frontiers | Frontiers in Ecology and Evolution

TYPE Review
PUBLISHED 14 July 2023
DOI 10.3389/fevo.2023.1201125

OPEN ACCESS

EDITED BY
Qianqian Wang,
Beijing Institute of Technology, China

REVIEWED BY
Shunchun Yao,
South China University of Technology, China
Jinping Liu,
Hunan Normal University, China
Yujin Zhang,
Shanghai University of Engineering Sciences, China

*CORRESPONDENCE
Qi Shen
sftshenqi@buu.edu.cn

RECEIVED 06 April 2023
ACCEPTED 22 June 2023
PUBLISHED 14 July 2023

CITATION
Lv J, Shen Q, Lv M, Li Y, Shi L and Zhang P (2023) Deep learning-based semantic segmentation of remote sensing images: a review. *Front. Ecol. Evol.* 11:1201125. doi: 10.3389/fevo.2023.1201125

Deep learning-based semantic segmentation of remote sensing images: a review

Jinna Lv[1], Qi Shen[2*], Mingzheng Lv[3], Yiran Li[1], Lei Shi[4] and Peiying Zhang[5]

[1]School of Information Management, Beijing Information Science and Technology University, Beijing, China, [2]Teacher's College, Beijing Union University, Beijing, China, [3]School of International Education, Shangqiu Normal University, Shangqiu, China, [4]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China, [5]College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China

Semantic segmentation is a fundamental but challenging problem of pixel-level remote sensing (RS) data analysis. Semantic segmentation tasks based on aerial and satellite images play an important role in a wide range of applications. Recently, with the successful applications of deep learning (DL) in the computer vision (CV) field, more and more researchers have introduced and improved DL methods to the task of RS data semantic segmentation and achieved excellent results. Although there are a large number of DL methods, there remains a deficiency in the evaluation and advancement of semantic segmentation techniques for RS data. To solve the problem, this paper surveys more than 100 papers in this field in the past 5 years and elaborates in detail on the aspects of technical framework classification discussion, datasets, experimental evaluation, research challenges, and future research directions. Different from several previously published surveys, this paper first focuses on comprehensively summarizing the advantages and disadvantages of techniques and models based on the important and difficult points. This research will help beginners quickly establish research ideas and processes in this field, allowing them to focus on algorithm innovation without paying too much attention to datasets, evaluation indicators, and research frameworks.

KEYWORDS

remote sensing, deep learning, convolutional neural network, semantic segmentation, satellite image

1 Introduction

Semantic segmentation is one of the most important problems in computer vision (CV) (Mo et al., 2022). The goal is to determine the class of each pixel in an image, which has great significance to the analysis and understanding of scene images. For RS data, semantic segmentation also plays a key role in a variety of geographic information applications, including urban planning (Zheng et al., 2020b; Abdollahi et al., 2021),

economic assessment (Song et al., 2022; Wang et al., 2022a), land resource management (Tong et al., 2020), precision agriculture (Weiss et al., 2020), and environmental protection (Subudhi et al., 2021; Li and Liu, 2023).

With the rise and evolution of deep neural networks, deep learning (DL) has made tremendous breakthroughs in artificial intelligence fields such as CV and natural language processing (NLP) (Kitaev et al., 2022; Ru et al., 2022; Zhang et al., 2022b; Noble et al., 2023). However, there are huge challenges for RS data, such as top-down data acquisition perspective, large image scale, different image resolution, and variable lighting conditions. DL methods on natural images can be directly employed in RS data. How to efficiently and accurately use the original RS data to obtain the required information and obtain more accurate segmentation results is difficult.

DL is gradually being applied to RS semantic segmentation (Piramanayagam et al., 2018; Sun et al., 2019). Early methods based on sliding windows and candidate regions are time-consuming and have a lot of redundant calculations (Davis et al., 1975; Özden and Polat, 2005; Senthilkumaran and Rajesh, 2009; Nowozin and Lampert, 2011; Ciresan et al., 2012). In recent years, there are more and more DL-based methods in RS, including U-Net methods and its variants (Yue et al., 2019; Foivos et al., 2020), multi-scale context aggregation networks (Liu et al., 2018a; Chen et al., 2020; Li C. et al., 2021; Xu H. et al., 2022), and multi-level feature fusion networks (Dong and Chen, 2021; Li et al., 2021b). The attention mechanism that pays attention to relevant information and ignores irrelevant information is frequently adopted with its advantages (Li et al., 2021a; Li et al., 2021b; Li YC. et al., 2021; Seong and Choi, 2021). Later, Transformer-based and generative adversarial network (GAN) methods no doubt are getting more and more attention (Shamsolmoali et al., 2020; Tian et al., 2021; Cui L. et al., 2022; He et al., 2022; Wang L. et al., 2022).

Some works of literature have reviewed the research methods in the field of semantic segmentation of RS data, classified and explored from different perspectives, including RS image analysis on general DL algorithms (Zhu et al., 2017; Tsagkatakis et al., 2019; Jiang et al., 2022), detection field methods (Asokan and Anitha, 2019; Li et al., 2022a), Transformer models (Aleissaee et al., 2022), and image registration methods (Zhang X. et al., 2021). Sebastian et al. (2022). explored and evaluated the strengths and weaknesses of various qualitative and quantitative image segmentation evaluation metrics used in RS applications. Lu et al. (2021) introduced and analyzed the studies and applications of satellite data from the perspective of semantics, and carried out analysis and discussions from the four research areas of semantic understanding, semantic segmentation, semantic classification, and semantic search. Li et al. (2018) discussed and performed a comparative analysis of DL models for semantic classification. Tsagkatakis et al. (2019) comprehensively reviewed DL methods for enhancing RS observations, focusing on key tasks including single- and multi-band super-resolution, denoising, restoration, pan-sharpening, and fusion. Most of the discussed methods are outdated and lack the interpretation of the latest research results and algorithms for semantic segmentation.

To promote the semantic segmentation methods of RS data, we focus on the latest research methods, recent open datasets, and evaluation methods, and look forward to the future direction. First, we describe the definition and research overview of semantic segmentation tasks. Second, we discuss different technical frameworks of DL and summarize their advantages and disadvantages. Then, we illustrate the public RS datasets from data collection and the comparison of experimental results with different methods. Finally, we summarize future challenges and research directions. The main contributions of this paper are as follows:

- Highlight approaches of DL in RS data for semantic segmentation tasks in the recent 5 years from different technical frameworks, including new architectures, DL components, and advantages and disadvantages.
- Detailed summaries of RS datasets. It contains the dataset name, description, classes, channel number, and URL. Statistical results of different methods on the same dataset are also summarized.
- This review not only summarizes the existing achievements but also points out some promising research directions for semantic segmentation. From this perspective, it helps potential readers find research points and motivates engineers to develop advanced application patterns.

2 Overview

This section provides an overview of methods for semantic segmentation in the RS field. First, the basic definition of semantic segmentation is described. Second, the traditional and mainstream methods are explained. Third, statistics are made on the semantic segmentation methods of RS images in the past 5 years, and the main published journals, quantity, and keyword visualization of papers are analyzed.

2.1 Definition and concept

Semantic segmentation is a very important direction in the CV. Unlike target detection and recognition, semantic segmentation achieves image pixel-level classification. It can divide a picture into multiple blocks according to the similarities and differences of categories. Semantically related pixels are annotated with the same label. The semantic segmentation algorithm can comprehensively complete the recognition, detection, and segmentation of visual elements in the scene, and improve the efficiency and accuracy of image understanding. Compared with image classification and target detection, the semantic segmentation results can provide richer information about image parts and details. Semantic segmentation algorithms have extensive applications and long-term development prospects. For example, in automatic driving technology, semantic segmentation algorithms can assist the automatic driving system to judge road conditions by segmenting roads, vehicles, and pedestrians. For RS images, semantic segmentation plays an irreplaceable role in disaster assessment, crop yield estimation, and land change monitoring.

2.2 Research overview

Through the Google Scholar search keys "semantic segmentation" and "remote sensing", the statistics of the number of papers published since 2015 are shown in Figure 1. We can see that semantic segmentation, as an important task in the field of remote sensing, has attracted the attention of researchers. Moreover, more and more new technologies and methods are emerging.

We collect the works from more than 100 articles based on RS semantic segmentation. This article counts them. The statistics of the research published are shown in Figure 2A (journals with less than three articles not shown). According to the number of journals published, the top four are, in order, *Remote Sensing, IEEE Transactions on Geoscience and Remote Sensing, Journal of Photogrammetry and Remote Sensing*, and *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. The year's distribution of the research papers is shown in Figure 2B. It can be found that we pay more attention to the latest studies of the past 2 years, which represent the current advanced technologies for semantic segmentation.

We analyze the keywords of the papers in the past 5 years. The keywords have substantial meaning for expressing the central content of the paper and can map the research content and direction in recent years. The statistical results are displayed in Figure 3, through the word cloud diagram. We can see that, in addition to the task keyword "semantic segmentation" and the data keyword "remote sensing", "attention", "convolutional neural", "Transformer", "GAN", and "unsupervised" are used more from the perspective of methods technology.

2.3 Method overview

All developments are a long technical accumulation. Early methods of semantic segmentation used traditional methods. With the emergence of DL, more and more new methods are emerging. There are also many excellent DL methods in the field of RS data.

Early semantic segmentation research focused on non-DL models, such as the threshold method (Davis et al., 1975), the clustering-based method (Özden and Polat, 2005), the edge detection method (Senthilkumaran and Rajesh, 2009), and conditional random fields (CRFs) (Nowozin and Lampert, 2011). These traditional methods have low efficiency and accuracy.

With the popularity of DL, many classic semantic segmentation models have been designed. The fully convolutional network (FCN) (Long et al., 2019) is the first applied DL to a semantic segmentation task. It changed the fully connected layer of CNN to a convolutional layer. However, the receptive field of FCN is fixed, and it is easy to lose detailed information. To solve this problem, the SegNet model (Badrinarayanan et al., 2017) was proposed. It can reduce the number of parameters by using pooling indices to save the contour information of the image. The U-Net network (Ronneberger et al., 2015) is an extension of FCN. Its main innovation is utilizing four layers of skip connections in the middle. DeepLab V1 (Chen et al., 2015) alleviated the down-sampling problem and makes the segmentation boundary clearer by replacing the traditional convolutional layer with porous convolution. It discarded the fully connected layer of the VGG16 and changed the last two pooling steps to one. Then, DeepLab v2 (Chen et al., 2017a), DeepLab v3 (Chen et al., 2017b), and DeepLab v3+ (Chen LC. et al., 2018) were proposed. The contribution of DeepLab v2 lies in the more flexible use of atrous convolution and Atrous Spatial Pyramid Pooling (ASPP). DeepLab v2 abandoned CRF, and improved ASPP, using dilated convolution to deepen the network. DeepLab v3+ modified the main network again and upgraded ResNet-101 to Xception.

In recent years, many improved methods based on classical DL semantic segmentation have been applied to remote sensing images (Zhou et al., 2018; Ding et al., 2020b; Pan et al., 2020; Bai et al., 2021; Huang et al., 2022). Bai et al. (2021) proposed an improved model HCANet and designed two Compact Atrous Spatial Pyramid Pooling (CASPP and CASPP+) modules. Huang et al. (2022) improved U-Net and U-net++ (Zhou et al., 2018) connections in one to four layers of U-Net. The advantage of this structure is that the network can learn the significance of characteristics from different depths and fuse them. There are many kinds of research with the latest technologies, for example, attention mechanism (Wang et al., 2022a), generative confrontation network (Pan et al., 2020), and Transformer (Ding et al., 2020b). These new methods have improved the performance of

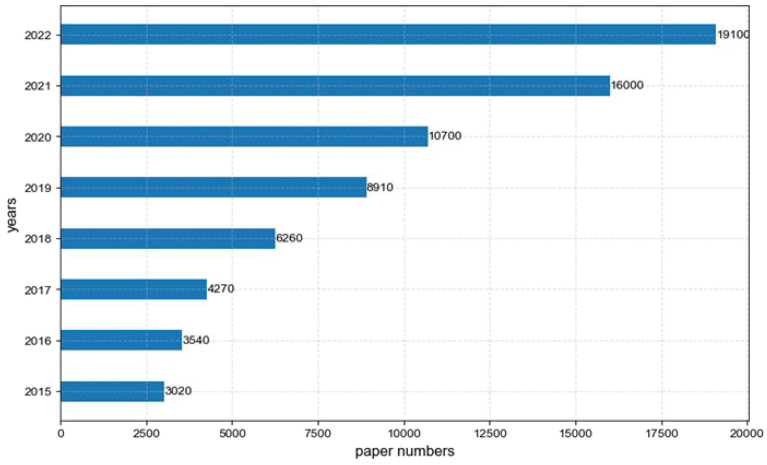

FIGURE 1
The number of papers based on remote sensing semantic segmentation since 2015.

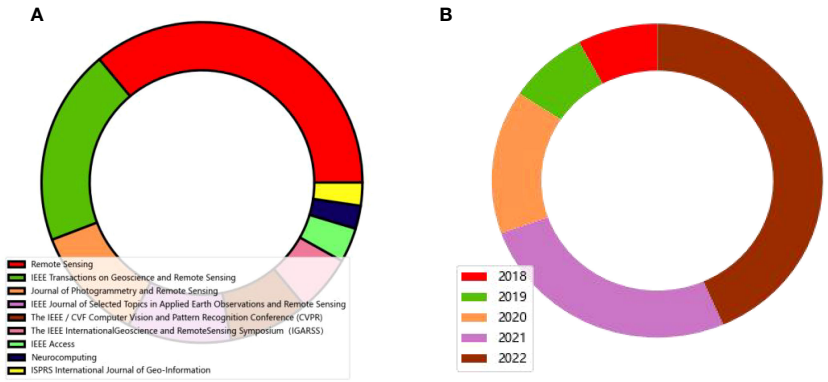

FIGURE 2
The distribution map of publications and years of analyzed articles. **(A)** The percentage distribution of publications. **(B)** The percentage distribution of years.

semantic segmentation tasks in RS data and have had an important impact on the evolution of this task.

From the perspective of the DL technology framework, this paper categorizes and outlines the semantic segmentation methods in RS data in the past 5 years by dividing them into six categories, namely, based on CNN, based on attention mechanism, multi-scale strategy, based on Transformer, based on GAN, and fusion-based methods. We display these network models in recent years in Figure 4.

3 Semantic segmentation framework

3.1 General CNN-based methods

There are many semantic segmentation studies based on the CNN of RS data. This section discusses these papers from the following categories: FCN-based, U-net-based, SegNet-based, DeepLab-based, and other convolutional network methods.

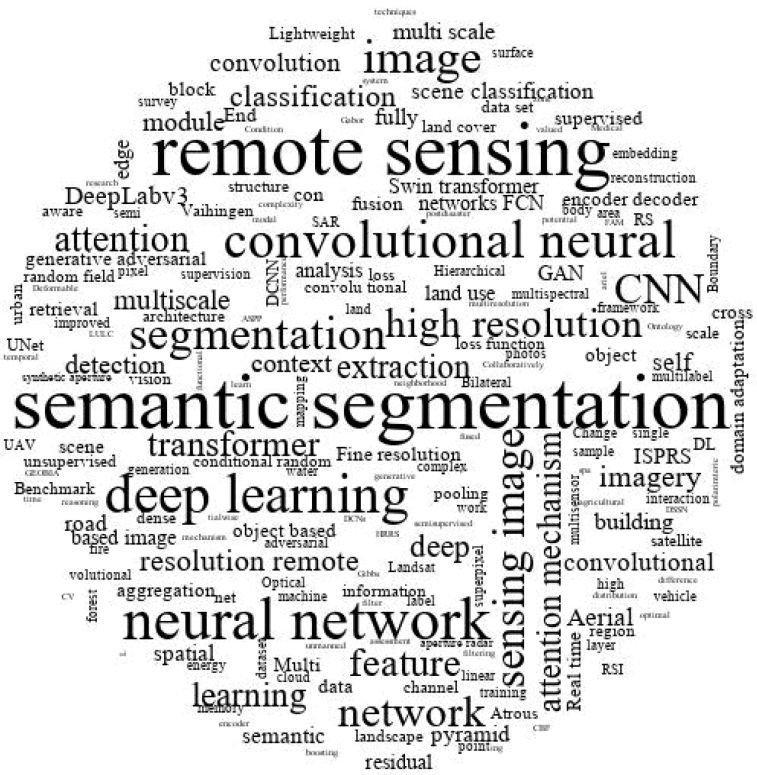

FIGURE 3
The cloud map of keywords statistics of analyzed papers.

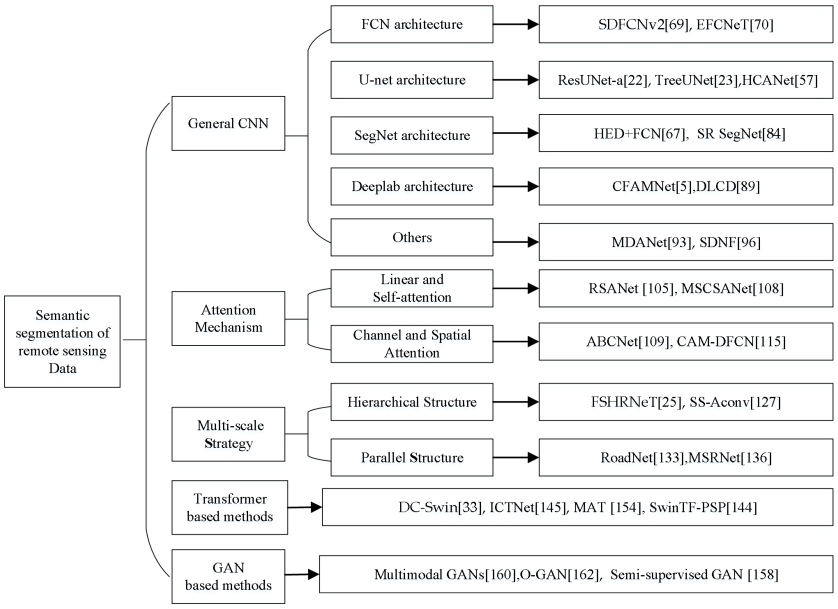

FIGURE 4
Semantic segmentation methods of remote sensing based on deep learning.

3.1.1 FCN architecture

In the FCN, 1×1 convolution replaces the full connection in the CNN. Then, the probability category value of each pixel is obtained through the softmax layer. FCN introduces the deconvolution shown in Figure 5. The true category of each pixel is the category with the largest corresponding probability value. Finally, a segmented image is obtained, whose size is the same resolution as the input image. The deconvolution uses known convolution kernels and convolutional output to restore images, thereby obtaining refined features. The reason that FCN is more efficient than CNNs is that computing convolutions are avoided one by one for each pixel block, in which adjacent pixel blocks are repeated.

Although the FCN has pushed great breakthroughs in the scene segmentation problem, it relies on a large-scale image recognition network, which is usually trained on a large number of images (Marmanis et al., 2016). However, in RS domains, label scarcity is a difficult problem. Kemker et al. (2018) were the first ones to apply the FCN to segment multispectral RS images. It used massive automatically labeled synthetic multispectral images and gained good results. Then, many FCN-based methods (Iglovikov et al., 2018; Shao et al., 2020; Wei et al., 2020; Chen et al., 2022) are introduced to improve the performance of the segmentation.

Owing to the characteristics of RS data that come from additional spectral bands set by multiple sensors, the commonly used RGB-based pre-training model cannot meet the requirements. According to the characteristics of RS data, some studies have improved the FCN-based method to achieve good semantic segmentation results (Liu et al., 2019; Chen G. et al., 2021; Chen L. et al., 2021). Liu et al. (2019) fused the RGB feature to obtain semantic labels from a DL framework with light detection and ranging (Li-DAR) features. Chen G. et al. (2021) proposed an improved structure, named SDFCNv2, to optimize the segmentation results of RS data. First, they designed a hybrid model basic convolutional block to obtain a larger receptive field.

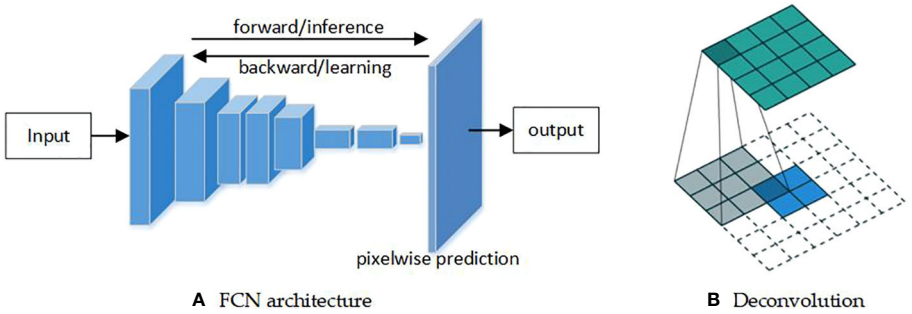

FIGURE 5
The FCN architecture and its deconvolution operator (Long et al., 2019). **(A)** FCN architecture. **(B)** Deconvolution.

Second, they develop the spatial channel fusion model to reduce training pressure and improve experimental results. The EFCNet (Chen L. et al., 2021) was an end-to-end network, which used a depth-variant block to learn the weights of different scale features. Transfer learning with FCN can improve segmentation accuracy (Wurm et al., 2019). Different convolutional blocks of the FCN network extract multi-scale information without the need for ensemble learning techniques (Pastorino et al., 2022a). Incorporating the features extracted from the FCN network and spatial information can obtain more accurate results (Pastorino et al., 2022b).

3.1.2 U-Net architecture

The U-Net architecture (Long et al., 2019) is the most widely utilized model in current semantic segmentation studies. It uses skip connections in addition to the traditional encoder–decoder layers to fuse low-level and high-level features in the expansion path to improve localization accuracy. Many variant methods have emerged later, such as Unet++ (Zhou et al., 2018), DC-Unet (Lou et al., 2021), and TransUNet (Chen J. et al., 2021), and methods based on the U-Net structure have been used for RS images (Huang et al., 2017; Huang et al., 2018; Tasar et al., 2019; Maxwell et al., 2020; Liu Z. et al., 2022; Priyanka et al., 2022; Wang K. et al., 2022) and show better performance.

Maxwell et al. (2020) experimented with a UNet-based approach on a large dataset of historical land surfaces from the US Geological Survey and reduced manual digitization operations. An incremental learning method (Tasar et al., 2019), which was a variant of U-Net, included an encoder as the first 13 convolutional layers of VGG16 and a decoder, and two central convolutional layers. ResUNet-a (Foivos et al., 2020) added residual connections in a U-Net backbone, which solved the problem of gradient disappearance and explosion. It employed multiple parallel atrous convolutions to extract object features at multiple scales. Priyanka et al. (2022) designed a DIResUNet model to combine the building blocks with the U-Net scheme by integrating initial modules, modified residual blocks, and Dense Global Space Pyramid Pooling (DGSPP). In this way, local and global related scenes are extracted in parallel by a dedicated processing operator, leading to more efficient semantic segmentation. HCANet (Bai et al., 2021) was similar to the encoder–decoder structure of U-Net. In HCANet, there are two modules, CASPP and CASPP+. The CASPP module substitutes the crop operations in U-Net and obtains multi-scale context information from ResNet with multi-scale features. To get aggregated context information, the HCANet method employed the CASPP+ module in the middle layer of the network. Yue et al. (2019) proposed TreeUNet, which connects a segmentation module and a Tree-CNN block.

3.1.3 SegNet architecture

The SegNet network includes an encoder network, a symmetric decoder network, and a classification layer pixel-wise. It has 13 convolutional layers that are the same as the VGG16. Up-pooling, which applies the index of Max Pooling, is used in the encoder to the decoder. It improves the recognition effect of the segmentation task on the segmentation boundary. As shown in Figure 6, the positions of the maximum values of the four colors are recorded. In the up-pooling block, these positions are marked, and the other positions are filled with zeros. In this way, the recognition effect of the segmentation task can be improved on the boundary.

Many studies utilized the method of combining SegNet with other operations to achieve semantic segmentation tasks (Marmanis et al., 2017), and some researchers used the idea of SegNet to design improved models (Weng et al., 2020; Zheng et al., 2020b). Marmanis et al. (2017) saved memory by adding boundary detection in the SegNet encoder–decoder architecture. Weng et al. (2020) proposed the SR SegNet to accomplish water segmentation. On the one hand, limit the number of parameters by adding an improved residual block and depth separable convolution into the encoder. Meanwhile, the dilated convolution can improve the

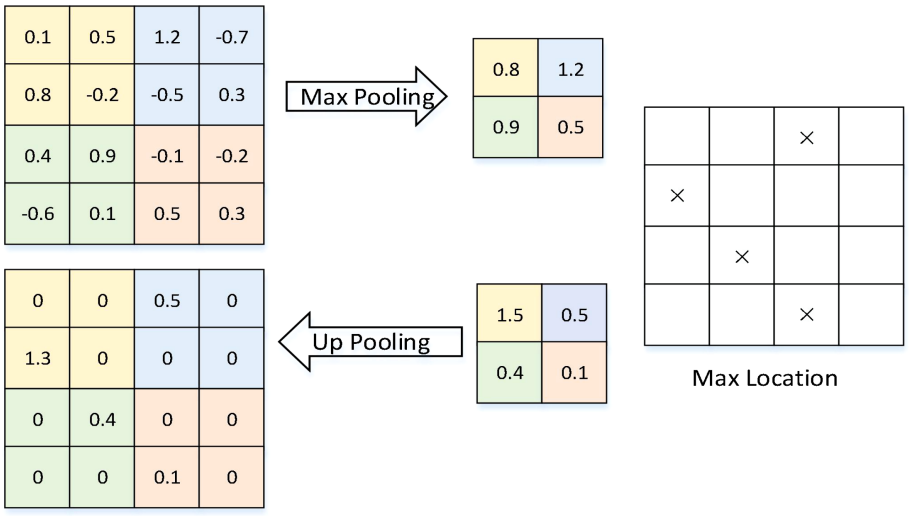

FIGURE 6
The up-pooling in SegNet architecture (Badrinarayanan et al., 2017).

ability of feature extraction. On the other hand, SR SegNet used more convolution kernels in the encoder network and employed a cascade method to combine different level features of images.

3.1.4 DeepLab architecture

The DeepLab series includes some semantic segmentation algorithms proposed by the Google team. DeepLab v1 was launched in 2014 and achieved second place in the segmentation task on the PASCAL VOC2012 dataset. Then, from 2017 to 2018, DeepLab v2, DeepLab v3, and DeepLab v3+ were successively established. The two innovations of DeepLab v1 are atrous convolution and fully connected CRF. The difference between DeepLab v2 is ASPP. DeepLab v3 further optimizes ASPP, including adding convolution and batch normalization operations. DeepLab v3+ is based on the structure of U-Net and adds an up-sampling decoder module to advance the accuracy of the edge.

In the field of RS, there are also many methods using DeepLab series structural models, such as those based on DeepLab (Chen K. et al., 2018; Hu et al., 2019; Venugopal, 2020; Wang Y. et al., 2021), and some using DeepLab v3 models (Du et al., 2014; Kong et al., 2021; Andrade et al., 2022; Wang et al., 2022a; Wang M. et al., 2022).

A dilated CNN method (Venugopal, 2020) was proposed based on DeepLab to catch the differences in images. Wang Y. et al. (2021) proposed a feature-regularized mask DeepLab model to alleviate the overfitting problem caused by small-scale samples. Chen K. et al. (2018) introduced a shuffling operator based on the DeepLab model to improve the convolutional network.

DeepLabv3+ extends DeepLabv3, which added an effective decoder module to refine segmentation results (Du et al., 2014; Kong et al., 2021; Andrade et al., 2022; Wang et al., 2022a; Wang M. et al., 2022). Wang et al. (2022a) combined features by an attention mechanism based on DeepLabv3+, named CFAMNet. First, a feature module based on attention focused on the correlation between different classes. Then, the multi-parallel space pyramid pool structure extracted features of different scales of the input data. To correctly handle the imbalance problem between different classes, Andrade et al. (2022) extended the original DeepLabv3+ model, which can improve the depiction quality of forest polygons. Wang M. et al. (2022) proposed an improved DeepLabv3+ semantic segmentation network, adopting style differences in the generalization RS data in the backbone network ResNet101 using the Instance Batch Normalization (IBN) module.

3.1.5 Other CNN methods

The power of CNN is that its multi-layer structure can automatically learn features (Li Y. et al., 2020; Mi and Chen, 2020; Ma et al., 2021; Zhang Y. et al., 2021; Cui H. et al., 2022; Ma J. et al., 2022). Cui H. et al. (2022) proposed a novel method, Hybrid DA Network (MDANet), for patch image adaptation. It reduced the difference in projection distribution of different patch images by placing them into the virtual center of the hybrid domain. Ma J. et al. (2022) designed a progressive reconstruction block based on the ASPP block and used different proportions of atrous convolutional layers to continuously process features of different resolutions. Super pixel-enhanced deep neural forests (SDNFs) (Mi and Chen, 2020) achieved better classification accuracy by combining deep convolutional neural networks (DCNNs) with decision forests.

Compared with the large-scale coverage areas of RS images, key objects such as cars and ships in HRS images usually only contain a few pixels. To address this issue, Ma et al. (2021) designed a semantic segmentation model of small objects, named foreground activation (FA), which is from the perspective of structure and optimization. Li Y. et al. (2020) coupled CNN and graph neural network (GNN) design models to discover the spatial topological relationship between visual elements. A novel activation function Hard-Swish in (Avenash and Viswanath, 2019) obtained better accurate results. Some new methods with the CNN network, for example, Yang and Ma (2022), proposed a sparse and complete latent structure *via* prototypes to solve the complex context of the background class. The weakly supervised method based on the CNN network can better solve tree species segmentation problems (Ahlswede et al., 2022).

3.1.6 Discussion

The advantage of the FCN-based method is that it can adapt any input image size. Although the effect of 8 times up-sampling is much better than that of 32 times, the result of up-sampling is still relatively blurred and smooth, and it is not sensitive to the details of images. The classification of each pixel does not fully consider the relationship between pixels. The spatial regularization ignored spatial consistency. Since the model based on the U-Net structure does not add pads during the convolution process, two pixels are reduced after each convolution. The SegNet network uses pooling indices to save the contour features of the input image, reducing parameters. The DeepLab series performs ASPP, which improves the positioning of the target boundary by using DCNN and reduces the positioning accuracy caused by the invariance of DCNN.

3.2 Attention mechanism-based methods

The attention mechanism is a prevalent technique in DL methods (Vaswani et al., 2017; Fu et al., 2019; Guo et al., 2022). Excellent semantic segmentation models are often complex and require massive computing resources. In particular, the frequently used FCNs rely on detailed spatial and contextual information, which hinders their practical application. In DANet (Fu et al., 2019), rich information relations can be obtained through a dot-product operator. Although attention technology greatly improves segmentation accuracy, the requirement of massive computation resources also hinders its application. In recent years, more and more improved methods have emerged, such as the self-attention mechanism and fusion attention mechanism. This section summarizes and discusses linear attention and sub-attention mechanisms, and channel and spatial attention mechanisms.

3.2.1 Self-attention and linear attention

The input received by the neural network is many vectors of different sizes, and there is some relationship among them, but actual training cannot fully utilize these relationships. To solve the problem that the fully connected neural network cannot establish correlations for multiple related inputs, the self-attention operator emerged. It requires the machine to recognize the correlation between different components.

RSANet (Zhao D. et al., 2021) was a region self-attention mechanism. Compared with the traditional methods, it can decrease the feature noise and the redundant features. Li C. et al. (2021) employed a layered self-attention embedded neural network with dense connections, which made full use of short- and long-range contextual features. Self-attention models were learned for automatic learning of channel and position weights (Chen Z. et al., 2021) and built a feature library and extract features of class-constrained (Deng et al., 2021). Li C. et al. (2021) proposed the Multi-Scale Context Self-Attention Network (MSCSANet). It combined the benefits of self-attention and the mechanism of CNN to improve the segmentation quality. Through the position and channel attention modules, the correlation within the feature map was calculated as well as the multi-scale contextual feature map and local features.

Linear attention is an optimization genre of self-attention, which can optimize the complexity from O (N^2) to O (N). The ith query feature is $q_i^T \in R^{D_k}$ and the ith key feature is $k_i \in R^{D_k}$. The first-order representation of the Taylor expansion is $e^{q_i^T k_j} \approx 1 + q_i^T k_j$, and guarantees $q_i^T k_j \geq -1$ using the L2 normal form for q_i and k_j.

$$sim(\mathrm{q}_i, \mathrm{k}_j) = 1 + \left(\frac{q_i}{\|q_i\|_2}\right)^T \left(\frac{k_j}{\|k_j\|_2}\right) \tag{1}$$

where the sim(.) measures the similarity between q_i and k_j. Therefore,

$$D(\mathrm{Q}, \mathrm{K}, \mathrm{V})_i = \frac{\sum_{j=1}^{N} \left(1 + \left(\frac{q_i}{\|q_i\|_2}\right)^T \left(\frac{k_j}{\|k_j\|_2}\right)\right) \mathrm{v}_j}{\sum_{j=1}^{N} \left(1 + \left(\frac{q_i}{\|q_i\|_2}\right)^T \left(\frac{k_j}{\|k_j\|_2}\right)\right)} \tag{2}$$

where Q is the corresponding query matrix, K is the key matrix, and V is the value matrix. The vector form is represented as follows:

$$D(\mathrm{Q}, \mathrm{K}, \mathrm{V})_i = \frac{\sum_j V_{i,j} + \left(\frac{Q}{\|Q\|_2}\right)\left(\left(\frac{K}{\|K\|_2}\right)^T V\right)}{N + \left(\frac{Q}{\|Q\|_2}\right) \sum_j \left(\frac{K}{\|K\|_2}\right)_{i,j}^T} \tag{3}$$

Li et al. (2021a) used a linear attention mechanism (LAM). They reconstruct the skip connections of the original U-Net and design a multi-stage method. Li et al. (2021c) designed a novel Attention Bilateral Context Network (ABCNet), which utilizes a lightweight CNN spatial path and contextual path for semantic segmentation of high-resolution RS images and used a LAM modeling the global contextual information. A2-FPN (Li R. et al., 2022) was proposed for attention aggregation. The model introduces a LAM and an attention aggregation module for a feature pyramid network to enhance multi-scale feature learning. Wang L. et al. (2021) utilized stacked convolution to build the texture path and to fuse dependency and texture features. Marsocci et al. (2021) proposed a combined self-supervised algorithm using an attention mechanism and a semantic segmentation algorithm based on a LAM for the shape of aerial images.

3.2.2 Channel and spatial attention

The semantic segmentation methods in RS data widely used attention mechanisms, such as channel and spatial attention. The channel attention focuses on feature learning in the important channel dimensions and weakens others. The spatial attention module emphasizes key areas and weakens the background. Most of the current research methods combine these two methods to improve the segmentation effect, and some methods use one of them separately.

3.2.2.1 Channel attention

Channel attention generates an attention mask in the channel domain to select important channels. Channel attention focuses on the channel dimension, which is shown in Figure 7A. A feature detector detected feature maps of each channel. For a feature map, the importance of each channel is calculated, and the weighted feature map is obtained by multiplying weights with the feature maps. Su et al. (2022) designed architecture similar to U-Net using wavelet frequency channel attention blocks as the attention mechanism. To select the most discriminative features, Panboonyuen et al. (2019) changed the weights of RS features at each stage to adaptively assign more weight values to important features. CFAMNet (Wang et al., 2022a) improved the deep DeepLabv3+ network. Its attention module obtained relevance between different categories. A multi-parallel ASPP extracted space relevance and obtained the context features of different scales.

3.2.2.2 Spatial attention

Spatial attention focuses on the space and which points on each channel are more important (Luo et al., 2019; Zhao Q. et al., 2021; Li et al., 2022b); thus, it is necessary to generate a spatial weight, which is shown in Figure 7B. First, average the values of different channels at the same plane space point (AvgPool) and take the maximum value (MaxPool) to obtain the weight. Then, a convolutional layer and a sigmoid function are used to obtain the final weight, and this weight is multiplied by each channel to achieve a weighted feature map in the spatial dimension. Owing to the size of the convolution kernel and the disappearing gradient, the data extracted from some buildings are inaccurate, and the information on some smaller buildings will be lost as the network deepens. A multi-scale spatial attention module (Li et al., 2022b) is designed to provide contextual information for the features obtained by this network model. A multi-scale spatial attention module provides contextual information for the features obtained by this network model. Zhao Q. et al. (2021) used a multi-scale module to advance the accuracy of high-resolution aerial labeling.

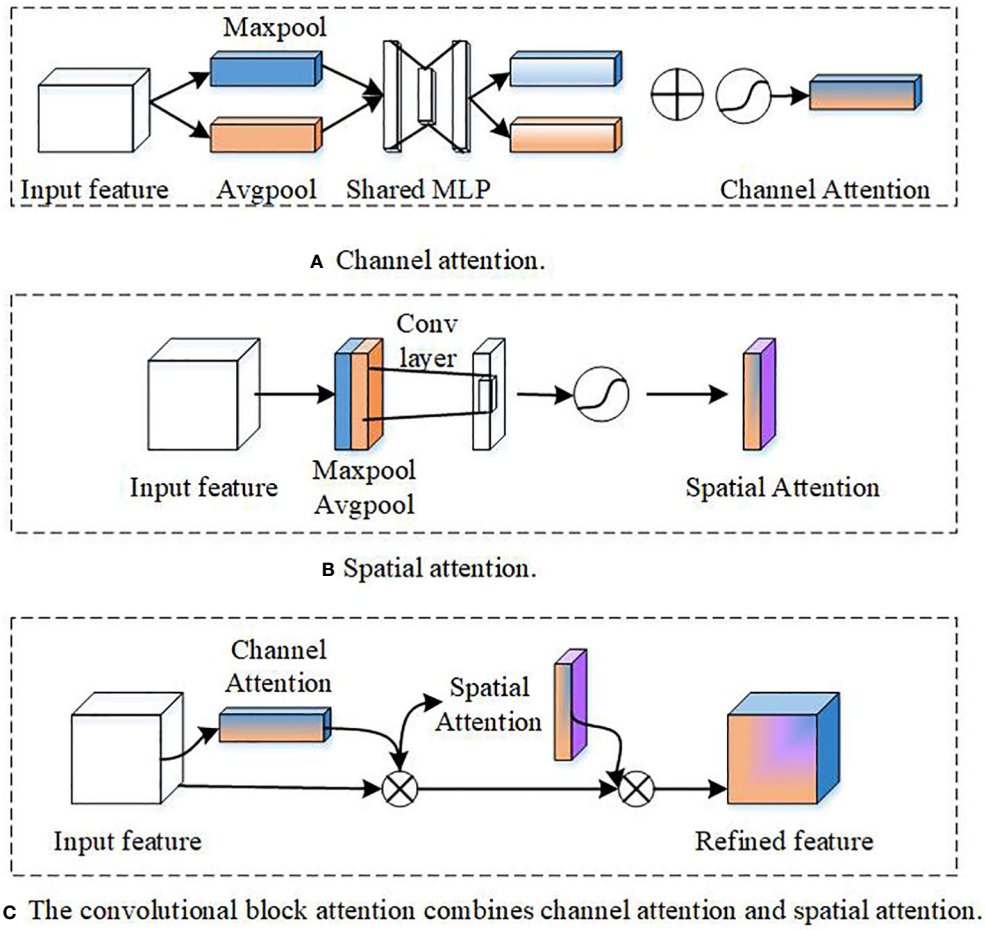

FIGURE 7
Channel attention and spatial attention (Woo et al., 2018). **(A)** Channel Attention. **(B)** Spatial attention. **(C)** The convolutional block attention combines channel attention and spatial attention.

3.2.2.3 Fusion attention mechanism

Many experiments prove that fusing channel and spatial attention can get better segmentation results (Ding et al., 2020a; Li H. et al., 2020; Sun et al., 2020; Seong and Choi, 2021; Fan et al., 2022; Liu R. et al., 2022), which is shown in Figure 7C. There are two ways for the integration of channel and spatial attention: (1) parallel model, where channel attention and spatial attention are paralleled, and (2) sequential model. First, let the feature map pass channel attention and then pass spatial attention or vice versa. Most experiments prove that it is better to pass channel attention first. The bilateral segmentation network (BiSeNetV2) (Sun et al., 2020) includes a detailed branch and a semantic branch. The detailed branch uses wide channels and shallow layers to capture low-level details and generate high-resolution feature representations. It takes the feature map $\{C_1, C_2, C_3, C_4, \text{and } C_5\}$ as input. C_1 contains rich spatial location information, which is concatenated with Conv 1×1 and C_2 to obtain feature map C_{12} through convolution operation. Next, the spatial boundary attention map $A_1=1/1(1+\exp(C_{12}))$ is obtained through the sigmoid operation. The channel attention gate assigns weights according to the importance of each channel, and the spatial attention gate assigns weights according to the importance of each pixel location for the entire channel. Ding et al. (2020a) represented features in two ways *via* augmentation. On the one hand, the attention module is utilized to enhance embedding attention based on contextual information computed by local stitching. On the other hand, local foci from high-level features are embedded by the attention embedding module. Fan et al. (2022) fused channel and spatial attention; the attention module is combined with dilated convolutional layers to form a new central region encoding and decoding, which improves the accuracy of river segmentation. Li H. et al. (2020) proposed an end-to-end semantic segmentation network that integrates lightweight spatial and channel attention modules to adaptively refine features. Global relationships between different spatial positions or feature maps can be learned and reasoned by relation-augmented representations (Mou et al., 2020).

3.2.3 Discussion

Self-attention is the weight given to each input depending on the relationship between the input data. Self-attention has the advantage of parallel computing when calculating. Linear attention is similar to dot-product attention, but it uses less memory and computation. Channel attention focuses on the importance of different channels, while spatial attention gates

focus on the importance of different pixel locations. In recent years, to improve semantic segmentation performance, most methods fuse channel and spatial attention mechanisms. However, researchers simply add or connect the attention results of the spatial and channel dimensions. How to identify the semantic segmentation of complex backgrounds is a problem that needs to be solved continuously. Therefore, it is necessary to design efficient fusion models to meet higher accuracy requirements.

3.3 Multi-scale strategy-based methods

RS images have high resolution and multi-scale variation characteristics. However, the receptive field size of the CNN is fixed. For the large-scale visual elements in the image, the receptive field can only cover its local area, which can easily cause wrong recognition results, and for the small-scale visual elements in the image. The challenge of exploiting multi-scale segmentation is to automatically select the best consecutive segmentation scale analysis (Zhang et al., 2020; Zhong et al., 2022a). Most methods are based on hierarchical structure or parallel structure, combined with an attention mechanism to achieve multi-scale feature fusion. This section discusses multi-scale semantic segmentation methods for RS images from hierarchical and parallel structures.

3.3.1 Hierarchical structure

The algorithm based on the hierarchical structure obtains multi-scale information through different stage features of the CNN, which is shown in Figure 8A. During the forward propagation process of the CNN, the receptive field increases continuously with the convolution and pooling operations. Multi-scale features from channel and spatial can be captured by fusing the features from CNN's different stages (Zheng et al., 2020a; Li Z. et al., 2021; Liu B. et al., 2022; Luo et al., 2022; Wang et al., 2022b; Zhao et al., 2022; Zheng et al., 2022).

High-resolution RS data have larger dimensions than typical natural images. Mou et al. (2020) studied a framework for object-specific optimization by identifying and fusing meaningful objects based on line segment tree models representing hierarchical multi-scale segmentation. Nodes in each path originate from leaf nodes. The EaNet model (Zheng et al., 2020b) is an edge-ware CNN. A kernel pyramid pooling (LKPP) module extracts different scale information. They designed a new loss function to optimize boundaries. Zheng et al. (2022) used a different scale input convolution module for extracting acceptable local information. Li Z. et al. (2021) extracted different features at multiple scales; SS AConv cascaded multi-scale structure (SCMS) transforms the SS AConv and residual correction scheme into a cascaded spatial pyramid by integrating different rates of SS AConv.

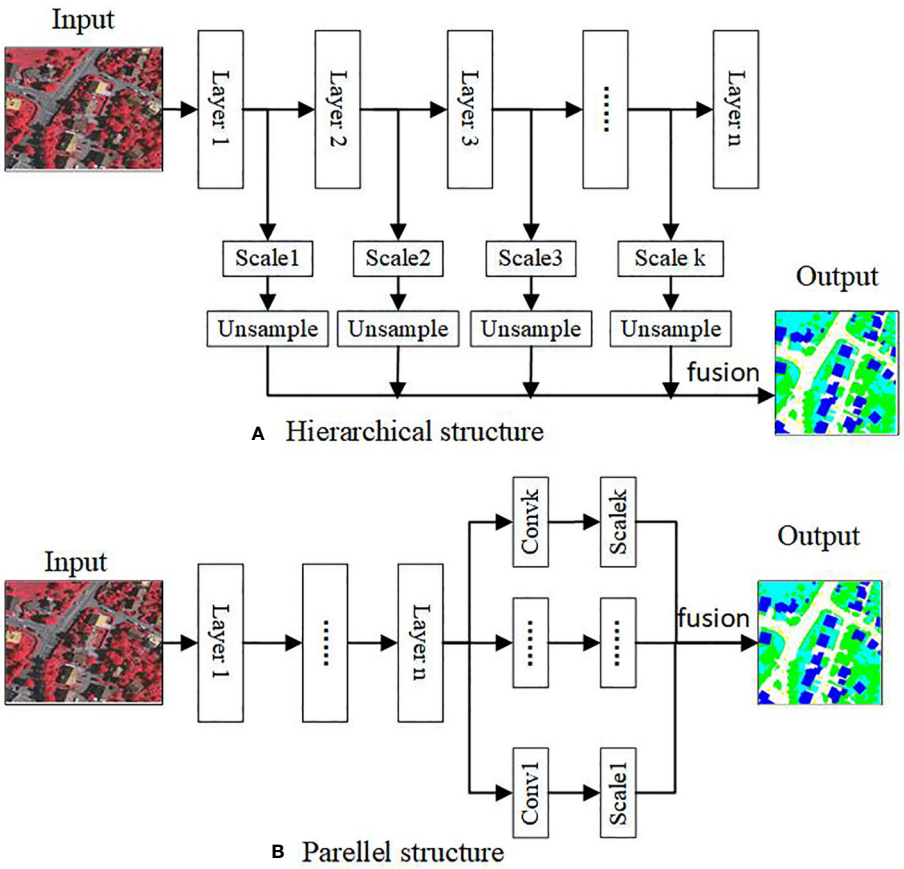

FIGURE 8
The hierarchical and parallel structures of the multi-scale strategy (Zhang and Li, 2020). **(A)** Hierarchical structure. **(B)** Parellel structure.

Xu H. et al. (2022) designed the FSHRNet using strong linear separability of high-resolution features to achieve multi-scale object segmentation in VHR images. Li et al. (2021c) proposed a layered self-attention model with dense connections. The method made full use of short and long contextual features. Inspired by transfer learning, Zhao et al. (2022) improved a multi-scale network that can advance the network's robustness. It learned scale-invariant and small objects context information. Liu B. et al. (2022) designed a method that can efficiently extract different scale features and generate maps, which helps to subdivide objects into small and different sizes. Luo et al. (2022) extracted categorical object representations from multi-scale pixel features. It can identify the similarities and differences between categories. The article by Zheng et al. (2020a) learned the symbiotic relationship between scenes through the foreground-scene relationship module. Relevant context-associated foreground augments foreground functionality, thereby reducing false positives. Wang et al. (2022b) used dynamic multi-scale dilated convolution to extract different scale features.

3.3.2 Parallel structure

The parallel structure algorithm connects multiple parallel branches with different receptive fields after the semantic feature map obtained by the convolution module to form a parallel structure to capture features of different scales, which is shown in Figure 8B. Liu et al. (2018b) automatically learned multi-scale and multi-level features, which are obtained from a deep supervision network to provide comprehensive direct supervision to deal with various scenarios and scales of the road. Liu et al. (2018a) captured different scale contexts in the output results of CNN encoders, and then continuously aggregate them in a self-cascading manner. Bello et al. (2022) proposed an efficient dense multi-scale segmentation network for accurate and specialized remote real-time segmentation of RS images. Wang et al. (2022 2022) designed a new backbone network, taking multi-scale problems as an entry point, which can focus on more important information of multi-scales.

Because of the size of the CNN kernel and the vanishing gradient, the data extracted from buildings are inaccurate, and the information of some smaller buildings will be lost as the network deepens. Duan and Hu (2019) proposed a new erasure attention module to cooperate with the multi-scale refinement scheme to efficiently perform feature embedding.

3.3.3 Discussion

The multi-scale strategy is a common technique for the semantic segmentation task of RS data. Since high-resolution images contain different object scales, it is necessary to combine the information of different scales of receptive fields to meet the requirements of the accurate segmentation of various objects. The FCN uses the same convolution operation on the entire image, without considering the multi-scale problem of visual elements, which damages the segmentation accuracy of larger-scale and smaller-scale visual elements. The multi-scale model generally builds a multi-scale RS image segmentation network first, then fuses multi-scale features, and finally predicts the results through convolution and up-sampling.

The method based on the hierarchical structure obtains multi-scale information through the features of different stages of the CNN. During the forward propagation process of the CNN, the receptive field expanded continuously using the pooling and convolution parts. The shallower feature map corresponds to a smaller receptive field, and the feature scale is also smaller. While the deep feature map corresponds to a larger receptive field, the feature scale is also larger. Therefore, different scale features can be obtained by fusing feature maps of different stages. The method based on parallel structure connects multiple parallel branches of different receptive fields after the semantic feature map obtained by the convolution module to form a parallel structure to capture features of different scales. These parallel branches are computed from the semantic feature map obtained by the convolution module, compared to the hierarchical algorithms, which are more suitable for learning semantic features.

3.4 Transformer-based methods

The Transformer was originally applied in the field of NLP. Each word is called a token in NLP, and in CV, the image is cut into non-overlapping patch sequences that are similar to tokens. SETR (Zheng et al., 2021) is the first representative model of semantic segmentation based on vision Transformer (ViT), which replaced the CNN encoder with a pure Transformer structure encoder. It drives the development of semantic segmentation in recent years.

Recently, Transformer technology makes significant contributions to improving semantic segmentation performance in the RS field (Li W. et al., 2022; Ma L et al., 2022; Sun et al., 2022). However, compared with the words in the text, the pixels in the image have a very high resolution, and the computational complexity of using a Transformer in CV is the square of the image scale, which will lead to an excessively large amount of calculation. To solve the above problems, the Swin Transformer (ST) (Liu et al., 2021) network was proposed, which is shown in Figure 9A. Its features are learned by moving the window. The moving window not only brings greater efficiency but also greatly reduces the sequence length. The advantage of the hierarchical structure is that it flexibly provides information on various scales. Because self-attention can calculate within the window, its computational complexity increases linearly with the size of the picture rather than quadratic. Therefore, in RS semantic segmentation, it is widely used (Panboonyuen et al., 2021; Xu et al., 2021; Feng et al., 2022; Gu et al., 2022; Liu Y. et al., 2022; Li X. et al., 2022; Meng et al., 2022; Xu Y. et al., 2022). ST first used the module to segment the data into many non-overlapping different patches. The state-of-the-art solutions for segmentation tasks in RS data are usually solved by CNN methods and Transformer technology. A pre-trained ST (SwinTF) (Panboonyuen et al., 2021) model with ViT was used as the backbone to weight downstream tasks by concatenating task layers on the pre-trained encoder. The original ST as the backbone of the encoder module contains a convolutional layer and attention operator. Li X. et al. (2022) utilized ST blocks and convolution blocks to advance the segmentation performance. Xu et al. (2021) argued that Transformer-based architectures usually face

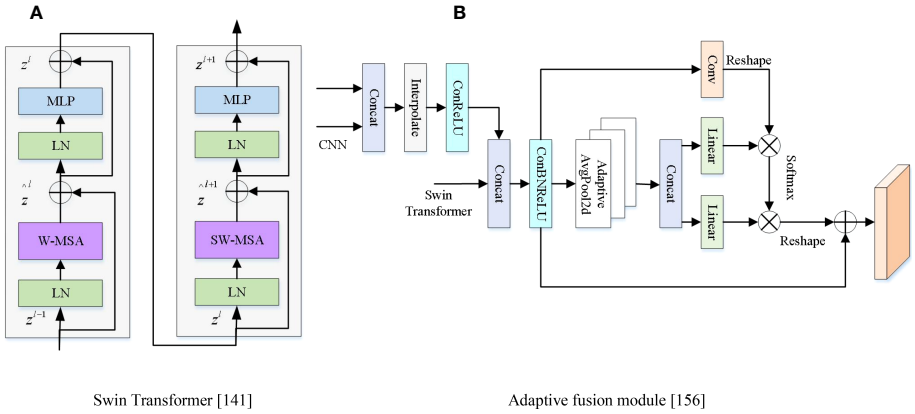

FIGURE 9
The Transformer unit and adaptive fusion module. **(A)** Swin Transformer (Liu et al., 2021). **(B)** Adaptive fusion module (Gao et al., 2021).

two main problems: massive computational and difficulty of edge segmentation. Therefore, the authors proposed a new model based on a Transformer network to achieve accurate edge detection and fewer parameters. Use an efficient Transformer backbone to improve ST to reduce computational load. Liu Y. et al. (2022) designed UPer head with ST to challenge the land-cover segmentation.

CNN cannot simulate global semantic correlation, and the Transformer model can be built with global features (Ghali et al., 2021). Combining CNN and Transformer can improve the performance of semantic segmentation (Zhao X. et al., 2021; Wang H. et al., 2022; Zhang C. et al., 2022; Zhang et al., 2022a). CNN obtained local detail features and the Transformer module obtained the global context features. Zhong et al. (2022b) designed a semantic segmentation network, which combined CNN and Transformer parts. It solved over-segmentation and the inaccurate edge detection problem, which was caused by small differences between lakes and complex texture features. StransFuse (Gao et al., 2021) was a new method combining both advantages of the Transformer and the CNN model. It can better improve the performance of various RS images, which is shown in Figure 9B. Multi-level Transformers can fuse features in different levels in each modality and high-level cross-modal features (Ma X. et al., 2022).

The Transformer breaks through the limitation that the CNN model cannot be calculated in parallel and can reasonably utilize GPU resources. The Transformer's ability to acquire local information is not as strong as CNN's. Therefore, combining Transformer and CNN can improve semantic segmentation. The ST improves the ordinary Transformer and can be flexibly modeled at various scales using a layered architecture. The sliding window feature of the ST enables it to compute self-attention in locally non-overlapping windows and allows cross-window connections.

3.5 GAN-based methods

Training neural networks, which largely depend on massive images with precise pixel-level annotation, is labor-intensive, especially for big-scale RS data. Segmenting multispectral images using supervised machine learning algorithms requires numerous pixel-level labeled data, which makes the task extremely challenging.

In recent years, some studies have introduced GAN into RS images for semantic segmentation tasks (Creswell et al., 2018; Kerdegari et al., 2019; Hong et al., 2020; Li D. et al., 2022). The GAN (Creswell et al., 2018) consists of generator (G) and discriminator (D) parts. The generator part can generate a fake image to fool the discriminator, and the discriminator distinguishes the fake image from the real image. The generator G transforms a random sample $z \in \mathbb{R}^d$ distribution γinto a generated sample G(z). The discriminator D discriminates them from the training samples from the distribution μ, while G tries to make the generated samples' distribution similar to that of the training samples. The adversarial target loss function is shown below:

$$V(\mathrm{D},\mathrm{G}) := \mathrm{E}_{x\sim\mu}[\log \mathrm{D}(\mathrm{x})] + \mathrm{E}_{z\sim\gamma}[\log(1 - D(\mathrm{G}(\mathrm{z})))] \quad (4)$$

Tian et al. (2021) proposed a combined GAN and FCN network and constructed an FCN-based segmentation network to enhance the deep semantic receptive field of the model. GAN is integrated into an FCN semantic segmentation network to synthesize global image feature information and then accurately segment and sense complex RS images. Hong et al. (2020) proposed plug-and-play units in two networks: a self-generative GANs module and mutual GANs module, to learn perturbation-insensitive feature representations and eliminate multimodality, yielding more efficient and robust information transfers, respectively. Sun et al. (2021) proposed a subdivision method based on GANs to reduce intra-class differences. The background and target should be generated separately *via* the Orthogonal GAN (O-GAN). The O-GAN works by adding new loss functions to their discriminators. To better extract architectural features, the drawing is based on the idea of fine-grained image classification through an O-GAN intermediate convolutional layer (SCDA) with selective convolutional descriptor aggregation.

Because of the cumbersome and difficult annotation for RS images, the exploration of unsupervised and semi-supervised models is difficult. The domain adaptive method using the confrontation generation network learns domain-invariant features through the confrontation between the generator and the

discriminator, which can effectively reduce the difference between domains. Most methods use GAN to generate RS images and combine them with network models such as CNN for semantic segmentation.

3.6 Fusion-based methods

As researchers continue to pursue the accuracy of semantic segmentation, a large number of fusion models of different technologies and structures have emerged, and have shown excellent results. Some fusion methods in recent years are listed in Table 1. First, the CNN network is the basis of most models. Adding an attention mechanism block is the most common way in the research of fusion models (Panboonyuen et al., 2019; Shamsolmoali et al., 2020; Kong et al., 2021; Liu Z. et al., 2022). Second, different targets have different scales on the image. Therefore, multi-scale methods often integrate other feature extraction methods to improve models, such as CNN and Transformer (Chen et al., 2020; Zheng et al., 2020b; Zhao Q. et al., 2021; Zhang et al., 2022a). Third, some complex models integrate more modules, such as GAN, ST, and multi-scale (Li Z. et al., 2021; Marsocci et al., 2021; Xu et al., 2021). However, complex models often require massive computing resources; thus, more models that balance computing resources and accuracy are needed.

4 Dataset description and experimental discussion

4.1 Dataset description

We describe some public RS datasets for semantic segmentation tasks in this section. The most frequently referenced datasets are the ISPRS Vaihingen and Potsdam datasets, followed by GID and WHDLD. The image samples and classifications of these four datasets are shown in Figure 10. We describe the datasets with more papers' references, which include the description, classes, channels, and URLs shown in Supplementary Table 1.

4.1.1 Satellite image datasets

In this section, we list a few datasets for semantic segmentation tasks, which are captured by satellites. Satellite images are obtained from the earth observation remote sensing instrument loaded on the satellite.

4.1.1.1 ISPRS Vaihingen

The ISPRS Vaihingen is a comparatively small village where there are many independent small buildings. The dataset contains 33 true orthophoto (TOP) images (GSD ~ 9 cm) with 2,500 × 2,000 pixels, which are of very high resolution. There are approximately 16 image tiles that are noted with pixel-wise labels. In addition, every pixel is split into one of six land categories, namely, impervious ground, architecture, low vegetation, tree, car, and clutter.

4.1.1.2 ISPRS Potsdam

The ISPRS Potsdam dataset covers an area of 3.42 km^2, which consists of 38 image tiles with a spatial resolution of 5 cm. All images are 6,000 × 6,000 pixels with four bands for near-infrared (NIR) and red (R), green (G), and blue (B) channels.

Similar to the Vaihingen region, it is also made up of three bands of RS TIFF files and a single band of digital surveying and mapping (DSM). RS images and DSM are defined on the same reference system (UTM WGS84) because of the same coverage size of each RS image. In particular, each image is decomposed into smaller images using radial transform files. The dataset also provides a tiff storage form for different channel combinations of the TOP image so that participants can select their respective desired data.

The dataset label is a semi-dense disparity map obtained by averaging DSM data matched by multiple sets of commercial software based on internal and external orientation elements. The dataset provides a normalized DSM that does not require manual annotation. Accordingly, it is not guaranteed that there is no false data here, which is to help researchers use high data without using absolute DSM.

4.1.1.3 GID

GID (Tong et al., 2020) covers 506-km^2 areas that are captured *via* the satellite Gaofen-2. This dataset includes 150 high-quality Gaofen-2 RS images with 7,200 × 6,800 pixels. The dataset has a rich diversity in spectrum, texture, and structure, which is very close to the real feature distribution characteristics. The GID dataset is divided into two parts: a large-scale set of labeling categories (GID-5) and a fine land cover set (GID-15). It contains five classes in GID-5. In addition, 150 image-level labeled Gaofen-2 satellite RS images are offered. Among them, there are 120 images in the training section; meanwhile, 30 images are included in the validation set.

4.1.1.4 WHDLD

The Wuhan dense labeling dataset (WHDLD) (Shao et al., 2020) is captured from an enormous image of the downtown area of Wuhan in the RS field. With a resolution of 2 m, this dataset provides 4,940 RGB images with 256 × 256 pixels. WHDLD is labeled with six categories. They are building, roads, sidewalks, vegetation, bare soil, and water.

4.1.1.5 DeepGlobe Land Cover

The dataset contains a space resolution of 0.5 m and is built of red, green, and blue bands. It is generated from a satellite with 2,448 × 2,448 pixels. Seven classes have been split into downtown area, farm land, range land, forest, water area, barren, and unknown.

4.1.1.6 GF-2

Based on the GF-2 satellite, this dataset has a space resolution of 0.8 m, with 2,000 × 2,000 pixels. With the help of ENVI, the image of GF-2 is preprocessed. These data are labeled by Matlab software with different colors and diverse image types.

TABLE 1 Different methods that integrated different models.

Papers	Year	Datasets	Methods
Panboonyuen et al. (2019)	2019	ISPRS Vaihingen, Landsat-8 dataset	CNN, Transfer learning, Attention mechanism
Luo et al. (2019)	2019	fg	CNN, Multi-scale, Self-attention
Zheng et al. (2020b)	2020	WHU building dataset, Cityscape, ISPRS Vaihingen	CNN, Object-specific optimization, Multi-scale
Li et al. (2021a)	2020	ISPRS Vaihingen	Attention mechanism, Multi-scale, ResU-Net
Duan and Hu (2019)	2020	GID	CNN, Multi-scale, Attention mechanism
Shao et al. (2020)	2020	WHDLD, DLRSD	FCN, Region convolutional features, Multi-scale
Sun et al. (2020)	2020	AIR-SEG, ISPRS Vaihingen	FCN, Boundary attention model, Channel-weighted, Multi-scale
Chen et al. (2020)	2021	ISPRS Potsdam, ISPRS Vaihingen	GAN, Multi-scale
Seong and Choi (2021)	2021	SpaceNet building datasets, GIS, WHU dataset	CNN, Attention mechanism, ResNet
Tian et al. (2021)	2021	ISPRS Vaihingen, ISPRS Potsdam, DeepGlobe Road	FCN, GAN
Panboonyuen et al. (2021)	2021	ISPRS Vaihingen	Feature pyramid network, CNN, Transformer
Ghali et al. (2021)	2021	Corsican Fire dataset	CNN, Transformer, TransUNet, U2Net Architecture
Chen Z. et al. (2021)	2021	WHU and Massachusetts Building datasets	U-Net, Self-attention, Multi-scale
Shamsolmoali et al. (2020)	2021	DeepGlobe Road Extraction Data Set	Feature pyramid, Multi-scale, Attention mechanism
Li C. et al. (2021)	2021	ISPRS Vaihingen, ISPRS Potsdam	CNN, Dense connection, Self-attention
Xu et al. (2021)	2021	ISPRS Vaihingen, ISPRS Potsdam	CNN, Swin Transformer
Kong et al. (2021)	2021	Sentinel-1 SAR images	Channel spatial Attention mechanism, DeepLabv3+, Multi-scale
Wang L. et al. (2022)	2022	ISPRS Potsdam, ISPRS Vaihingen	Transformer, Multi-scale
Feng et al. (2022)	2022	GID	CNN, Swin Transformer, Multi-scale
Gu et al. (2022)	2022	WHDLD, LoveDA	CNN, Swin Transformer, U-Net, Multi-scale, A deformable adaptive patch merging layer
Meng et al. (2022)	2022	ISPRS Vaihingen, ISPRS Potsdam	FCN, Swin Transformer
Zhang et al. (2022a)	2022	ISPRS Potsdam, WHU Building, dataset	CNN, Transformer, Depthwise channel self-attention
Liu Z. et al. (2022)	2022	ISPRS Vaihingen, ISPRS Potsdam	DCNN, Attention mechanism
Li X. et al. (2022)	2022	DeepGlobe Land Cover Dataset, ISPRS Vaihingen, ISPRS Potsdam	CNN, Transformer, Multi-scale
Li et al. (2021b)	2022	ISPRS Vaihingen, ISPRS Potsdam	CNN, Multi-attention network, Multi-scale,
Zhao et al. (2022)	2022	ISPRS Potsdam	Collaborative enhanced fusion, Attention mechanism, Multi-scale
Luo et al. (2022)	2022	ISPRS Potsdam, GID	Feature pyramid, cross-attention, Transformer, Multi-scale
Zheng et al. (2022)	2022	GID	Multi-scale, Transformer, Attention mechanism, semi-supervised, Pyramid scene parsing network
Liu Y. et al. (2022)	2022	ISPRS Vaihingen, ISPRS Potsdam	CNN, Swin Transformer, Multi-scale, Dynamic attention pyramid head
He et al. (2022)	2022	ISPRS Vaihingen, ISPRS Potsdam	CNN, Swin Transformer, UNet, Spatial interaction module
Wang et al. (2022b)	2022	SSS image datasets	CNN, Attention mechanism, Dynamic Multi-scale Dilated Convolution, Adaptive Receptive Field Mechanism
Cui L. et al. (2022)	2023	24 remote sensing city-scale images of Yushu city and Beichuan city after the Yushu and Wenchuan earthquakes	CNN, Swin Transformer, Convolutional block attention module

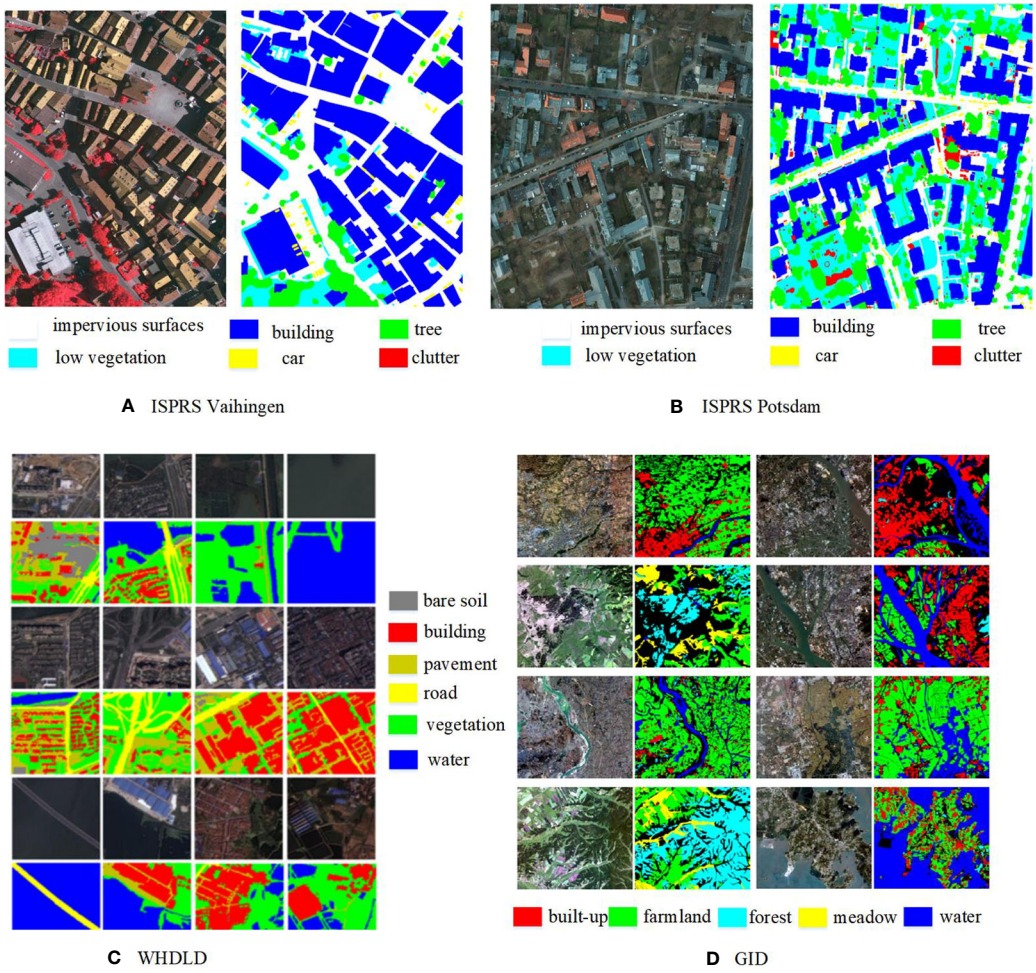

FIGURE 10
Visualization of the four common datasets. **(A)** ISPRS Vaihingen. **(B)** ISPRS Postdam. **(C)** WHDLD. **(D)** GID.

4.1.1.7 RSSCN7

RSSCN7 (Qin et al., 2015) consists of 2,800 RS images. Collected from Google Earth, each class is equipped with 400 images with 400 × 400 pixels. This dataset is split into seven different classes, namely, grass land, forest, farm land, parking lots, residential region, industrial region, and rivers/lakes.

4.1.1.8 LoveDA

The LoveDA dataset (Wang J. et al., 2021) collects different images of different cities and villages from Nanjing, Changzhou, and Wuhan, China. Along with a spatial resolution of 3 m, this dataset offers 5,987 RS images. Each picture has a resolution of 1,024 × 1,024. This dataset provides six categories, namely, building, roads, water, infertile soil, forest, and agriculture.

4.1.2 Aerial image datasets

We review a few RS semantic segmentation datasets captured by aircraft in this section. These data have the following characteristics: high definition, large scale, small area, and high visibility.

4.1.2.1 Landcover

The Landcover aerial image labeling dataset consists of images from Poland's rural areas, from which there are 39.51 km^2 with a size of 50 cm/pixel and 176.76 km^2 with a resolution of 25 cm/pixel. These images are labeled with four classes. They are forests, water, building, and others.

4.1.2.2 UAVid

UAVid (Ye et al., 2020) is a UAV semantic segmentation dataset revolving around city street scenes with a resolution of 4,096 × 2,160 and 3,840 × 2,160. It contains 300 images intensively labeled with eight classes to cope with the semantic labeling task. The eight classes are architecture, urban road, tree, low vegetation, moving car, static car, human, and clutter/background. UAV is a quite challenging field due to the high resolution of images and the elaboration of scenes.

4.1.2.3 ISAID

This dataset is designed for instance segmentation (Zheng et al., 2020a), offering 2,806 high-resolution RS images from

approximately 800 × 800 pixels to approximately 4,000 × 13,000 pixels with 15 foreground classes and 1 background class.

4.1.2.4 Massachusetts road datasets

The Massachusetts road dataset covers 2,600 km^2 of Massachusetts. This dataset consists of aerial images with a size of at least 1,500 × 1,500 and a resolution of 1 m. In addition, this dataset also provides seven pixels of ground segmentation truth collected from OpenStreetMap.

4.1.2.5 DLRSD

With a spatial size of 256 × 256, DLRSE (Shao et al., 2020) consists of 2,100 RGB images and a resolution of 0.3 m. The dataset is labeled based on the UCMerced LandUse dataset with 17 categories, namely, airplanes, bare soil, architecture, car, chaparral, courthouse, dock, field, grass, mobile house, sidewalk, sand, marine, ship, tank, trees, and water.

4.2 Experimental comparison

Semantic segmentation methods for RS images are most commonly used for experimental comparisons on datasets ISPRS Vaihingen and ISPRS Potsdam. This paper summarizes the referenced RS semantic segmentation papers in the experimental comparison of the two as shown in Table 2, using the indicators mF1, mIoU, and OA.

The values cannot rank the performance of the methods, because the training set and test size of different papers are

TABLE 2 Comparison of different methods on ISPRS Potsdam and Vaihingen datasets.

Methods	Models	ISPRS Potsdam			ISPRS Vaihingen		
		mF1 (%)	mIoU (%)	OA (%)	mF1 (%)	mIoU (%)	OA (%)
Based on CNN	EFCNet (Chen L. et al., 2021)	79.74	65.7	80.72	81.87	70.14	85.46
	SDFCNv2 (Chen G. et al., 2021)	–	67.82	85.03	–	–	–
	EGCAN (Liu Z. et al., 2022)	93	–	91.4	89.7	–	91
	HCANet (Bai et al., 2021)	88.07	–	88.92	88.94	–	89.71
Attention Mechanism	MANet (Li et al., 2021b)	92.9	86.95	91.32	90.41	82.71	90.96
	ABCNet (Li et al., 2021c)	92.7	86.5	91.3			
	A2-FPN (Li R. et al., 2022)	92.4	86.1	91.1	90.1	82.2	91
	LANet (Ding et al., 2020a)	91.95	–	90.84	88.09	–	89.83
	SCAttNet (Li H. et al., 2020)	–	68.31	87.97	–	–	–
	CAM-DFCN (Luo et al., 2019)	89.43	–	90.26	88.55	–	90.41
	DSPCANet (Li YC. et al., 2021)	–	77.66	90.13	–	72.56	87.32
	MARE (Marsocci et al., 2021)	–	–	–	87.95	90.35	81.76
	MAResUNet (Li et al., 2021a)	–	–	–	90.28	83.3	90.86
	EaNet (Zheng et al., 2020b)	–	–	–	90.3		90.8
Multi-scale Strategy	FSHRNet (Xu H. et al., 2022)	90.67	83.16	89.82	86.66	88.38	76.86
	SMAFNet (Chen et al., 2020)	88.18	71.31	86.77	86.91	65.28	88.45
	MFNet (Li et al., 2021b)	–	–	91.65	88.24	77.05	91.47
	DGPRNet (Zhang Y. et al., 2021)	–	77.05	85.69	–	82.36	90.43
Based on Transformer	DC-Swin (Wang L. et al., 2022)	93.25	87.56	92	90.71	83.22	91.63
	DHT-E (Zhang et al., 2022a)	–	81.7	89.3	–	–	–
	ICTNet (Li X. et al., 2022)	93	–	91.57	92.34	–	90.14
	MAT (Zhao X. et al., 2021)	91.59	84.82	–	88.7	79.93	–
	SUDNet (Xu Y. et al., 2022)	92.57	86.4	92.98	89.49	81.26	90.95
	CG-Swin (Meng et al., 2022)	93.29	87.61	91.93	90.81	83.39	91.68
	SSAtNet (Zhao Q. et al., 2021)	–	–	–	–	76.4	88.01
	SwinTF-PSP (Panboonyuen et al., 2021)	–	–	–	94.83	90.98	–
Based on GAN	Semi-supervised GAN (Kerdegari et al., 2019)	88.57	–	87.89	87.08	–	88.34

different during the experiments. However, according to the comparison of different methods in the table, the overall performance of the method based on the attention mechanism and the Transformer mechanism is better than others.

Attention mechanisms are widely used in RS semantic segmentation, combining channel and spatial attention or multi-scale features to improve segmentation performance. The Transformer can perceive the global information of the input sequence, which is a huge advantage of the Transformer over CNN. In CNN, information can only start locally, and as the number of layers increases, the area that can be perceived gradually increases. However, the Transformer starts from the input, and each layer structure can see all the information and establish the association between the basic units, so it can handle more complex problems.

4.3 Discussion

The benefits and drawbacks of typical techniques are analyzed through the experimental outcomes combined with their characteristics, as shown in Table 3. Researchers can use the strengths and weaknesses of the methods as a research reference to carry out future work.

5 Conclusion and future direction

This paper reviews the state-of-the-art progress in semantic segmentation of RS images, which summarizes them from the angle of DL framework and technology. The earliest CNN-based classical methods were applied to the semantic segmentation task in the field of RS data and achieved good experimental results. Next, with burgeoning technologies such as attention mechanism, multi-scale, Transformer, and GAN, the performance of high-pixel semantic segmentation is improved. Integrating multiple techniques is a wise choice for researchers, which enables the progress of both the accuracy and efficiency of the segmentation.

After an in-depth study of semantic segmentation techniques, we found that although researchers have made effective efforts, there are still many challenges in this research, and further efforts are required in future work.

TABLE 3 The advantages and disadvantages analysis and selection guidance for the existing methods.

Methods	Advantages	Disadvantages
Chen G. et al. (2021)	The number of model parameters is small; it can excavate deep generalized features.	Rely on a large number of training datasets.
Chen et al. (2022)	It takes much less running time.	Need to focus on unsupervised learning.
Abdollahi et al. (2021)	It takes forward and backward dependencies into account and considers all the information.	Need to do multi-object segmentation from remote sensing data simultaneously.
Foivos et al. (2020).	Tanimoto loss results in balanced gradients can be used for regression problems	Due to the original image being reduced, the fine details of the trees cannot be recognized
Weng et al. (2020)	It reduces a large number of parameters; The training speed is high.	Missed detections and false alarms, achieved poor water-body extraction results without complete water-body boundaries.
Du et al. (2014)	It can alleviate the retention of accurate boundary information on ground objects.	The recognition accuracy of objects with large scale is not high.
Li et al. (2021c)	It can obtain detailed spatial and contextual information. It reduces the parameter number.	It is dependent on fully convolutional networks.
Li Y. C. et al. (2021)	It can extract effective spectral and spatial enhancement features.	Need to focus on the multi-scale convolution in different topologies.
Zheng et al. (2022)	It combines the advantages of merging Transformer and CNN to get local and global features.	Obtains more refined object information
Li Z. et al. (2021)	It draws contextual information and refines objects at dense multi-scales.	It leads to a decreased performance in the recovery of edges of very thin semantics.
Li et al. (2022b)	It can better identify dense buildings and small targets.	Need to automatic enhancement of training data.
Ma L. et al. (2022)	It has effective attention weight enhancement and edge convolutions for powerful local feature encodings.	Missing validation results on other remote sensing datasets.
Xu et al. (2021)	It can better solve the problem of high computing load and blurred edges.	Boundary detection is not well resolved.
Gao et al. (2021)	Avoiding gradient disappearance and feature map information loss.	The algorithm structure is complex.
Pan et al. (2020)	The model can generate ground-truth data by controlling the numeral and scope of samples.	Need to use supervised training data to fit the parameters.
Hong et al. (2020)	It eliminates the gap between modalities and obtains a smoother and more detailed appearance in urban scene parsing.	Massive labeled RS images are required for its training.

- High-resolution RS images require manual pixel labeling, which is arduous and labor-intensive. Therefore, the problem of insufficient samples still exists. Future work can be improved in the following aspects: (1) how to construct multi-angle, multi-tone, and other sample analysis models; (2) exploring approaches to achieve more promising performance, rarely using fine annotation or rough brands, and reducing training samples; and (3) merging datasets and combining different optical and SAR datasets. Robust Transformer models can be explored for multi-source RS data, which comprise aerial and satellite images with diverse spatial and spectral resolutions.
- Optimize and improve the semantic segmentation models. Semantic segmentation technology can directly promote the development of smart cities, resource monitoring, and other fields. These tasks generate a higher demand for models. (1) How to better capture more differentiated features and context information for its high-resolution images. (2) How to design unsupervised learning models for improving the performance of high-resolution images, including weakly supervised and semi-supervised methods, which do not require a large amount of labeled data. (3) Change the number or types of convolutions in convolutional models. (4) How to replace the edge-guided context aggregation method and use better edge extractors in explicit augmentation methods.
- Reduce the computational complexity and improve the robustness of the model. It is important to improve the performance and quality of the existing models, which are large and computationally intensive and hinder their wide application. How to balance the performance and computer power of semantic segmentation is a future research direction. (1) Build real-time semantic segmentation models with less model size and computational complexity. (2) Design a more efficient and concise feature extraction method. (3) Reduce latency.
- Research on more complex actual scenarios. Many experiments are only implemented on specific datasets. Therefore, how to design new methods that can be suitable for actual complex scenarios remains to be studied.
- Research on small target segmentation. Owing to the small proportion of the pixel area of the small target, a certain amount of detailed information will be lost after multiple down-sampling, which will give rise to an accuracy decrease to a certain extent. In the future, we can start with small targets and improve accuracy with methods such as residual connections, attention mechanisms, and pyramid structures.

Unfortunately, since semantic segmentation of RS images is a hot research field, a large number of research methods have emerged in recent years and are constantly updated, so it is difficult for us to find all semantic segmentation methods. In the future, researchers' attention should be directed to new methods and theories for semantic segmentation of RS images.

Author contributions

Conceptualization and methodology, JL; investigation and resources, JL, ML, and LS; data curation, YL and PZ; writing—original draft preparation, JL, ML, and YL; writing—review and editing: QS. All authors contributed to the article and approved the submitted version.

Funding

This work was partially supported by the R&D Program of Beijing Municipal Education Commission (No. KM202211417014), the Academic Research Projects of Beijing Union University (No. ZK20202215), the Natural Science Foundation of Shandong Province under Grant ZR2022LZH015 (ZR2020MF006), the Industry–University Research Innovation Foundation of Ministry of Education of China under Grant (2021FNA01001), and the Shandong Provincial Natural Science Foundation, China under Grant ZR2020MF006 and ZR2022LZH015.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2023.1201125/full#supplementary-material

References

Abdollahi, A., Pradhan, B., Shukla, N., Chakraborty, S., and Alamri, A. M. (2021). Multi-object segmentation in complex urban scenes from high-resolution remote sensing data. *Remote Sens.* 13, 3710. doi: 10.3390/RS13183710

Ahlswede, S., Madam, N. T., Schulz, C., Kleinschmit, B., and Demir, B. (2022). "Weakly supervised semantic segmentation of remote sensing images for tree species classification based on explanation methods," in *Proceedings of the IGARSS*, Kuala Lumpur, Malaysia. IEEE, Vol. 2022. 4847–4850. doi: 10.1109/IGARSS46834.2022.9884676

Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G. S., et al. (2022). Transformers in remote sensing: a survey. *arXiv* 2209, 01206. doi: 10.3390/rs15071860

Andrade, R. B., Mota, G. L. A., and da Costa, G. A. O. P. (2022). Deforestation detection in the Amazon using DeepLabv3+ semantic segmentation model variants. *Remote. Sens.* 14 (19), 4694. doi: 10.3390/rs14194694

Asokan, A., and Anitha, J. (2019). Change detection techniques for remote sensing applications: a survey. *Earth Sci. Inf.* 12, 143–160. doi: 10.1007/s12145-019-00380-5

Avenash, R., and Viswanath, P. (2019). "Semantic segmentation of satellite images using a modified CNN with hard-swish activation function," in *Proceedings of the VISIGRAPP 2019*. Prague, Czech Republic. 413–420. doi: 10.5220/0007469604130420

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE T. Pattern. Anal.* 39, 2481–2495. doi: 10.1109/TPAMI.2016.2644615

Bai, H., Cheng, J., Huang, X., Liu, S. Y., and Deng, C. J. (2021). HCANet: a hierarchical context aggregation network for semantic segmentation of high-resolution remote sensing images. *IEEE Geosci. Remote. Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2021.3063799

Bello, I. M., Zhang, K., Su, Y., Wang, J. Y., and Aslam, M. A. (2022). Densely multiscale framework for segmentation of high resolution remote sensing imagery. *Comput. Geosci.* 167, 105196. doi: 10.1016/j.cageo.2022.105196

Chen, L., Dou, X., Peng, J., Li, W. B., Sun, B. Y., Li, H. F., et al. (2021). EFCNet: ensemble full convolutional network for semantic segmentation of high-resolution remote sensing images. *IEEE Geosci. Remote. Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2021.3076093

Chen, K., Fu, K., Yan, M., Gao, X., Sun, X., and Wei, X. (2018). Semantic segmentation of aerial images with shuffling convolutional neural networks. *IEEE Geosci. Remote. Sens. Lett.* 15, 173–177. doi: 10.1109/LGRS.2017.2778181

Chen, G., He, C., Wang, T., Zhu, K., Liao, P. Y., and Zhang, X. D. (2022). A superpixel-guided unsupervised fast semantic segmentation method of remote sensing images. *IEEE Geosci. Remote. Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2022.3198065

Chen, Z., Li, D., Fan, W., Guan, H. Y., Wang, C., and Li, J. (2021). Self-attention in reconstruction bias U-net for semantic segmentation of building rooftops in optical remote sensing images. *Remote. Sens.* 13 (13), 2524. doi: 10.3390/rs13132524

Chen, J., Lu, Y., Yu, Q., Luo, X., and Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* 2102, 4306. doi: 10.48550/arXiv.2102.04306

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2015). "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proceedings of the 2015 ICLR*, San Diego, CA, USA, Vol. 4. 357–361.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017a). Deeplab: semantic image segmentation with deep convolutional nets,atrous convolution, and fully connected crfs. *IEEE T. Pattern. Anal.* 40, 834–848. doi: 10.1109/TPAMI.2017.2699184

Chen, L. C., Papandreou, G., Schroff, F., and Adam, H. (2017b). Rethinking atrous convolution for semantic image segmentation. *arXiv*. doi: 10.48550/arXiv.1706.05587

Chen, G., Tan, X., Guo, B., Zhu, K., Liao, P., Wang, T., et al. (2021). SDFCNv2: an improved FCN framework for remote sensing images semantic segmentation. *Remote. Sens.* 13, 4902. doi: 10.3390/rs13234902

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 ECCV 2018*, Munich, Germany, Springer, Vol. 2. 801–818. doi: 10.1007/978-3-030-01234-2_49

Chen, J., Zhu, J., Sun, G., Li, J. H., and Deng, M. (2020). SMAF-net: sharing multiscale adversarial feature for high-resolution remote sensing imagery semantic segmentation. *IEEE Geosci. Remote. Sens. Lett.* 18, 1921–1925. doi: 10.1109/LGRS.2020.3011151

Ciresan, D., Giusti, A., Gambardella, L., and Schmidhuber. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. *NIPS* 2012, 2825–2860.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: an overview. *IEEE Signal Proc. Mag.* 35, 53–65. doi: 10.1109/MSP.2017.2765202

Cui, L., Jing, X., Wang, Y., Huan, Y. X., Xu, Y., and Zhang, Q. Q. (2022). Improved Swin Transformer-based semantic segmentation of postearthquake dense buildings in urban areas using remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* 16, 369–385. doi: 10.1109/JSTARS.2022.3225150

Cui, H., Zhang, G., Qi, J., Li, H. F., Tao, C., Li, X., et al. (2022). MDANet: unsupervised, mixed-domain adaptation for semantic segmentation of remote sensing images. *IEEE Geosci. Remote. Sens. Lett.* 19, 1–5. doi: 10.3390/rs13030454

Davis, L. S., Rosenfeld, A., and Weszka, J. S. (1975). Region extraction by averaging and thresholding. *IEEE T. Syst. Man. CY-S* 1975, 3, 383–3, 388. doi: 10.1109/tsmc.1975.5408419

Deng, G., Wu, Z., Wang, C., Xu, M. Z., and Zhong, Y. F. (2021). CCANet: class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation. *IEEE Trans. Geosci. Remote. Sens.* 60, 1–20. doi: 10.1109/TGRS.2021.3055950

Ding, L., Tang, H., and Bruzzone, L. (2020a). LANet: local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote. Sens.* 59, 426–435. doi: 10.1109/TGRS.2020.2994150

Ding, L., Zhang, J., and Bruzzone, L. (2020b). Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture. *IEEE Trans. Geosci. Remote. Sens.* 58, 5367–5376. doi: 10.1109/TGRS.2020.2964675

Dong, S., and Chen, Z. (2021). A multi-level feature fusion network for remote sensing image segmentation. *Sensors* 21, 1267. doi: 10.3390/s21041267

Du, S., Du, S., Liu, B., and Zhang, X. Y. (2014). Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. *Int. J. Digit. Earth* 14, 357–378. doi: 10.1080/17538947.2020.1831087

Duan, L., and Hu, X. (2019). Multiscale refinement network for water-body segmentation in high-resolution satellite imagery. *IEEE Geosci. Remote. Sens. Lett.* 17, 686–690. doi: 10.1109/LGRS.2019.2926412

Fan, Z. Y., Hou, J. M., Zang, Q., Chen, Y. J., and Yan, F. (2022). River segmentation of remote sensing images based on composite attention network. *Complex* 1–13. doi: 10.1155/2022/7750281

Feng, D., Zhang, Z., and Yan, K. (2022). A semantic segmentation method for remote sensing images based on the Swin Transformer fusion gabor filter. *IEEE Access* 10, 77432–77451. doi: 10.1109/ACCESS.2022.3193248

Foivos, I., Diakogiannis, F. W., Peter, C., and Chen, W. (2020). ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm.* 162, 94–114. doi: 10.1016/j.isprsjprs.2020.01.013

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y. J., Fang, Z. W., et al. (2019). "Dual attention network for scene segmentation," in *Proceedings of the CVPR 2019*. Long Beach, CA, USA, IEEE, 3146–3154. doi: 10.1109/CVPR.2019.00326

Gao, L., Liu, H., Yang, M., Chen, L., Wan, Y. L., Xiao, Z. Q., et al. (2021). STransFuse: fusing Swin Transformer and convolutional neural network for remote sensing image semantic segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 14, 10990–11003. doi: 10.1109/JSTARS.2021.3119654

Ghali, R., Akhloufi, M. A., Jmal, M., Mseddi, W. S., and Attia, R. (2021). Wildfire segmentation using deep vision Transformers. *Remote. Sens.* 13, 3527. doi: 10.3390/rs13173527

Gu, X., Li, S., Ren, S., Zheng, H. B., Fan, C. C., and Xu, H. L. (2022). Adaptive enhanced Swin Transformer with U-net for remote sensing image segmentation. *Comput. Electr. Eng.* 102, 108223. doi: 10.1016/j.compeleceng.2022.108223

Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., et al. (2022). Attention mechanisms in computer vision: a survey. *Comput. Visual Media* 8, 331–368. doi: 10.1007/s41095-022-0271-y

He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., and Xue, Y. (2022). Swin Transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* 60, 4408715. doi: 10.1109/TGRS.2022.3144165

Hong, D., Yao, J., Meng, D., Xu, Z. B., and Chanussot, J. (2020). Multimodal GANs: toward crossmodal hyperspectral–multispectral image segmentation. *IEEE Trans. Geosci. Remote. Sens.* 59, 5103–5113. doi: 10.1109/TGRS.2020.3020823

Hu, H., Cai, S., Wang, W., Zhang, P., and Li, Z. Y. (2019). "A semantic segmentation approach based on deepLab network in high-resolution remote sensing images," in *Proceedings of the ICIG 2019*, Beijing, China. Springer, 292–304. doi: 10.1007/978-3-030-34113-8_25

Huang, H., Lin, L., Tong, R., Hu, H. J., Zhang, Q. W., and Iwamoto, Y. (2022). "UNet 3+: a full-scale connected UNet for medical image segmentation," in *Proceedings of the ICASSP 2020*, Barcelona. IEEE, 1055–1059. doi: 10.1109/ICASSP40776.2020.9053405

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the CVPR 2017*, Honolulu, Hawaii, USA. 4700–4708. doi: 10.1109/CVPR.2017.243

Huang, B., Lu, K., Audeberr, N., Khalel, A., Tarabalka, Y., Malof, J., et al. (2018). "Large-Scale semantic classification: outcome of the first year of inria aerial image labeling benchmark," in *Proceedings of the IGARSS 2018*, Valencia, Spain. 6947–6950. doi: 10.1109/IGARSS.2018.8518525

Iglovikov, V., Seferbekov, S., Buslaev, A., and Shvets, A. (2018). "Ternausnetv2: fully convolutional network for instance segmentation," in *Proceedings of the CVPR 2018*, Munich, Germany. IEEE, 233–237. doi: 10.1109/CVPRW.2018.00042

Jiang, H., Peng, M., Zhong, Y., Xie, H. F., Hao, Z. M., Lin, J. M., et al. (2022). A survey on deep learning-based change detection from high-resolution remote sensing images. *Remote. Sens.* 14, 1552. doi: 10.3390/rs14071552

Kemker, R., Salvaggio, C., and Kanan, C. (2018). Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm.* 145, 60–77. doi: 10.1016/j.isprsjprs.2018.04.014

Kerdegari, H., Razaak, M., Argyriou, V., and Remagnino, P. (2019). "Urban scene segmentation using semi-supervised GAN," in *Proceedings of the Image and Signal Processing for Remote Sensing*, Denver, USA. 477–484. doi: 10.1117/12.2533055

Kitaev, N., Lu, T., and Klein, D. (2022). "Learned incremental representations for parsing," in *Proceedings of the ACL 2022*, Xiangcheng, China. 3086–3095. doi: 10.18653/v1/2022.acl-long.220

Kong, Y., Liu, Y., Yan, B., Leung, H., and Peng, X. Y. (2021). A novel deeplabv3+ network for sar imagery semantic segmentation based on the potential energy loss function of gibbs distribution. *Remote. Sens.* 13, 454. doi: 10.3390/rs13030454

Li, Z. Q., Chen, X., Jiang, J., Han, Z., Li, Z. H., Fang, T., et al. (2021). Cascaded multiscale structure with self-smoothing atrous convolution for semantic segmentation. *IEEE Trans. Geosci. Remote. Sens.* 60, 1–13. doi: 10.1109/TGRS.2021.3088902

Li, Y., Chen, R., Zhang, Y., Zhang, M., and Chen, L. (2020). Multi-label remote sensing image scene classification by combining a convolutional neural network and a graph neural network. *Remote Sens.* 12, 4003. doi: 10.3390/rs12234003

Li, W., Gao, H., Su, Y., and Momanyi, B. M. (2022). Unsupervised domain adaptation for remote sensing semantic segmentation with Transformer. *Remote. Sens.* 14, 4942. doi: 10.3390/rs14194942

Li, Y. C., Li, H. C., Hu, W. S., and Yu, H. L. (2021). DSPCANet: dual-channel scale-aware segmentation network with position and channel attentions for high-resolution aerial images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* 14, 8552–8565. doi: 10.1109/JSTARS.2021.3102137

Li, C., Li, X., Xia, R., Li, T., Lyu, X., Tong, Y., et al. (2021). Hierarchical self-attention embedded neural network with dense connection for remote-sensing image semantic segmentation. *IEEE Access* 9, 126623–126634. doi: 10.1109/ACCESS.2021.3111899

Li, W., and Liu, X. (2023). System dynamics simulation and regulation of human-water system coevolution in Northwest China. *Front. Ecol. Evol.* 10. doi: 10.3389/FEVO.2022.1106998

Li, D., Liu, J., Liu, F., Zhang, W. H., Zhang, A. D., Gao, W. F., et al. (2022). "A dual-fusion semantic segmentation framework with gan for sar images," in *Proceedings of the IGARSS*, Vol. 2022. 991–994. doi: 10.1109/IGARSS46834.2022.9884931

Li, H., Qiu, K., Chen, L., Mei, X. M., Hong, L., and Tao, C. (2020). SCAttNet: semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. *IEEE Geosci. Remote. Sens. Lett.* 18, 905–909. doi: 10.1109/LGRS.2020.2988294

Li, Z., Wang, Y., Zhang, N., Zhang, Y. X., Zhao, Z. K., Xu, D. D., et al. (2022a). Deep learning-based object detection techniques for remote sensing images: a survey. *Remote Sens. Remote. Sens.* 14, 2385. doi: 10.3390/rs14102385

Li, R., Wang, L., Zhang, C., Duan, C. X., and Zheng, S. Y. (2022). A2-FPN for semantic segmentation of fine-resolution remotely sensed images. *Remote Sens.* 43 (3), 1131–1155. 16. doi: 10.1080/01431161.2022.2030071

Li, X., Xu, F., Xia, R. L., Li, T., Chen, Z. Q., Wang, X. Y., et al. (2022). Encoding contextual information by interlacing Transformer and convolution for remote sensing imagery semantic segmentation. *Remote. Sens.* 14, 4065. doi: 10.3390/rs14164065

Li, Z., Zhang, Z., Chen, D., Zhang, L. Q., Zhu, L., Wang, Q., et al. (2022b). HCRB-MSAN: horizontally connected residual blocks-based multiscale attention network for semantic segmentation of buildings in HSR remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* 15, 5534–5544. doi: 10.1109/JSTARS.2022.3188515

Li, Y., Zhang, H. K., Xue, X. Z., Jiang, Y., Shen, Q., Li, Y., et al. (2018). Deep learning for remote sensing image classification: a survey. *WIREs Data Min. Knowl. Discov.* 8, e1264. doi: 10.1002/widm.1264

Li, R., Zheng, S. Y., Duan, C. X., Su, J. L., and Zhang, C. (2021a). Multistage attention ResU-net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geosci. Remote. Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2021.3063381

Li, R., Zheng, S. Y., Zhang, C., Duan, C. X., Su, J. L., Wang, L. B., et al. (2021b). Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–13. doi: 10.1109/TGRS.2021.3093977

Li, R., and Duan, C. X. (2021c). ABCNet: attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS J. Photogramm.* 181, 84–98.46. doi: 10.1016/j.isprsjprs.2021.09.005

Liu, Y., Fan, B., Wang, L., Bai, J., Xiang, S., and Pan, C. (2018a). Semantic labeling in very high resolution images *via* a self-cascaded convolutional neural network. *ISPRS J. Photogramm.* 145, 78–95. doi: 10.1016/j.isprsjprs.2017.12.007

Liu, B., Hu, J. W., Bi, X. L., Li, W. S., and Gao, X. B. (2022). PGNet: positioning guidance network for semantic segmentation of very-High-Resolution remote sensing images. *Remote. Sens.* 14, 4219. doi: 10.3390/rs14174219

Liu, Z. Q., Li, J. J., Song, R., Wu, C. X., Liu, W., Li, Z., et al. (2022). Edge guided context aggregation network for semantic segmentation of remote sensing imagery. *Remote. Sens.* 14, 1353. doi: 10.3390/rs14061353

Liu, Z., Lin, Y. T., Cao, Y., Hu, H., Wei, Y. X., Zhang, Z., et al. (2021). "Swin Transformer: hierarchical vision Transformer using shifted windows," in *Proceedings of the CVPR 2021*, Kuala Lumpur, Malaysia. IEEE, 10012–10022. doi: 10.1109/ICCV48922.2021.00986

Liu, Y. H., Mei, S. H., Zhang, S., Wang, Y., He, M. Y., and Du, Q. (2022). "Semantic segmentation of high-resolution remote sensing images using an improved Transformer," in *Proceedings of the IGARSS 2022*. Kuala Lumpur, Malaysia, IEEE, 3496–3499. doi: 10.1109/IGARSS46834.2022.9884103

Liu, Y., Piramanayagam, S., Monteiro, S. T., and Saber, E. (2019). Semantic segmentation of multisensor remote sensing imagery with deep ConvNets and higher-order conditional random fields. *J. Appl. Remote Sens.* 13, 1. doi: 10.1117/1.JRS.13.016501

Liu, R. R., Tao, F., Liu, X. T., Na, J. M., Leng, H. J., Wu, J. J., et al. (2022). RAANet: a residual ASPP with attention framework for semantic segmentation of high-resolution remote sensing images. *Remote. Sens.* 14, 3109. doi: 10.1109/JSTARS.2020.3021098

Liu, Y. H., Yao, J., Lu, X. H., Xia, M. H., Wang, X. B., and Liu, Y. (2018b). RoadNet: learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images. *IEEE Trans. Geosci. Remote. Sens.* 57, 2043–2056. doi: 10.1109/TGRS.2018.2870871

Long, J., Shelhamer, E., and Darrell, T. (2019). Fully convolutional networks for semantic segmentation. *IEEE T. Pattern. Anal.* 39, 640–651. doi: 10.1109/TPAMI.2016.2572683

Lou, A., Guan, S., and Loew, M. (2021). "DC-UNet: rethinking the U-net architecture with dual channel efficient CNN for medical image segmentation," in *Proceedings of the Medical Imaging Kunming*, China, Vol. 11596. 758–768. doi: 10.1117/12.2582338

Lu, H., Liu, Q., Liu, X. D., and Zhang, Y. H. (2021). A survey of semantic construction and application of satellite remote sensing images and data. *Organ. End User Comput.* 33, 1–20. doi: 10.4018/JOEUC.20211101.oa6

Luo, H. F., Chen, C. C., Fang, L. N., Zhu, X., and Lu, L. J. (2019). High-resolution aerial images semantic segmentation using deep fully convolutional network with channel attention mechanism. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* 12, 3492–3507. doi: 10.1109/JSTARS.2019.2930724

Luo, Y. Y., Wang, J. N., Yang, X. K., Yu, Z. Y., and Tan, Z. X. (2022). Pixel representation augmented through cross-attention for high-resolution remote sensing imagery segmentation. *Remote. Sens.* 14, 5415. doi: 10.3390/rs14215415

Ma, A. L., Wang, J. J., Zhong, Y. F., and Zheng, Z. (2021). Factseg: foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* 60, 1–16. doi: 10.1109/TGRS.2021.3097148

Ma, X. P., Zhang, X. K., Pun, M., and Liu, M. (2022). "MSFNET: multi-stage fusion network for semantic segmentation of fine-resolution remote sensing data," in *Proceedings of the IGARSS 2022*. Kuala Lumpur, Malaysia, IEEE, 2833–2836. doi: 10.1109/IGARSS46834.2022.9883789

Ma, J. B., Zhou, W. J., Qian, X. H., and Yu, L. (2022). Deep-separation guided progressive reconstruction network for semantic segmentation of remote sensing images. *Remote. Sens.* 14, 5510. doi: 10.3390/rs14215510

Ma, L. F., Li, J., Guan, H. Y., Yu, Y. T., and Chen, Y. P. (2022). STN: saliency-guided Transformer network for point-wise semantic segmentation of urban scenes. *IEEE Geosci. Remote. Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2022.3190558

Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M., and Stilla, U. (2017). Classification with an edge: improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm.* 135, 158–172. doi: 10.1016/j.isprsjprs.2017.11.009

Marmanis, M., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., and Stilla, U. (2016). "Semantic segmentation of aerial images with an ensemble of CNSS," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. 3. 473–480.

Marsocci, V., Scardapane, S., and Komodakis, N. (2021). MARE: self-supervised multi-attention REsu-net for semantic segmentation in remote sensing. *Remote. Sens.* 13 (16), 3275.8. doi: 10.3390/rs13163275

Maxwell, A. E., Bester, M. S., Guillén, L. A., Ramezan, C. A., Carpinello, D. J., Fan, Y. T., et al. (2020). Semantic segmentation deep learning for extracting surface mine extents from historic topographic maps. *Remote. Sens.* 12, 4145. doi: 10.3390/rs12244145

Meng, X. L., Yang, Y. C., Wang, L. B., Wang, T., Li, R., and Zhang, C. (2022). Class-guided Swin Transformer for semantic segmentation of remote sensing imagery. *IEEE Geosci. Remote. Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2022.3215200

Mi, L., and Chen, Z. (2020). Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation. *ISPRS J. Photogramm.* 159, 140–152. doi: 10.1016/j.isprsjprs.2020.08.015

Mo, Y., Wu, Y., Yang, X., Liu, F., and Liao, Y. (2022). Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing* 493, 626–646. doi: 10.1016/J.NEUCOM.2022.01.005

Mou, L., Hua, Y., and Zhu, X. X. (2020). Relation matters: relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. *IEEE Trans. Geosci. Remote Sens.* 58 (11), 7557–7569. doi: 10.1109/TGRS.2020.2979552

Noble, P. J., Seitz, C., Lee, S. S., Manoylov, K. M., and Chandra, S. (2023). Characterization of algal community composition and structure from the nearshore environment, Lake Tahoe. *Front. Ecol. Evol.* 10, 1053499. doi: 10.3389/fevo.2022.1053499

Nowozin, S., and Lampert, C. H. (2011). Structured learning and prediction in computer vision. *Found. Trends. Comput.* 6, 185–365. doi: 10.1561/0600000033

Özden, M., and Polat, E. (2005). "Image segmentation using color and texture features," in *Proceedings of the EUSIPCO 2005*. Antalya, Turkey, IEEE, 1–4.

Pan, X., Zhao, J., and Xu, J. (2020). Conditional generative adversarial network-based training sample set improvement model for the semantic segmentation of high-resolution remote sensing images. *IEEE Trans. Geosci. Remote. Sens.* 59, 7854–7870. doi: 10.1109/TGRS.2020.3033816

Panboonyuen, T., Jitkajornwanich, K., Lawawirojwong, S., Srestasathiern, P., and Vateekul, P. (2019). Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning. *Remote. Sens.* 11, 83. doi: 10.3390/rs11010083

Panboonyuen, T., Jitkajornwanich, K., Lawawirojwong, S., Srestasathiern, P., and Vateekul, P. (2021). Transformer-based decoder designs for semantic segmentation on remotely sensed images. *Remote. Sens.* 13, 5100. doi: 10.3390/rs13245100

Pastorino, M., Moser, G., Serpico, S. B., and Zerubia, J. (2022a). Fully convolutional and feedforward networks for the semantic segmentation of remotely sensed images. *Proc. ICIP 2022*, 1876–1880. doi: 10.1109/ICIP46576.2022.9897336

Pastorino, M., Moser, G., Serpico, S. B., and Zerubia, J. (2022b). "Semantic segmentation of sar images through fully convolutional networks and hierarchical probabilistic graphical models," in *Proceedings of the IGARSS 2022*. Kuala Lumpur, Malaysia, IEEE, 1047–1050. doi: 10.1109/IGARSS46834.2022.9883111

Piramanayagam, S., Saber, E., Schwartzkopf, W., and Koehler, F. (2018). Supervised classification of multisensor remotely sensed images using a deep learning framework. *Remote. Sens.* 10, 1429. doi: 10.3390/rs10091429

Priyanka, Sravya, N., Shyam, L., Nalini, J., Chintala, S. R., and Fabio, D. (2022). DIResUNet: architecture for multiclass semantic segmentation of high resolution remote sensing imagery data. *Appl. Intell.* 52, 15462–15482. doi: 10.1007/s10489-022-03310-z

Qin, Z., Ni, L. H., Zhang, T., and Wang, Q. (2015). Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote. Sens. Lett.* 12, 1–5. doi: 10.1109/LGRS.2015.2475299

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *Proceedings of the MICCAI 2015*, Munich, Germany, Springer, Vol. 2015. 234–241.

Ru, L. X., Zhan, Y. B., Yu, B. S., and Du, B. (2022). "Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with Transformers," in *Proceedings of the CVPR 2022*, New Orleans, Louisiana. IEEE, 16846–16855. doi: 10.1109/CVPR52688.2022.01634

Sebastian, S., Rohith, G., and Kumar, L. S. (2022). Significant full reference image segmentation evaluation: a survey in remote sensing field. *Multim. Tools Appl.* 81, 17959–17987. doi: 10.1007/s11042-022-12769-4

Senthilkumaran, N., and Rajesh, R. (2009). "Image segmentation-a survey of soft computing approaches," in *Proceedings of 2009 International Conference on Advances in Recent Technologies in Communication and Computing. IEEE*, Vol. 2009. 844–846. doi: 10.1109/ARTCom.2009.219

Seong, S., and Choi, J. (2021). Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates. *Remote. Sens.* 13, 3087. doi: 10.3390/rs13163087

Shamsolmoali, P., Zareapoor, M., Zhou, H., Wang, R., and Yang, J. (2020). Road segmentation for remote sensing images using adversarial spatial pyramid networks. *IEEE Trans. Geosci. Remote. Sens.* 59, 4673–4688. doi: 10.1109/TGRS.2020.3016086

Shao, Z. F., Zhou, W. X., Deng, X. Q., Zhang, M. D., and Cheng, Q. M. (2020). Multilabel remote sensing image retrieval based on fully convolutional network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* 13, 318–328. doi: 10.1109/JSTARS.2019.2961634

Song, M. X., Li, B. C., Wei, P. J., Shao, Z. H., Wang, J., and Huang, J. (2022). "DMF-CL: dense multi-scale feature contrastive learning for semantic segmentation of remote-sensing images," in *Proceedings of Pattern Recognition and Computer Vision*, Shenzhen, China. Springer, 152–164. doi: 10.1007/978-3-031-18916-6_13

Su, Y. C., Liu, T. J., and Liuy, K. H. (2022). "Multi-scale wavelet frequency channel attention for remote sensing image segmentation," in *Proceedings of IVMSP 2022*, Nafplio, Greece, IEEE, 1–5. doi: 10.1109/IVMSP54334.2022.9816247

Subudhi, S., Narayan, R., Biswal, P. K., and Dell'acqua, F. (2021). A survey on superpixel segmentation as a preprocessing step in hyperspectral image analysis. *IEEE J-STARS* 14, 5015–5035. doi: 10.1109/JSTARS.2021.3076005

Sun, S. T., Mu, L., Wang, L. Z., Liu, P., Liu, X. L., and Zhang, Y. W. (2021). Semantic segmentation for buildings of large intra-class variation in remote sensing images with O-GAN. *Remote. Sens.* 13, 475. doi: 10.3390/rs13030475

Sun, X., Shi, A. J., Huang, H., and Mayer, H. (2020). BAS4Net: boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* 13, 5398–5413. doi: 10.1109/JSTARS.2020.3021098

Sun, Y., Tian, Y., and Xu, Y. (2019). Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: structural stereotype and insufficient learning. *Neurocomputing* 330, 297–304. doi: 10.1016/j.neucom.2018.11.051

Sun, Z. Y., Zhou, W. P., Ding, C., and Xia, M. (2022). Multi-resolution Transformer network for building and road segmentation of remote sensing image. *ISPRS Int. J. Geo Inf.* 11, 165. doi: 10.3390/ijgi11030165

Tasar, O., Tarabalka, Y., and Alliez, P. (2019). Incremental learning for semantic segmentation of large-scale remote sensing data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* 12, 3524–3537. doi: 10.1109/jstars.2019.2925416

Tian, L., Zhong, X., and Chen, M. (2021). Semantic segmentation of remote sensing image based on GAN and FCN network model. *Sci. Program.* 11, 1–11. doi: 10.1155/2021/9491376

Tong, X. Y., Xia, G. S., Lu, Q., Shen, H., Li, S., You, S., et al. (2020). Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sens. Environ.* 237, 111322. doi: 10.1016/J.CAGEO.2021.104969

Tsagkatakis, G., Aidini, A., Fotiadou, K., Giannopoulos, M., Pentari, A., and Tsakalides, P. (2019). Survey of deep-learning approaches for remote sensing observation enhancement. *Sensors* 19, 3929. doi: 10.3390/s19183929

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Aidan, N., et al. (2017). "Attention is all you need," in *Proceedings of the NIPS 2017*. Long Beach, CA, USA, Vol. 2017. 5998–6008.

Venugopal, N. (2020). Automatic semantic segmentation with DeepLab dilated learning network for change detection in remote sensing images. *Neural Process. Lett.* 51, 2355–2377. doi: 10.1007/s11063-019-10174-x

Wang, H., Chen, X. Z., Zhang, T. X., Xu, Z. Y., and Li, J. Y. (2022). CCTNet: coupled CNN and Transformer network for crop segmentation of remote sensing images. *Remote. Sens.* 14, 1956. doi: 10.3390/rs14091956

Wang, M., Du, H. B., and Xu, S. Q. (2022). "Remote sensing image segmentation of ground objects based on improved Deeplabv3+," in *Proceedings of the ICIT 2022*, Shanghai, China. IEEE, 1–6. doi: 10.1109/ICIT48603.2022.10002795

Wang, K., Fan, X., and Wang, Q. (2022). "FPB-UNet++: semantic segmentation for remote sensing images of reservoir area *via* improved UNet++ with FPN," in *Proceedings of the ICIAI 2022*, Guangzhou, China. ACM, 100–104. doi: 10.1117/12.2582338

Wang, Y. H., Gao, L., Hong, D. F., Sha, J. J., Liu, L., Zhang, B., et al. (2021). Mask DeepLab: end-to-end image segmentation for change detection in high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinform.* 104, 102582. doi: 10.1016/j.jag.2021.102582

Wang, L. B., Li, R., Duan, C. X., Zhang, C., and Meng, X. L. (2022). A novel Transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote. Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2022.3143368

Wang, L., Li, R., Wang, D., Duan, C. X., and Meng, X. L. (2021). Transformer meets convolution: a bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote. Sens.* 13 (16), 3065.37. doi: 10.3390/rs13163065

Wang, Z. M., Wang, J. S., Yang, K., Wang, L. M., Su, F. J., and Chen, X. Y. (2022a). Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+. *Comput. Geosci.* 158, 104969. doi: 10.1016/J.CAGEO.2021.104969

Wang, G., Zhai, Q., and Lin, J. (2022). Multi-scale network for remote sensing segmentation. *IET Image Process.* 16, 1742–1751. doi: 10.1049/ipr2.12444

Wang, Z., Zhang, S. W., Gross, L., Zhang, C. L., and Wang, B. H. (2022b). Fused adaptive receptive field mechanism and dynamic multiscale dilated convolution for side-scan sonar image segmentation. *IEEE Trans. Geosci. Remote. Sens.* 60, 1–17. doi: 10.1109/TGRS.2022.3201248

Wang, J. J., Zheng, Z., Ma, A. L., Lu, X. Y., and Zhong, Y. F. (2021). LoveDA: a remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv* 2110, 8733. doi: 10.48550/arXiv.2110.08733

Wei, Y., Zhang, K., and Ji, S. (2020). Simultaneous road surface and centerline extraction from Large-scale remote sensing images using CNN-based segmentation and tracing. *IEEE Trans. Geosci. Remote Sens.* 99, 1–13. doi: 10.1109/TGRS.2020.2991733

Weiss, M., Jacob, F., and Duveiller, G. (2020). Remote sensing for agricultural applications: a meta-review. *Remote Sens. Environ.* 236, 111402. doi: 10.1016/j.rse.2019.111402

Weng, L. G., Xu, Y. M., Xia, M., Zhang, Y. H., Liu, J., and Xu, Y. Q. (2020). Water areas segmentation from remote sensing images using a separable residual segnet network. *ISPRS Int. J. Geo Inf.* 9, 256. doi: 10.3390/ijgi9040256

Woo, S., Park, J. C., Lee, J. Y., and Kweon, I. S. (2018). CBAM: convolutional block attention module. *Proceedings of the ECCV* (Munich, Germany), 3–19.

Wurm, M., Stark, T., Zhu, X. X., Weigand, M., and Hannes., T. (2019). Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS J. Photogrammetry Remote Sens.* 150, 59–69. doi: 10.1016/j.isprsjprs.2019.02.006

Xu, H., Tang, X. M., Ai, B., Yang, F. L., Wen, Z., and Yang, X. M. (2022). Feature-selection high-resolution network with hypersphere embedding for semantic segmentation of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15. doi: 10.1109/TGRS.2022.3183144

Xu, Z., Zhang, W. C., Zhang, T. X., Yang, Z. F., and Li, J. Y. (2021). Efficient Transformer for remote sensing image segmentation. *Remote. Sens.* 13, 3585. doi: 10.3390/rs13183585

Xu, Y., Zhou, S., and Huang, Y. (2022). Transformer-based model with dynamic attention pyramid head for semantic segmentation of VHR remote sensing imagery. *Entropy* 24, 1619. doi: 10.3390/e24111619

Yang, F., and Ma, C. (2022). "Sparse and complete latent organization for geospatial semantic segmentation," in *Proceedings of the CVPR 2022*, IEEE, 1809–1818. doi: 10.1109/CVPR52688.2022.00185

Ye, L., Vosselman, G., Xia, G. S., Yilmaz, A., and Yang, M. Y. (2020). UAVid: a semantic segmentation dataset for UAV imagery. *ISPRS J. Photogramm.* 165, 108–119. doi: 10.1016/j.isprsjprs.2020.05.009

Yue, K., Yang, L., Li, R., Hu, W., Zhang, F., and Li, W. (2019). TreeUNet: adaptive tree convolutional neural networks for subdecimeter aerial image segmentation. *ISPRS J. Photogramm.* 156, 1–13. doi: 10.1016/j.isprsjprs.2019.07.007

Zhang, Y., Gao, X. Y., Duan, Q. Y., Yuan, L., and Gao, X. B. (2022a). DHT: deformable hybrid Transformer for aerial image segmentation. *IEEE Geosci. Remote. Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2022.3222916

Zhang, C., Jiang, W. S., Zhang, Y., Wang, W., Zhao, Q., and Wang, C. J. (2022). Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* 60, 1–20. doi: 10.1109/TGRS.2022.3144894

Zhang, X. Y., Leng, C. C., Hong, Y. M., Pei, Z., Cheng, I., and Basu, A. (2021). Multimodal remote sensing image registration methods and advancements: a survey. *Remote. Sens.* 13, 5128. doi: 10.3390/rs13245128

Zhang, Y., Liu, T., Cattani, C., Cui, Q., and Liu., S. (2021). Diffusion-based image inpainting forensics *via* weighted least squares filtering enhancement. *Multim. Tools Appl.* 80, 30725–30739. doi: 10.1007/s11042-021-10623-7

Zhang, R., and Li, J. T. (2020). A survey algorithm research of scene parsing based on deepLearning. *J. Com. Res. Develop.* 57, 859–875. doi: 10.7544/issn1000-1239.2020.20190513

Zhang, X., Xiao, P., and Feng, X. (2020). Object-specific optimization of hierarchical multiscale segmentations for high-spatial resolution remote sensing images - science direct. *ISPRS J. Photogramm.* 159, 308–321. doi: 10.1016/j.isprsjprs.2019.11.009

Zhang, Y., Yu, L., Fang, Z., Xiong, N. N., Zhang, L., and Tian, H. (2022b). An end-to-end deep learning model for robust smooth filtering identification. *Future Gener. Comp. Sy* 127, 263–275. doi: 10.1016/j.future.2021.09.004

Zhao, X., Guo, J., Zhang, Y., and Wu, Y. R. (2021). Memory-augmented Transformer for remote sensing image semantic segmentation. *Remote. Sens.* 13, 4518. doi: 10.3390/rs13224518

Zhao, Q., Liu, J. H., Li, Y. W., and Zhang, H. (2021). Semantic segmentation with attention mechanism for remote sensing images. *IEEE Trans. Geosci. Remote. Sens.* 60, 1–13. doi: 10.1109/TGRS.2021.3085889

Zhao, D. P., Wang, C. X., Gao, Y., Shi, Z. W., and Xie, F. Y. (2021). Semantic segmentation of remote sensing image based on regional self-attention mechanism. *IEEE Geosci. Remote. Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2021.3071624

Zhao, J. Q., Zhang, D., Shi, B. Y., Zhou, Y., Chen, J. Y., Yao, R., et al. (2022). Multi-source collaborative enhanced for remote sensing images semantic segmentation. *Neurocomputing* 493, 76–90. doi: 10.1016/j.neucom.2022.04.045

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., et al. (2021). "Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers," in *Proceedings of the CVPR 2021*, Nashville, TN, USA. IEEE, 6881–6890. doi: 10.1109/CVPR46437.2021.00681

Zheng, Y. L., Yang, M. Y., Wang, M., Qian, X. J., Yang, R., Zhang, X., et al. (2022). Semi-supervised adversarial semantic segmentation network using Transformer and multiscale convolution for high-resolution remote sensing imagery. *Remote. Sens.* 14, 1786. doi: 10.3390/rs14081786

Zheng, Z., Zhong, Y. F., Wang, J. J., and Ma, A. L. (2020a). "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proceedings of the CVPR 2020*, Washington, Seattle. IEEE, 4096–4105. doi: 10.1109/CVPR42600.2020.00415

Zheng, X. W., Huan, L. X., and Gong, J. Y. (2020b). Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss. *ISPRS J. Photogramm.* 170, 15–28. doi: 10.1016/j.isprsjprs.2020.09.019

Zhong, H. F., Sun, H. M., Han, D. N., Li, Z. H., and Jia, R. S. (2022a). Lake water body extraction of optical remote sensing images based on semantic segmentation. *Appl. Intell.* 52, 1–16. doi: 10.1007/s10489-022-03345-2

Zhong, H. F., Sun, Q., Sun, H. M., and Jia, R. S. (2022b). NT-Net: a semantic segmentation network for extracting lake water bodies from optical remote sensing images based on Transformer. *IEEE Trans. Geosci. Remote. Sens.* 60, 1–13. doi: 10.1109/TGRS.2022.3197402

Zhou, Z. W., Siddiquee, M. M., Tajbakhsh, N., and Liang, J. M. (2018). "Unet++: a nested u-net architecture for medical image segmentation," in *Proceedings of the DLMIA 2018*, Granada, Spain. doi: 10.1007/978-3-030-00889-5_1

Zhu, X. X., Devis, T., Mou, L. C., Xia, G. S., Zhang, L. P., Xu, F., et al. (2017). Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geosc. Rem. Sen. M.* 5, 8–36. doi: 10.1109/MGRS.2017.2762307