

# Multi-class Segmentation Based on the B-UNet Model

ZhenXiong Xu

School of Computer and  
Information Engineering  
Fuyang Normal University  
Anhui, China  
17201766330@163.com

HuaiMeng Xiao

School of Computer and  
Information Engineering  
Fuyang Normal University  
Anhui, China  
sduxmh@163.com

Ya Wang\*

School of Computer and  
Information Engineering  
Fuyang Normal University  
Anhui, China  
fync\_wy80@163.com  
\*corresponding author

WenJing Gao

School of Computer and  
Information Engineering  
Fuyang Normal University  
Anhui, China  
2812791469@qq.com

**Abstract**—To solve the problem that UNet performs poorly in multi-class segmentation on large datasets, this paper proposes a new neural network model, B-UNet. UNet has some problems, such as poor feature processing method and large splicing span of features, leading to its poor feature extraction ability for complex images. Although UNet++ solves the problem of large feature span, the number of parameters in the model is greatly increased. Based on UNet, the B-UNet model obtained by introducing bridge features not only solves the problem that UNet has poor performance in multi-class segmentation on large datasets, but also reduces the number of parameters compared with UNet++. In order to make the introduced bridge features play a greater role in the process of feature splicing, B-UNet model introduces an attention mechanism, so that the model can focus more on the features that contribute more to the segmentation, and the reference of residual module also enhances the stability of the model. On VOC2007 dataset, MIoU, MPA and Accuracy were used as evaluation indexes, Experimental results show that B-UNet model is more suitable for multi-class segmentation on larger datasets than UNet model.

**Keywords**—neural network, multi-class segmentation, bridge feature, residual module, attention mechanism

## I. INTRODUCTION

Image segmentation has wide applications in multiple fields. In computer vision, it is the basis of object recognition, target tracking, face recognition and other tasks. In medical image processing, image segmentation can be used for auxiliary diagnosis, surgical planning and lesion detection. In addition, in robotics and agricultural technology, image segmentation also plays an important role, such as autonomous navigation, target recognition, plant detection and pest identification. It plays an indispensable role in multiple fields, providing a basis for subsequent image analysis and understanding. Therefore, the research of image segmentation has great application value[1].

Commonly used models for image segmentation include DeepLab, SegNet, PSPNet, UNet, etc., among which UNet is widely used in the medical field. There are a large number of improved models based on UNet in the medical field, such as the latest MSS-UNet model, which improves UNet to perform better in the medical field[2].

Through its unique U-shaped structure, UNet can effectively fuse shallow features and deep features, making the model more accurate when extracting image features. Medical images tend to have a relatively simple and fixed structure, and their semantic information is relatively simple. The design of UNet network structure enables it to make full use of the characteristics of medical images, so it shows significant advantages in the medical field image segmentation[3]. Fig.1 shows an example of UNet segmentation in the medical field.

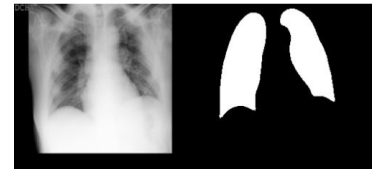


Fig.1. Medical domain segmentation example diagram

UNet's unique U-shaped structure can effectively integrate features at all levels, and it is outstanding on small sample datasets in the medical field. However, the effect of UNet on multi-class segmentation on large datasets is poor, which is partly due to the large span of its feature concatenation, which is proved by U++ network[4]. U++ network model inherits the U-shaped structure of UNet in structure, and realizes feature extraction and image reconstruction through two main parts: encoder and decoder. The encoder part is responsible for extracting multi-scale features of the input image layer by layer, while the decoder part fuses and reconstructs these features to generate the final segmentation result. Compared with UNet, U++ network model has made important improvements in feature fusion. It introduces richer feature fusion strategies, so that different levels of feature information can be fused together more effectively[5]. This improvement helps the model to better capture details and context information in the image, thus improving the accuracy of segmentation.

However, the U++ model reduces the feature span by frequent sampling and concatenation, which solves the problem of large feature concatenation span, but increases the number of parameters and training difficulty of the model. Reference [6] proposes to apply the trained VGG network to the feature extraction of the model, which not only improves the performance of the model, but also reduces the number of parameters of the model. Image feature extraction has a great

impact on the segmentation results, and the trained VGG network can optimize the feature extraction method of the model[7].

The VGG network, known for its concise and deep structure, excels in high-level feature extraction through stacking 3x3 convolutional layers and 2x2 pooling layers, making it easy to understand and implement. Combining VGG with UNet enhances feature extraction and representation, addressing UNet's struggles with large datasets and multi-class segmentation tasks. Despite its numerous parameters, VGG is simple to train, and its integration with UNet boosts model performance and clarity. Additionally, using pre-trained VGG models, which are widely utilized in various image tasks, accelerates convergence and improves generalization. Consequently, the VGG-UNet model outperforms the original UNet, though it still has room for improvement regarding large feature splicing spans in the U-shaped structure.

This paper is based on the UNet model to improve. The unique U-shaped structure is maintained, the Bridge features are generated by using U++ up-sampling to reduce the feature span, and a new model B-UNet (Bridge-UNet) is proposed by using VGG network to optimize the feature extraction method.

## II. RELEVANT MODELS AND TECHNOLOGIES

### A. UNet model structure

Through its unique U-shaped structure, UNet can effectively fuse shallow features and deep features, making the model more accurate in extracting image features, especially on small sample datasets[8]. The model structure of UNet is shown in Fig. 2.

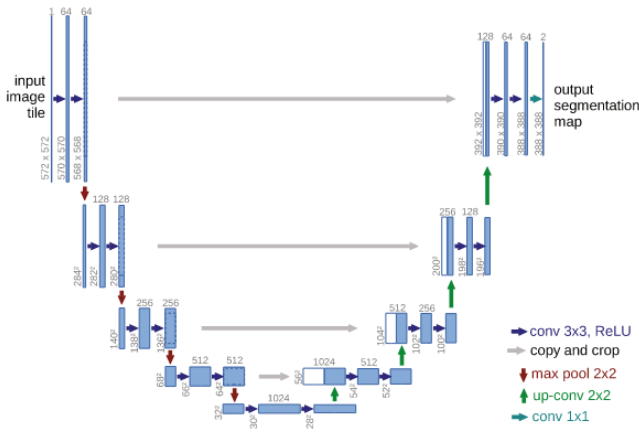


Fig.2.UNet model structure[9]

It is precisely because the U-shaped structure of UNet can comprehensively consider different levels of features, and the lightweight UNet has great improvement space. By improving the extraction and processing operations of features of UNet, B-UNet can be improved on the basis of maintaining the original U-shaped structure of UNet. At the same time, the U-shape structure allows B-UNet to perform feature optimization using skip connections like UNet[10].

### B. Skip connection

The jump connection of UNet can fuse the feature information of different network layers. In the structure of UNet, the encoder part is responsible for extracting the low-level features of the image, such as edges and textures, While the decoder part is responsible for combining these features for pixel-level classification. Skip connections allow low-level features in the encoder to be passed directly to the corresponding layer in the decoder, enabling efficient fusion of feature information. This fusion helps the model to better capture the details of the image and improve the accuracy of segmentation[11].

Jump connections help improve gradient flow and avoid gradient disappearance during training[12]. Gradient disappearance is a common problem in deep neural networks, which can cause the model to fail to update parameters effectively during training. By establishing a direct connection between different network layers, the jump connection of UNet helps to pass gradient information from higher layers to lower layers, thereby improving gradient flow and making the model easier to train.

In a U-shaped structure, as the level of skip connections increases, the corresponding feature span becomes larger. U++ network solves this problem by intensive upsampling, which greatly increases the model complexity and calculation amount.

### C. Attention and residual module

To further optimize features and improve model performance, B-UNet also introduces attention mechanisms and residual connections.

Attention Mechanism stems from the study of human vision and is an important concept in the field of deep learning. In cognitive science, due to bottlenecks in information processing, humans selectively focus on a subset of all information while ignoring other visible information, a mechanism often referred to as the attention mechanism. In computer vision, attention mechanisms improve the accuracy and performance of models by weighting important regions of input data in tasks such as image recognition, object detection, and semantic segmentation[13]. Specifically, it helps models focus on important areas and objects while processing images, ignoring background and other irrelevant information. For example, in an image classification task, the attention mechanism can help the model focus on important areas in the image, thus improving the accuracy of the classification. Previous studies have shown that applying attention mechanisms to UNet networks improves the model's performance. Fig.3 is a schematic of the attention mechanism used in this paper.

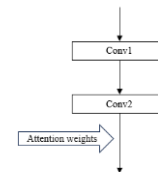


Fig.3.Attention mechanism

The residual module is a special connection method in neural network, which aims to establish a direct relationship between input and output, so that the neural network can learn the residual mapping between input and output[14]. The core idea of this module is to solve the problem of gradient disappearance and gradient explosion in deep network training by introducing Skip Connection. In a traditional neural network, information needs to be passed through multiple layers to reach the following layer, but in the residual module, information can be passed directly to the following layer by jumping connections, thus maintaining the integrity of information. Specifically, the residual module calculates the difference between the input and the features passed over by the jump connection, and adds this difference to the input to get the output. This difference is equivalent to a correction of the input, making the features learned by the module more accurate. The residual network is applied to UNet, which makes the model training more stable and the segmentation results more accurate[15]. The structure of the residual module used in this paper is shown in Fig.4.

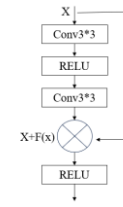


Fig.4.Residual module

#### D. B-UNet model structure

B-UNet is based on UNet, referring to the feature splicing mode of U++, introducing bridge features to reduce feature span, using VGG network to extract features. Adding attention mechanism and residual module to improve model performance, this paper proposes a B-UNet network model. The model structure is shown in Fig.5.

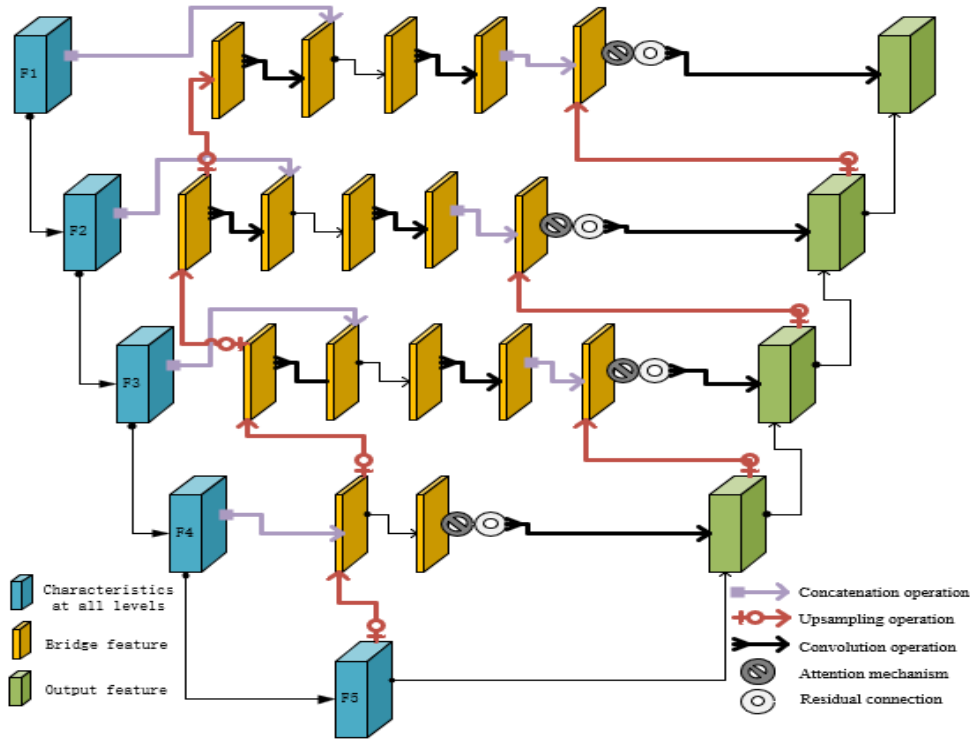


Fig.5.B-UNet model structure diagram

B-UNet generally retains the U-shaped structure of the original UNet, so that the model uses the features obtained by jump connection optimization like UNet. B-UNet uses a single bridge feature to reduce the span between features, in which F1-F5 features are extracted by VGG network, and F5 features are in the middle of the convolution layer, which is the most balanced feature between global and local features among all features. Bridge features are obtained by blending the features extracted by VGG network with F5 features. The span of feature splicing is greatly reduced. The overall framework structure of the model is shown in Fig. 6.

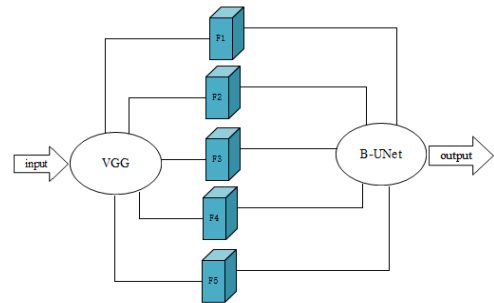


Fig.6.Overall model frame diagram

In the above figure, the image to be segmented is input into the VGG network to obtain various levels of features (such as F1, F2,... F5), and then the obtained features are provided to the B-UNet network, and finally the B-UNet network outputs the segmented image. By referencing VGG, B-UNet not only reduces the training difficulty of the model, but also improves the model performance.

The unique structure of U-shaped network enables features at different levels to be fused. B-UNet introduces bridge features to perfect fusion, but also increases the number of fused features. By introducing attention mechanism, the model can focus more on key features. The use of residual module also solves the problem of gradient disappearing and gradient explosion in the process of model training.

### III. EXPERIMENT AND RESULT

#### A. Dataset

This experiment uses VOC2007(Visual Object Classes) which is a widely used benchmark dataset in the field of computer vision. This dataset is mainly used to evaluate the performance of visual object classification, detection and image segmentation[16]. The dataset covers 20 different object categories, including humans, animals, vehicles, and common indoor and outdoor objects. These categories are broad and diverse, making the dataset suitable for a wide variety of computer vision tasks. Each image is labeled with the location and category of multiple objects by professional taggers, which provides rich training and test data for many computer vision tasks, such as object detection, semantic segmentation, and image classification. The VOC2007 dataset contains 9,963 images from a number of different sources, including 5,011 images in the training and validation sets and 4,952 images in the test set. Compared with the commonly used medical dataset, it has larger specifications and more segmentation types.

#### B. Evaluation indicators

In image segmentation, MIoU, MPA and Accuracy are commonly used evaluation indexes, which can evaluate the performance of the model in image segmentation tasks. The following is a detailed introduction of these three evaluation indicators and their calculation formulas.

##### 1) MIoU (Mean Intersection over Union)

MIoU, or average intersection ratio, is a commonly used accuracy measure in semantic segmentation. It calculates the ratio of the intersection to the union between the true label and the predicted result, and then averages the IoU for all classes. The value of MIoU ranges from 0 to 1, with the closer to 1 indicating that the predicted result is more similar to the real label. The calculation formula is (1).

$$MIoU = \frac{1}{n} \sum_{i=1}^n IoU_i \quad (1)$$

where, n is the total class number,  $IoU_i$  is the intersection ratio of the i class,  $I_{pre}$  represents the prediction graph,  $I_{gt}$  represents GT, and  $I_{pre} \cap I_{gt}$  represents the overlap between the prediction graph and GT graph. The calculation formula is (2).

$$IoU_i = \frac{I_{pre} \cap I_{gt}}{I_{gt}} \quad (2)$$

##### 2) MPA (Mean Pixel Accuracy)

Average pixel accuracy (MPA) is one of the commonly used evaluation indexes in image segmentation tasks. It measures the number of correctly classified pixels as a proportion of the total number of pixels in the image. The MPA provides a simple but informative metric for evaluating the overall performance of a segmentation model, revealing the model's prediction accuracy for each pixel class over the entire image. MPA is especially useful when there is a class imbalance or when evaluating models trained on datasets with different object sizes and shapes. The calculation formula of MPA is (3).

$$MPA = \frac{1}{n} \sum_{i=1}^n PA_i \quad (3)$$

where PA is equal to the ratio of predicted number of correct pixels and the predicted total of pixels, the calculation formula is (4).

$$PA_i = \frac{\sum_{i=0}^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}} \quad (4)$$

##### 3) Accuracy

Accuracy is one of the most commonly used evaluation indexes in classification problems. In image segmentation, it represents the ratio of the number of correctly classified pixels to the total number of pixels, and the calculation formula is (5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

where TP represents the detected and correct target, TN represents the correctly identified background. For specific calculations, you can iterate over all the predicted values, compare the predicted value to the label value (the true value) of the dataset, count to 1 when the two are equal, and then sum and divide by the total number of pixels.

These evaluation indicators provide a way to quantify the performance of the models, so that we can evaluate the performance of different models on image segmentation tasks more objectively.

#### C. Analysis of experimental results

##### 1) Experimental setting

This experiment was implemented in PyTorch version 1.12, using the NVIDIA GeForce RTX 2080 Ti graphics card to train and test the model. The network is iterated 100 times, batch\_size is set to 16, the initial learning rate lr is set to 1e-4, and the learning rate is adjusted adaptively during the iteration process, and the Adam optimizer is used to update the gradient. During training and testing, the size of the input image remains the same.

##### 2) Comparative test

This experiment was conducted on VOC2007 dataset, and MIoU, MPA and Accuracy were taken as evaluation indexes, with specific values shown in Table 1.

TABLE 1. COMPARISON TABLE OF METRICS

Model Name	MIoU	MPA	Accuracy
UNet	7.8	6.9	64.9
UNet++	8.12	11.15	67.95
VGG-UNet	41.69	52.44	82.46
B-UNet	<b>43.95</b>	<b>55.77</b>	<b>84.88</b>

When UNet performs segmentation on VOC2007 dataset, its MIoU is only 7.8, MPA and Accuracy are 6.9 and 64.9. The upsampling of UNet++ reduces the feature span, resulting in MIoU of 8.12, which is slightly improved compared with UNet. Its MPA is 6.9, an increase of nearly 5 percentage points, and Accuracy is 67.95, an increase of nearly 3 percentage points. By introducing the VGG-UNet model of VGG network optimization feature extraction mode, MIoU is nearly 5 times higher than UNet compared with MPA, reaching 41.69 and 52.44, and its Accuracy is also nearly 15 percentage points higher than UNet, reaching 82.46. B-UNet introduced bridge features on the basis of VGG-UNet, and optimized the model with attention mechanism and residual module. Its MIoU increased by nearly three percentage points on the basis of VGG-UNet, reaching 43.95, and MPA increased by nearly 4 percentage points to 55.77. At the same time, the Accuracy reached 84.88, which was nearly 3 percentage points higher than that of VGG-UNet.

The experimental results show that B-UNet has obvious advantages over the original UNet model in image segmentation on a larger scale and more kinds of datasets such as VOC2007. By improving the UNet feature extraction method and maintaining the unique U-shaped structure of UNet, B-UNet achieves excellent performance compared with UNet when performing multi-class segmentation on large datasets.

### 3) Split the instance

In order to more intuitively compare the segmentation performance of different models for different categories of images, the dataset images are divided into three categories, which are single-category segmentation for simple background, single-category segmentation for complex background and multi-class segmentation for complex background. This demonstration is divided into three categories of images, one for each. Fig.7 is the segmentation example diagram.

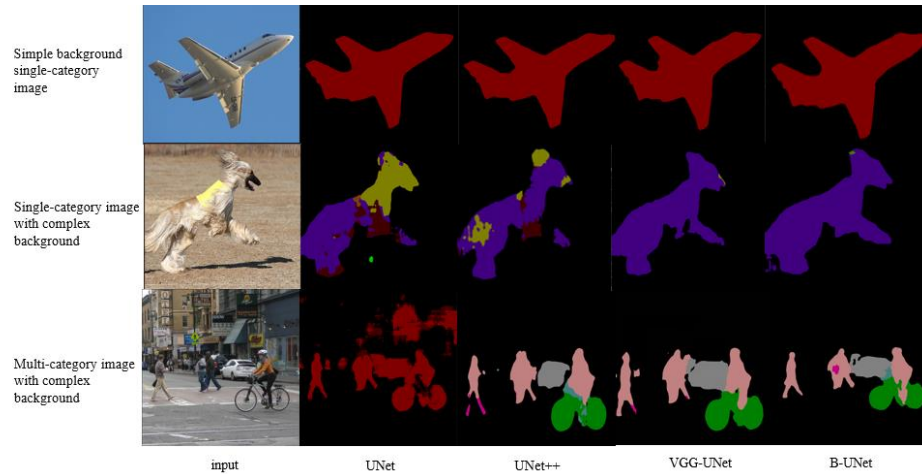


Fig.7. Split example diagram

As shown in the figure above, UNet and UNet++ models perform well in simple background segmentation with fewer categories. But when the segmentation types increase and the background becomes complex, these models perform less. The experimental results show that the UNet model and the UNet++ model perform poorly on such datasets. Compared with UNet, UNet++ and VGG-UNet, B-UNet has greater advantages in the segmentation of multiple types of complex scenes, which not only improves the segmentation accuracy but also greatly improves the training efficiency of the model.

### 4) Loss analysis

The optimizer of B-UNet model uses Adam, which can adaptively adjust the learning rate of each parameter and perform well in dealing with large-scale data and high-dimensional parameter space. Its adaptive learning rate

mechanism enables it to quickly adapt to various datasets and model complexity, so that it is easier to converge to the global optimal solution[17]. Adam generally has high computational efficiency in practice. Although it has some additional computational overhead (such as the moving average of the first moment estimate and the second moment estimate of the computed gradient), these additional computations can usually be reduced by parallelization and efficient implementation. Specifically, the Adam optimizer calculates gradients of the model parameters based on the loss values, and updates the model parameters based on these gradients, so that the loss values are gradually reduced until convergence or minimum values are reached. Fig. 8 shows the comparison between train\_loss and val\_loss of B-UNet and UNet on VOC2007 dataset.

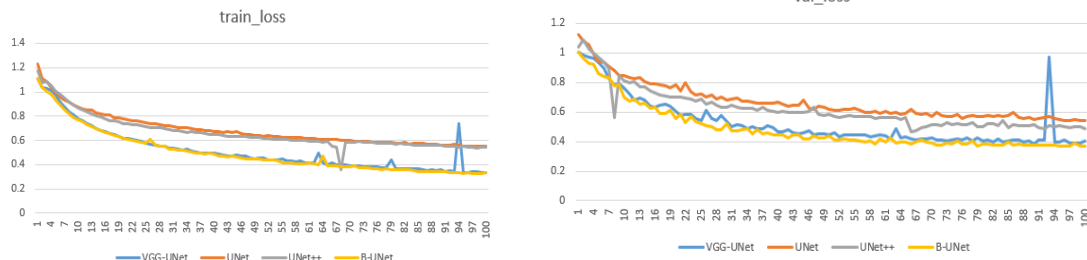


Fig.8. Loss comparison chart

Through the introduction of VGG network optimization feature extraction, the training and verification losses of B-UNet and VGG-UNet are kept at a low value, and the training efficiency and convergence speed of the model are faster. It can be seen from the above figure that the training and verification loss of VGG-UNet model fluctuates greatly. B-UNet, which not only introduces VGG network but also uses bridge features to reduce feature span, has stronger model stability than VGG-UNet.

#### IV. SUMMARY AND OUTLOOK

To address UNet's poor performance on large datasets, this paper proposes the B-UNet model, which integrates feature extraction from the VGG network and introduces bridge features to the UNet model. Training and testing on the VOC2007 dataset demonstrate that B-UNet significantly improves MIOU, MPA, and Accuracy compared to UNet. These results indicate that B-UNet excels in multi-class segmentation tasks on large datasets, enhancing both segmentation quality and model stability.

The core idea of B-UNet is to reduce feature span by introducing bridge features from the network's middle layer, which are processed upward to support the U-shaped structure. Future work will focus on improving these bridge features by training a specialized network to extract them, thereby reducing the feature splicing span. This dedicated bridge feature extraction network will provide better bridge features for model splicing.

#### ACKNOWLEDGMENT

This research was funded by An-Hui Provincial Graduate Innovation and Entrepreneurship Practice Project (Grant No. 2022xcysj189), Anhui Provincial Natural Science Research Project (Grant No. 2022AH051324), The Ministry of Education's Industry University Cooperation Collaborative Education Project (Grant No. 220602279172021), Anhui Provincial Graduate Online and Offline Mixed Course (Grant No. 2022hhsfkc042), Fuyang Normal University Graduate Innovation and Entrepreneurship Practice Project (Grant No. FNU2023cysj003).

#### REFERENCES

- [1] Z. Zhiwei, T. hui, X. Zhenshun, B. Yun and W. Jie, "Application of a pyramid pooling Unet model with integrated attention mechanism and Inception module in pancreatic tumor segmentation," *Journal of Applied Clinical Medical Physics*, vol. 24, no. 12, pp. e14204-e14204, 2023.
- [2] Z. Wenhao, T. Jiya, C. Mingzhi, C. Lingna and C. Junxi, "MSS-UNet: A Multi-Spatial-Shift MLP-Based UNet for Skin Lesion Segmentation," *Computers in biology and medicine*, vol. 168, pp. 107719-107719, 2023.

- [3] T. Hui, W. Ming, Y. Qiushi, Z. Jiayi, L. Liantao and W. Nan, "Root image segmentation method based on improved UNet and transfer learning," *Smart Agriculture*, vol. 5, pp. 96-109, 2023. (in Chinese)
- [4] Z. Zongwei, M. M. Siddiquee, N. Tajbakhsh, and J. Li, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1-1, 2019.
- [5] Q. Haitao and S. Cao, "LM-UNet: Lateral MLP Augmented U-Net for Medical Image Segmentation," *Computer Systems & Applications*, pp. 1-8, doi:10.15888/j.cnki.csa.009510. (in Chinese)
- [6] K. Han, J. Li, and R. Tao, "Coal Mine Key Position Personnel Unsafe Behavior Recognition Based on Improved YOLOv7 and ByteTrack," *Industrial and Mine Automation*, pp. 1-11, doi:10.13272/j.issn.1671-251x.2024030015. (in Chinese)
- [7] V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation," *CoRR*, vol. abs/1801.05746, 2018.
- [8] L. Wu, Q. Zhang, Q. Zhen, S. Shao, and C. Cui, "Steel Bridge Corrosion Detection Method Based on Fusion of Adaptive Illumination Preprocessing and Deep Learning," *China Journal of Highway and Transport*, vol. 37, no. 02, pp. 110-124, doi:10.19721/j.cnki.1001-7372.2024.02.010. (in Chinese)
- [9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *CoRR*, vol. abs/1505.04597, 2015.
- [10] W. Huang, C. Qu, and Y. Yan, "Retinal Vasculature Segmentation Based on TransUNet with Integrated Attention Mechanism," *Optics and Precision Engineering*, vol. 31, no. 23, pp. 3482-3489, 2023. (in Chinese)
- [11] H. Nanjun, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-Connected Covariance Network for Remote Sensing Scene Classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1-14, 2019.
- [12] S. Deva Kumar, V. Sistla, and K. V. Kishore, "HRUNET: Hybrid Residual U-Net for Automatic Severity Prediction of Diabetic Retinopathy," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 11, no. 3, pp. 530-541, 2023.
- [13] Y. Yuan, H. Liu, and J. Wang, "Using the Wide-Range Attention U-Net for Road Segmentation," *Remote Sensing Letters*, vol. 10, no. 5, pp. 506-515, 2019.
- [14] Md. Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari, "Recurrent Residual Convolutional Neural Network Based on U-Net (R2U-Net) for Medical Image Segmentation," *ArXiv preprint arXiv:1802.06955*, 2018.
- [15] W. Shuai, D. Yuhong, Z. Shuaijie, H. Jinhua, and G. Lian, "Research on the Construction of Weaponry Indicator System and Intelligent Evaluation Methods," *Scientific Reports*, vol. 13, no. 1, pp. 19370-19370, 2023.
- [16] D. Xiangyu and L. Shanshan, "An Improved SSD Object Detection Algorithm Based on Attention Mechanism and Feature Fusion," *Journal of Physics: Conference Series*, vol. 2450, no. 1, 2023.
- [17] A. Sadeghnejad Barkousaraie, O. Olalekan, S. Jiang, and D. Nguyen, "A Fast Deep Learning Approach for Beam Orientation Optimization for Prostate Cancer Treated with Intensity-Modulated Radiation Therapy," *Medical Physics*, vol. 47, no. 3, pp. 880-897, 2020.