Image Segmentation Based On U-Net and Adjusted U-Nets

Jixun Pan *

New York University, New York, United States

* Corresponding Author Email: jp6418@nyu.edu

Abstract. Image segmentation, the process of dividing an image into various areas is quite crucial recently. Numerous industries, including robotics, remote sensing, and medical imaging, have applications for this task. In recent years, deep learning techniques, especially U-shaped networks (U-nets), have shown remarkable success in solving image segmentation problems. This paper provides an overview of image segmentation using neural networks, introduces different types of adjusted U-nets used for this task, including the implementation of attention gates and the use of residual neural network as the encoder path based on the original encoder-decoder structure, and then uses U-net and adjusted U-nets to conduct image segmentation on the black sea sprat. The study uses dice similarity coefficient and binary cross entropy loss function to compare the model training results and further judges the functionality of the models by the predicted segmented images. According to the test results, the Res34-UNet with attention gates performs most efficiently in segmenting this image dataset, although it's more unstable compared to the basic U-Net.

Keywords: U-shaped network; image segmentation; attention gates.

1. Introduction

Image segmentation has received great attention recently and is used in various industries, such as treating ailments more effectively in medical imagery and utilizing in remote sensing (RS) photos to pinpoint particular objects. Companies also aim to present products to clients online without background distractions.

After machine learning has gradually been used in a large number of applications, researchers have realized that neural networks can be an efficient and accurate way to solve image segmentation. Ciresan et al. [1] use Deep Neural Network (DNN) to segment and make the implementation on GPU over 50 times faster than on microprocessors. Long et al.[2] reference the idea of Matan et al. [3] of fully convolutional computation and construct the fully convolutional network (FCN), which can take arbitrary-sized inputs and compute the outputs of the same size. They use deconvolution layers for upsampling, and design a skip architecture to improve precision [2]. FCN has reached the top of accuracy and training speed for semantic segmentation in the chosen datasets, compared with VGG net, AlexNet, and GoogleLeNet [2]. Afterwards, the encoder-decoder architecture has made further advances in the accuracy and efficiency of image segmentation [4], most notably the U-shaped network (U-net).

Based on FCN, Ronneberger et al. [5] adjust and broaden its structure and build U-net, which is able to deal with few training images and produce more accurate results. They add more feature channels in the upsampling part to enhance the propagation of information to higher resolution layers [5]. The U-net employs the valid portion of each convolution and infers the missing context by reflecting the input content in the absence of fully connected layers [5]. The structure reduces GPU memory constraints when segmenting large images [5], and addresses the limitation of FCNs by enabling the network to learn global context information [6]. U-net brings a significant progress in image segmentation, taking less than 1 second to segment a 512x512 image on a recent GPU, and with better accuracy than other networks [5]. Due to its superior performance, U-net and advanced models based on U-net have become a widely used deep learning technique in medical imaging, such as retinal fundus imaging, microscopy, ultrasound, X-Ray, and fighting for COVID-19 [6].

Using the U-net as the base model structure, researchers have proposed models that may further improve accuracy or efficiency. Çiçek et al. [7] build 3D U-net, which changes all of the 2D operations into 3D operations, enabling the model to handling 3D volumetric segmentation. Utilizing

attention mechanisms, Oktay et al. [8] build the attention U-net, every single layer of which has an attention gate in the decoder route. Using the attention gate's mechanism, the feature map is weighted in accordance with each class, allowing the network to concentrate on a certain class [6, 9]. The U-net architecture allows distinct network segments to focus on the segmentation of various objects without adding an excessive amount of computing complexity [6]. Diakogiannis et al. [10] design ResUNet, introducing Residual Network (ResNet) that He et al. [11] propose for image recognition into the U-net's encoder-decoder architecture. To avoid degradation and enhance the capacity of networks with many more layers, ResNet incorporates residual connections, which involves incorporating the feature maps from one layer into a deeper layer [6, 10]. Modified ResNets constitute the encoder path of the structure of ResUNet. According to the results of Diakogiannis et al. [10], the ResUNet conditioned multitasking performs excellent for segmenting remotely sensed images compared with other models.

In this paper, the author adapts U-net, Attention U-net, ResUNet, and combines attention gate with ResUNet, to conduct image segmentation of black sea sprat, a common commodity in the marketplace, and analyze the effectiveness of each model.

2. Method

2.1. Dataset

The dataset utilised for the research is the black sea sprat of the large fish dataset, which is collected from a supermarket's seafood department in Izmir, Turkey [12]. It contains 1000 pairs of images of the black sea sprat and respectively its masks. Each image's format is PNG, with 590 width, 445 height.

2.2. Data Preprocessing

Combine the fish image file and masks file into one dataset. Through inspection, there is no significantly incorrectly labeled data.

2.2.1 Image decode.

Decode the PNG file into the tensors that can be operated in TensorFlow.

2.2.2 Image resize

Resize all images into 256 width, 256 width.

2.2.3 Image normalization

Divide the pixel values by the maximum value, which is 255 in this case. The procedure converts the initial values to ranges of 0 to 1.

2.3. Data Augmentation

In image segmentation, data augmentation is frequently utilized to enhance the capability to generalize of deep learning models and prevent overfitting. However, it also costs large memory and might lead to underfitting. In this study, the model trained without data augmentation is compared with the model trained with horizontal flip and contrast augmentation.

2.4. Models

In the research, the author trains U-net, Attention U-net, ResUNet, and Attention ResUNet to conduct image segmentation. All models are implemented in TensorFlow with Adam [13] optimizer.

2.4.1 U-net.

Based on FCN, U-net has symmetrical downsampling layers and upsampling layers, making the architecture look like U. Between the levels of the encoder and decoder, it further adds skip connections.

In the path of the encoder, every block comprises two consecutive 3 × 3 convolutions, succeeded by a Rectified Linear Unit (ReLU) activation unit, which has been proved to perform better than sigmoid function, and a max-polling layer. The setup is used repeatedly four times. The encoder path is followed by a bridging section, consisting of the same convolutions and activation function as the encoder path.

Each stage engages in upsampling the feature map utilizing 2 × 2 up-convolutions in the decoder path. The encoder layer's feature map is cropped and fused with the corresponding upsampled feature map. Two continuous 3 × 3 convolutions and a ReLU activation come after the fusion. This configuration is also applied four times repeatedly. After these processes, finally, to compress the feature map to 3 channels, an extra 1 × 1 convolution operation is performed, ultimately generating the segmented image. Since the contracting path and expansive path are basically symmetrical, the structure of the model is similar to the U shape. The structure of U-net is shown in Figure 1.

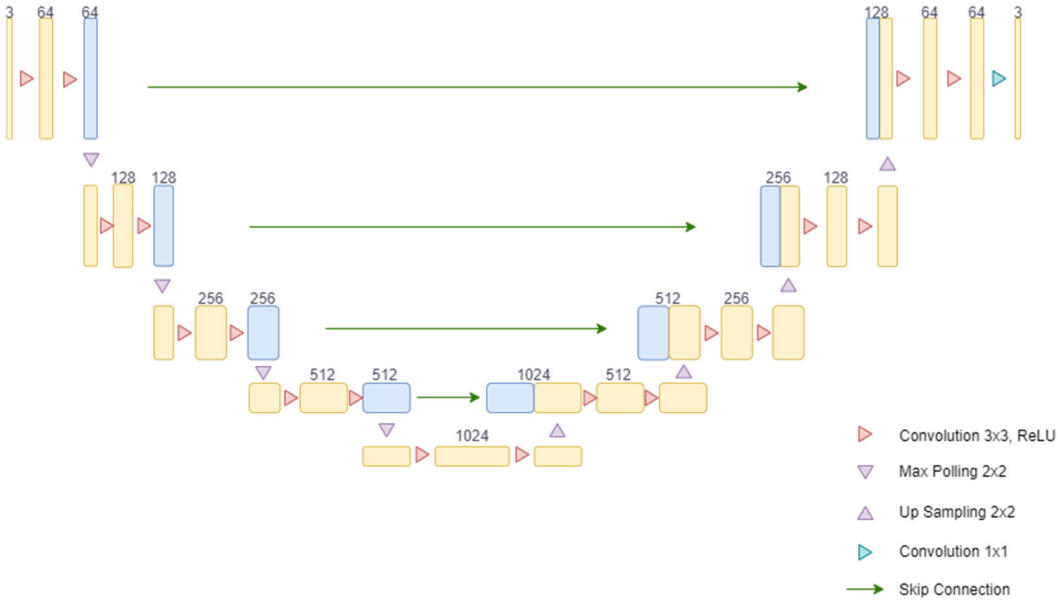


Fig. 1 U-net architecture

2.4.2 Attention U-net.

Attention U-net adds attention gates to every layer of the decoder path, while other parts are the same as vanilla U-net. The attention gate firstly performs a 1x1 convolution operation on each of the two input tensors to perform a feature transformation on the input tensor, then applies a batch normalization operation to each of the new tensors formed, then sums the two feature tensors and activates them via ReLU, then conducts a 1x1 convolution operation on the previous result, and finally, restricts the outputs between 0 and 1 by activating the outputs using a Sigmoid activation function [14]. Features transmitted across the skip connections are scrutinized by the attention gates. The structure of Attention U-net is shown in Figure 2.

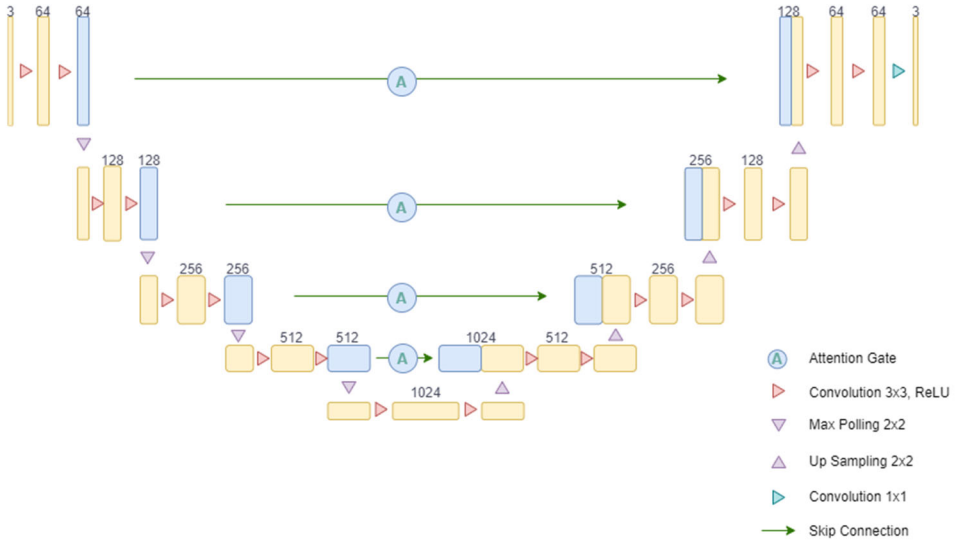


Fig. 2 Attention U-net architecture

2.4.3 Res34-UNet.

ResNet 34 makes up the encoder portion of the encoder-decoder architecture in the Res34-UNet model. The structure of Res34-UNet is as Figure 3. In the encoder part, ResNet34 performs five downsamplings, consisting of one 7×7 convolution operation and four convolution blocks. The four convolution blocks in order have 7, 8, 12, and 6 layers. Skip connections are added between each two layers of ResNet 34. Each ResNet 34 block's feature map is cropped and fused with the corresponding upsampling feature map.

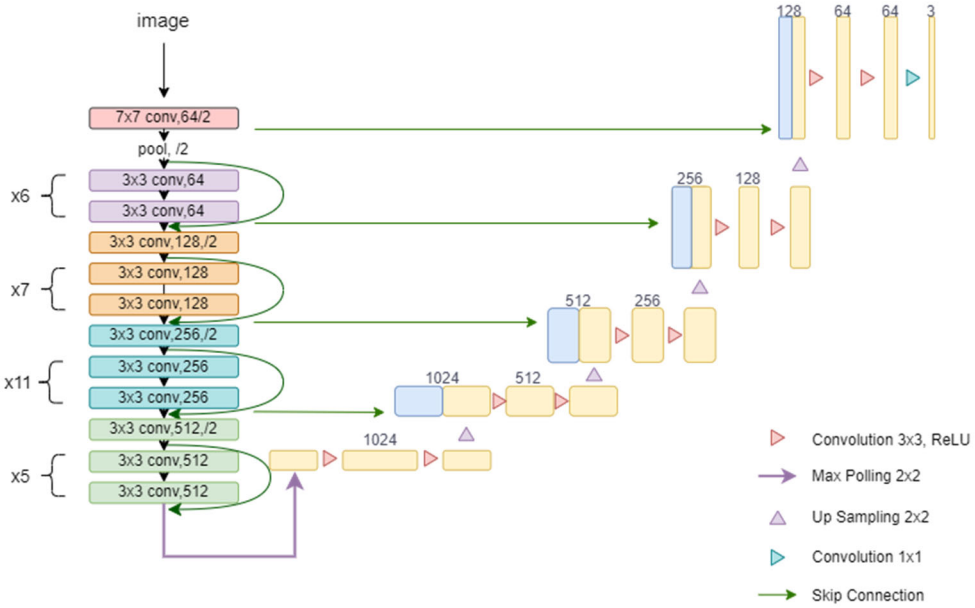


Fig. 3 Res34-UNet architecture

2.4.4 Attention Res34-UNet.

Attention Res34-UNet adds attention gates to the expanding path of Res34-UNet, filtering features communicated through skip connections between encoder levels and decoder layers. The structure of Attention Res34-UNet is shown in Figure 4.

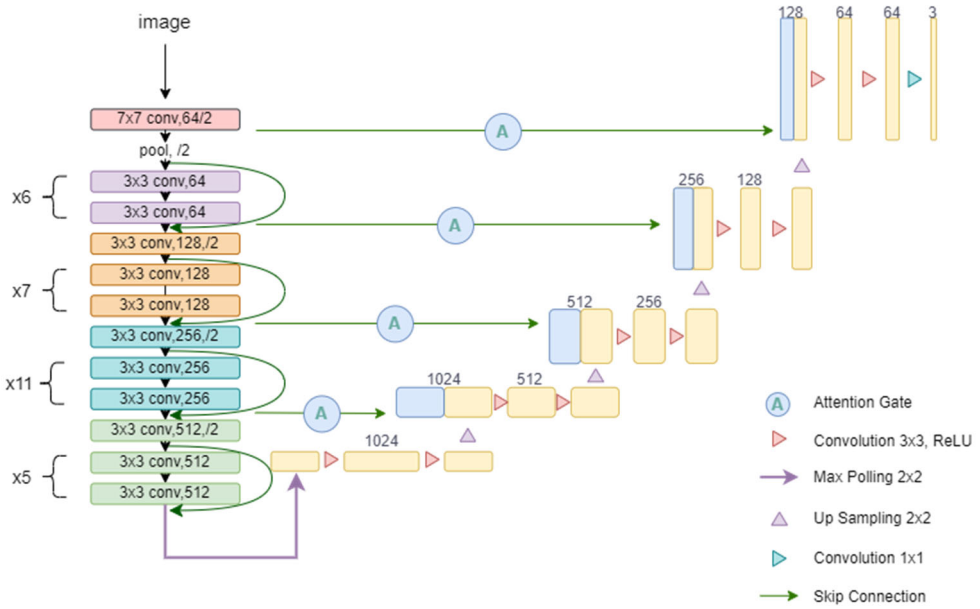


Fig. 4 Attention Res34-UNet architecture

2.5. Evaluation Criteria

2.5.1 Loss

The research adapts binary cross entropy (BCE) loss function to evaluate the results, which is used to calculate the gap between the actual labels and predicted results, widely used in image classification and segmentation works [15].

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i)\log(1 - p_i) \tag{1}$$

In this function, y represents the actual label image, y_i represents an individual component within that label, and p_i represents the corresponding prediction of the output image. The BCE loss function penalizes the model more severely when it makes confident incorrect predictions. When the result of the loss function is close to 0, the model tends to have good accuracy.

2.5.2 Dice similarity coefficient (DSC)

DSC evaluates how closely X and Y sets resemble one another. The sizes of each set are denoted by the letters |X| and |Y|, and the amount to which X and Y overlap is denoted by the symbol $|X| \cap |Y|$. The value of DSC varies between 0 to 1. When the DSC is closer to 1, it means that the predicted mask has more overlap with the original mask, and the closer it is to 0, it indicates that the overlap is less.

$$DSC = \frac{2 \times Area\ of\ overlap}{Total\ Area} = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{2}$$

2.5.3 Intersection over Union (IoU)

IoU also computes the intersection between two data sets. Similar to DSC, it has a 0-1 scale, and closer to 1 means more overlaps between the original mask and the predicted result. What makes it different is that it calculates the proportion of overlap of |X| and |Y| to the union of |X| and |Y|, which is $|X| \cup |Y|$. The DSC and IoU are positively linked, while IoU penalizes more on under and over segmentation [16].

$$IoU = \frac{Area\ of\ overlap}{Area\ of\ union} = \frac{|X \cap Y|}{|X| \cup |Y|} \tag{3}$$

3. Result

3.1. Data Augmentation

Figure 5 displays examples of the training dataset after data augmentation.

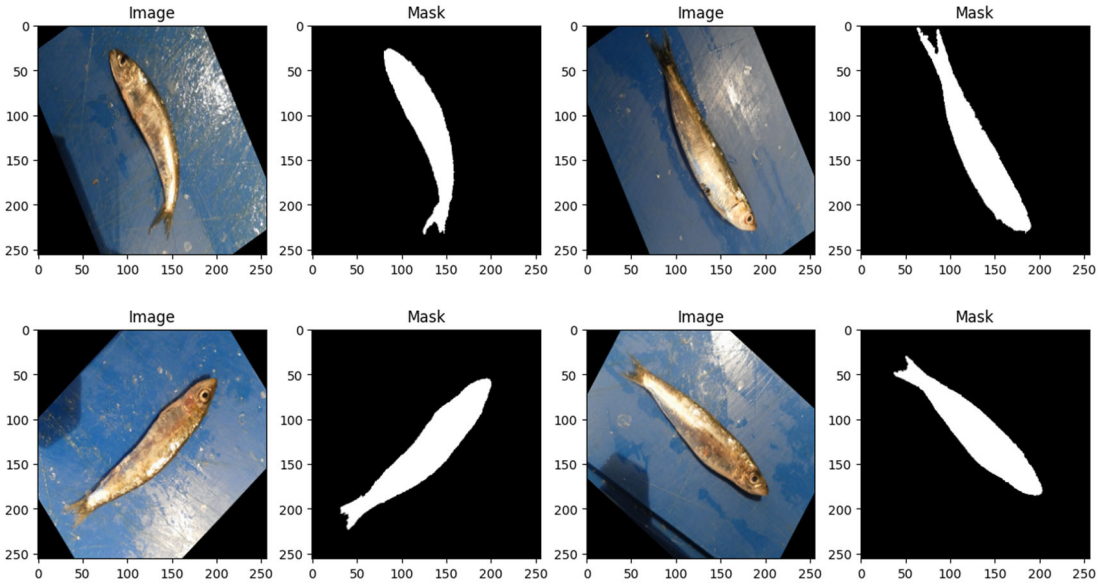


Fig. 5 Training data set after data augmentation

3.2. U-Net

The U-net training outcomes after 20 epochs of training in training sets with and without data augmentation are shown in Figure 6.

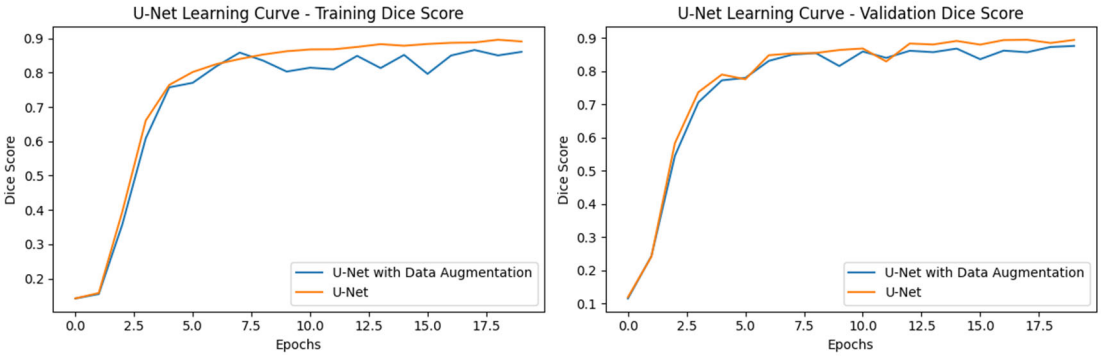


Fig. 6 Learning curves of U-net with and without data augmentation

After training 20 epochs, U-net without data augmentation has gotten a higher score in validation set. After training7 epochs, U-net with data augmentation has reached its local maximum value on training set, but later the value goes down. Meanwhile, the learning curve of the U-net without data augmentation on training set is smoother, while the model with data augmentation fluctuates. On the validation set, both models have encountered two fluctuations with a sudden drop in DSC. In general, the data-augmented U-net performs with worse accuracy than the non-data-augmented U-net, both on the training and test sets. Hence the author trains U-net without data augmentation for 50 epochs. The result is as following Table 1 and Figure 7.

Table 1. Results of U-net

	Loss	DSC	IoU
Training	0.0197	0.9385	0.4692
Validation	0.0159	0.9489	0.4744

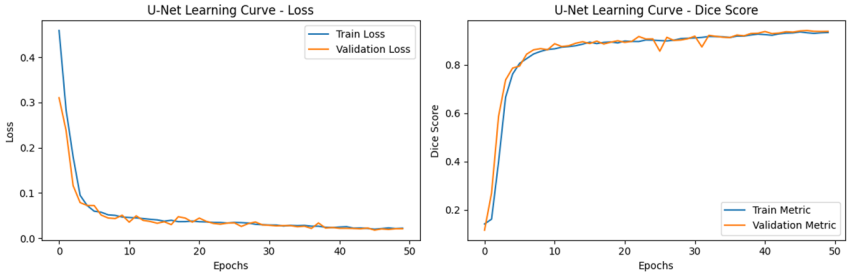


Fig. 7 Learning curves of U-net during 50 epochs

According to the Figure 7 and Table 1, the U-net performs with high precision after training 50 epochs. The value of loss is less than 0.05 on both sets, close to 0. The DSC score has also reached 0.9489. It indicates that the model is not underfitting. From the learning curve, it's clear that the training curve is very close to the validation curve and the results gradually improve with increased training, which indicates that the model is not overfitting. Despite two notable sudden drops in DSC on the validation set, all curves are generally smooth.

3.3. Attention U-Net

After training 20 epochs in training sets with data augmentation and without augmentation, the learning curves of Attention U-net are as Figure 8.

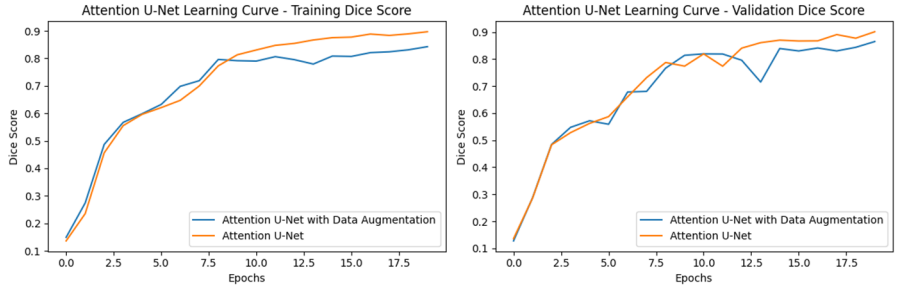


Fig. 8 Learning curves of Attention U-net with and without data augmentation.

According to the Figure 8, on the training set, the model without data augmentation eventually overtakes Attention U-net with data augmentation after achieving better DSC values over the first seven epochs. Two curves overlap more on the validation set, but the model without data augmentation achieves higher DSC scores after 12 epochs and maintains this trend, and the magnitude of fluctuations is smaller than that of the model with data augmentation.

Train Attention U-net without data augmentation for 50 epochs.

Table 2. Results of Attention U-net

	Loss	DSC	IoU
Training	0.0162	0.9500	0.4750
Valdation	0.0154	0.9499	0.4749

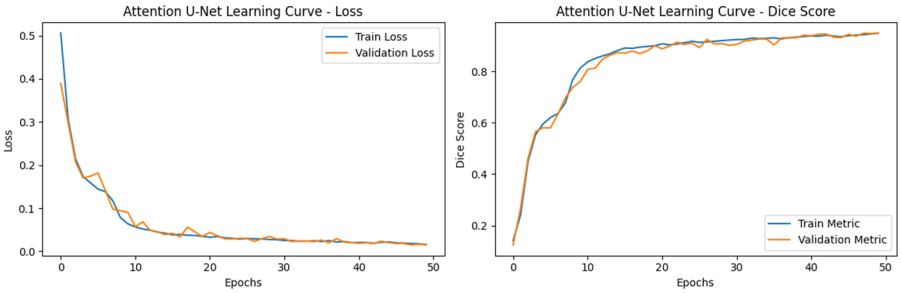


Fig. 9 Learning curves of Attention U-net during 50 epochs.

According to the Figure 9 and Table 2, the Attention U-net curves on the validation set shows some relative fluctuations, but the magnitude is smaller than that of U-net. The curves on the training and prediction sets are also close to overlapping. The final loss value is less than that of U-net and DSC value is greater than that of U-net, demonstrating enhanced precision on the validation set conducted by trained Attention U-net.

3.4. Res34-UNet

In the first 20 epochs training, the trending of Res34-UNet with and without augmentation are not very clear. After 50 epochs in training sets with data augmentation and without augmentation, the learning curves of Res34-UNet are as Figure 10.

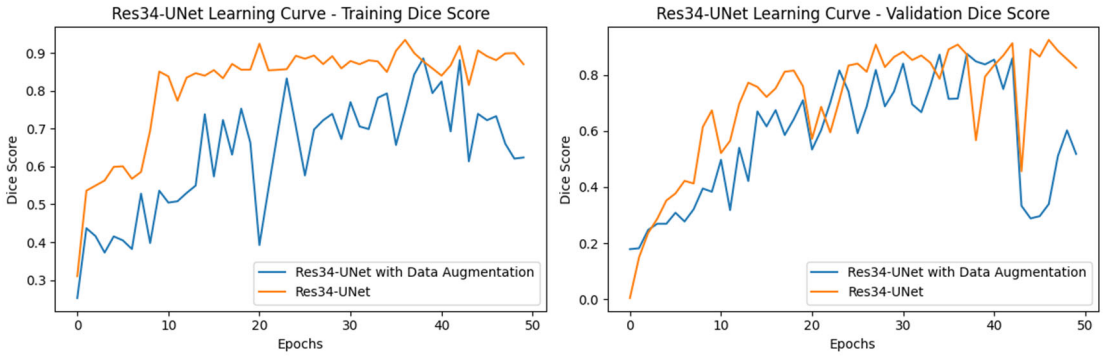


Fig. 10 Learning curves of Res34-UNet with and without data augmentation.

From Figure 10, the curves are fluctuating dramatically, with the data-augmented curve being more variable and having lower minimum values, while the maximum values are still smaller than the model without data enhancement.

Table 3. Results of ResNet34-UNet

	Loss	DSC	IoU
Training	0.0637	0.8699	0.4350
Valdation	0.0581	0.8251	0.4126

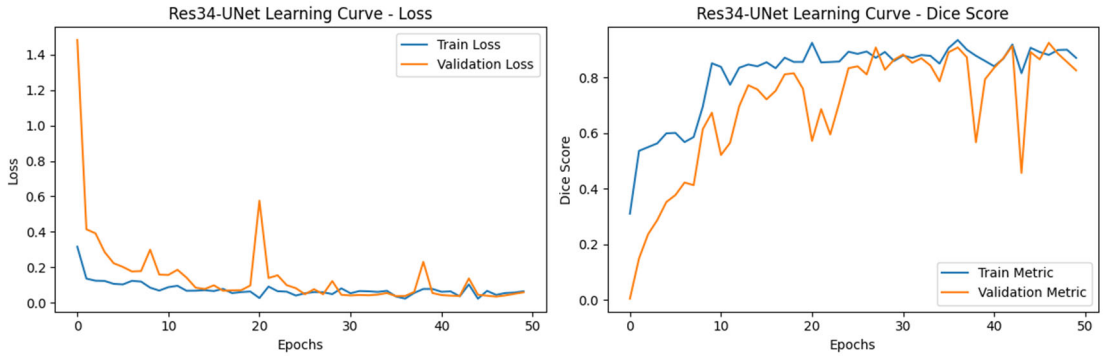


Fig. 11 Learning curves of Attention U-net during 50 epochs.

According to Table 3 and Figure 11, the learning curve of Res34-UNet is significantly more fluctuating than that of the two models mentioned before, showing a sudden increase in the loss five times and a significant decrease in the DSC value seven times. With more training epochs, the abrupt reduction in DSC value's magnitude grows. The final loss value is higher and the DSC value is lower than other models after 50 training epochs.

3.5. Attention Res34-UNet

After 20 epochs in training sets with data augmentation and without augmentation, the learning curves of Attention Res34-UNet are as Figure 12.

Fig. 12 Learning curves of Attention Res34-UNet with and without data augmentation

From Figure 12, the DSC results for the Attention Res34-UNet without data augmentation are significantly higher than the model with data augmentation on the training set, which fluctuates extremely after 20 training epochs. On the valdiation set, although the DSC of the model with data augmentation reaches the local maximum at the 8th and 15th epochs, even slightly higher, it is extremely unstable and drops rapidly.

Train the Attention Res34-Unet for 50 epochs. The results are as Table 4 and Figure 13.

Table 4. Results of Attention Res34-UNet

	Loss	DSC	IoU
Training	0.0136	0.9578	0.4789
Validation	0.0138	0.9593	0.4796

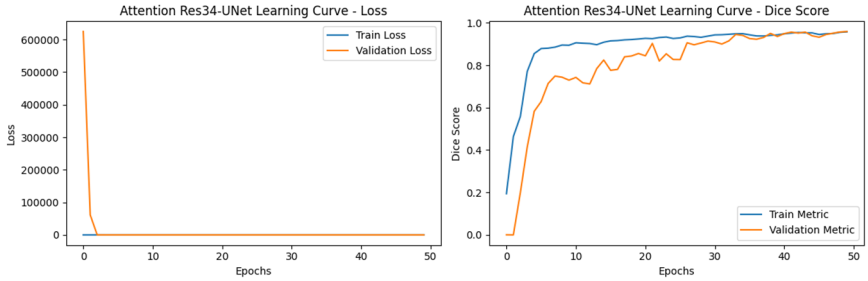


Fig. 13 Learning curves of Attention Res34-UNet during 50 epochs

According to Figure 13 and Table 4, the initial loss value is very high but drops off quickly. The training curve of the DSC values is smooth. The curve of the validation set has a wide gap between the curves of the training set and fluctuates a lot at the beginning, but gradually converges to the training curve and basically overlaps at the end of the training.

3.6. Other Models

The author also builds and trains FCN, which reaching loss value of 0.2057, DSC value of 0.3802, IoU value of 0.1901 on the validation set after 20 epochs, significantly more underfitting than the U-net and adjusted U-nets. ResNet 34 and ResNet 50 are also less accurate on the validation set. When training the U-net with pretrained ResNet50 as encoder, the model appears to be severely overfitting and the issue cannot be resolved by data augmentation. Therefore, those models are out of the description and comparison of this paper.

3.7. Summary the Results

The results of four models after 50 training epochs are as Figure 14.

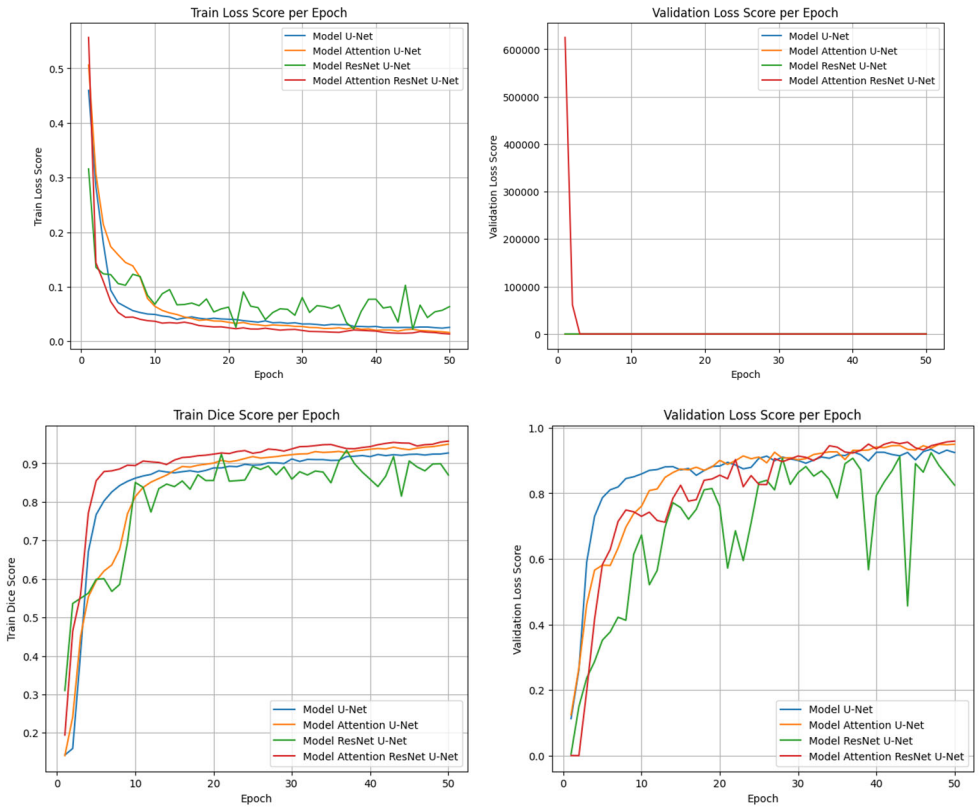
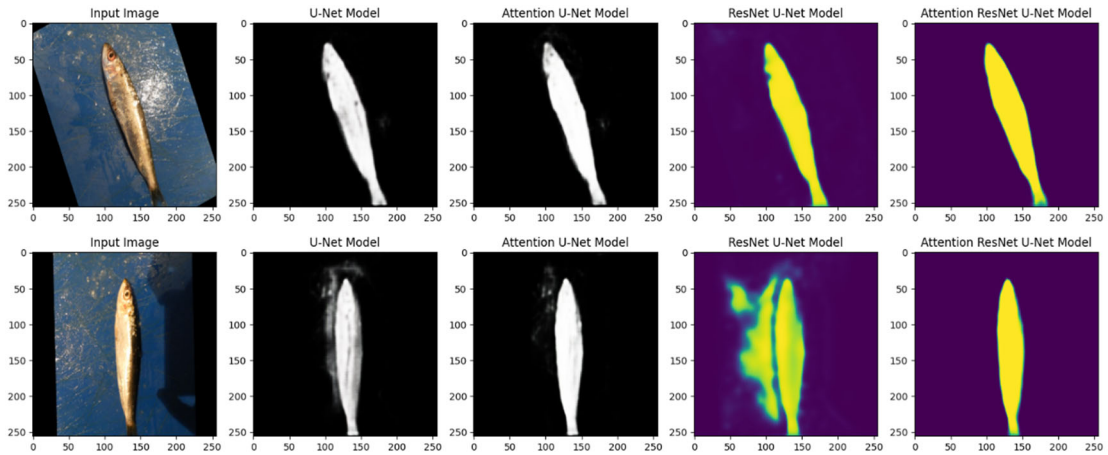


Fig. 14 Learning curves of U-net, Attention U-net, Res34-UNet, Attention Res34-UNet during 50 epochs.

According to the Figure 14 and tables listed above, Attention Res34-UNet has won the best score of DSC, IoU and loss after training 50 epochs. It has a smooth training learning curve similar to that of U-net, while it achieves better loss and DSC scores finally on the validation set though it shows more fluctuations than other three models. Compared with the rapid rise in the DSC value of U-net, DSC value of Attention U-net rises more slowly, but exceeds U-net after 15 epochs on both the training and validation sets, and possesses higher scores than U-net at the end of the training, close to Attention Res34-UNet.

3.8. Prediction

Make predictions using trained U-net, Attention U-net, Res34-UNet, and Attention Res34-UNet. The outcomes of the prediction are shown in Figure 15.



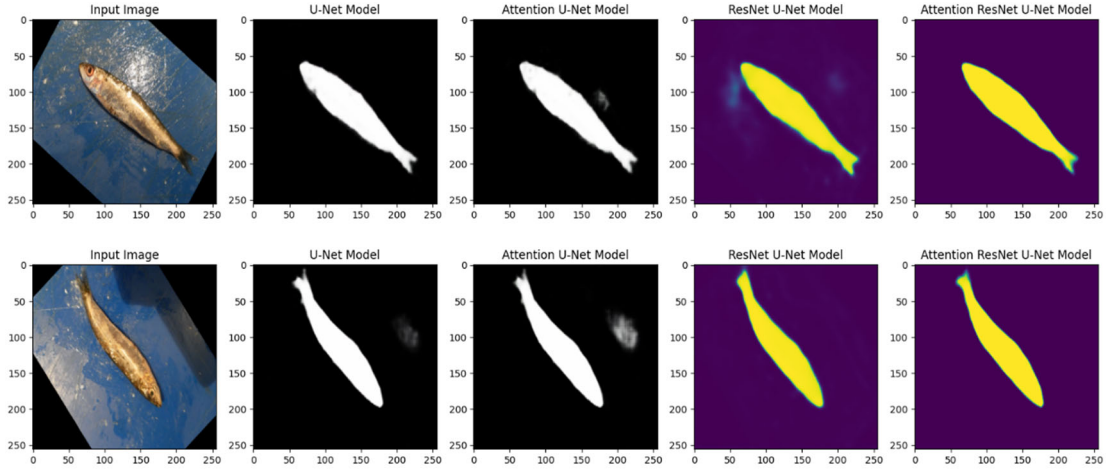


Fig. 15 Prediction Results of each model

From the plots, it is clear that Res34-UNet don't have precise result to distinguish fish and its background. But with attention gates, the improved Attention Res34-UNet performs better than other models, avoiding background interference to a greater extent.

4. Discussion

Comparing the results of four models, Attention Res34-UNet has the highest value of DSC, reaching 0.9593 on the validation set, while the DSC values of U-net, Attention U-net, Res34-UNet are 0.9489, 0.9499, and 0.8251 in order. Res34-UNet, however, has the lowest value of DSC, and there are large areas of discrepancy between the predicted results and the actual mask. Attention gates also improve the performance of U-net, but not as significantly as it improves Res34-UNet. From the prediction results, both U-net with attention gates and vanilla U-net present incomplete segmentation of commodity and background with gray areas. And Attention Res34-UNet solves these problems better.

Meanwhile, the learning curves of Attention U-net and vanilla U-net and is generally smooth. But the Res34-UNet's learning curves has meet extremely large fluctuations and a tendency to have greater fluctuations with increased training in terms of DSC. By adding Attention Res34-UNet, however, the learning curve of DSC on training set is as smooth as U-net. Though the learning curve of DSC on validation set also has fluctuations, these fluctuations are relevantly small and the curve trend continues to keep rising. It might indicate that attention gates improve model performance more dramatically on improved U-net with deeper layers.

During the research, data augmentation doesn't improve the accuracy of models. The author also adapts other methods of data augmentation, such as brightness, only gets lower results. How to use data augmentation to improve the accuracy still need to be found. It is also notable that all models don't win good enough IoU score, all lower than 0.5. It might result from over segmentation, which is penalized more on IoU.

5. Conclusion

The research tests U-net and other models on segmenting photos of black sea sprat, a common commodity. Attention gates improve the efficiency of U-net and Res34-UNet, especially the Res34-UNet. ResNet34 encoder doesn't guarantee efficiency, but has the potential for continued improvement.

During the research, the data augmentation doesn't bring better results on the research. There are more methods of data augmentation need to be complied. Meanwhile, the mask file in the dataset is not very accurate, which may have contributed to inaccuracies in the model.

Considering that the results of U-net are already very accurate, the author has tried to improve the results of U-net through various means, and finally found that the Res34-UNet with attention gates can achieve the best results without data enhancement. Though the Res34-UNet, the U-net architecture with ResNet 34 encoder, doesn't perform well enough compared with other models, after improved the model with attention gates, the Attention Res34-UNet has become the most effective model during the test, getting the DSC value of 95.93% on the validation set and good segmentation results.

References

[1] Ciresan, D., Giusti, A., Gambardella, L., & Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. Advances in neural information processing systems, 25.

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

[3] Matan, O., Burges, C. J., LeCun, Y., & Denker, J. (1991). Multi-digit recognition using a space displacement neural network. Advances in neural information processing systems, 4.

[4] Asgari Taghanaki, S., Abhishek, K., Cohen, J. P., Cohen-Adad, J., & Hamarneh, G. (2021). Deep semantic segmentation of natural and medical images: a review. Artificial Intelligence Review, 54, 137-178.

[5] Ronneberger, O., Fischer, P., & Brox, T. (2015). 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 (pp. 234-241). Springer International Publishing.

[6] Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). U-net and its variants for medical image segmentation: A review of theory and applications. Ieee Access, 9, 82031-82057.

[7] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-net: learning dense volumetric segmentation from sparse annotation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19 (pp. 424-432). Springer International Publishing.

[8] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

[9] He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., & Xue, Y. (2022). Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-15.

[10] Diakogiannis, F. I., Waldner, F., Caccetta, P., & Wu, C. (2020). ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 162, 94-114.

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

[12] Ulucan, O., Karakaya, D., & Turkan, M. (2020). A Large Scale Fish Dataset. Kaggle. https://www.kaggle.com/datasets/crowww/a-large-scale-fish-dataset

[13] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[15] Ruby, U., & Yendapalli, V. (2020). Binary cross entropy with deep learning technique for image classification. Int. J. Adv. Trends Comput. Sci. Eng, 9(10).

[16] Müller, D., Soto-Rey, I., & Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. BMC Research Notes, 15(1), 1-8.