varying sizes and shapes.
3) A dilated convolution bridging module is introduced to expand the receptive field, improving the model's ability to capture contextual information.
4) A multi-scale training and single-scale inference strategy is proposed, improving inference efficiency while maintaining segmentation accuracy.

These innovations enhance both segmentation accuracy and inference speed, enabling real-time clinical application.

II. Related Work

Medical image segmentation, a key computer vision application in healthcare, has seen significant deep learning advances. Existing methods are categorized by architecture and optimization strategies as follows:

A. CNN-Based Medical Image Segmentation

1) CNN Architecture Optimization: UNet [2], as a benchmark model for medical image segmentation, adopts a symmetric encoder-decoder structure and employs skip connections to achieve multi-scale feature fusion, demonstrating outstanding performance in cell and organ segmentation tasks. U2Net [9] further enhances segmentation precision for pathological slices (e.g., neuroimaging) by introducing a dual U-shaped structure with residual connections and attention mechanisms. 3DFCN [10] employs a two-stage, coarse-to-fine strategy, improving the Dice score for liver, spleen, and pancreas segmentation from 68.5% to 82.2%. Additionally, the modified DeepLabv3+ [11] integrates atrous convolution residual networks for multi-scale feature extraction, achieving a best Dice coefficient of 0.95 in colon polyp segmentation. PSPNet [12], utilizing a pyramid scene parsing structure, has achieved a segmentation accuracy of 0.9865 in prostate MRI segmentation tasks.

2) Attention Mechanism Optimization: Recently, attention mechanisms have become increasingly prominent in medical segmentation. SwinUnet [6], which adopts a Transformer-based U-shaped structure, excels in multi-organ and cardiac segmentation tasks. SINet [13], based on the YOLOv5 framework, incorporates the GAM attention mechanism to enhance feature extraction and employs a semantic segmentation head for parallel detection and segmentation, achieving 97.9% mean accuracy (mAP). OCENet [14] introduces a cross-attention feature fusion (CaFF) module to optimize skip connections and integrates a dual-branch pooling fusion (DBPF) module to reduce spatial information loss, significantly improving segmentation performance in vascular segmentation tasks.

B. Medical Image Segmentation Combining Frequency and Spatial Features

To further improve segmentation accuracy, researchers have explored the joint modeling of frequency-domain information and spatial features. FEUNet [15] is built on SAM2 as the backbone and employs HieraLarge as the pretraining block, proposes a wavelet-guided spectral pooling module (WSPM) to enhance and balance frequency-domain features. Additionally, a frequency-enhanced receptive field block (FERFB) is introduced to extract rich frequency-domain information, ensuring high segmentation precision while enhancing generalization capability. GFUNet [16] combines Fourier transform with the UNet structure to optimize the efficiency of the encoding and decoding processes. FDFUNet [17] employs a channel attention-enhanced network to model inter-channel feature relationships, integrating depthwise separable convolutions and frequency-domain filters to improve high-order multi-scale feature extraction.

C. Transformer-Based Medical Image Segmentation

Transformer-based medical image segmentation has made significant progress in hybrid architecture innovations and lightweight optimizations.

1) Hybrid Architectures: Recently, Transformers have excelled in medical image segmentation tasks. SwinUnet [6] and TransUNet [7] both adopt Transformer-based U-shaped architectures, excelling in multi-organ and cardiac segmentation tasks. TransUNet combines a Transformer encoder with the UNet structure to restore local spatial information and enhance fine-grained segmentation. MedSAM [18], built upon the Segment Anything Model (SAM), introduces Sync-SAM with a synchronized dual-branch encoder to enhance medical image encoding capability while incorporating a multi-scale dual-branch decoder to preserve image details, achieving remarkable results in zero-shot medical image segmentation tasks. DFormer [19] proposes a local-global range module and dynamic positional encoding, reaching 92.29% Dice in 3D cardiac segmentation. UTNetV2 [20] integrates bidirectional Transformer blocks with depthwise separable convolutions, achieving a Dice score of 92.14% for 3D inputs.

2) Lightweight Model Optimization: Alongside improving segmentation accuracy, researchers have also explored lightweight optimization strategies. H2Former [21] combines CNN, multi-scale channel attention, and Transformer advantages to propose an efficient hierarchical hybrid vision Transformer. It improves segmentation accuracy by 2.29% compared to TransUNet while reducing parameter count to 30.77%. IMedSAM [22] incorporates implicit neural representation (INR) and adopts an uncertainty-guided sampling strategy to further enhance medical image segmentation. AFTerUNet [23] employs an axial fusion mechanism, achieving 92.32% Dice on the Thorax85 dataset with only 41.5M parameters. EGUnet [24] introduces an edge-guided module that fuses all-layer features and optimizes concatenation and skip connections in later stages to effectively address fuzzy boundary issues in medical image segmentation. LeViTUNet [25] replaces positional encoding with attention bias, improving Dice to 78.53% on the Synapse dataset. UCTransNet [26] enhances skip connections with a channel-cross fusion Transformer, achieving a 2% Dice improvement over SwinUnet while reducing parameters by 20%.

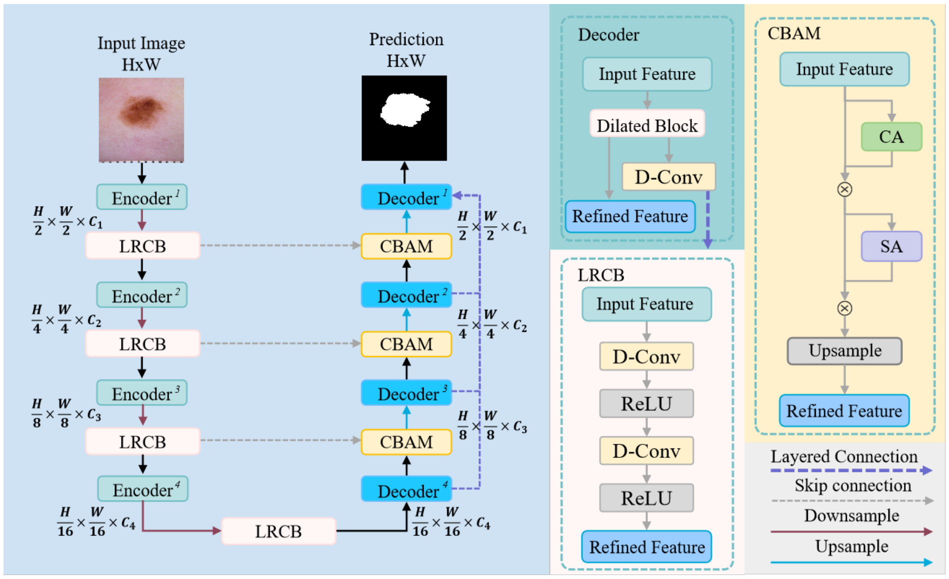


Fig. 1: Overview of the proposed MAR-NET architecture.

III. MAR-Net Architecture

MAR-Net seeks to enhance medical segmentation accuracy and robustness by combining attention mechanisms, multi-scale feature fusion, and contextual modeling. As in Fig. 1, MAR-Net uses an encoder-decoder framework with key module optimizations to boost feature representation. Specifically, the architecture employs a pre-trained ResNet-34 as the encoder, combined with CBAM-based feature enhancement and a multi-scale decoding strategy to effectively capture the boundary information of target regions and global semantic information.

A. Lightweight Residual Convolution Block

To improve computational efficiency while maintaining feature extraction capabilities, MAR-Net introduces the Lightweight Residual Convolution Block (LRCB). This module swaps standard convolutions for depthwise separable ones, sharply cutting parameters and computation.

LRCB consists of two depthwise separable convolutions, each followed by Batch Normalization and ReLU activation. The feature transformation process is formulated as:

$$\begin{aligned} F' &= \mathrm{ReLU}\big(\mathrm{BN}(\mathrm{Conv}^{\mathrm{dw}}_{3\times3}(F_{\mathrm{in}}))\big) \\ F'' &= \mathrm{BN}\big(\mathrm{Conv}_{1\times1}(F')\big) \\ F_{\mathrm{LRCB}} &= \mathrm{ReLU}(F'' + F_{\mathrm{res}}) \end{aligned} \tag{1}$$

where $\mathrm{Conv}^{\mathrm{dw}}_{3\times3}$ represents a depthwise convolution, and $\mathrm{Conv}_{1\times1}$ is a pointwise convolution for channel-wise feature integration. The residual connection ensures stable gradient propagation.

The output of LRCB follows two branches:

1) **Main branch**: Passes enhanced features to the next stage for further processing.
2) **Residual branch**: Preserves spatial details and stabilizes gradient flow.

By replacing standard convolutions with depthwise separable convolutions, LRCB reduces computation by approximately $\frac{1}{C_{\mathrm{out}}}$ while maintaining segmentation accuracy, making it ideal for lightweight medical image analysis.

B. Encoder

The encoder of MAR-Net adopts ResNet-34 as the backbone, leveraging multiple residual connections to ensure strong feature representation. For efficiency, all standard convolutions are substituted with depthwise separable ones, reducing complexity without sacrificing performance. The encoder consists of four stages (layer1 to layer4), each employing stride convolution for spatial downsampling. This progressively reduces the feature map size while increasing channel depth, enabling more discriminative feature extraction. A final downsampling operation reduces the feature map size to $1/32$ of the original input, enhancing high-level feature representation.

The computation is as follows:

$$H_{\mathrm{out}} = \frac{H_{\mathrm{in}}}{2}, \quad W_{\mathrm{out}} = \frac{W_{\mathrm{in}}}{2}, \quad C_{\mathrm{out}} = 2C_{\mathrm{in}} \tag{2}$$

C. Attention-Based Feature Fusion

To focus on key regions, MAR-Net incorporates CBAM in decoding for feature selection/enhancement. CBAM handles multi-scale features from residual/decoder pathways via channel (CA) and spatial (SA) attention.

a) Channel Attention (CA): CA calculates feature weights via global average pooling (GAP) and max pooling (GMP):

$$M_c = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))) \tag{3}$$

where M_c is the channel attention map, σ is the sigmoid function.

b) Spatial Attention (SA): SA computes spatial attention weights using a 7×7 convolution:

$$M_s = \sigma(\text{Conv}_{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])) \quad (4)$$

M_s is the spatial attention weight. After spatial attention computation, CBAM applies an upsampling operation to restore feature map resolution and refine segmentation results:

$$H_{\text{out}} = 2H_{\text{in}}, \quad W_{\text{out}} = 2W_{\text{in}} \quad C_{\text{out}} = \frac{1}{2}C_{\text{in}} \quad (5)$$

D. Multi-Scale Decoder

MAR-Net's decoder integrates multi-scale features to restore spatial resolution and improve segmentation accuracy. The core components include:

- **Dilated Block**: Fuse shallow edge features with deep semantic features to enhance fine-grained segmentation.
- **Pyramid Pooling**: Extracts global context information via multi-scale pooling to improve adaptability to different target sizes.
- **Layered Feature Transmission**: Different-scale encoded features feed into decoder^4 via hierarchical connections to preserve high-res info during decoding.

E. Multi-Scale Training and Loss Function

To further enhance generalization, MAR-Net employs a multi-scale training strategy, where feature maps from different encoder layers contribute to loss computation, optimizing feature representation. Specifically, the loss is computed as follows:

$$\begin{aligned}\mathcal{L} = \lambda_1\text{BCE}(F_1, G) + \lambda_2\text{BCE}(F_2, G) + \\ \lambda_3\text{BCE}(F_3, G) + \lambda_4\text{BCE}(F_4, G)\end{aligned} \quad (6)$$

where:

- F_1, F_2, F_3, F_4 are predicted feature maps at different scales,
- G is the ground truth,
- λ_i are weight coefficients for different scale losses.

Compared to traditional single-scale supervision, this strategy offers several advantages:

1) Enhances hierarchical feature modeling, ensuring optimal utilization of both low-level details and high-level semantics.
2) Improves training stability, mitigating gradient vanishing issues and allowing multi-scale features to contribute effectively to loss optimization.
3) Enhances segmentation accuracy, ensuring robust detection of targets at different scales, which is particularly beneficial for medical images with varying object sizes.

F. Conclusion

MAR-Net optimizes medical image segmentation by integrating a ResNet-34 encoder with depthwise separable convolutions for efficiency and downsampling, a Lightweight Residual Convolution Block to enhance feature extraction capability, a CBAM attention mechanism for improved channel and spatial feature selection, and a multi-scale decoder for fine feature reconstruction. Additionally, MAR-Net employs a multi-scale training strategy combined with a multi-BCE loss fusion method, improving the model's adaptability to varying target sizes and fully leveraging multi-level feature information, thereby enhancing segmentation accuracy and generalization performance.

IV. Experiments

A. Datasets

To evaluate the method, experiments used two public datasets: the ISIC dataset (for skin cancer diagnosis) and the CholecSeg8K dataset (for laparoscopic surgery). Table I summarizes the dataset splits and key statistics.

1) ISIC Dataset: The ISIC dataset [28] serves as a benchmark for skin cancer diagnosis, comprising 1279 dermoscopic images with expert annotations, which are split into 621 training, 279 validation, and 379 test images. Each image is labeled for lesion types (e.g., melanoma, nevus) and malignancy levels. The original images exhibit significant variations in resolution (ranging from 722×542 to 4288×2848 pixels) and illumination conditions, posing challenges for model robustness under diverse imaging conditions.

2) CholecSeg8K Dataset: The CholecSeg8K dataset [27] builds upon the Cholec80 laparoscopic surgery dataset, designed for semantic segmentation of laparoscopic cholecystectomy images. It includes 8,080 frames from 17 video sequences, with pixel-level annotations for 13 classes (e.g., liver, gallbladder, surgical instruments). The dataset splits into 674 training, 78 validation, and 28 test images. All images have a fixed resolution of 854×480 pixels and cover diverse surgical scenarios, including tissue texture variations and occlusions by surgical instruments.

B. Implementation Details

All experiments were conducted on the PyTorch platform using multiple NVIDIA 3090 GPUs (24GB) with CUDA 12.1. Seven models trained for 500 epochs on both datasets, with best results chosen pre-overfitting. Initial lr=1×10^{-4}, decay=0.9. IoU and Dice were primary metrics.

C. Comparative Analysis

Fig. 2a and 2b illustrate the training and validation performance of six models on the ISIC and CholecSeg8K datasets, respectively. The visualization results of semantic segmentation masks for various models among different datasets are shown in 3a and 3b.These results depict the evolution of the IoU and Dice metrics across different epochs. Although all models were trained for 500 epochs, the analysis focuses on the best-performing results prior to overfitting, providing insights into each model's learning capacity and convergence efficiency.

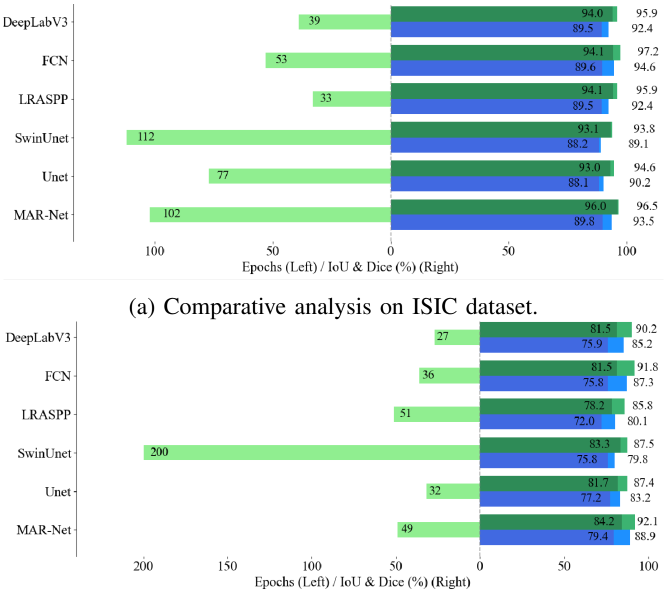
Fig. 2a illustrates the performance of MAR-Net on the ISIC skin lesion segmentation task. At epoch 125, MAR-Net achieved peak performance with training IoU and Dice of 0.935 and 0.965 and validation scores of 0.898 and 0.960, showing strong generalization. The variant MAR-Net1 (without certain augmentations) peaked at epoch 102 with

TABLE I: Dataset Classification and Training Splits

Dataset	Total	Train	Validation	Classes	Application
ISIC	1,279	895	255	129	Dermoscopic Diagnosis
CholecSeg8K	8,080	5656	1616	808	Laparoscopic Surgery

Data sources: ISIC 2017 Challenge [27], CholecSeg8K [28]

training IoU and Dice of 0.915 and 0.953, and validation scores of 0.894 and 0.939, slightly lower than MAR-Net. This suggests that augmentation strategies contribute to improved generalization. Among conventional segmentation models, U-Net (77 epochs) and Swin-Unet (112 epochs) performed well, achieving training IoU scores of 0.902 and 0.891, with validation IoU scores of 0.881 and 0.882, respectively, highlighting their stability in this task. The lightweight model LRASPP (33 epochs) demonstrated efficient learning, achieving high training and validation IoU scores of 0.924 and 0.895 within a short training period. Additionally, FCN and DeepLabV3, peaking at 53 and 39 epochs, respectively, exhibited strong generalization, both attaining validation IoU scores around 0.895.

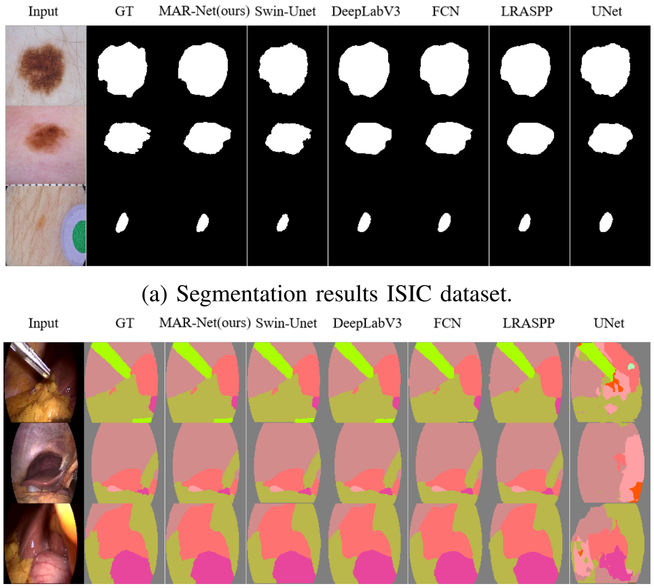


(a) Comparative analysis on ISIC dataset.

(b) Comparative analysis on CholecSeg8K dataset.

Fig. 2: Comparison of seven models on ISIC and CholecSeg8K datasets, showcasing an overview of training and validation accuracy across epochs.

Fig. 2b presents the results on the CholecSeg8K endoscopic surgery segmentation task. Unlike the ISIC dataset, MAR-Net exhibited a significantly different performance trend, achieving rapid convergence in just 22 epochs. It obtained training IoU and Dice scores of 0.840 and 0.884, with validation IoU and Dice scores of 0.800 and 0.847. MAR-Net-L (49 epochs), a variant of MAR-Net without multi-scale training and loss computation strategies, exhibited a higher training IoU (0.8886) but a lower validation IoU (0.794) than MAR-Net, indicating that these strategies enhance generalization in CholecSeg8K. U-Net (32 epochs) and Swin-Unet (200 epochs) reached training IoU scores of 0.832 and 0.798, with validation IoU scores of 0.772 and 0.758, respectively, suggesting that Swin-Unet requires extended training for effective feature learning in this task. FCN (36 epochs) and DeepLabV3 (27 epochs) achieved high training IoU scores (0.873 and 0.852, respectively), but their validation IoU scores remained below 0.76, indicating weaker generalization compared to their performance on the ISIC dataset.



(a) Segmentation results ISIC dataset.

(b) Segmentation results of CholecSeg8K dataset.

Fig. 3: Visual comparison of segmentation results on ISIC and CholecSeg8K datasets.

Overall, MAR-Net and its variants achieved superior performance on ISIC, while MAR-Net demonstrated strong generalization on CholecSeg8K due to its multi-scale training strategy and loss computation method. The optimal epoch selection for each model revealed differences in learning efficiency, with some models such as LRASPP and MAR-Net reaching high performance within fewer epochs, whereas models like Swin-Unet and U-Net required longer training to achieve optimal feature learning. These results highlight the influence of model architecture, training strategies, and enhancement techniques on generalization performance across different segmentation tasks.

V. Conclusion

MAR-Net is proposed as a medical segmentation model that improves performance via joint optimization of key components. By integrating CBAM attention, multi-scale feature fusion, and efficient training strategies, MAR-Net enhances segmentation accuracy and generalization. Experimental results on ISIC and CholecSeg8K datasets demonstrate its superiority over mainstream methods, achieving higher accuracy and recall, confirming its effectiveness in complex segmentation tasks.

References

[1] Cong Wang, Witold Pedrycz, ZhiWu Li, and MengChu Zhou. Residual-driven fuzzy c-means clustering for image segmentation. *IEEE/CAA Journal of Automatica Sinica*, 8(4):876–889, 2021.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[5] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[6] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[7] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[8] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[9] Jie Shao, Kun Zhou, Ye-Hua Cai, and Dao-Ying Geng. Application of an improved u2-net model in ultrasound median neural image segmentation. *Ultrasound in Medicine & Biology*, 48(12):2512–2520, 2022.

[10] Holger R Roth, Hirohisa Oda, Xiangrong Zhou, Natsuki Shimizu, Ying Yang, Yuichiro Hayashi, Masahiro Oda, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. An application of cascaded 3d fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics*, 66:90–99, 2018.

[11] Shweta Gangrade, Prakash Chandra Sharma, Akhilesh Kumar Sharma, and Yadvendra Pratap Singh. Modified deeplabv3+ with multi-level context attention mechanism for colonoscopy polyp segmentation. *Computers in Biology and Medicine*, 170:108096, 2024.

[12] Lingfei Yan, Dawei Liu, Qi Xiang, Yang Luo, Tao Wang, Dali Wu, Haiping Chen, Yu Zhang, and Qing Li. Psp net-based automatic segmentation network model for prostate magnetic resonance imaging. *Computer Methods and Programs in Biomedicine*, 207:106211, 2021.

[13] Zhenzhong Liu, Yifan Zhou, Laiwang Zheng, and Guobin Zhang. Sinet: A hybrid deep cnn model for real-time detection and segmentation of surgical instruments. *Biomedical Signal Processing and Control*, 88:105670, 2024.

[14] Tian Feng, Jiaheng Wang, Junao Shen, Qiangguo Jin, Zhiyuan Zhu, and Xinyu Wang. Retinal vessel segmentation via cross-attention feature fusion. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[15] Guohao Huo, Ruiting Dai, Ling Shao, and Hao Tang. Fe-unet: Frequency domain enhanced u-net with segment anything capability for versatile image segmentation. *arXiv preprint arXiv:2502.03829*, 2025.

[16] Penghui Li, Rui Zhou, Jin He, Shifeng Zhao, and Yun Tian. A global-frequency-domain network for medical image segmentation. *Computers in Biology and Medicine*, 164:107290, 2023.

[17] Yufeng Chen, Xiaoqian Zhang, Lifan Peng, Youdong He, Feng Sun, and Huaijiang Sun. Medical image segmentation network based on multi-scale frequency domain filter. *Neural Networks*, 175:106280, 2024.

[18] Sihan Yang, Haixia Bi, Hai Zhang, and Jian Sun. Sam-unet: Enhancing zero-shot segmentation of sam for universal medical images. *arXiv preprint arXiv:2408.09886*, 2024.

[19] Yixuan Wu, Kuanlun Liao, Jintai Chen, Jinhong Wang, Danny Z Chen, Honghao Gao, and Jian Wu. D-former: A u-shaped dilated transformer for 3d medical image segmentation. *Neural Computing and Applications*, 35(2):1931–1944, 2023.

[20] Jiacheng Ruan, Suncheng Xiang, Mingye Xie, Ting Liu, and Yuzhuo Fu. Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1150–1156. IEEE, 2022.

[21] Along He, Kai Wang, Tao Li, Chengkun Du, Shuang Xia, and Huazhu Fu. H2former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(9):2763–2775, 2023.

[22] Xiaobao Wei, Jiajun Cao, Yizhu Jin, Ming Lu, Guangyu Wang, and Shanghang Zhang. I-medsam: Implicit medical image segmentation with segment anything. In *European Conference on Computer Vision*, pages 90–107. Springer, 2024.

[23] Xiangyi Yan, Hao Tang, Shanlin Sun, Haoyu Ma, Deying Kong, and Xiaohui Xie. After-unet: Axial fusion transformer unet for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3971–3981, 2022.

[24] Xiufeng Zhang, Yansong Liu, Shengjin Guo, and Zhao Song. Eg-unet: Edge-guided cascaded networks for automated frontal brain segmentation in mr images. *Computers in Biology and Medicine*, 158:106891, 2023.

[25] Guoping Xu, Xuan Zhang, Xinwei He, and Xinglong Wu. Levit-unet: Make faster encoders with transformer for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 42–53. Springer, 2023.

[26] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2441–2449, 2022.

[27] Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. Analysis of the isic image datasets: Usage, benchmarks and recommendations. *Medical image analysis*, 75:102305, 2022.

[28] W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.