Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.028225
Article

Tech Science Press

Sea-Land Segmentation of Remote Sensing Images Based on SDW-UNet

Tianyu Liu[1,3,4], Pengyu Liu[1,2,3,4,*], Xiaowei Jia[5], Shanji Chen[2], Ying Ma[2] and Qian Gao[1,3,4]

[1]The Information Department, Beijing University of Technology, Beijing 100124, China
[2]School of Physics and Electronic Information Engineering, Qinghai Minzu University, Xining, 810000, China
[3]Advanced Information Network Beijing Laboratory, Beijing, 100124, China
[4]Computational Intelligence and Intelligent Systems Beijing key Laboratory, Beijing, 100124, China
[5]Department of Computer Science, University of Pittsburgh, 15260, USA
*Corresponding Author: Pengyu Liu. Email: liupengyu@bjut.edu.cn

Received: 05 February 2022; Accepted: 07 April 2022

Abstract: Image segmentation of sea-land remote sensing images is of great importance for downstream applications including shoreline extraction, the monitoring of near-shore marine environment, and near-shore target recognition. To mitigate large number of parameters and improve the segmentation accuracy, we propose a new Squeeze-Depth-Wise UNet (SDW-UNet) deep learning model for sea-land remote sensing image segmentation. The proposed SDW-UNet model leverages the squeeze-excitation and depth-wise separable convolution to construct new convolution modules, which enhance the model capacity in combining multiple channels and reduces the model parameters. We further explore the effect of position-encoded information in NLP (Natural Language Processing) domain on sea-land segmentation task. We have conducted extensive experiments to compare the proposed network with the mainstream segmentation network in terms of accuracy, the number of parameters and the time cost for prediction. The test results on remote sensing data sets of Guam, Okinawa, Taiwan, San Diego, and Diego Garcia demonstrate the effectiveness of SDW-UNet in recognizing different types of sea-land areas with a smaller number of parameters, reduces prediction time cost and improves performance over other mainstream segmentation models. We also show that the position encoding can further improve the accuracy of model segmentation.

Keywords: Sea-land segmentation; UNet; depth-wise separable convolution; squeeze-excitation; position encoding

1 Introduction

Earth-observing satellites have been widely used for land use and land cover monitoring, natural resource management, and national defense. The remote sensing data collected by these satellites provide a great promise for data-driven modeling given their wide observation coverage, high imaging resolution, rich spectral information, and high acquisition frequency. In particular, remote sensing imagery of sea-land contains information of coastal landform, land surface and sea surface, which can reflect the change of coastal areas. Rapid and accurate segmentation of sea-land is critically needed for a variety of important tasks such as coastline extraction, marine environmental monitoring and near-shore target detection. Such technique can also be of far-reaching significance to development and management of China's coastal zone, the construction of national defense, and the monitoring of foreign coastal zone.

With the rapid development of deep learning [1–3], and its initial success in image discrimination, classification and segmentation, deep learning-based approaches (e.g., convolutional neural networks) has become the mainstream method for image-related tasks. One of the earliest attempts convolution-based image segmentation is the FCN method [4], which differs from classical convolution network, it use deconvolutional layers to achieve pixel-level classification of image. SegNet [5] is another segmentation network that using the VGG-16 backbone with encoder and decoder structure. It also has more detailed edge information of feature maps. UNet [6] proposed by Ronneberger et al. consists of symmetric encoder and decoder structure and layer-skipping connections. This structure also makes the network less demanding on the number of data samples. Attention UNet [7] proposed by Oktay et al. further adds attention gate mechanism to UNet. The attention gate between up-down sampling feature maps is performed to improve the relationship of up-down sampling. Chen et al. proposed the Deeplab [8,9] family of convolution networks to further optimize the feature fusion of multi-scale feature maps. There are also other works that leverage these segmentation techniques for specific applications. For example, Peng et al. used image segmentation to complete segmentation of cardiac magnetic resonance images [10]. Jiang et al. used adversarial segmentation to process multispectral fundus images [11]. Yang et al. adopted deep learning techniques for segmenting Landsat-8 OLI image [12].

Remote sensing image has multi-spectral information which will appear that the same spectrum corresponds to different objects, it is difficult to obtain accurate semantic information even for UNet with tightly linked information of up-down sampling, and it will lead to misjudgment of the same spectral target. Moreover, the large number of parameters in UNet and its time cost for prediction remain a major concern when it is applied to large regions. Tanalysis address these issue, this paper extends UNet by leveraging the squeeze-excitation [13] and the depth-wise separable convolution [14]. A new donwsampling convolution module SDW1 (Squeeze Depth-Wise separable conv1) and an upsampling convolution module SDW2 (Squeeze Depth-Wise separable conv2) are created, and they are used to modify sampling structure of Unet. The final proposed model SDW-UNet consists of staked layers of the modified sampling structure. By using the squeeze-excitation and depth-wise separable convolution, SDW-UNet allows weight optimization and allocation, improves the capability of information extraction, reduces network parameters, and improves execution efficiency. Furthermore, given the promise of self-attentive-transformer in image processing [15,16], we design an additional experiment in which the position encoding obtained from the transformer is used to embed the feature map position information and enhance the restoration of image information.

Experiments on remote sensing data sets of Guam, Okinawa, Taiwan, San Diego and Diego Garcia show that SDW-UNet improves the capability of feature extraction to remote sensing images, and reduces the network parameters and prediction time. Besides the improved segmentation results, we also demonstrate that the segmentation accuracy can be further promoted by leveraging the position encoding.

2 Pre-Work

Image segmentation [17–19] aims to classify the pixel points into different categories. In short, different objects are represented by different color masks. The mature network structures take use of the feature extraction module to obtain feature maps of different sizes and fuse these feature maps together, then use a discriminator to determine the classes of pixels in the feature maps. The feature extraction module is the core of the network, we explore this part in order to obtain module with high extraction power and low computational consumption. In addition, we also investigate the fusion part and propose a position encoding module, which is expected to better preserve the pixel information of each layer feature maps.

We have considered two neural network structures, as shown in Fig. 1. The former achieves segmentation through feature-extraction layers and the following aggregation layers of feature maps processed by different dilation factors [20]. The latter is accomplished by an encoder-decoder structure. Both constructions can perform the segmentation tasks well, and we would like to introduce the position encoding for further experiments, while the second network construction is similar to the encoder-decoder structure of transformer, which the feature map information of each stage can be processed well, so the latter is chosen for exploration.

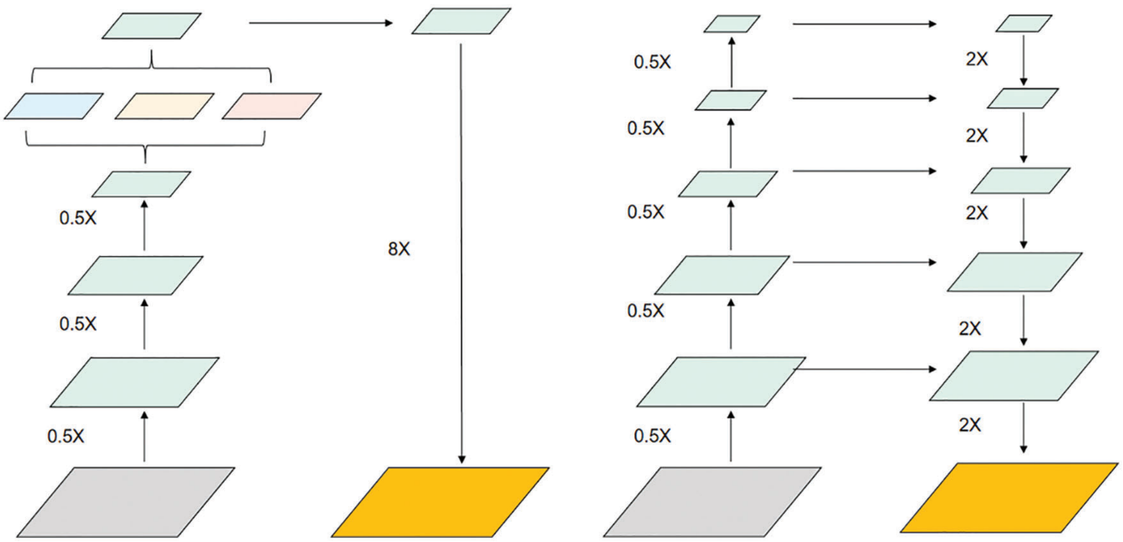

Figure 1: Two structures

Position encoding is a method that uses location information to provide a secondary representation of each word in sequence, and is represented for each pixel point in the feature map when it is used to image. Position encoding is first used in NLP domain to obtain the information about the position of word in sequence, as shown in Fig. 2. It is important for text sequences, where the meaning of a word may deviate from the whole sentence if it is in a different position or in a different order of arrangement. Recent research has shown that accuracy gains can also be obtained by using positional coding on images [21]. Therefore, we add position encoding to the feature map based on SDW-UNet in order to explore the performance improvement.

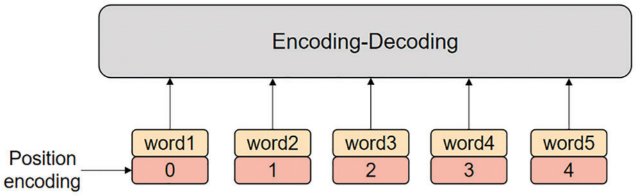

Figure 2: Position encoding in NLP

3 Methods

In this section, we will introduce the basic principles of depth-wise separable convolution and squeeze-excitation in Section 3.1 and Section 3.2. Then we will describe the specific structure of convolution modules SDW1 and SDW2 in Section 3.3. Finally, Section 3.4 gives the network framework of SDW-Net, and Section 3.5 introduces the application of position encoding on feature maps.

3.1 Depth-Wise Separable Convolution

To solve the problem of excessive amount of parameters in UNet, we use depth-wise separable convolution to construct convolution modules, as shown in Fig. 3. Here each feature map channel is processed using a single convolution kernel channel(lower part of Fig. 3). Compared with the standard convolution operation that uses multi-channel convolution kernel to process one feature map(upper part of Fig. 3), depth-wise separable convolution is better optimized for matrix multiplication using a smaller number parameters and improves the computational efficiency.

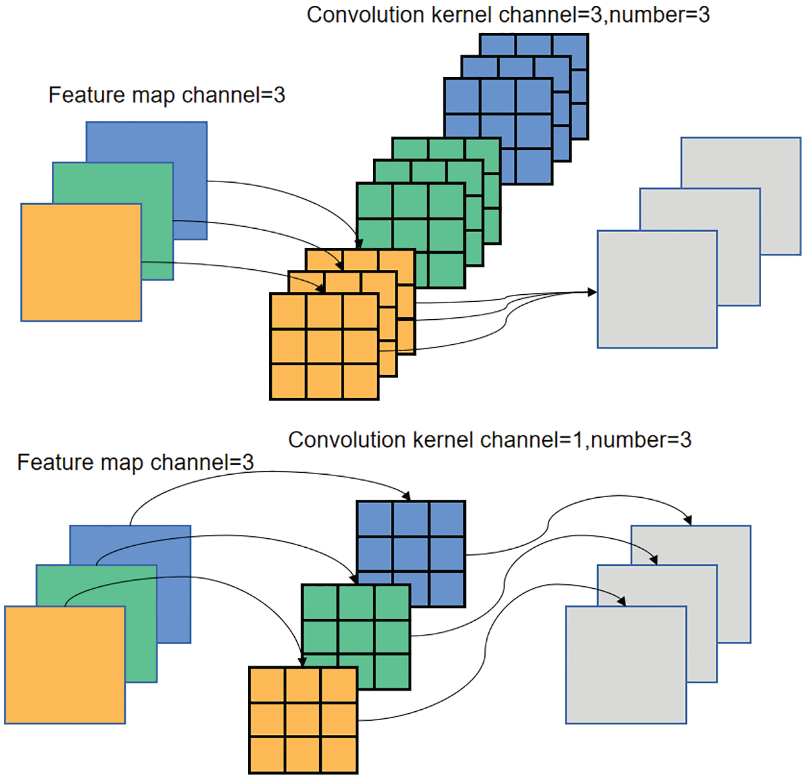
Convolution kernel channel=3,number=3
Feature map channel=3
Convolution kernel channel=1,number=3
Feature map channel=3

Figure 3: Depth-wise separable convolution

3.2 Squeeze-Excitation

We adopt a similar attention mechanism as used in Attention UNet, which uses attention gate for up-down sampling feature maps, as shown in Fig. 4. Squeeze-excitation performs average pooling for each channel of the feature map to obtain a one-dimensional vector. Then it uses two additional fully connected layers to get the attention weights, which indicates the weight relationship for each channel of the feature map and assigns higher weights to important channels. Finally, each feature channel is multiplied by the corresponding attention weight to obtain the new feature map. Squeeze-excitation can effectively improve semantic information of feature maps, enhance the feature representation and segmentation performance of network.

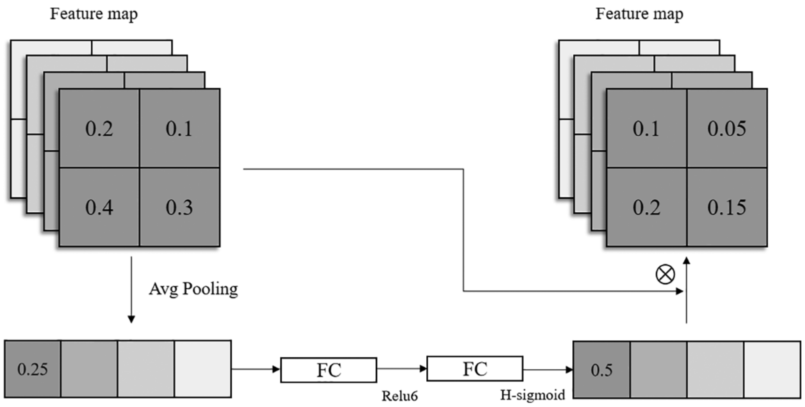

Figure 4: Squeeze-excitation

3.3 Structure of Convolution Modules

In this paper, we propose an enhanced UNet model using SDW1 and SDW2. SDW1 is designed to replace the downsampling process of UNet, as shown in Fig. 5 left. Similar to the inverse residual structure of MobileNetV2 [22], in the lower part, the number channel is increased by 1 × 1 convolution with a multiplication of α, then the Depth-Wise separable convolution and the Squeeze-excitation module are used to extract information and improve the weight distribution, and finally the number of channel is reduced by 1 × 1 convolution. In the upper part, the model achieves extract information by Depth-Wise separable convolution, and then uses Squeeze-excitation to improve weight assignment and uses a 1 × 1 convolution to refine information. Finally a shortcut branching similar to ResNet [23] is used to skip-connect feature maps from different layers and achieve feature fusion. SDW2 is proposed to replace some structures of upsampling, as shown in Fig. 5 right. The depth-wise separable convolution and Squeeze-excitation module are used to improve network performance, then the number of channel is reduced by 1 × 1 convolution. This design of SDW1 and SDW2 reduce the number of parameters greatly while maintaining the capability of information extraction and restoration. It is also noteworthy that our proposed SDW1 and SDW2 modules are applicable to many other segmentation architectures that involve downsampling and upsampling processes.

In order to speed up the convergence of network during training, the Batch Normalization [24] method is used after each convolution layer in sampling, so that the feature maps satisfied distribution law with mean 0 and variance 1. For better extracting low-dimensional feature information, the ReLu activation function Eq. (1) which used in UNet is replaced with the ReLu6 activation function Eq. (2).

$$y = max(0,\ x) \tag{1}$$

$$y = min(max(0,\ x),\ 6) \tag{2}$$

Compared with the ReLu activation function, ReLu6 activation function sets the part larger than 6 to 6, as shown in Fig. 6. It can effectively moderate the phenomenon that the ranges of weights differ too much and reduce the loss of low-dimensional information.

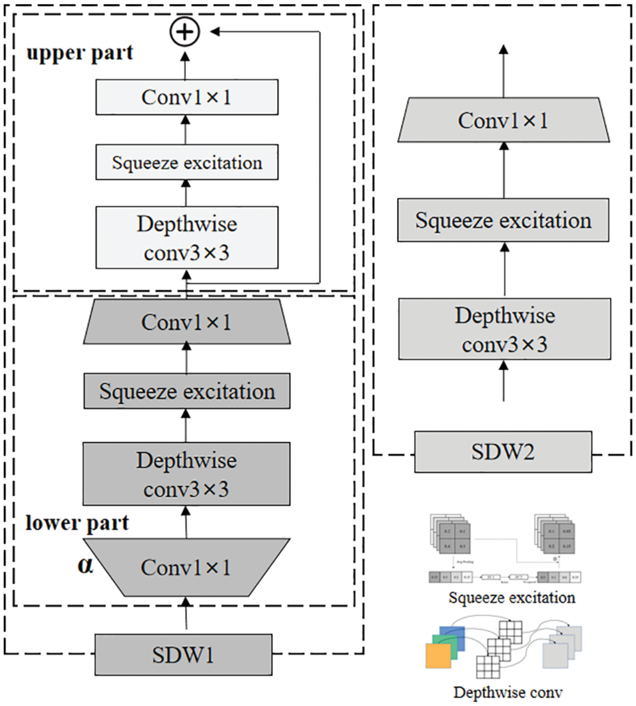

Figure 5: SDW1 and SDW2

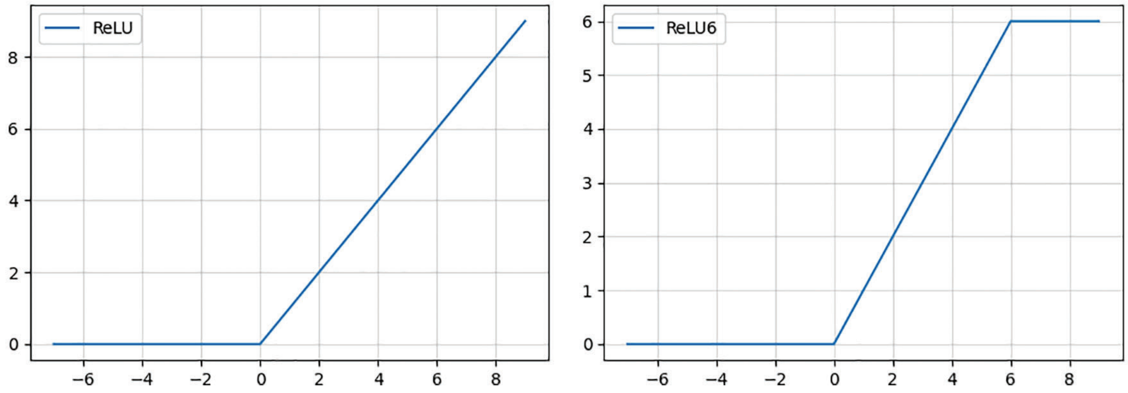

Figure 6: ReLu and ReLu6

3.4 Structure of SDW-UNet

We combine the new structure into UNet. First, two 3 × 3 convolutions are used to enhance the channel dimension, then the maximum pooling layer is stacked with SDW1 to build a downsampling structure to achieve information extraction, the upsampling convolution is stacked with SDW2 to build an upsampling structure to achieve information reduction. The layer-skipping connection is performed between up-down sampling feature maps at the same stage, in order to enhance connection of sampling. Lastly we use 1 × 1 convolution to output the result. The network structure of SDW-UNet is shown in Fig. 7.

3.5 Structure of Position Encoding

Positional encoding is often used to record the positional information of the original data in the field of NLP, which is important for text sequences, where the meaning of a word may deviate from the whole sentence if it is in a different position or in a different order of arrangement. Previous research has demonstrated that positional coding can also be applied to the image to record the location information of pixel points. There are three attributes of length-width-channel for satellite imagery and thus it requires the modeling of the spatial information when designing position encoding. Fig. 8 shows how we design the position encoding. We generate a zero matrix with its length, width, and channel numbers same with the corresponding feature map, and we assign a specific number for each matrix entry. Then we take the specific number values from relative position bias table and fill them into the zero matrix. These feature matrices contain position information and are fused with the corresponding feature maps for enhancing the restoration ability of image information. These matrices remain trainable during the model optimization process are trainable.

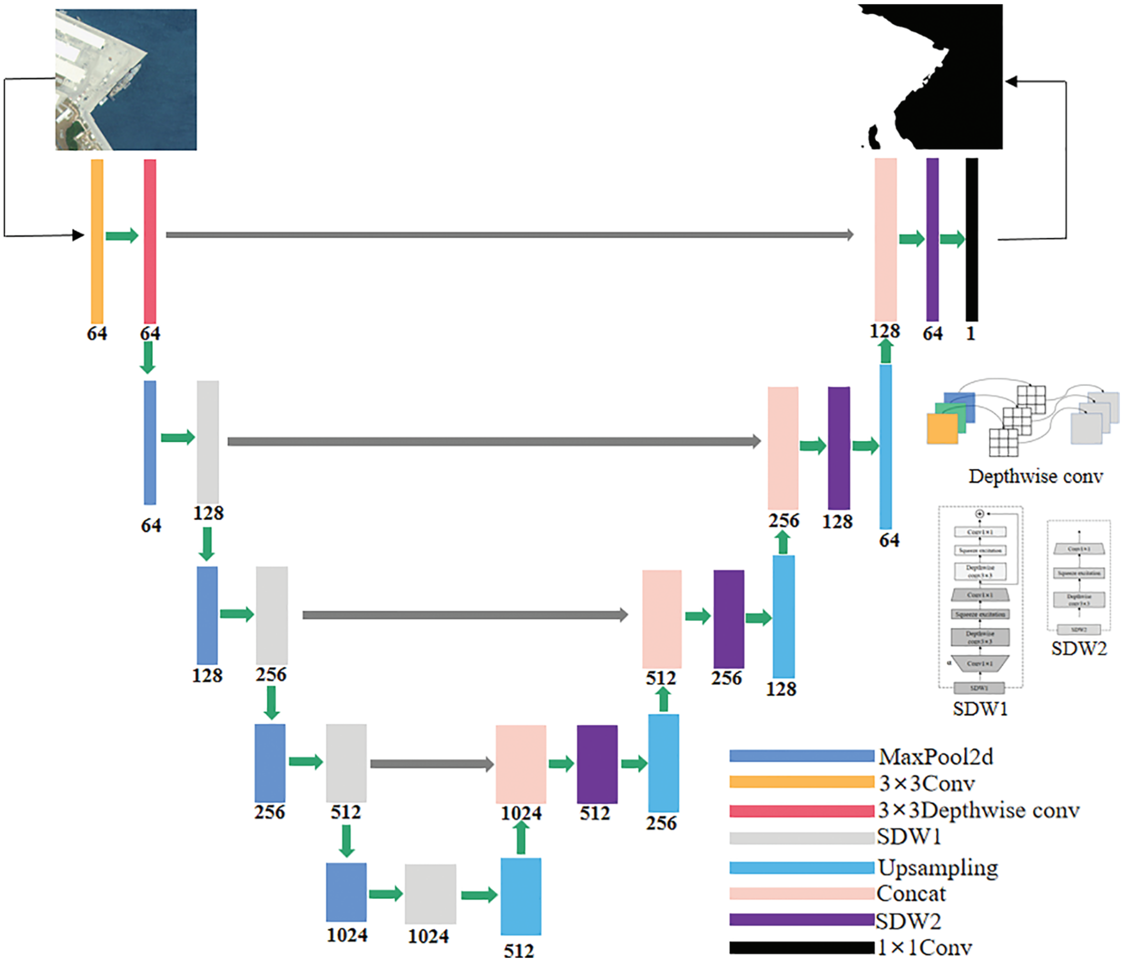

Figure 7: Structure of SDW-UNet

4 Experiment and Evaluation

4.1 Data Sets

The data sets used for experiment are from Beijing II remote sensing satellite, which covering Okinawa, Taiwan, Guam, San Diego and Diego Garcia, with a resolution of 0.8 m, and captured in years 2019, 2020 and 2021. These images are cropped and filtered to 512 × 512 pixel images, then annotated them to generate labels with white represents land and black represents ocean amount 320. The data sets are divided into training sets and testing sets in ratio of 9:1, the training sets are sent to network for training in order to build model, and the testing sets are used to achieve comparison and analysis of experiment.

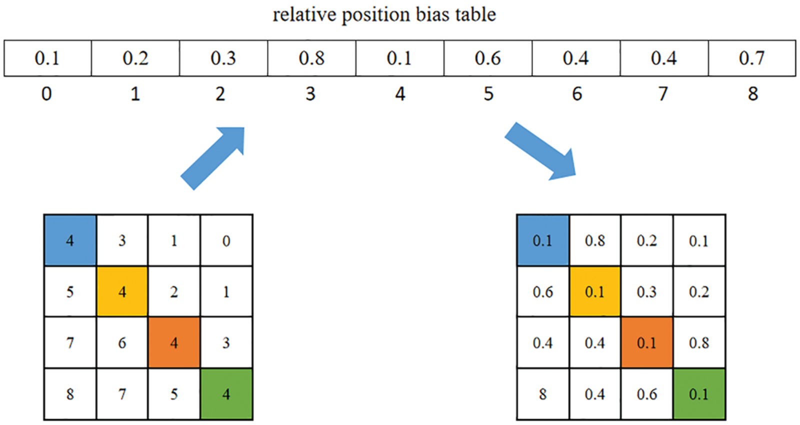

Figure 8: Position encoding

4.2 *Experimental Environment*

The experiments in this paper are trained by RTX3090 24GB GPU, the programming environment is python3.6, the deep learning framework is Pytorch1.6+torchvision0.7. We use Adam as optimizer, the initial size of learning rate is 5×10^{-4}, and learning rate is adaptive adjusted by ReduceLROnPlatea with decay factor of 0.9. The Batchsize is set to 8, and image size is uniformly scaled to 256×256. Considering limited amount of data, the epoch of training is set to 40 to prevent over-fitting. The loss function is selected as MSELoss, final we save the lowest loss model at the end of training.

4.3 *Experimental Results and Analysis*

To demonstrate the advantages of SDW-UNet, we select four convolution segmentation networks as comparison: UNet, Attention UNet, SegNet and Deeplabv3+(xception16). We use 32 images in our tests and we consider the metrics including IOU (Intersection-over-Union) Eq. (3), network parameters, and the prediction cost.

$$IoU = TP/(TP + FP + FN) \tag{3}$$

IOU has comprehensive evaluation criteria, where TP represents positive prediction and positive true value, FP represents positive prediction but negative true value, and FN represents negative prediction but positive true value. Network parameters refers to how many parameters the network contains, which directly determines the size of model file and also affects the memory usage during model predict. Network prediction time cost refers to the relative time consumed when using model to predict images, representing the efficiency of network.

Tab. 1 shows the performance comparison between SDW-UNet (boosting multiplier $\alpha = 2$) and other networks. It can be seen that SDW-UNet has the highest IOU on 32 test images, reaching over 95%, and has the lowest number of parameters only 1/3 of UNet. It is slightly slower than SegNet and Deeplabv3+ in terms of network prediction time and faster than UNet and Attention UNet. It can be concluded that SDW-UNet can extract sea-land information more effectively with high efficiency and smaller size.

We select six different types of images to visualize the results, these images contain complex and simpler scene information with different sea features, as shown in Fig. 9. By observing the results of different networks, it can be seen that SDW-Unet has better segmentation performance for both complex and simple scenes, especially in the position of red rectangle marked box, which reduces the false detection of similar features on sea-land.

Table 1: Comparison of different network performance metrics

	IOU	Network parameters	Network predict times
UNet	0.9462	34.53 m	4.78 s
Attention UNet	0.9457	34.88 m	5.11 s
SegNet	0.9452	29.44 m	3.42 s
Deeplabv3+(xception16)	0.9441	54.61 m	**3.40 s**
Our Method(α = 2)	**0.9520**	**12.62 m**	4.31 s

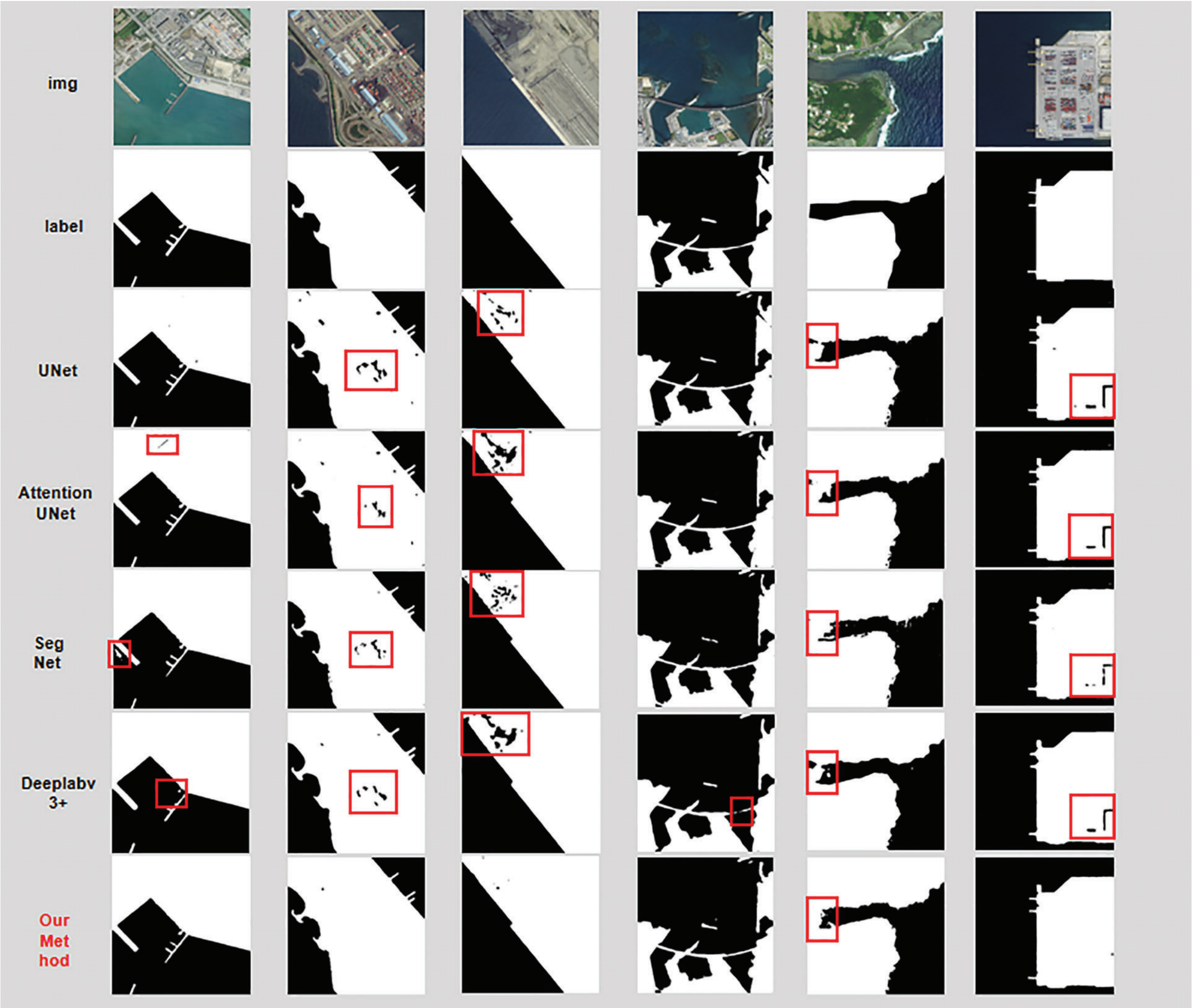

Figure 9: Visualization of results

In this paper, SDW1 is designed by adding channel boosting multiplier α in the first layer of 1 × 1 boosting convolution. To investigate the effect of channel's number on network performance metrics, the value of α is modified from 1 to 5, and the results obtained are shown in Tab. 2 and Fig. 10. Results show that with the increasing value of multiplicity α, the network parameters and prediction times are also increased, which in turn leads to higher IOU values. α allows the convolutional layer maps image to higher dimension, which means that more feature information can be extracted. But the high dimension makes the mapping process more complex and time-consuming. When the multiplicity α changes from 1 to 2, the larger IOU increase is obtained with smaller parameters and time increase.

Table 2: Comparison of different α performance indicators

	IOU	Network parameters	Network predict times
$\alpha = 1$	0.9500	11.04 m	4.17 s
$\alpha = 2$	0.9520	12.62 m	4.31 s
$\alpha = 3$	0.9529	14.55 m	4.52 s
$\alpha = 4$	0.9541	16.83 m	4.69 s
$\alpha = 5$	0.9554	19.95 m	4.91 s

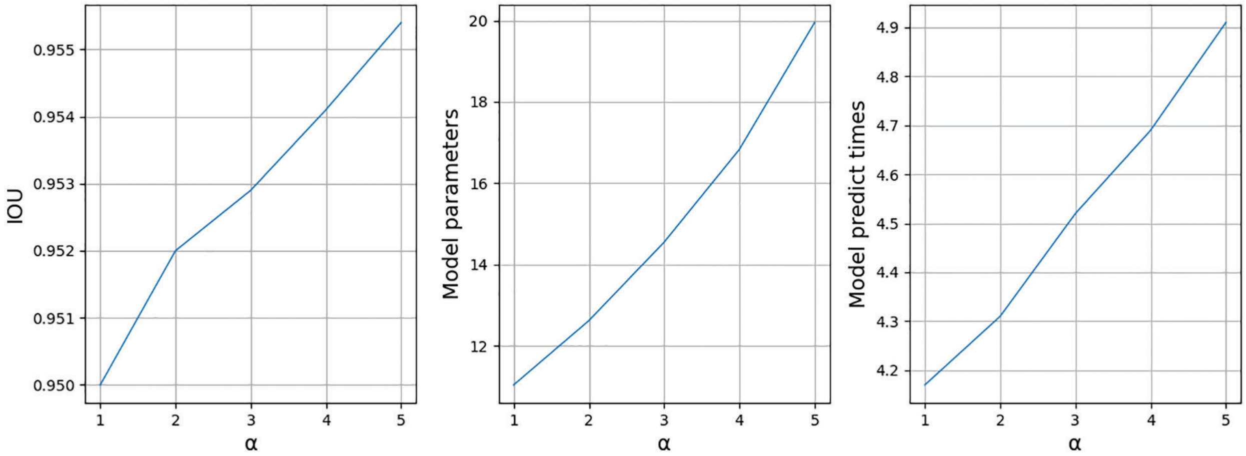

Figure 10: Comparison of different α performance indicators

Finally we explore effect of the proposed position encoding on IOU, the number of parameters and prediction time cost. The network structure is shown in Fig. 11. We chose SDW-UNet with α = 2 as basic framework and add position encoding to the corresponding feature map. The experiment results are shown in Tab. 3. These results indicate that position encoding can retain pixel location information of feature maps during feature fusion and improve the ability of image information reproduction effectively, but it brings a larger number of parameters and thus increases the prediction time. This needs to be considered for use when it comes to mobile deployments and other lightweight requirements.

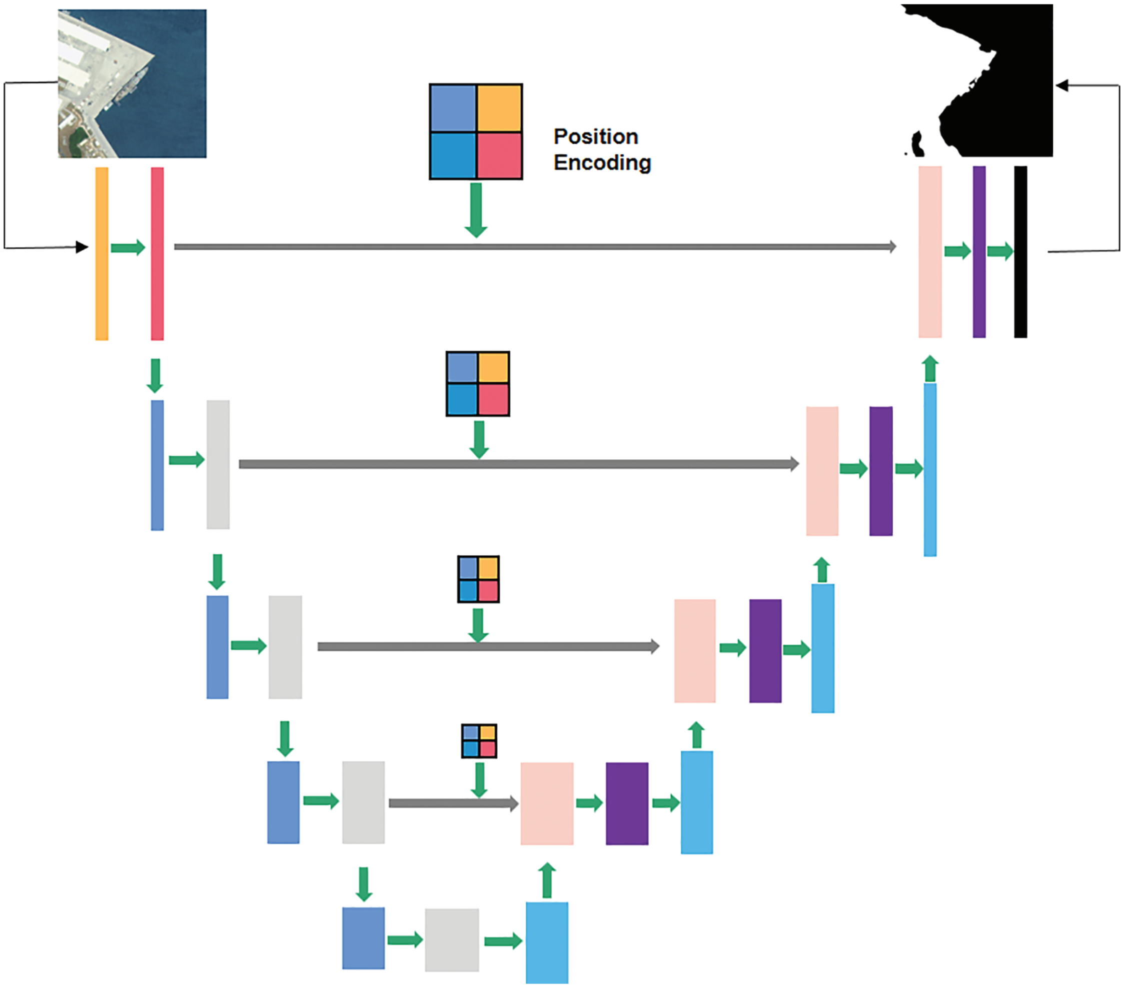

Figure 11: SDW-UNet with position encoding

Table 3: SDW-UNet with position encoding

α = 2	IOU	Network parameters	Network predict times
SDW-UNet	0.9520	12.62 m	4.31 s
SDW-UNet+position encoding	0.9556	44.08 m	4.52 s

5 Conclusion

This paper proposes a new deep segmentation model SDW-UNet network for sea-land remote sensing image segmentation tasks. It introduces the standard UNet as the base model, leverages the SDW1 and SDW2 modules to create a new sampling structure, in order to extract sea-land information and improve the segmentation capability of network. Moreover, these proposed modules reduct the number of model parameters significantly, which mitigate the need for large training data and improve the run time efficiency. We design three types of experiments: (1) comparing the SDW-UNet's performance metrics with mainstream classical networks; (2) testing the effect of boosting multiplier α on SDW-UNet's performance, and exploring the optimal multiplication factor meanwhile; (3) exploring the effect of the proposed position encoding on IOU, the number of parameters and prediction time. Experiments on remote sensing data sets show that SDW-UNet has better recognition ability in sea-land segmentation, possessing lower network parameters and less prediction time. Meanwhile the proposed position encoding improves the ability of image information reproduction.

Funding Statement: This paper is supported by the following funds: The National Key Research and Development Program of China (2018YFF01010100), The Beijing Natural Science Foundation (4212001), Basic Research Program of Qinghai Province under Grants No. 2021-ZJ-704, and Advanced information network Beijing laboratory (PXM2019_014204_5000 29).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] X. R. Zhang, X. Sun, W. Sun, T. Xu and P. P. Wang, “Deformation expression of soft tissue based on BP neural network,” *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 1041–1053, 2022.

[2] V. Krishna, N. Pappa and S. Rani, “Realization of deep learning based embedded soft sensor for bioprocess application,” *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 781–794, 2022.

[3] F. Bi, X. Ma, W. Chen, W. Fang, H. Chen *et al.,* “Review on video object tracking based on deep learning,” *Journal of New Media*, vol. 1, no. 2, pp. 63–74, 2019.

[4] E. Shelhamer, J. Long and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[5] V. Badrinarayanan, A. Kendall and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[6] O. Ronneberger, P. Fischer and T. Brox, “UNet: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015*, Munich, Germany, pp. 234–241, 2015.

[7] O. Oktay, J. Schlemper, L. L. Folgocs, M. Lee, M. Heinrich *et al.,* “Attention UNet: Learning where to look for the pancreas,” in *Medical Imaging with Deep Learning, MIDL 2018*, Amsterdam, Netherlands, 2018.

[8] L. C. Chen, Y. K. Zhu, G. Papandreou, F. Schroff and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *European Conf. on Computer Vision, ECCV 2018*, Munich, Germany, pp. 801–818, 2018.

[9] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[10] J. Peng, C. Xia, Y. Xu, X. Li, X. Wu *et al.,* “A Multi-task network for cardiac magnetic resonance image segmentation and classification,” *Intelligent Automation & Soft Computing*, vol. 30, no. 1, pp. 259–272, 2021.

[11] Y. Jiang, Y. Zheng, X. Sui, W. Jiao, Y. He *et al.,* “Asrnet: Adversarial segmentation and registration networks for multispectral fundus images,” *Computer Systems Science and Engineering*, vol. 36, no. 3, pp. 537–549, 2021.

[12] T. Yang, S. L. Jiang, Z. H. Hong, Y. Zhang, Y. L. Han *et al.,* “Sea-land segmentation using deep learning techniques for landsat-8 OLI imagery,” *Marine Geodesy*, vol. 43, no. 1, pp. 105–133, 2020.

[13] J. Hu, L. Shen and G. Sun, “Squeeze-and-excitation networks,” in *Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, USA, pp. 7132–7141, 2018.

[14] F. Chollet, “Xception: Deep learning with depthwise separable convolution,” in *Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, USA, pp. 1251–1258, 2017.

[15] V. Ashish, S. Noam, P. Niki, U. Jakob, J. Llion *et al.,* “Attention is all you need,” in *Neural Information Processing Systems, NIPS 2017*, Long Beach, USA, 2017.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai *et al.,* "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. on Learning Representations, ICLR 2021*, Vienna, Austria, 2021.

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler *et al.,* "The cityscapes dataset for semantic urban scene understanding," in *Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, USA, pp. 3213–3223, 2016.

[18] K. M. He, X. Y. Zhang, S. Q. Ren and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[19] G. Feng, Z. W. He, L. H. Zhang and H. C. Lu, "Encoder fusion network with co-attention embedding for referring image segmentation," in *Computer Vision and Pattern Recognition, CVPR 2021*, Online Meeting, pp. 15506–15515, 2021.

[20] R. Mottaghi, X. Chen, X. Liu, N. G. Cho, S. W. Lee *et al.,* "The role of context for object detection and semantic segmentation in the wild," in *Computer Vision and Pattern Recognition, CVPR 2014*, Columbia, USA, pp. 891–898, 2014.

[21] Z. Liu, Y. T. Lin, Y. Cao, H. Hu, Y. X. Wei *et al.,* "Swin transformer: Hierarchical vision transformer using shifted windows," in *Computer Vision and Pattern Recognition, CVPR 2021*, Online Meeting, pp. 10012–10022, 2021.

[22] M. Sandler, A. Howard, M. L. Zhu, A. Zhmoginov and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, USA, pp. 4510–4520, 2018.

[23] K. M. He, X. Y. Zhang, S. Q. Ren and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, USA, pp. 770–778, 2016.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. on Machine Learning, ICML 2015*, Lille, France, pp. 448–456, 2015.