Proceedings of CONF-FMCE 2025 Symposium: Semantic Communication for Media Compression and Transmission
DOI: 10.54254/2755-2721/2025.GL24843

Image Recognition and Enhancements of the U-Net Model

Zhaoxuan Liu

Naval Architecture and Ocean Engineering College, Dalian Maritime University, Dalian, China
lzx030328@outlook.com

Abstract: Image segmentation, a core computer vision task, aims to partition digital images into semantically distinct regions. To accomplish these tasks, the U-Net architecture—a deep learning neural network enabling automated, high-precision segmentation—was developed. While U-Net derivatives have expanded into interdisciplinary domains like remote sensing, meteorological monitoring, agricultural disease detection, and geological exploration, the original architecture no longer satisfies the precision demands of modern medical segmentation and remote sensing applications, yet unexplored architectural innovations offer further potential for maximizing segmentation precision. This study investigates improvements to the baseline U-Net model using a dermatological image dataset, conducting rigorous metric evaluation to advance segmentation accuracy. The study employed U-Net's fundamental encoder-decoder convolutional structure. Primary innovations included implementation of comprehensive data augmentation on the original dataset and Integration of the Convolutional Block Attention Module (CBAM) into the model architecture to enhance robustness and performance. The experimental procedure included baseline evaluation, the standard U-Net model was executed on dermatological lesion imagery, yielding suboptimal binary segmentation masks as quantified by evaluation metrics; then enhanced methodology, applied data augmentation to improve dataset robustness and incorporated CBAM attention mechanisms to enhance focus on ambiguous boundary regions. Comparative analysis of both approaches demonstrated significant improvements in critical metrics (mIoU, Loss) for the augmented and attention-enhanced model. Through controlled comparison between standard and refined U-Net architectures, this research empirically validates that targeted enhancements—specifically data augmentation and CBAM integration—substantially elevate segmentation precision and enhance model robustness. These contributions represent notable innovations in U-Net-based image segmentation methodologies.

Keywords: U-Net, Image Recognition, CBAM, data Aug, mIoU

1. Introduction

Image segmentation is a fundamental task in the computer vision, aiming to partition digital images into multiple semantically meaningful regions, thereby providing structured information representation for image understanding. This technology plays an indispensable role in diverse fields such as medical diagnosis, remote sensing monitoring, environmental perception, and industrial inspection. Particularly within the medical image analysis, the precise segmentation of lesions and identification of organs directly impact the accuracy of clinical decision-making. However, image

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.GL24843

segmentation consistently faces significant challenges, including the ambiguity of object boundaries, interference from complex backgrounds, scale diversity, and the scarcity of high-quality annotated data, and these challenges are especially pronounced in the domain of medical imaging [1, 2].

Traditional image segmentation methods primarily relied on algorithms such as threshold segmentation, edge detection, and region growing. These approaches, however, often require manually engineered features that are prone to inaccuracy and inefficiency, exhibit sensitivity to noise, and possess limited generalization capabilities. The advent of deep learning ushered in new pathways for image segmentation through Convolutional Neural Networks (CNNs). Nevertheless, Fully Convolutional Networks (FCNs) demonstrated notable limitations in spatial detail recovery, as their coarse upsampling processes struggled to reconstruct fine object boundaries. In 2015, Ronneberger introduced the U-Net network, building upon the FCN architecture. Its innovative U-shaped symmetric topology, featuring an encoder-decoder structure coupled with skip connections, enabled the deep fusion of multi-scale features. Achieving superior performance in the ISBI cell tracking challenge, U-Net marked the advent of a new era in medical image segmentation [1].

The core breakthrough of U-Net lies in its biologically inspired design. The encoder simulates the feature abstraction process of the human visual system, progressively extracting semantic information through hierarchical convolutions and pooling. The decoder restores spatial resolution via upsampling operations. Crucially, the skip connections integrate shallow localization information with deep semantic features, effectively overcoming the bottleneck of detail reconstruction inherent in traditional FCNs. This architecture exhibits exceptional robustness in small-sample scenarios, rapidly establishing U-Net as the benchmark model for medical image segmentation. Subsequent research has extended U-Net-derived architectures to interdisciplinary domains, including remote sensing image analysis, meteorological monitoring (such as aerosol vertical structure identification), agricultural disease detection, and geological exploration [3-5].

However, the traditional U-Net structure can no longer meet the escalating accuracy requirements of modern applications such as medical segmentation and remote sensing image recognition. Enhanced variants like U-Net++, architectures incorporating attention mechanisms, and techniques employing sophisticated data augmentation strategies are increasingly prevalent, yielding more precise results. Despite these advances, numerous novel architectural concepts remain underutilized, and the ultimate precision potential of image segmentation has yet to be fully realized, necessitating continued exploration and research. This study will focus on investigating improvements (including CBAM Attention-U-Net and data enhancement) to the traditional U-Net model using a dermatological image dataset, conducting exploratory research and performance metric evaluation (including mIoU, mPrecision, mPA and mRecall) to enhance the segmentation accuracy of the U-Net model, thereby aiming to advance the field of image segmentation.

2. Method

2.1. Theoretical basis

The fundamental architecture of UNet employs an encoder-decoder design, and the specific configuration is illustrated in Figure 1. The encoder is composed of repeated convolutional layers and downsampling layers, progressively extracting high-level semantic features. Conversely, the decoder utilizes transposed convolutions or deconvolutions to incrementally restore spatial resolution, culminating in a pixel-level classification output. A key innovation of this structure lies in the introduction of skip connections. These connections perform channel-wise concatenation of the feature maps from corresponding stages in the encoder with those in the decoder. This mechanism enables the network to simultaneously leverage the precise localization capabilities inherent in the shallow layers and the high-level semantic understanding captured within the deeper layers [1]. This

DOI: 10.54254/2755-2721/2025.GL24843

design proves particularly valuable in medical imaging, where it significantly enhances the segmentation accuracy of small target structures, such as blood vessels and pathological cells [1].

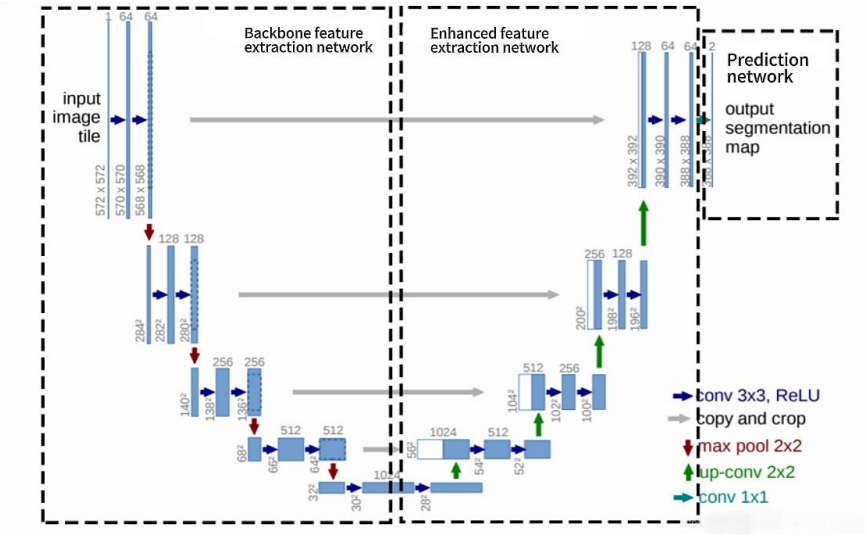


Figure 1: The structure of the U-Net [6]

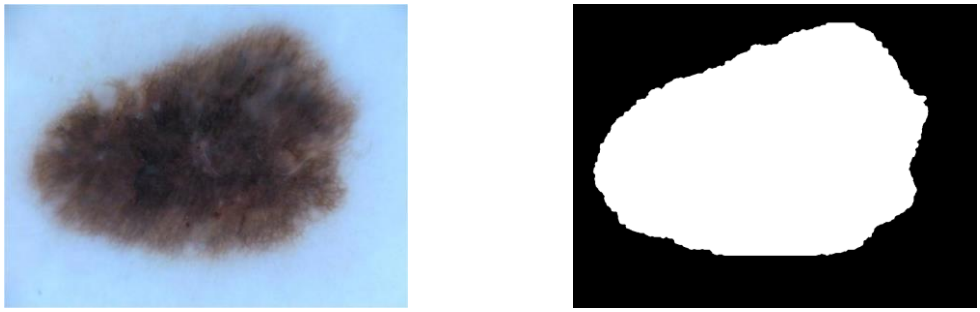
The Convolutional Block Attention Module (CBAM) employs a dual-attention mechanism—channel attention and spatial attention—to dynamically recalibrate the weight distribution of feature maps. This process enhances the model's focus on salient regions within the input data [7]. Furthermore, CBAM facilitates the extraction of multi-scale feature representations. Critically, its spatial attention sub-module, utilizing sequential global pooling operations followed by convolutional layers, demonstrates robust efficacy in identifying regions characterized by edge ambiguity or inadequate contrast [8]. Consequently, the integration of the CBAM mechanism within the U-Net architecture, via dynamic feature recalibration across both channel and spatial dimensions, yields statistically significant improvements in segmentation performance under these challenging conditions. The principal advantages are quantifiably demonstrated through enhanced segmentation accuracy, evidenced by reductions in loss metrics and elevated Intersection over Union (IoU) scores, along with improvements in parameter efficiency and training stability.

2.2. Research process

2.2.1. Dataset acquisition and preparation

During the research process, a foundational dataset comprising over 1,200 dermatological lesion images was acquired from the Kaggle platform [9]. This substantial sample size was selected to ensure dataset sufficiency and enhance the robustness of the subsequent results. Each image underwent manual segmentation and annotation using the labelme software tool, generating corresponding JSON files delineating the lesion boundaries. These JSON annotations were then converted into pixel-wise label maps by executing the corresponding script, resulting in corresponding PNG format label images. Within this annotation schema, the defined classes were background and sick, representing the healthy background and pathological regions, respectively. Representative examples of the original images and their corresponding label maps are illustrated in Figure 2 (a) and Figure 2 (b).

DOI: 10.54254/2755-2721/2025.GL24843



(a) (b)

Figure 2: (a) Representative image; (b) label image [10]

2.2.2. Data augmentation

To further bolster result robustness, the original JPEG images and the generated PNG label maps were subjected to a comprehensive data augmentation pipeline following the manual annotation phase. This pipeline incorporated stochastic geometric transformations (including random rotation and flipping) and photometric transformations (such as contrast enhancement and Gaussian blurring). This process yielded an augmented dataset with increased diversity and size, thereby improving model generalization capabilities.

2.2.3. Training processing

In the training pipeline configuration, following dataset preparation, the corresponding script was executed to automatically generate distinct training and validation sets in TXT format. The corresponding ratio parameter was configured to establish a 9:1 ratio between the training and validation data partitions. Subsequently, the `train.py` script was employed to initiate model training using both datasets. Training utilized RGB-format original images alongside 8-bit color-indexed label maps, adhering to the Pascal VOC data format. Input images possessed dimensions of [512, 512, 3] (height × width × channels), while corresponding labels had dimensions of [512, 512]. The model architecture employed VGG16 as the backbone network with classes equaling 2. Additionally, in the weight initialization strategy, to ensure effective feature extraction and mitigate the detrimental effects of excessively randomized initialization in the backbone network, pre-trained weights were loaded from the logs path binary file. This strategy eliminated the need for separate pre-training. During it, the training regimen comprised two distinct phases, including a freezing phase (initial layers fixed) followed by an unfreezing phase (full network fine-tuned). Model weights were saved every 5 epochs, and a comprehensive evaluation was conducted similarly every 5 epochs. This balanced approach ensured precise performance monitoring while optimizing computational efficiency. Importantly, a hybrid loss function, integrating Dice Loss and Focal Loss, was implemented to address key segmentation challenges. Dice Loss ensured global structural similarity by equally weighting all pixels, thereby preventing small targets from being overwhelmed [8]; Focal Loss optimized the learning for hard-to-classify pixels, particularly local boundaries, by assigning higher weights to low-confidence, small-target pixels [10]. Synergistically, this combination enhanced robustness against class imbalance and boundary ambiguity. Crucially, Focal Loss acted as a stabilizer, mitigating the gradient oscillation inherent in Dice Loss under extreme predictions, resulting in a more stable training trajectory. Upon completion of training, the process yielded 20 epoch-specific weight files, alongside `best` and `last` weight files, providing ready-to-use model checkpoints for future U-Net inference and prediction tasks, also yielded mean Intersection over Union (mIoU) and loss curves, generated for both the original and augmented datasets, accompanied

DOI: 10.54254/2755-2721/2025.GL24843

by real-time TXT log files which serve as quantitative metrics for comparative analysis and model evaluation.

2.2.4. U-net predicting process

Subsequently, processing the prediction step, the `predict.py` script was executed to generate pure binary segmentation masks (black-and-white) for each input image. During this inference process, zero-padding was dynamically applied to maintain the original aspect ratio of the input images. This padding was subsequently removed from the output masks to ensure dimensional fidelity and prevent any alteration to the final prediction results. Additionally, to augment the prediction robustness of the neural network architecture, a Convolutional Block Attention Module (CBAM) was integrated into the U-Net framework in the study. Lastly, following model inference, the file paths of all input images were specified via the terminal command-line interface. This facilitated the batch generation of pure binary PNG prediction masks corresponding to each original JPEG input image. Representative examples of these output segmentation results are presented in Figure 3.



Figure 3: Output binary image (photo credit: original)

The computational experiments were conducted within a dedicated Anaconda environment configured for PyTorch development. This environment incorporated GPU-accelerated computing support via CUDA and cuDNN libraries. The software components and Python packages utilized are shown in the following Table 1.

Table 1: APP version

Component	Version	Category
Anaconda	(Platform)	Environment Management
labelme	3.16.7	Annotation Tool
Visual Studio Code	(IDE)	Development Environment
PyTorch(torch)	(GPU-enabled)	Deep Learning Framework
torchvision	-	Computer Vision Library
tensorboard	-	Visualization Toolkit
scipy	1.2.1	Scientific Computing
numpy	1.17.0	Numerical Operations
matplotlib	3.1.2	Data Visualization
opencv-python	4.1.2.30	Image Processing
tqdm	4.60.0	Progress Monitoring
Pillow	8.2.0	Image Handling
h5py	2.10.0	HDF5 Data Format Support

DOI: 10.54254/2755-2721/2025.GL24843

3. Evaluation index

Key Evaluation Metrics value for Image Segmentation Performance are very important such as Mean Intersection over Union (mIoU), Loss, Mean Pixel Accuracy (mPA), Mean Precision (mPrecision), and Mean Recall (mRecall) constitute critical quantitative indicators for assessing the segmentation efficacy of U-Net models. These metrics provide multidimensional perspectives on predictive performance and are essential for rigorous data analysis and result evaluation, and this study also took them as evaluation metrics.

Within these, IoU quantifies the degree of overlap between predicted segmentation masks and ground truth annotations, directly measuring segmentation accuracy at the pixel level, and mIoU represents the mean IoU value across all semantic classes, providing a comprehensive assessment of model performance on each category. A higher mIoU signifies superior segmentation quality, indicating greater alignment between predictions and manually verified labelme annotations; The Loss metric quantifies the discrepancy between model predictions and ground truth labels, serving as a fundamental indicator of predictive fidelity. During training, the loss function provides directional feedback to optimize model parameters by identifying areas requiring improvement; mPA computes the average pixel-wise accuracy per class, calculated as the ratio of correctly classified pixels to total pixels within each category, then averaged across classes. This metric is particularly relevant for class-balanced segmentation tasks, reflecting the model's per-class classification precision. Precision measures the proportion of correctly identified pixels among all pixels predicted as belonging to a specific class. Therefore, mPrecision denotes the class-averaged precision, evaluating the model's prediction reliability. this metric is crucial in class-imbalance scenarios, revealing the trustworthiness of positive class predictions; Recall calculates the proportion of actual positive pixels correctly detected by the model, indicating its sensitivity to target regions. mRecall represents the class-averaged recall, assessing the model's ability to identify all relevant positive samples (e.g., lesion boundaries). It is vital for evaluating coverage completeness in class-imbalance contexts, particularly for avoiding critical false negatives.

Comprehensively, in practice, class-wise averages of Precision, Recall, and Pixel Accuracy yield mPrecision, mRecall, and mPA, respectively. These aggregated metrics furnish a comprehensive, class-agnostic evaluation of overall model performance, enabling robust comparative analysis across segmentation tasks. During the experimental phase, a comparative analysis was conducted between two distinct methodological configurations:

Baseline Configuration: Utilizing the original dataset without data augmentation and employing the standard U-Net architecture without CBAM integration.

Enhanced Configuration: Employing the augmented dataset and incorporating the CBAM attention mechanism within the U-Net framework.

Both configurations were executed under identical training protocols. This comparative approach yielded grouped fluctuation profiles for the mIoU (Mean Intersection over Union) and Loss (including train and val) metrics across the two experimental conditions. These diagnostic curves, illustrating metric convergence and stability throughout training, are presented in Figures 4 (a) through 4 (d).

DOI: 10.54254/2755-2721/2025.GL24843

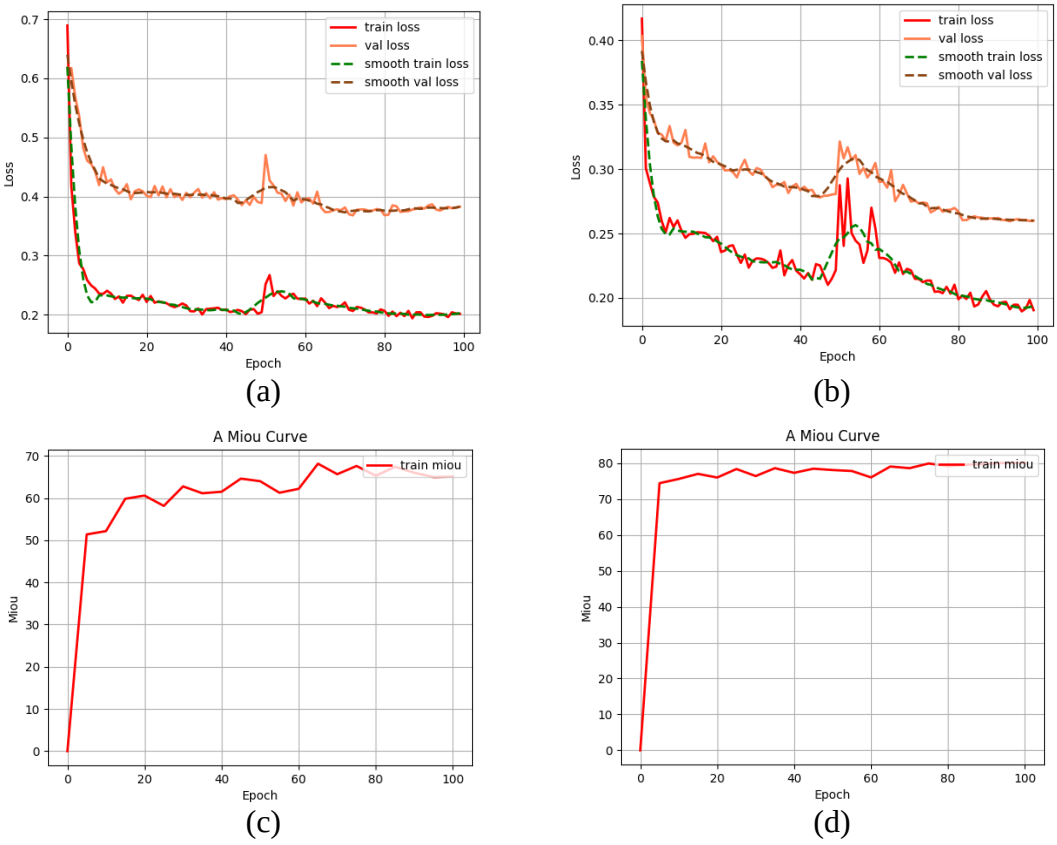


Figure 4: (a) Base loss; (b) Aug loss; base mIoU; (d) Aug mIoU (photo/picture credit: original)

As evidenced by the diagnostic curves (Figures 4a-4d), through the quantitative analysis of performance improvements, the integration of data augmentation and the CBAM attention mechanism yielded statistically significant performance enhancements across all evaluated metrics:

1. Loss Metrics: Training Loss decreased from convergence near 0.20 to convergence near 0.18, representing a measurable reduction in optimization error.

$$\frac{0.20-0.18}{0.20} \times 100\% = 10\% \tag{1}$$

Validation Loss decreased from convergence near 0.39 to convergence near 0.26, indicating enhanced generalization capability.

$$\frac{0.39-0.26}{0.39} \times 100\% \approx 33.33\% \tag{2}$$

2. mIoU Metric: Training mIoU increased from convergence near 65% to convergence near 80%, demonstrating substantial improvement in segmentation accuracy.

$$\frac{80-65}{65} \times 100\% \approx 23.08\% \tag{3}$$

The systematic reduction in loss values coupled with the significant elevation in mIoU provides empirical validation that the combined methodology of data augmentation and CBAM integration enhances model robustness. These quantitative results conclusively demonstrate the efficacy of the proposed architectural and data-processing enhancements. While the integrated methodology

DOI: 10.54254/2755-2721/2025.GL24843

demonstrated significant performance gains, diagnostic curves reveal persistent limitations requiring further refinement:

1. Loss Metric Instability. Although both training and validation losses exhibited quantifiable reductions post-enhancement, the optimized Loss curves displayed pronounced oscillations around epoch 50. This heightened volatility may be attributed to an excessively high learning rate、suboptimal batch size configuration, and incipient gradient explosion phenomena. Future work will implement hyperparameter tuning, like adaptive learning rate schedulers and optimized data loading pipelines, to stabilize convergence dynamics.

2. mIoU Convergence Behavior. Despite substantial mIoU improvement, the enhanced metric profile exhibited premature stabilization with attenuated fluctuations during later training stages. This plateau suggests potential convergence to local optima、ineffective learning rate decay protocols, and class imbalance effects undermining gradient diversity. Subsequent iterations will incorporate dynamic learning rate adjustment, real-time gradient monitoring, and rigorous overfitting diagnostics like stratified k-fold validation to escape suboptimal solutions.

Furthermore, this study employed the 'get_mIoU.py' script to compute class-specific segmentation metrics for the augmented dataset incorporating the CBAM attention mechanism. Quantitative evaluations of background and skin lesion regions were systematically derived for four key performance indicators: mIoU, Mpa, mPrecision, and mRecall. The resultant class-wise metric distributions are visually presented in Figures 5(a) through 5(d), providing granular performance characterization across semantic categories.

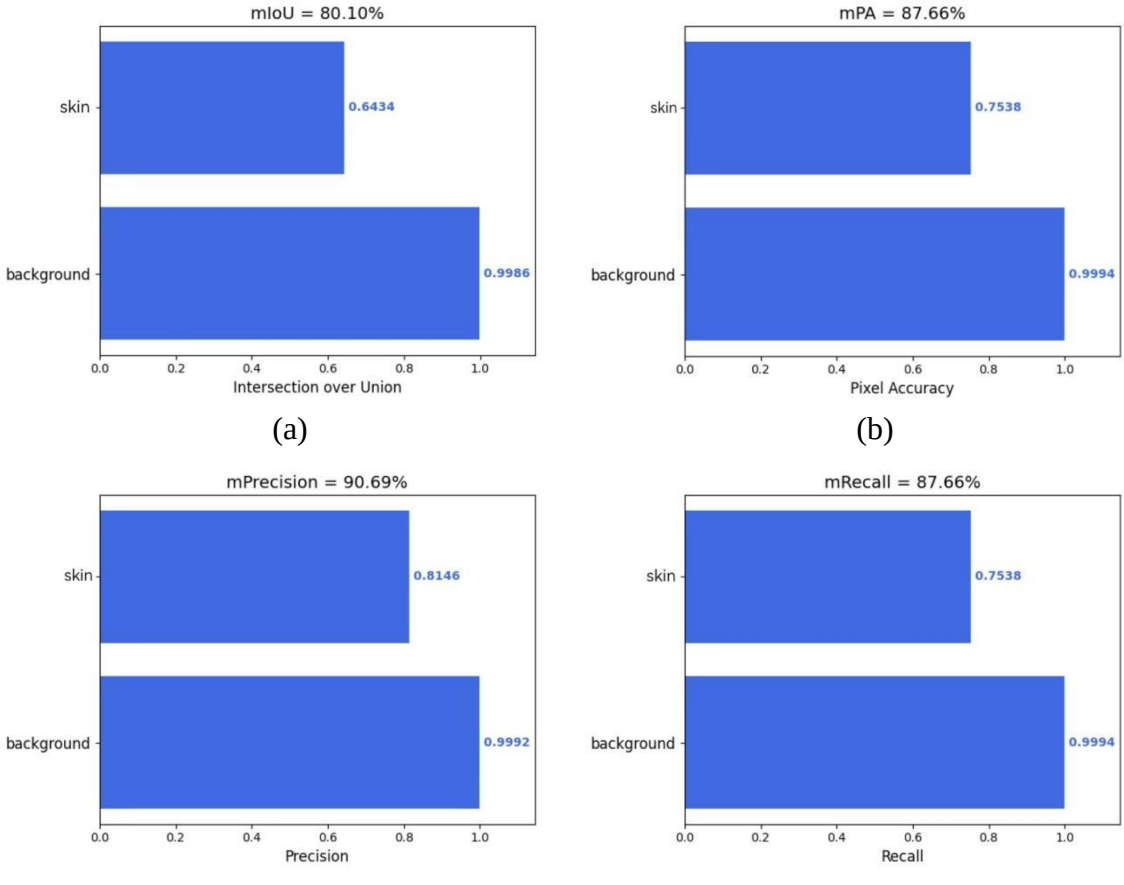


Figure 5: (a) mIoU; (b) Mpa; (c) mPrecision; (d) mRecall (photo/picture credit: original)

DOI: 10.54254/2755-2721/2025.GL24843

4. Conclusion

This study addresses the limitations of traditional manual segmentation, namely its low accuracy and inefficiency, by proposing the use of a U-Net model for automated and efficient segmentation, concurrently enhancing segmentation precision. However, the conventional U-Net model has proven inadequate for scenarios demanding high precision, such as medical imaging or remote sensing, where targets are often small or exhibit blurred boundaries. Despite widespread application, existing modifications to the standard U-Net model (e.g., U-Net++, ResU-Net, Attention U-Net, and data augmentation techniques) have yet to yield substantial advancements, particularly lacking widespread adoption. Motivated by this gap, the present research undertakes an empirical investigation into the enhancement of the traditional U-Net model. Given its predominant application in the medical domain, this study utilizes an open-source dermatological lesion dataset sourced from Kaggle for experimentation.

Leveraging the dermatological lesion dataset, this work explores and quantitatively evaluates specific modifications to the standard U-Net architecture: namely, the implementation of data augmentation (including random rotation, flipping, contrast adjustment, and Gaussian blurring) and the integration of the Convolutional Block Attention Module (CBAM) spatial attention mechanism. Performance is rigorously assessed using established metrics: mean Intersection over Union (mIoU), mean Precision (mPrecision), mean Recall (mRecall), and mean Pixel Accuracy (mPA). These enhancements demonstrably improved segmentation accuracy and model robustness, achieving a 23% improvement in mIoU and reductions in loss of 10% and 33%, respectively. Critically, high mIoU values were maintained concurrently with high mPrecision and mRecall, demonstrating the potential to contribute significantly to advancements in image segmentation fidelity.

Looking forward, the application of U-Net models for image segmentation is anticipated to expand considerably, encompassing interdisciplinary fields such as remote sensing image analysis, meteorological monitoring, agricultural disease detection, and geological exploration. It is envisaged that the improvements demonstrated in this study will facilitate the deployment of refined U-Net models in domains requiring even higher precision (e.g., medical diagnostics, remote sensing object recognition), thereby making substantial contributions to both academia and society.While U-Net-based image recognition in medicine is relatively mature, future developments hold promise for broader applicability in disciplines like mechanical engineering and, crucially, for fostering interdisciplinary innovation. For instance, integrating U-Net models for polar sea ice remote sensing image recognition with discrete element simulations could yield significant contributions to understanding polar climate dynamics and maritime navigation. Ultimately, this research underscores the potential for further refinement to unlock the full potential of U-Net models.

References

[1] *Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of MICCAI (pp. 234–241).*

[2] *Fang, C., Yang, Z., Jiang, C., Tao, F., Fa, T., & Zhang, J. (2025). Aerosol recognition based on Attention-U-Net neural network. China Laser, 52(10), 217–227.*

[3] *Intelligence-Based Medicine. (2025). Attention-DenseUNet for breast cancer ultrasound image segmentation. Ebioweb.*

[4] *Ding, J., et al. (2025). CMSAF-Net: Integrative network design with enhanced decoder for precision segmentation of pear leaf diseases. Plant Methods, 4(7).*

[5] *Yuan, Y. (2023). Improved UNet for post-stack seismic fault identification. Chinese Journal of Computational Physics.*

[6] *Kaggle. (n.d.). Kaggle. https://www.kaggle.com*

[7] *Hong, D., Xu, J., & Wen, F. (2024). Study on image segmentation of acute pancreatitis based on attention mechanism. International Journal of Biomedical Engineering, 47(2), 141–148.*

DOI: 10.54254/2755-2721/2025.GL24843

[8] Ying, L., Xue, H., Shanyang, L., Hui, Y., & Jiali, W. (2024). Attention-guided high resolution image change detection in full scale connection networks. Journal of Remote Sensing, 28(4), 1052–1065.

[9] Wang, S. (Shaoyu), Chen, Q., & Chen, H. (2025). Segmentation of colorectal cancer section images based on VMamba-CNN hybrid. Modeling and Simulation, 14(4), 799–810.

[10] Ming, J., Quan, R., & Shuo, L. (2023). Semantic segmentation of strawberry diseases based on improved UNet with attention mechanism. Application of Computer Systems, 32(6), 251–259.