AttU-Net: A Submarine Image Segmentation Model Based on Hybrid Attention Mechanism and Residual Structure

Wanxin Liang[1], Yicheng Sheng[2*], Zefeng Zhao [3]

[1]Beijing Institute of Technology, Zhuhai, China

[2] EMGO-TECH TECHNOLOYGY CO., LTD. Zhuhai, China

[3]China Faculty of Data Science, City University of Macau, Zhuhai, China

*** Correspondence:**
shengyicheng87@gmail.com

Keywords: Semantic Segmentation$_1$, U-Net$_2$, Attention$_3$, Residual Structure$_4$, Deep Learning$_5$.

Abstract

We construct a submarine image segmentation model based on a hybrid attention mechanism and a residual structure to address the problems of low imaging resolution, low visibility, and spatial location mismatch due to a wide variety of objects and large differences in target object sizes caused by imaging machines and equipment, ocean environment, and light in underwater scenes. We fuse the hybrid attention mechanism and the residual network on the U-Net neural network architecture to ensure the preservation of the shallow semantic information describing the objects, and also enable the model to extract the deeper semantic information characterizing the object categories, and the feature information of the images can be better preserved and the model has higher accuracy and precision. Our work is applied to the SUIM underwater image segmentation dataset and compared with U-Net (VGG16),U-Net (Resnet50), PSPNet (Mobilenet), PSPNet (Resnet50), Deeplabv3 for experimental results: the average cross-merge ratio of our proposed AttU-Net is higher than the above networks by The average intersection ratio of our proposed AttU-Net is 5.3%, 2.59%, 8.45%, 3.7%, and 4.88% higher than the above networks; the average pixel accuracy values are 4.38%, 2.71%, 6.54%, 3.07%, and 3.89% higher, respectively.

1 INTRODUCTION

Semantic segmentation is to segment the image into regional blocks with certain semantic meaning by certain methods, and identify the semantic category of each regional block to realize the semantic inference process from the bottom to the top level, and finally get a segmented image with pixel-by-pixel semantic annotation. Semantic segmentation of the seafloor can help us to improve the seafloor topographic data system, which can help researchers to carry out ocean mapping and ocean exploration.

Due to the complex environment of the seafloor, the large size difference between different categories, the variety of objects, and the fact that seawater is a complex mixture of chemical composition, microorganisms in water, suspended particulate matter and refraction of light by seawater, etc., lead to poor quality of seafloor imaging images, increase the complexity of image processing, make the segmentation of target objects difficult, and make it difficult to carry out research work.

The main disadvantages of traditional methods are listed： There are traditional methods and convolutional neural network-based methods for image semantic segmentation, among which the traditional semantic segmentation methods can be divided into statistical-based methods and geometry-based methods. With the development of deep learning, the semantic segmentation technology has been greatly improved. The biggest difference between the semantic segmentation method based on convolutional neural network and the traditional semantic segmentation method is that the network can automatically learn the image development features and perform end-to-end classification learning, which greatly improves the accuracy of semantic segmentation.

Compared to traditional medical image segmentation, the semantic segmentation task for images of underwater scenes suffers from problems such as drastic changes in the scale of the target object, which is a great challenge for the pixel-level prediction task. In order to be able to extract deep semantic information characterizing the object class, the model constantly performs downsampling convolution operations, resulting in the loss of shallow semantic information describing the object color, texture, size, etc. of the image.

In the past decade, deep learning methods have been increasingly used in computer vision. FCN (Jonathan Long et al, 2014) networks can accept image inputs of arbitrary size, avoiding the duplicate storage and computation problems caused by using pixel blocks, but the results obtained by the models are not accurate enough, insensitive to the details of the images, do not consider pixel-to-pixel relationships, and lack spatial consistency; U-net(Ronneberger et al, 2015) simply stitches the encoder feature maps to the upsampled feature maps of the decoder at each stage, forming a trapezoidal structure with a jump-connected architecture that allows the decoder to learn correlations lost in encoder pooling and capture multi-scale objects for finer semantic segmentation using multiple perceptual fields; Deeplab(Chen et al, 2018) uses a null convolution operation that adds fully connected conditional random fields, but it obtains results only of the original input 1/8 size, which cannot be consistent with the original image size; PSPNet(Zhao et al, 2017) proposes a pyramid module to aggregate background information, uses additional loss, and it uses four different pyramid pooling modules, which require higher detail processing.

The hopping link structure of U-Net framework is able to capture multi-scale objects for finer semantic segmentation using multiple sensory fields, but the simple replication problem is not able to solve the spatial matching problem of target object features. The semantic feature information at the bottom layer is relatively small, but the target location is accurate; the semantic feature information at the top layer is richer, but the target location is coarse. We need to bridge the semantic and resolution gap between low-level and high-level features by more effective feature fusion and multi-scale channel cross-attention to capture more complex channel dependencies.

Based on the U-Net structure, we propose AttU-Net, which integrates several advanced deep learning methods on the u-type architecture, including attention mechanism, residual connectivity, dropout learning, and freeze training. Compared with the general U-net model, our method has the following four advantages：

(a) The residual structure deepens the depth of the model and helps us capture more high-level semantic information characterizing the target object class.

(b) The hopping links of the U-shaped architecture retain the shallow semantic information of the images, which can provide more shallow detail information when we recover the image features.

(c) The inclusion of the hybrid attention mechanism allows the network to pay more attention to the spatial location matching of target objects and learn more features in the case of large differences in the size of the imaged targets.

(d)The model has both shallow information retention and deep feature extraction, making a balance between the two, so that the more lightweight U-Net neural network structure can also handle semantic segmentation tasks with a large variety of segmentation.

2 METHODS

Our proposed AttU-net model is shown in **Figure 1**. We use the U-Net neural network as the base architecture and incorporate three advanced computer vision methods, the residual module, the channel attention mechanism and the spatial attention mechanism module, into this architecture(Wang et al, 2017).

The AttU-Net model consists of three main components: backbone feature extraction network, enhanced feature extraction and prediction. We first resize the input incoming images uniformly, and we resize the images to the size of [512, 512], which solves the problem of drastic changes in object size due to camera differences in the input images, solves unnecessary troubles from the perspective of data set specification, and improves the operation efficiency of the model.

The backbone feature extraction part of AttU-Net consists of convolution + maximum pooling + channel attention mechanism + spatial attention mechanism stacking, and using the backbone feature extraction network we can obtain five preliminary effective feature layers, which can be used for feature fusion in the jump layer connection part.

The enhanced feature extraction part of AttU-Net then uses the five initial effective feature layers of the first part for up-sampling and stacking operations to obtain an effective feature layer that fuses all the features. At this point, the prediction part can use the final valid features to obtain prediction results.

We use Resnet50(He et al, 2016) as the backbone of the model, fusing the attention mechanism with residual connectivity, which makes the network depth vary, the encoder extracts the semantic information at the local bottom and global top layers through layer-by-layer convolution and pooling operations, and the decoder recovers the low-resolution features to enhance the model's extraction of features. Our model fuses the feature mapping of the encoder with the feature mapping of the corresponding scale of the decoder processed by the attention mechanism, which ensures the global macroscopic semantic information of the model and preserves the local microscopic information as well.

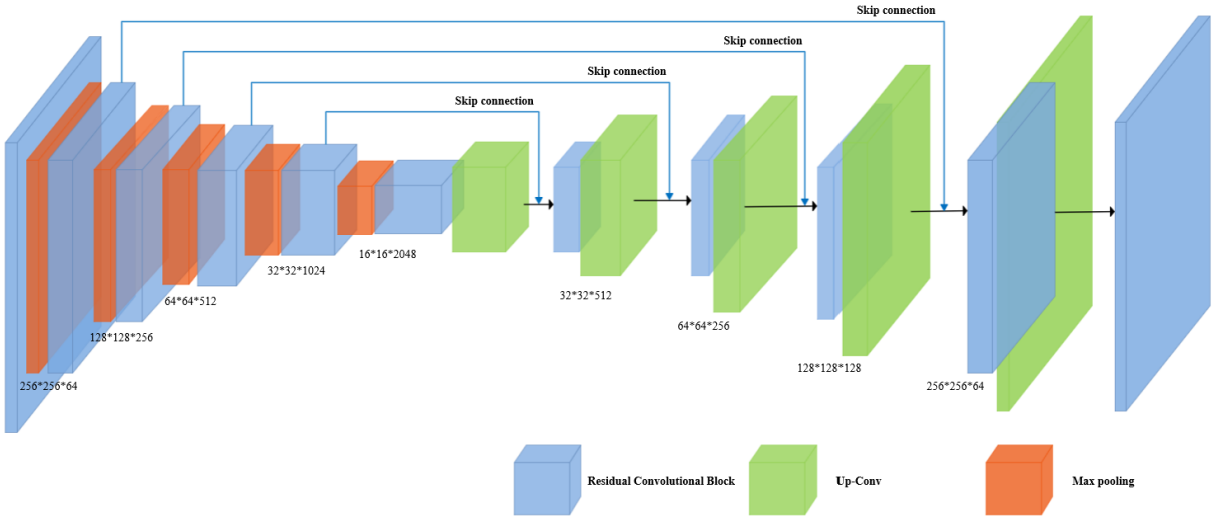

Figure 1.AttU-Net Architecture

2.1 Attentional Mechanisms

Attention mechanism is a special structure that people embed in machine learning models to automatically learn and calculate the contribution size of input data to output data, which is a data processing method for machine learning, and it is widely used in different types of machine learning tasks such as natural language processing, image processing and speech recognition.

The attention mechanism not only tells us where to focus on, but also improves the representation of the attention points. The goal is to increase the representation by using the attention mechanism to focus on important features and suppress unnecessary ones. To emphasize meaningful features in both dimensions, spatial and channel, CBAM(Woo et al, 2018) is the successive integration of the channel attention module and the spatial attention module, which tells us what and where to learn to focus on in the channel and spatial dimensions, respectively, by understanding that the information to be emphasized or suppressed also contributes to the flow of information within the network. After integrating the two modules to obtain channel attention(Hu et al, 2018) and spatial attention(Jaderberg et al, 2015), the original feature map is refined with information using a broadcast mechanism to finally obtain the feature map after targeted training.

Our model inserts the attention mechanism in the front and back layers of the backbone feature extraction network. The hybrid attention module is shown in **Figure 2**:

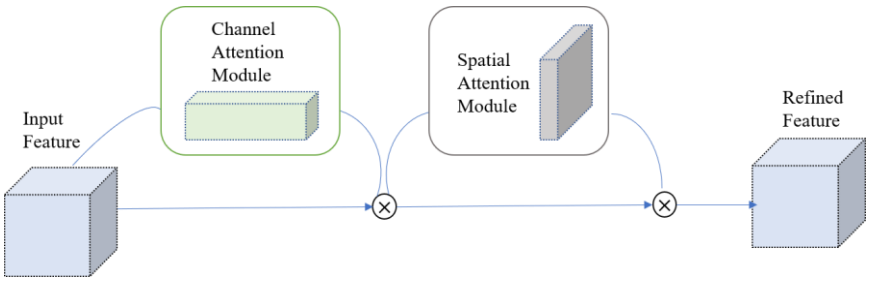

Figure 2.Convolutional Block Attention Module

2.2 Residuals Module

Applying Resnet50 as the backbone feature extraction network to the U-shaped network architecture, the application of the residual module can well enhance the feature grasping ability of the model, alleviate the degradation problem of the neural network, and utilize the structural characteristics of the U-shaped neural network architecture, i.e., the hopping layer linking part, to fuse target information of different depths and scales, so that the network retains most of the shallow feature information of the underwater segmented images and uses the shallow features to guide the extraction of deeper features, which helps the model to complete the multi-classification task without excessive loss of shallow feature information while obtaining high-level semantic information.

The bottleneck residual module used by Resnet50 is shown in **Figure 3**.

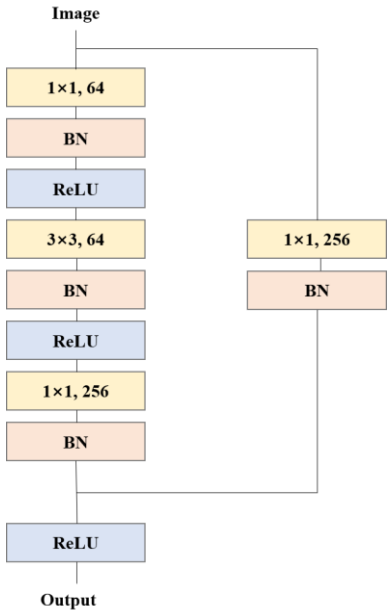

Figure 3. Bottle neck residual module display diagram.

2.3 Freeze Training

Freeze learning (Lester et al, 2021) is a method of migration learning that allows retraining on additional new data and fine-tuning network parameters on top of the initial training. Freeze learning can help us save a lot of computational resources and time. The training parameters of freeze learning during the training in this paper are: Freeze_epoch = 50, Freeze_batch_size = 2.

3 MANUSCRIPT FORMATTING

3.1 Dataset

Our work is applied to experiment on the SUIM(Islam et al, 2020) underwater image segmentation dataset, which is an underwater image segmentation dataset with a training image to test image ratio of about 14:1, containing 1525 training images with semantic annotation and 110 test images waiting for semantic annotation, divided into 8 semantic categories, as shown in **Table 1**:

Table 1-The code for pixel annotations in the object category SUIM dataset.

Object category	Code
Background(waterbody)	BW
Human divers	HD
Aquatic plants and sea-grass	PF
Wrecks or ruins	WR
Robots(AUVs/ROVs/instruments)	RO
Reefs and invertebrates	RI
Fish and vertebrates	FV
Sea-floor and rocks	SR

3.2 Loss Function

The equations should be inserted in editable format from the equation editor.

The loss used in this paper consists of two parts: Cross entropy loss(Zhang et al, 2018) and Dice loss(Li et al, 2019).Cross entropy loss is the ordinary cross entropy loss, which is used when the semantic segmentation platform classifies pixel points using softmax operation.

X is the prediction graph, Y is the split graph, and the dice loss is calculated as follows：

$$\boldsymbol{s} = \frac{\mathbf{2}|\boldsymbol{X}| \cap |\boldsymbol{Y}|}{|\boldsymbol{X}| + |\boldsymbol{Y}|} \quad (\mathbf{1})$$

3.3 Evaluation Indicators

In this paper, we use two evaluation metrics commonly used in semantic segmentation tasks to measure the accuracy, namely the average intersection ratio and the average pixel accuracy. Both the average intersection ratio and the average pixel accuracy are based on the confusion matrix. The confusion matrix is the classification result of the statistical classification model.

The pixel accuracy(Garcia-Garcia et al, 2018) represents the ratio of the number of correctly classified pixels to the total number of pixels, which can measure the segmentation performance of the model to a certain extent, but it is difficult to objectively evaluate the model performance for unbalanced data sets.

The formula is shown below:

$$mPA = \frac{Sum(P)}{N} (P\ for\ the\ pixel\ accuracy\ per\ category) \quad (2)$$

The average intersection and merge ratio measures the summed average of the intersection and merge ratios of the predicted segmentation region and the true segmentation region for each category.

The formula is shown below:

$$\mathbf{mIoU} = \frac{\mathbf{Sum(IoU_i)}}{\mathbf{N}} (\mathbf{N\ for\ the\ num\ of\ categories}) \quad (3)$$

3.4 Comparison Method

In this study, we take an in-depth look at some advanced segmentation models in the field of computer vision semantic segmentation over the years, and finally we select three advanced models, **U-Net**, **PSPNet**, and **Deeplab**, for analysis and comparison. In order to quantify the impact of backbone networks and attention mechanisms on the performance of segmentation models, we select **VGG16**(**Simonyan** et al, 2014), **Resnet50**, **Mobilenet**(**Howard** et al, 2017) three classical backbone networks used in computer domain segmentation models were selected for ablation experiments. We used the official models from the related literature with the optimal parameters recommended by the authors, selected a suitable gradient descent method, trained them using the same semantic segmentation dataset of undersea images, and kept the same training parameters for each model. During the training process, the training results of different batches of models are kept for comparison, and we select the model with the best model performance for prediction from the training process.
During testing, we evaluated each model using the same dataset of semantic segmentation of undersea images and selected 3 given evaluation metrics to draw line graphs for comparative analysis of the models. We selected six typical seafloor images as the analytical proof of the experimental results and compared them with the segmentation results of manually segmented, different segmentation models.

4 RESULTS

4.1 Performance of Our Model

The results show that our model outperforms other models on the semantic segmentation of seafloor images dataset (SUIM dataset). The experimental results of our comparative analysis are as follows: the detected mean intersection ratio (MIoU) value is 56.98%, the mean pixel accuracy (MPa) value is 67.15%, and the pixel accuracy is (Accracy) value is 78.75%. background, Wrecks or ruins , Aquatic plants and sea-grass, Human divers, Robots (AUVs/ROVs/instruments), Reefs and invertebrates, Fish and vertebrates, Sea-floor and rocks were tested and the average cross-merge ratio values were 81.73%, 46.94%, 15.98%, 71.32%, 75.06%, 67.17%, 60.41%, and 37.21%, respectively, and the tested pixel accuracies were 89.6%,48.78%, 19.39%, 82.3%, 91.51%, 89.55%, 66.71%, and 49.37%. Our model achieves good segmentation results on most segmentation categories, and also obtains better matching of spatial location information in the case of large differences in the scale of segmented objects. Our model achieves better performance on both complex background images and simple background images, and has a high overlap with the manually labeled segmentation labels.

4.2 Comparison With State-of-the-Art Models

The network model proposed in this paper is based on U-shaped network architecture, and the hybrid attention mechanism and residual network are fused into the U-Net neural network, comparing our model with U-Net (VGG16),U-Net (Resnet50), PSPNet (Mobilenet), PSPNet (Resnet50), and Deeplabv3 (Mobilenet)， PSPNet (Resnet50), Deeplabv3 (Mobilenet), the average cross-merge ratio

of our proposed network is improved by 5.3%, 2.59%, 8.45%, 3.7%, and 4.88%, respectively. As shown in **Table 2**:

Table 2-The valiation results of mIoU(%) in the proposed method. The bold values in each column means the best entries.

Model	Backbone	PF	HD	RO	RI	SR	**Combined**	Growth
Unet	Vgg16	6.47	43.99	40.45	64.18	37.11	51.68	5.3
Unet	Resnet50	12.66	67.97	74.16	65.71	40.15	54.39	2.59
PSPnet	Mobilenet	4.72	59.17	35.44	64.2	35.89	48.53	8.45
PSPnet	Resnet50	13.87	53.71	39.48	66.91	**41.03**	53.28	3.7
Deeplabv3	Mobilenet	8.66	39.61	40.25	63.91	35.54	52.10	4.88
AttU-Net	Resnet50	**15.98**	**71.32**	**75.06**	**67.17**	37.21	**56.98**	/

Applying the evaluation metric of pixel accuracy (mPa), our model improves 4.38%, 2.71%, 6.54%, 3.07% and 3.89% respectively compared to the above networks. The segmentation accuracy of different species in the SUIM dataset is explained in detail in this paper, compared to other networks, our model has significant segmentation capability on the SUIM underwater image segmentation dataset The average pixel accuracy of single species segmentation is 91.51%, 89.55%, and 82.3%, respectively, as shown in **Table 3**:

Table 3-The valiation results of mPa(%) in the proposed method. The bold values in each column means the best entries.

Model	Backbone	PF	HD	RO	RI	SR	**Combined**	Growth
Unet	Vgg16	6.9	49.58	43.88	**90.54**	48.56	62.73	4.38
Unet	Resnet50	14.62	77.46	91.12	90.08	52.29	64.4	2.71
PSPnet	Mobilenet	5.3	68.57	37.38	90.34	45.33	60.57	6.54
PSPnet	Resnet50	16.0	64.54	43.49	89.6	**55.0**	64.04	3.07
Deeplabv3	Mobilenet	10.42	43.2	44.58	89.33	48.68	63.22	3.89
AttU-Net	Resnet50	**19.39**	**82.3**	**91.51**	89.55	49.37	**67.11**	/

We selected six typical segmentation image cases to show the segmentation results, as shown in **Figure 4**. From the images, we can intuitively know that our model achieves better results in segmentation

accuracy and segmentation type misclassification rate, and also achieves better edge segmentation results and spatial location alignment for object categories such as humans and machines that require deep information for feature classification.

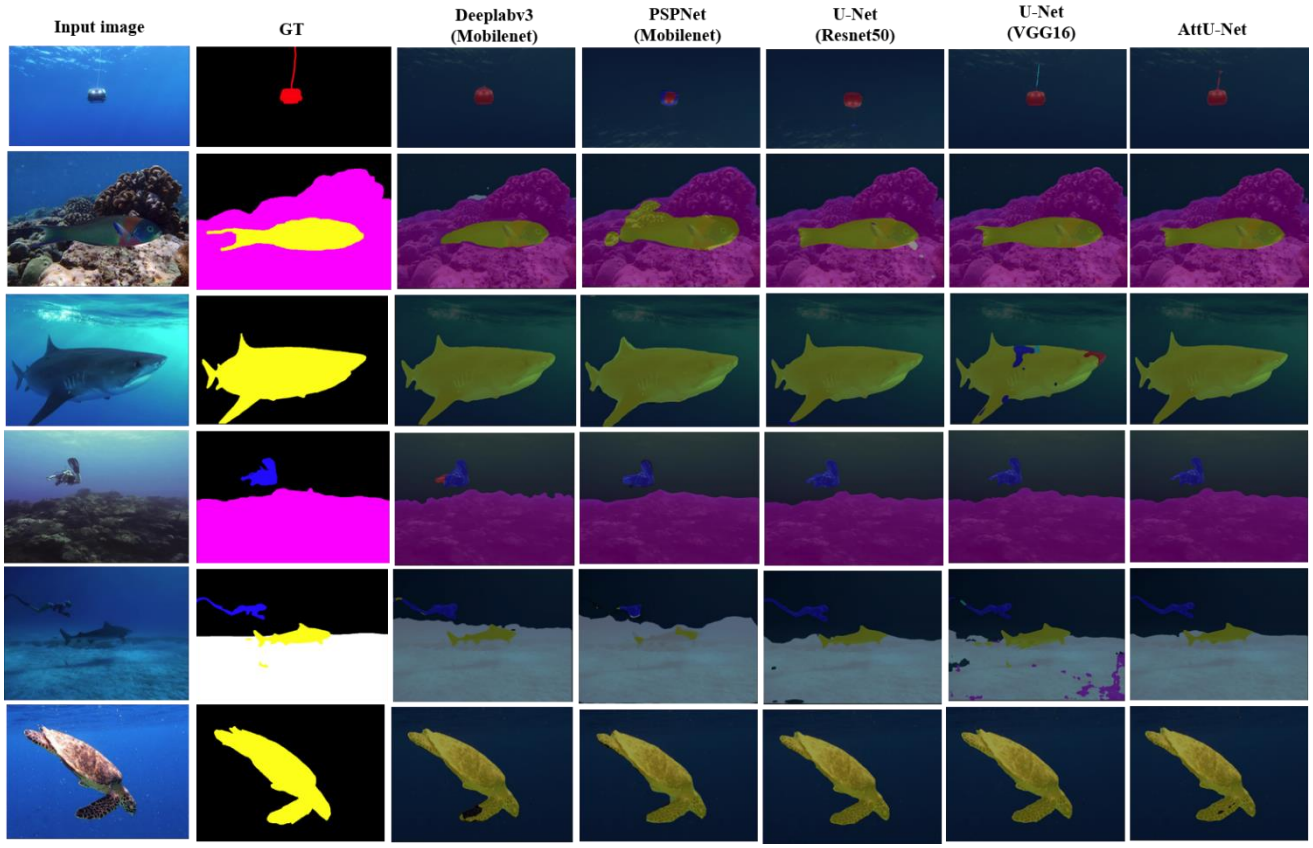

Figure 4 .Image of underwater image segmentation results for the SUIM dataset, showing the segmentation results of our model versus the four models used for comparison.

We recorded the average cross-merge ratios for the five models of the comparison experiment at batches of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100, respectively, as shown in **Figure 5**. From this line graph, we can visualize the degree of excellence of the segmentation results of different models.

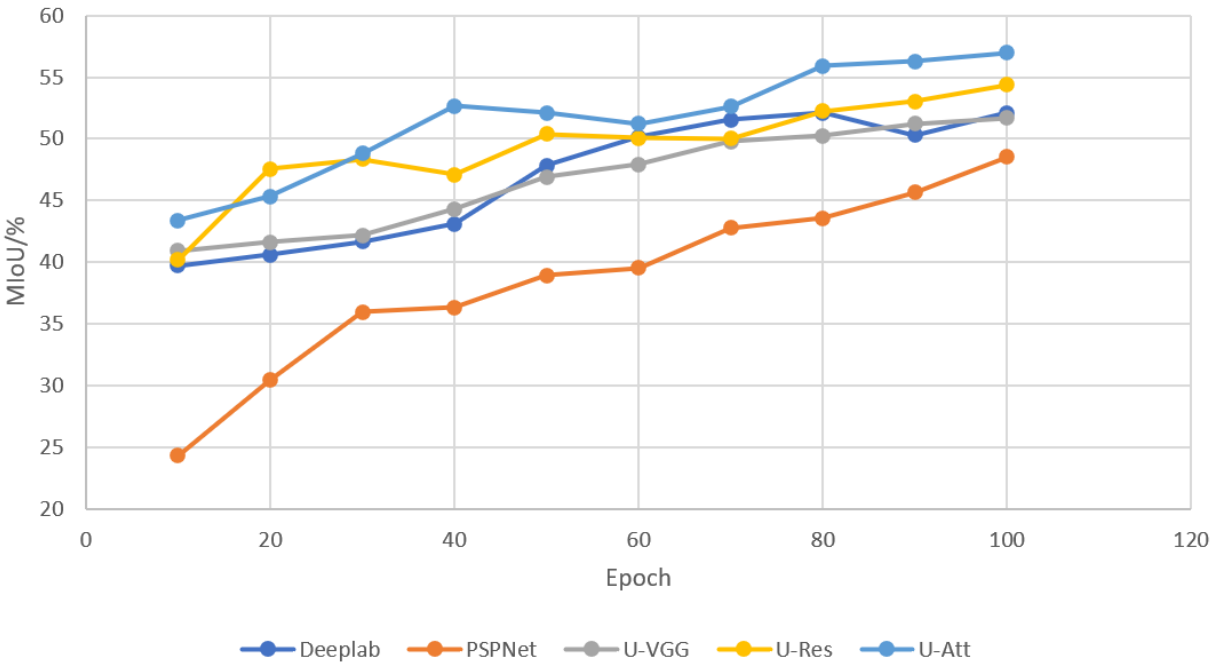

Figure 5. Presentation of training data for different batches of 5 comparison models.

5 DISCUSSION

In this study, we design an underwater image segmentation network incorporating a hybrid attention mechanism and residual connectivity, and experiments show that the combination of these features improves the accuracy and precision of the segmentation model and maintains good segmentation performance despite the complexity of the imaging environment. The addition of the residual network makes the model learn more deep semantic features characterizing object categories, and the extraction of deep features is crucial on multi-classification tasks. In this paper, the comparison of ablation experiments reveals that our model achieves 91.51% and 89.55% pixel accuracy for human and machine segmentation effects; and the addition of hybrid attention makes the spatial location of features more instructive, and the U-Net hopping layer The link structure ensures that the shallow semantic information is well preserved while guiding the extraction of deep features, which in turn improves the segmentation accuracy of the model.

However, the experiments found that although the model in this paper has better advantages for complex classification tasks, its performance on segmenting simple-shaped objects is less impressive. Although the segmentation effect is better, the U-Net model needs to store more data and parameters during the training process, so it is not suitable for tasks with high real-time performance. It is necessary to make a comprehensive consideration between guaranteeing the model segmentation accuracy and running speed for selection.

DATA AVAILABILITY STATEMENT

The dataset analyzed for this study can be found in the SUIM SUIM dataset | Minnesota Interactive Robotics and Vision Laboratory (umn.edu).

AUTHOR CONTRIBUTIONS

Sheng gave revisions and assistance with the research methodology and training experiments for this thesis, and Liang performed the experiments, analysis, and report writing for this thesis. All authors were involved in revising the manuscript, proofreading, and approving the submitted version.

FUNDING

This research was funded the Science and Technology Program of Social Development, Zhuhai, 2022(grant number 2220004000195).

REFERENCES

Long, J., Shelhamer, E., and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* ,pp. 3431-3440.

Ronneberger, O., Fischer, P., and Brox, T. ,2015, October. U-net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical image computing and computer-assisted intervention* ,pp. 234-241.

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. ,2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *In Proceedings of the European conference on computer vision* ,pp. 801-818.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. ,2017. Pyramid scene parsing network. *In Proceedings of the IEEE conference on computer vision and pattern recognition* ,pp. 2881-2890.

He, K., Zhang, X., Ren, S., and Sun, J. ,2016. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* ,pp. 770-778.

Woo, S., Park, J., Lee, J. Y., and Kweon, I. S. ,2018. Cbam: Convolutional block attention module. *In Proceedings of the European conference on computer vision*, pp. 3-19.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., ... and Tang, X. ,2017. Residual attention network for image classification. *In Proceedings of the IEEE conference on computer vision and pattern recognition* ,pp. 3156-3164.

Hu, J., Shen, L., and Sun, G. ,2018. Squeeze-and-excitation networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition* ,pp. 7132-7141.

Jaderberg, M., Simonyan, K., & Zisserman, A. ,2015. Spatial transformer networks. *Advances in neural information processing systems, 28.*

Lester, B., Al-Rfou, R., & Constant, N. ,2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Islam, M. J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., ... and Sattar, J. ,2020, October. Semantic segmentation of underwater imagery: *Dataset and benchmark. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*,pp. 1769-1776.

Simonyan, K., and Zisserman, A. ,2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... and Adam, H. ,2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., and Li, J. ,2019. Dice loss for data-imbalanced NLP tasks. *arXiv preprint arXiv:1911.02855.*

Zhang, Z., and Sabuncu, M. ,2018. Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems, 31.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., and Garcia-Rodriguez, J. ,2017. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857.*