A Novel Hybrid Model Based on CNN and Multi-scale Transformer for Extracting Water Bodies from High Resolution Remote Sensing Images

Qi Zhang[1], Xiangyun Hu[1,2,3,*] , Yao Xiao[4]

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, 430079, China - zhangqi_whursgcm@whu.edu.cn; huxy@whu.edu.cn
[2] Hubei Luojia Laboratory, Wuhan University, Wuhan, 430079, China.
[3] Institute of Artificial Intelligence in Geomatics, Wuhan University, Wuhan, 430079, China.
[4] Wuhan Geomatics Institute, Wuhan, China - xywhu2014@whu.edu.cn

KEY WORDS: Water Body Extraction, Remote sensing images, Deep learning, Convolutional neural networks, Transformer, Multi-scale features.

ABSTRACT:

Extracting water bodies from high-resolution remote sensing images has always been a challenging and hot task in the field of remote sensing. Considering that the accuracy and reliability of water body extraction still have some room for improvement, this paper proposes a hybrid network model based on CNN and multi-scale transformer for water body extraction from high-resolution remote sensing images. Specifically, the proposed network first uses a CNN model to extract a series of multi-scale features from shallow to deep from remote sensing images. These multi-scale features are then fed into a designed multi-scale transformer module to extract global contextual association information of water bodies. Afterwards, the water separability in the new multi-scale features output from the multi-scale transformer module is evaluated separately, and the features at different scales are adaptively weighted and fused according to their water separability. Subsequently, the network adaptively refines the fused features with the aid of a hybrid attention model to generate refined features that can effectively distinguish between water bodies and non-water bodies. Finally, these refined features are input into the prediction head to generate the final water body extraction results. The proposed network integrates the ability of CNN to capture local detail features and the ability of transformer to model global contextual semantic associations in a large range. Therefore, it can more accurately identify water bodies in remote sensing images, and the extracted water body boundaries have high accuracy and continuity. Finally, water body extraction experiments on the public dataset demonstrate the effectiveness of the proposed network. Moreover, the results of comparative experiments also show that compared with existing networks or methods such as U-Net, FCN8s, DeepLabv3+, and MSFA-Net, the proposed network has certain advantages in terms of water body extraction accuracy.

1. INTRODUCTION

The spatial distribution information of surface water bodies plays an important role in many applications such as water resources management and protection, flood monitoring, and sustainable development.

With the rapid development of remote sensing technology, water body extraction based on remote sensing images has become the mainstream way to obtain the spatial distribution information of surface water bodies. In particular, high-resolution remote sensing images have become the main data source for water body extraction due to their increasing availability. For instance, Liu et al. (2023) proposed a multi-scale features extraction network for water extraction from optical high-resolution remote sensing imagery. Zhang et al. (2021) also used the high-resolution remote sensing images as a data source for fine-grained tidal flat water body extraction.

However, as deep learning continues to make breakthroughs in various related fields, water body extraction from high-resolution remote sensing images based on deep learning has gradually become the focus of attention. Correspondingly, many deep learning-based methods have been proposed for water body extraction tasks based on remote sensing images. As one of the mainstream deep learning models, Convolutional Neural Network (CNN) is widely used in these existing methods. This is because CNN can effectively extract rich features of objects in the image for better distinction between water bodies and non-water bodies, including shallow detailed features and deep abstract features. For example, Lu et al. (2022) proposed a neighbor feature aggregation network for weakly supervised water extraction from high-resolution remote sensing imagery. Duan et al. (2021) proposed a new lightweight CNN named Lightweight Multi-Scale Land Surface Water Extraction Network (LMSWENet) to extract the land surface water information from GaoFen-1D satellite images. Hu et al. (2022) also adopted the CNN model to extract rich features for water body segmentation. Some similar studies on CNN-based water body extraction can be found in literature (Chen et al., 2018; Tao et al., 2020; Liu et al., 2021), etc.

However, although these CNN-based methods can achieve higher accuracy in water body extraction than traditional methods, there are still some problems such as many false detections, inaccurate water body boundaries, and poor continuity in the generated water body maps.

In recent years, transformer has demonstrated excellent performance in the field of image processing, so it has gradually been introduced into water body extraction tasks based on remote sensing images. For example, Zhong et al. (2022) proposed a transformer-based water extraction network called NT-Net. Song et al. (2023) used the swin transformer network for water extraction from remote sensing images. Since transformer has a larger receptive field and stronger context modeling capabilities, it can obtain semantic dependence information of water bodies in a wide range in remote sensing images. Therefore, water bodies segmented from remote sensing images using these transformer-based methods or models tend to have higher boundary accuracy and continuity than CNN-based methods.

[*] Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W1-2023-889-2023 | © Author(s) 2023. CC BY 4.0 License.

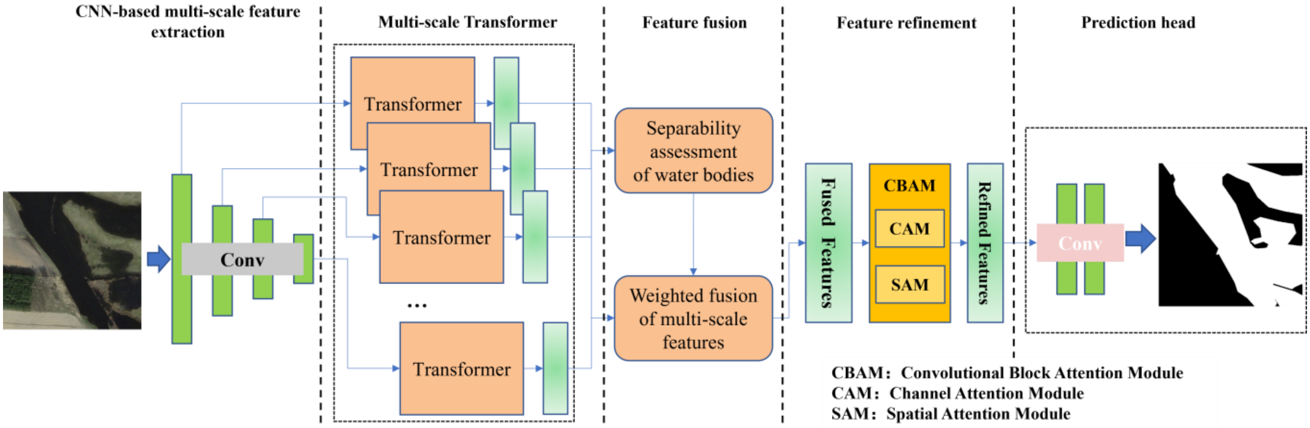
CNN-based multi-scale feature extraction
Multi-scale Transformer
Feature fusion
Feature refinement
Prediction head
Conv
Transformer
Transformer
Transformer
...
Transformer
Separability assessment of water bodies
Weighted fusion of multi-scale features
Fused Features
CBAM
CAM
SAM
Refined Features
Conv
CBAM: Convolutional Block Attention Module
CAM: Channel Attention Module
SAM: Spatial Attention Module

Figure 1. The overall architecture of the proposed network.

Generally speaking, these existing methods can achieve acceptable results in water body extraction tasks. However, it is undeniable that there is still room for improvement in the accuracy of water body extraction, especially in terms of the accuracy and continuity of water body boundaries, and when faced with complex surface scenes. Therefore, under the current research background, exploring how to effectively integrate the advantages of CNN and transformer to further improve the accuracy of water body extraction from remote sensing images is a direction worthy of research.

In view of the above considerations, this paper proposes a hybrid network model based on CNN and multi-scale transformer for water body extraction from high-resolution remote sensing images. The proposed network first uses a CNN backbone to extract the multi-scale features from shallow to deep from remote sensing images. These multi-scale features are then fed into a designed multi-scale transformer module to extract global contextual association information of water bodies. Afterwards, the separability of water bodies in all features was evaluated separately, and the features at different scales are adaptively weighted and fused guided by water separability. Subsequently, the network adaptively refines the fused features with the aid of a hybrid attention model to generate refined features that can effectively distinguish between water bodies and non-water bodies. Finally, these refined features are input into the prediction head to generate the final water body extraction results.

The rest of this paper is organized as follows. Section 2 describes the details of proposed network. Section 3 presents the experimental results and analysis. Finally, the conclusions are drawn in Section 4.

2. METHODOLOGY

Figure 1 shows the overall architecture of the proposed network. As can be seen from Figure 1, the proposed network model mainly includes five stages: 1) Multi-scale feature extraction based on CNN; 2) Global contextual association information extraction based on multi-scale transformer; 3) Multi-scale feature fusion guided by water separability; 4) Feature refinement based on attention mechanism; 5) Water body prediction. Overall, the proposed network first uses a CNN backbone to extract a series of multi-scale features from shallow to deep from the remote sensing images. Then, these extracted features are respectively input into the corresponding processing channels in the multi-scale transformer module for the extraction of global association information at different scales. Correspondingly, a series of new features containing global dependence information of water bodies at different scales are generated. Subsequently, the water separability is evaluated on the features output by each processing channel of the multi-scale transformer module, and the features of different scales are weighted and fused according to the water separability. Afterwards, a hybrid attention model is adopted to adaptively refine the fused features and generate the refined features that can effectively distinguish water bodies from non-water bodies. Finally, these refined features are input into the prediction head to generate the final water body extraction results. Considering that rich local and global features have been extracted from the imagery in the previous modules, a shallow FCN is used for water body prediction in the prediction head (Chen et al., 2022).

2.1 Multi-scale feature extraction based on CNN

CNN has shown excellent performance in image feature extraction. In particular, U-Net, a modified fully CNN, can combine high-level semantic information with low-level texture information through skip connections to achieve feature extraction and detail recovery. The U-Net has been widely adopted in the segmentation or classification tasks of remote sensing images. Therefore, the proposed network uses the U-Net as the backbone for multi-scale feature extraction.

2.2 Global contextual association information extraction based on multi-scale transformer

In the proposed network, the purpose of the multi-scale transformer module is to obtain the global dependence information of water bodies at different scales in the image to enhance the accuracy of water body extraction, especially to improve the accuracy and continuity of the extracted water body boundaries. Obviously, a series of features extracted from images by the CNN contain different levels of semantic information. Using the transformer to model the global semantic dependencies separately at different scales will more effectively extract the spatial features and contextual semantic associations of water bodies in the image.

The core of the transformer lies in its multi-head self-attention mechanism. Figure 2 shows the structure of self-attention, which is actually the scaled dot-product attention mechanism (Ma et al., 2022). The self-attention head function can be represented as follows:

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{1}$$

where Q, K, V represent Query, Key, and Value respectively, which are obtained from the input data X through three linear mapping layers (Dosovitskiy et al., 2020). And d denotes the dimension of K. softmax(·) represents the softmax function to generate attention scores.

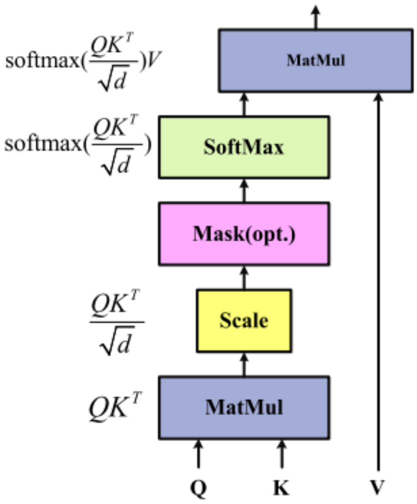
$\text{softmax}(\frac{QK^T}{\sqrt{d}})V$
MatMul
$\text{softmax}(\frac{QK^T}{\sqrt{d}})$
SoftMax
Mask(opt.)
$\frac{QK^T}{\sqrt{d}}$
Scale
QK^T
MatMul
Q
K
V

Figure 2. The self-attention mechanism in transformer.

2.3 Multi-scale feature fusion guided by water separability

A series of new multi-scale features containing the global dependence information of the water body are obtained from each processing channel of the multi-scale transformer. Apparently, the separability of water bodies and non-water bodies in each feature differs to varying degrees in different features. Those features with high water separability are more helpful to improve the accuracy and reliability of water extraction. Therefore, in this section, the proposed network first evaluates the water separability of features at different scales, and then performs water separability-guided adaptive weighted fusion of features at different scales. Correspondingly, a fused feature set can be obtained.

Suppose a total of M processing channels are used in the multi-scale transformer module, which correspond to M different scales. At the scale m (m=1,2,...,M), the generated series of features is denoted as F^m={ f_1、f_2、……、f_N }, where N represents the total number of features output by the corresponding channel. For the n-th feature f_n (n=1, 2, ..., N) in F^m, let μ_i, μ_j and δ_i, δ_j denote the mean and standard deviation of water and non-water samples in this feature, respectively. Then the water separability S_n in the feature f_n can be expressed by the following Equation (2):

$$S_n=\frac{|\mu_i-\mu_j|}{\sigma_i+\sigma_j} \tag{2}$$

Furthermore, the total separability TS^m of water bodies and non-water bodies in the feature set F^m corresponding to scale m can be expressed by the following Equation (3):

$$TS^m=\frac{1}{N}\cdot\sum_{n=1}^{N}S_n \tag{3}$$

After evaluating the total separability of the feature sets at all scales, the features of different scales can be weighted and fused according to the following Equation (4) to generate a new fused feature set F'. Obviously, in the weighted fusion process of features, features with higher water separability have higher weights.

$$F'=\sum_{m=1}^{M}TS^m\cdot F^m \tag{4}$$

2.4 Feature refinement based on attention mechanism

To eliminate redundant features in the newly generated feature set F' and further improve the difference between water bodies and non-water bodies in the features, this paper adopts the hybrid attention model CBAM to adaptively refine the feature set F'. As shown in Figure 3, for the input feature map $F'\in R^{C\times H\times W}$, CBAM first uses the CAM module to generate a 1D channel attention map $M_c\in R^{C\times1\times1}$, and then uses the SAM module to generate a 2D spatial attention map $M_s\in R^{1\times H\times W}$. And these attention maps are multiplied to the input feature maps to achieve the purpose of adaptive feature refinement (Woo et al., 2018). This process can be expressed by the following Equations (5) and (6).

$$F''=M_c(F')\otimes F' \tag{5}$$

$$F'''=M_s(F'')\otimes F'' \tag{6}$$

where $\otimes$ represents element-wise multiplication. F'' represents the intermediate features generated after processing the input features using the CAM module, and F''' represents the final refined features generated after CBAM processing.

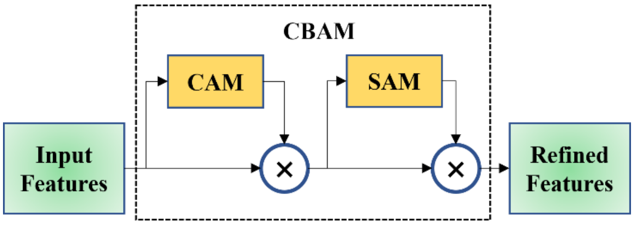

Figure 3. The structure of the CBAM.

2.5 Water body prediction

In the above CNN backbone and multi-scale transformer modules, rich local and global features have been mined from remote sensing images. Therefore, after obtaining the refined feature F''', the proposed network adopts a shallow FCN as the prediction head to generate the final water body map.

3. EXPERIMENT AND ANALYSIS

To verify the effectiveness of the proposed network, we test it on a publicly available dataset. This section first introduces the dataset used and related experimental settings, and then presents and analyzes the experimental results in detail.

3.1 Data Set

A popular and publicly available dataset GID (Tong et al., 2018) with 150 images is used in this study, 120 of which are used as training set and the rest as test set. In the GID dataset, the image size is 6800 pixels × 7200 pixels, and the spatial resolution is 4 meters.

In addition to the unknown class, GID includes five classes: built-up, farmland, forest, meadow, and waters. In order to make it suitable for water segmentation tasks, we divide built-up, farmland, forest, and meadow into one class, which is the non-water class. Unknown data are not considered in the training and testing processes. Figure 4 shows a part of the experimental data and its ground truth.

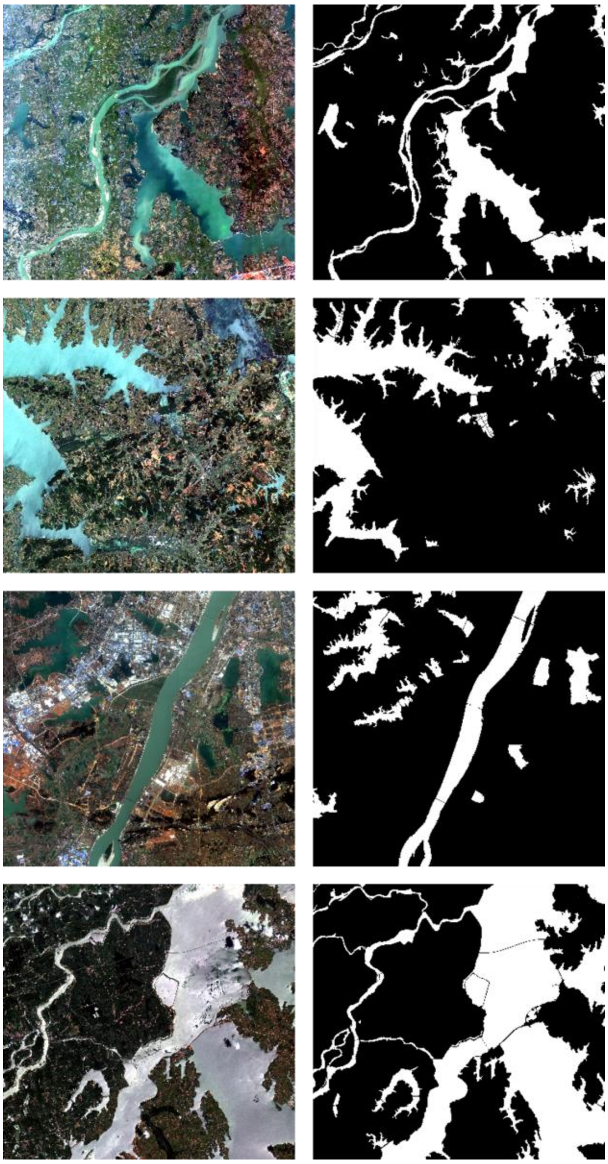

Figure 4. Experimental data and their ground truths (partial).

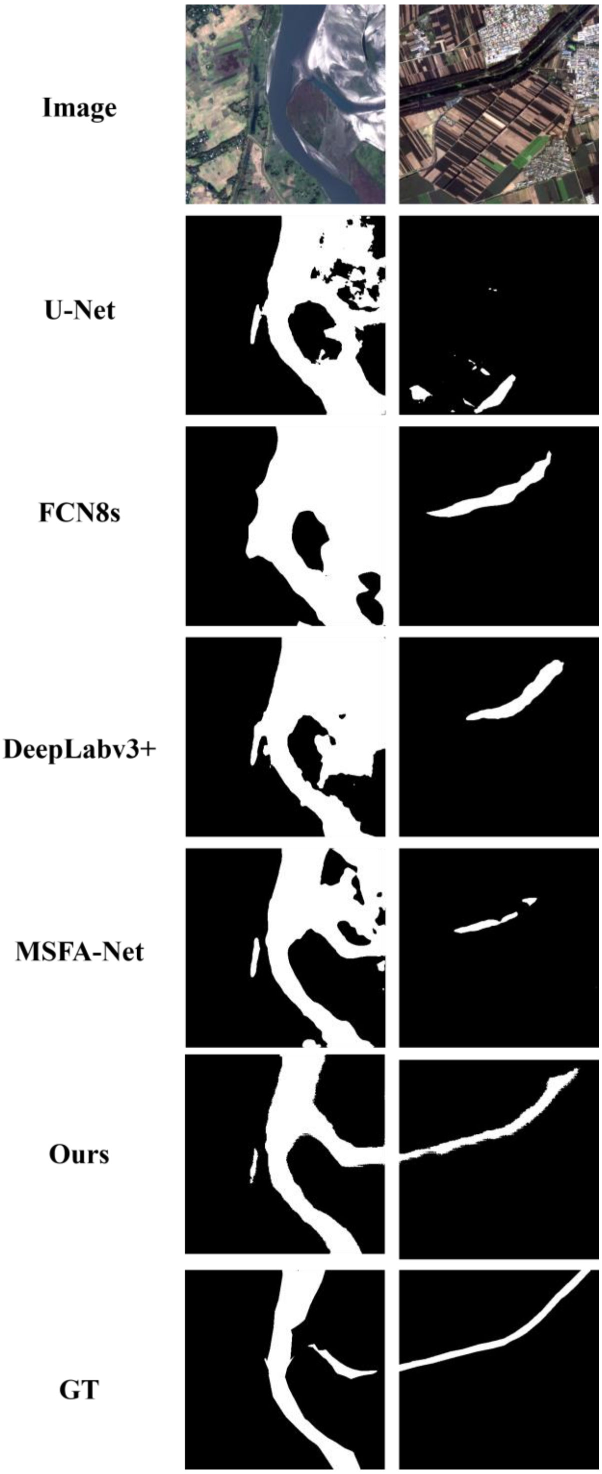

Figure 5. Partial results of water body extraction by different methods on the GID dataset.

3.2 Experimental settings and evaluation indicators

To verify the effectiveness of the proposed network, a recently proposed water segmentation network MSFA-Net (Hu et al., 2022) and several deep learning network models commonly used in water segmentation tasks were used for comparison experiments, including U-Net (Ronneberger et al., 2015) , FCN8s (Shelhamer et al., 2017), and DeepLabv3+ (Chen et al., 2018), etc. At the same time, in order to quantitatively compare the water extraction results of different methods, several commonly used evaluation indexes were adopted, namely intersection on union (IoU), precision, recall, F1-score, and kappa coefficient (KC)(Carletta, 1996; Duan and Hu, 2020).

3.3 Experimental results and analysis

Figure 5 shows some results of water body extraction by different methods on the GID dataset. As can be seen from Figure 5, the water body extraction results of the proposed network (Ours) have a high consistency with the ground truth, and the water body regions extracted by the proposed network have good continuity and high boundary accuracy. However, in the water body extraction results generated by U-Net, FCN8s, and DeepLabv3+, the false detection is serious. The proposed network can also accurately identify narrow water bodies, while U-Net and MSFA-Net have poor performance in such scenarios. Meanwhile, it can be seen from Figure 5 that the networks based on the pure CNN structure cannot accurately locate the boundary position of the water body, resulting in low boundary accuracy of the extracted water body.

Table 1 shows the water body extraction accuracy of different methods on the GID dataset. From Table 1, it can be seen that the proposed network exhibits the optimal water body extraction performance as a whole. Specifically, the proposed network is superior to other models in the four indicators of IoU, Recall, F1-score and KC, only its Precision is slightly lower than that of DeepLabv3+. Such experimental results prove the effectiveness of the network proposed in this paper. At the same time, it also proves that compared with existing network models such as U-Net, FCN8s, DeepLabv3+, and MSFA-Net, the proposed network has certain advantages in terms of water body extraction accuracy.

Table 1. Water body extraction accuracies of different methods on the GID dataset.

Method	**IoU**	**Precision**	**Recall**	**F1-score**	**KC**
U-Net	91.26	94.92	95.95	95.43	0.922
FCN8s	94.25	97.82	96.27	97.04	0.950
DeepLabv3+	94.84	**98.28**	96.45	97.35	0.956
MSFA-Net	95.46	98.14	97.22	97.68	0.961
Ours	**95.98**	97.49	**98.42**	**97.95**	**0.966**

4. CONCLUSION

To more accurately extract water bodies from high-resolution remote sensing images, this paper proposes a new hybrid network model based on CNN and multi-scale transformer. The proposed network integrates the ability of CNN to capture local detail features and the ability of transformer to model global contextual semantic associations in a large range. Therefore, it can more accurately identify water bodies in remote sensing images, and the extracted water body boundaries have high accuracy and continuity. In this study, the proposed network is tested on the publicly available dataset, and the results of comparative experiments demonstrate its effectiveness and superiority.

ACKNOWLEDGEMENTS

This work was supported by the Special Fund of Hubei Luojia Laboratory under grant 220100028.

REFERENCES

Carletta, J. (1996). "Assessing agreement on classification tasks: the kappa statistic." J Comput. Linguist. **22**(2): 249–254.

Chen, H., Z. Qi and Z. Shi (2022). "Remote Sensing Image Change Detection With Transformers." IEEE Transactions on Geoscience and Remote Sensing **60**: 1-14.

Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff and H. Adam (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. European Conference on Computer Vision.

Chen, Y., R. Fan, X. Yang, J. Wang and A. Latif (2018) "Extraction of Urban Water Bodies from High-Resolution Remote-Sensing Imagery Using Deep Learning." Water **10** DOI: 10.3390/w10050585.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. J. A. Houlsby (2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." **abs/2010.11929**.

Duan, L. and X. Hu (2020). "Multiscale Refinement Network for Water-Body Segmentation in High-Resolution Satellite Imagery." IEEE Geoscience and Remote Sensing Letters **17**(4): 686-690.

Duan, Y., W. Zhang, P. Huang, G. He and H. Guo (2021) "A New Lightweight Convolutional Neural Network for Multi-Scale Land Surface Water Extraction from GaoFen-1D Satellite Images." Remote Sensing **13** DOI: 10.3390/rs13224576.

Fengyu, Y., F. Tao, X. Ganyang and C. Ying (2020). "Applied method for water-body segmentation based on mask R-CNN." Journal of Applied Remote Sensing **14**(1): 014502.

Hu, K., M. Li, M. Xia and H. Lin (2022) "Multi-Scale Feature Aggregation Network for Water Area Segmentation." Remote Sensing **14** DOI: 10.3390/rs14010206.

Liu, B., S. Du, L. Bai, S. Ouyang, H. Wang and X. Zhang (2023). "Water extraction from optical high-resolution remote sensing imagery: a multi-scale feature extraction network with contrastive learning." GIScience & Remote Sensing **60**(1): 2166396.

Liu, W., X. Chen, J. Ran, L. Liu, Q. Wang, L. Xin and G. Li (2021) "LaeNet: A Novel Lightweight Multitask CNN for Automatically Extracting Lake Area and Shoreline from Remote Sensing Images." Remote Sensing **13** DOI: 10.3390/rs13010056.

Lu, M., L. Fang, M. Li, B. Zhang, Y. Zhang and P. Ghamisi (2022). "NFANet: A Novel Method for Weakly Supervised Water Extraction From High-Resolution Remote-Sensing Imagery." IEEE Transactions on Geoscience and Remote Sensing **60**: 1-14.

Ma, C., J. Jiang, H. Li, X. Mei and C. Bai (2022) "Hyperspectral Image Classification via Spectral Pooling and Hybrid Transformer." Remote Sensing **14** DOI: 10.3390/rs14194732.

Ronneberger, O., P. Fischer and T. Brox (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Cham, Springer International Publishing.

Shelhamer, E., J. Long and T. Darrell (2017). "Fully Convolutional Networks for Semantic Segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4): 640-651.

Sihan, H., G. Lingjia and J. Mingda (2022). Research on water body extraction based on a joint probability model of convolution neural network and spectral information. Proc.SPIE.

Song, J. and X. Yan (2023) "The Effect of Negative Samples on the Accuracy of Water Body Extraction Using Deep Learning Networks." Remote Sensing **15** DOI: 10.3390/rs15020514.

Tong, X.-Y., G. Xia, Q. Lu, H. Shen, S. Li, S. You and L. J. R. S. o. E. Zhang (2018). "Land-cover classification with high-resolution remote sensing images using transferable deep models."

Woo, S., J. Park, J.-Y. Lee and I.-S. Kweon (2018). CBAM: Convolutional Block Attention Module. European Conference on Computer Vision.

Zhang, L., Y. Fan, R. Yan, Y. Shao, G. Wang and J. Wu (2021) "Fine-Grained Tidal Flat Waterbody Extraction Method (FYOLOv3) for High-Resolution Remote Sensing Images." Remote Sensing **13** DOI: 10.3390/rs13132594.

Zhong, H. F., Q. Sun, H. M. Sun and R. S. Jia (2022). "NT-Net: A Semantic Segmentation Network for Extracting Lake Water Bodies From Optical Remote Sensing Images Based on Transformer." IEEE Transactions on Geoscience and Remote Sensing **60**: 1-13.