Journal of the Indian Society of Remote Sensing (September 2023) 51(9):1865–1878
https://doi.org/10.1007/s12524-022-01496-w

RESEARCH ARTICLE

Intelligent Image Semantic Segmentation: A Review Through Deep Learning Techniques for Remote Sensing Image Analysis

Baode Jiang[1] · Xiaoya An[2] · Shaofen Xu[1] · Zhanlong Chen[1]

Received: 2 November 2021 / Accepted: 2 January 2022 / Published online: 20 January 2022

Abstract

Image semantic segmentation is an important part of fundamental in image interpretation and computer vision. With the development of convolutional neural network technology, deep learning-based image semantic segmentation methods have received more and more attention and research. At present, many excellent semantic segmentation methods have been proposed and applied in the field of remote sensing. In this paper, we summarized the semantic segmentation methods used for remote sensing image, including the traditional remote sensing image semantic segmentation methods and the methods based on deep learning, we emphasize on summarizing the remote sensing image semantic segmentation algorithms based on deep learning and classify them into different categories, and then we introduce the datasets that commonly used and data preparation methods including pre-processing and augmentation techniques. Finally, the challenges and future directions of research in this domain are analyzed and prospected. It is hoped that this study can widen the frontiers of knowledge and provide useful literature for researchers interested in advancing this field of research.

Keywords Deep learning · Image semantic segmentation · Remote sensing image · Computer vision

Introduction

Image semantic segmentation is the process of splitting an image into distinct sections that have comparable properties but do not overlap (Badrinarayanan et al., 2017). It is one of the fundamental issues in computer vision and image processing. Deep learning-based image semantic segmentation technology has advanced significantly with the development of convolutional neural network technology, and is now used in a variety of scenarios that require accurate and efficient semantic segmentation, such as automatic driving, indoor navigation, virtual reality, and augmented reality. Semantic segmentation is a computer vision problem that includes grouping comparable components of an image that belong to the same class together. Several steps are used to perform semantic segmentation, such as localizing and classification, in which classification is the process of classifying a certain object in the image and the object detection and bounding box drawing are processed with the help of localizing. Virtual reality (VR) is a computer-generated environment that blends realistic-looking visuals and objects to give the spectator the feeling of being entirely immersed in their surroundings. It is most commonly used in business, education, and entertainment. Augmented reality (AR) is an exciting experience in a physical situation in which real-world elements are complemented with software sensory input, potentially encompassing several sensations including optical, aural, emotional, sensorimotor, as well as aromatic.

Similarly, in the realm of remote sensing image processing, an increasing number of academics are focusing

✉ Xiaoya An
science_xya2001@hotmail.com

Baode Jiang
pauljiang27@163.com

Shaofen Xu
xvshaofen@163.com

Zhanlong Chen
Chenzhanlong2005@126.com

[1] State Key Laboratory of Geo-Information Engineering, School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, Hubei, China

[2] Xi'an Research Institute of Surveying and Mapping, Xi'an 710054, Shaanxi, China

on image semantic segmentation (Kattenborn et al., 2019; Mohanty et al., 2016; Kussul et al., 2017).

Traditional remote sensing image semantic segmentation methods are usually pixel-based, edge-detection-based, or region-based (Yu et al., 2007). These methods have various problems, such as edge-detection-based methods, are difficult to form closed regions, and region-based methods are difficult to accurately segment edges (Wang et al. 2020a, 2020b, 2020c). A metaheuristic is a high-level algorithmic framework that provides a collection of recommendations or strategies for constructing heuristic optimization algorithms that is independent of the task. There are also mathematical theory-based methods and meta-heuristic algorithm-based methods (Lopez et al., 2019; Wang et al. 2020a, 2020b, 2020c), which may take into account image context and have significant learning and adaptability, but these algorithms are complex to develop, time-consuming, and require a big quantity of label data, making it challenging to produce improved processing outcomes for high-resolution remote sensing images. Therefore, in order to obtain better remote sensing image segmentation results, researchers have tried to combine traditional segmentation methods to overcome the shortcomings of single segmentation methods, and good research results have been achieved. With the continued development of high-resolution remote sensing images and the increasing influence of “the same object with different spectra” and “the different object with the same spectra” in high-resolution remote sensing images, the segmentation accuracy that can be achieved by traditional remote sensing image segmentation methods is poor, and the increasing application demands and increasing remote sensing data have placed higher demands on the segmentation.

Deep learning (Lecun et al., 2015; Gao et al., 2020) is widely used in computer vision, and deep learning-based image segmentation methods have achieved a good application effect, and by increasing the depth of the model, the performance and accuracy of the algorithm can be improved. Deep learning can quickly and automatically extract image features from very large datasets, and using complex models iteratively to improve the accuracy of regression algorithms. Regression methods anticipate expected output related to statistical input files qualities. The normal procedure is for the program to build a replica depending on the characteristics of testing phase or use that modeling to anticipate the impact of new information. Deep learning has recently been used to various elements of remote sensing research, such as plant identification (Kattenborn et al., 2019), plant disease identification (Mohanty et al., 2016), crop-type classification (Krogh Mortensen et al., 2016; Kussul et al., 2017) natural disaster prediction (Ghorbanzadeh et al., 2019), land cover classification (Zhu et al., 2017), and so on. Convolutional neural network (CNN) (He et al., 2016; Krizhevsky et al., 2017; Simonyan & Zisserman, 2014) is the most used deep learning model for image segmentation, which uses a convolutional layer to extract image features, followed by a nonlinear layer for function modeling, a pooling layer to reduce the spatial resolution and training parameters, and finally a fully connected layer to output the classification score of the image. Fully convolutional network (FCN) (Long et al., 2015) outputs the label of each pixel, namely the segmentation result of the image, by replacing the fully connected layer of CNN with a fully convolutional layer, and finally, FCN allows images of arbitrary size as input and achieves high segmentation accuracy on standard datasets (PASCAL VOC) (Everingham et al., 2010). However, remote sensing image segmentation is more challenging than traditional image segmentation, so researchers have also made some improvements to the image segmentation method of FCN. FCN is the abbreviated term of fully connected network. It is a type of neural network, which can perform only subsampling and upsampling in the convolution operation.

This paper reviews and analyzes the semantic segmentation methods used in remote sensing images, and summarizes the commonly used remote sensing image datasets and data processing methods. The rest of the paper is organized as follows: “Deep learning-based RSISS methods” section illustrates literature on deep learning-based semantic segmentation of remote sensing image semantic segmentation (RSISS). In “Remote sensing datasets and data processing” section provides remote sensing image data sets that commonly used and data preparation methods. In “Discussion” section presents discussion of existing problems and future research directions, and in “Conclusion” section concludes the paper.

Deep Learning-Based RSISS Methods

Convolution neural network is the sort of deep neural network; it is used to determine and classify the certain features from the images, also used for visual image analyzing. Further, three types of layers are utilized to generate convolutional neural network, i.e., convolutional layers, fully connected layers (FCN) besides pooling layer. The CNN architecture is formed by stacking these layers. Meanwhile, dropout layer and the activation function are the two important parameter that are used in addition to layers. The major applications of CNN are computer vision, image classification, image and video recognition, natural language processing, and medical image analysis. An input layer, hidden layers, besides an output layer make up the three layers of a convolutional neural network. Since the perceptron and eventual combination conceal their output signals, most generally represented in a feed-forward CNN architecture are considered concealed.

Fully Convolutional Network (FCN)

Long et al. (2015) presented the completely convolutional network (FCN) as a convolutional neural network design that replaces the fully connected layer with a fully convolutional layer, hence integrating a non-fixed size input. The model combines the feature map of the last layer with the feature map of the preceding layer using jump connection and upsampling to produce the spatial segmentation map of the original image pixel-by-pixel.

Because remote sensing images are much complex as compared to natural images, the impact of applying FCN directly to the semantic segmentation of remote sensing images is weak. Fu et al. (2017) employed Atrous convolution to optimize the FCN model and conditional random fields (CRFs) to post-process the segmentation data, resulting in dramatically increased segmentation accuracy when compared to conventional networks. Atrous convolution is one of the types of convolution layer; it is mainly used as an alternative for down sampling layer. Also, it helps to maximize the receptive field whilst in order to maintain the feature map spatial dimension. Chen et al. (2018a, 2018b) employed the overlaid semantic segmentation framework SNFCN and SDFCN approaches to increase algorithm accuracy and remove noise effect. The accuracy and recall rate of RSISS may be enhanced by using image feature information such as infrared images (Zhang & Hu, 2017) and digital surface models (DSM) (Peng et al., 2019; Mangalraj et al., 2019). RSISS is the abbreviated form of "Remote sensing image semantic segmentation". Feature extraction in high-resolution spatial data photos aims to assign conceptual descriptors to every pixel location. With the rapid improvement of spatial data imaging techniques, exceptionally high remotely sensed data pictures with a ground sampling distance (GSD) of 5–10 cm are now attainable. To address the issue of FCN's lack of previous information guidance. Edge-FCN, reported by He et al. (2020), employs edge information gathered through a holistically nested edge detection (HED) network to correct FCN segmentation findings. Edge detection is an image processing technique that detects discontinuities or abrupt changes in image brightness in a digital image. The image's edges are the places where the brightness of the image varies dramatically. The region-based model employs a specific region explanatory approach as a contour guide to identify each area, whereas the edge-based model uses edge information for image segmentation. To overcome the constraints of the canny edge detector, holistically nested edge detection (HED) employs an end-to-end deep neural network. This network accepts an RGB image as input and generates an edge map. Parallel inception design was used by Zhang et al. (2019) and Liu et al. (2020) to simplify the training process and increase the network's operating efficiency. The FCN-based automated segmentation approach for remote sensing images is critical for large-scale land cover mapping to be realized (Han et al., 2020) and the rapid production of farmland maps for agricultural automation (Osco et al., 2021).

Graph-Based Models and Dilated Convolution

Contextual scenario-level semantics, which are important for segmentation, are not taken into account during the pixel-by-pixel segmentation process utilizing FCN. As a result, researchers developed a number of methods for adding probabilistic graph models into deep learning network design [such as conditional random field (CRF) and Markov random field (MRF)]. Conditional random fields (CRFs) are a type of numerical modeling tool that is commonly used for structured prediction in pattern recognition and machine learning. A predicted class a classification for a specific subset without taking into account "neighboring" observations, whereas a CRF can take into account contextual. A graphical description of a joint probability distribution is a Markov random field (MRF), and it is made up of nodes that represent random variables in an undirected graph. The set of random variables associated with the set of nodes is denoted by S. A node n is independent of all other nodes in the network if it has a set of neighbors. Chen et al. (2014) developed a semantic segmentation approach that makes use of CNNs and fully linked CRFs. They noticed that the responses from the last layer of deep CNNs aren't well-localized enough for successful object separation, so they added a fully connected CRF to the final CNN layer and observed that this model can better localize the borders.

In the segmentation of remotely sensed images, segmentation methods that take into account of contextual semantics are particularly important because geographically the objects are closely connected to the surrounding scenes. Fully connected CRF can synthesis spatial information from remote sensing images to produce spatially consistent segmentation results and enhance coarse prediction outcomes, and most researchers have post-processed and optimized the segmentation results with the help of CRF to improve the accuracy of segmentation boundaries. Li et al. (2019) added fully connected CRF to the back end of a deep learning model by defining the potential function and using the computation of the mean approximate field CRF to make the boundaries and details clearer when extracting water bodies. Xia et al. (2021a, 2021b) similarly used CRF in post-processing of water body information extraction by modeling the image with a Gaussian kernel potential function using pixels as nodes, thus reducing segmentation errors for complex water bodies. A Gaussian blur is the result of blurring an image with a Gaussian function in image processing. It is a common effect in graphics software, generally used to minimize visual noise and detail. The method of using CRF to segment and refine the boundary has also been applied to the extraction of architectural footprint map (Li et al., 2020; Zhu et al., 2020) and agricultural and forestry applications (e.g., large-scale oil palm plantation detection (Dong et al., 2020). When CRF is used for post-processing optimization, it is mostly trained alone, while He et al. (2019a, 2019b) enables end-to-end network training by combining a jump-connected coding-decoding network architecture with CRF, thus allowing the architecture of CRF to guide the training of CRF to take into account of more information for improving the segmentation results. Pan et al. (2020a, 2020b) have presented an end-to-end, localized post-processing (ELP) technique by limiting the CRF's processing range and determining the iteration termination condition, thus avoiding over-correction due to the global processing of the CRF, which can effectively correct the classification results and improve the classification accuracy compared with the traditional methods.

DeepLabv1 (Chen et al., 2014) and DeepLabv2 (Chen et al. 2018a, 2018b) is a very effective image segmentation method. It performs image segmentation by using dilated convolution (a.k.a. "atrous" convolution) and combining CRF, in which atrous convolution refers to filling the dilated convolution kernel with 0 according to a certain expansion rate, so as to expand the receptive field under the condition of using a few parameters, thus achieving an expanded perceptual field with the use of a small number of parameters, and thus obtaining more contextual semantic information. The Atrous-convolution space pyramid pooling (ASPP) proposed by DeepLabv2 can capture the context of objects and images at multiple scales to better segment objects, and combine CRF with CNN to better locate the segmentation boundaries. Chen also proposed DeepLabv3 (Chen et al., 2017) and DeepLabv3+ (Chen et al. 2018a, 2018b) in 2017 and 2018, respectively, the former adding a 1*1 convolutional layer to ASPP and using batch normalization, and the latter using an encoding–decoding architecture based on DeepLabv3. The references/pointers to services and configuration information used/needed by other objects are encapsulated in a context object. It permits the items in a context to view the world outside of it. Objects that live in a different environment have a distinct perspective on the world. In a number of contexts, such as road extraction (He et al., 2019a, 2019b) (increasing the performance of the road extraction network by incorporating ASPP), DeepLab-based semantic segmentation algorithms have been employed for semantic segmentation of remote sensing images and coding-decoding networks), automatic vegetation extraction for multi-context and multi-scale land cover (Zhan et al., 2020), and change detection of multi-temporal hyperspectral images (Venugopal, 2020).

Encoder–Decoder Architecture Network

Another prevalent deep learning model for image semantic segmentation is the convolutional neural system, which is dependent on the encoder–decoder architecture. An Encoder–Decoder architecture was created in which a complete input sequence was read and encoded to a fixed-length internal representation. The internal representation was then employed by a decoder network to produce words until reach the end of the sequence token. Most deep learning-based semantic segmentation techniques employ encoder–decoder architecture. Two well-known encoder–decoder networks are SegNet (Badrinarayanan et al., 2017) and U-Net (Badrinarayanan et al., 2017; Ronneberger & colleagues, 2015). SegNet's network design conducts nonlinear upsampling in the decoder stage using the combined index established during the maximum pooling phase of the corresponding encoder stage, minimizing the number of parameters required in the training process. U-Net was created to help with image segmentation in biological microscopy. It consists of two sections: a contracting path for gathering context and a symmetric expanding path for identifying precise position.

In RSISS, the encoder–decoder network architecture is commonly utilized. Li et al. (2018b) presented DeepUNet, which uses DownBlocks instead of convolution layers in the contracting path and UpBlocks in the expanding path, based on the U-Net network, UNet is a convolutional neural network architecture derived from the CNN design with some modifications. It was created to deal with biological images in which the goal is to not only categorize whether or not there is an infection, but also to determine the region of illness. DeepUNet, on the other hand, is a deep fully convolutional network for pixel-level sea-land segmentation and applied it in the sea-land segmentation of high-resolution remote sensing images, while Bona et al. used Landsat-8 images to conduct image segmentation on a coastal area with significant water turbidity, utilizing a U-Net architecture with ResNet connectivity (Bona et al., 2019). A down block is one that has an angle toward the interior. The down blocker generally targets one of two landmarks, which are usually selected based on the scouting report or the defender's technique on the previous five scrimmage downs. Cui et al. (2019) presented MSRIN, which employs FCN and UNet networks to segment the same image concurrently on feature images of different sizes to construct a multi-scale hierarchy, and then uses LSTM algorithms to analyze the image, which can achieve both semantic segmentation and end-to-end spatial relationship recognition of remote sensing objects. Cheng et al. (2020) proposed a hybrid convolution U-Net (HDCUNet) is a semantic segmentation network that combines U-Net with hybrid dilated convolution (HDC) to further increase the receptive field while avoiding gridding., and the method has achieved certain success in the problem of how to quickly and accurately extract coastal aquaculture areas. The encoder–decoder network based RSISS has also been applied to precision agriculture (Zhao et al., 2018), urban landscape extraction with small data sets (Song & Kim, 2020), water body information extraction (Xia et al., 2021a), and other application areas. Yang et al. (2019) used SegNet networks for monitoring of farmland plastic mulch, and Song et al. (2020) improved the SegNet network by adopting skip connection, separable convolution, and conditional random fields to achieve rapid detection of sunflower lodging, which helps to cope with extreme and destructive weather events. Weng et al. (2020) proposed a separable residual SegNet (SR-SegNet) for water segmentation of remotely sensed images, and experiments showed that the segmentation effect of the method was significantly improved compared with networks such as FCN and conventional SegNet.

Feature Pyramid Network

The feature pyramid network (FPN) proposed by Lin et al. (2017a, 2017b) is one of the most famous models in multiscale neural networks which is mainly for target detection and later also for image segmentation. Deep CNNs' intrinsic multi-scale, pyramidal structure was leveraged to build feature pyramids for a minimal extra cost. The FPN is made up of a bottom-up pathway, a top-down pathway, and lateral connections to combine low- and high-resolution information. Except for FPN, there are various network architectures to achieve better segmentation results by merging multi-scale feature maps, such as ASPP in DeepLabv2 mentioned above, and pyramid scene parsing network (PSPN) proposed by Zhao et al. (2017), Lin et al. (2017a, 2017b) proposed RefineNet, a multi-path optimization network, and so on.

There are many small target objects in remote sensing images, which only retain feature information in the high-resolution feature map, and feature information will be lost after the downsampling operation, resulting in large errors in image segmentation results for small target object extraction, while the multiscale network architecture enables the deep learning process to consider both the detailed information contained in high-resolution images and the global information contained in low-resolution images. Liu et al. (2019) proposed a new pyramidal loss-enhanced fully convolutional network (PLFCN) that explores multi-scale spatial context information by introducing deep pyramidal supervision to improve semantic segmentation performance while combining the advantages of multi-scale architecture and auxiliary loss to maintain efficiency. Shang et al. (2020) proposed an end-to-end multiscale context extraction module (MCM) that uses 2 layers of atrous convolution with dissimilar expansion rates as well as global average pooling to remove contextual data at multiple scales in equivalent in a multiscale adaptive feature fusion network (MANet) for segmenting high-resolution remote sensing images.

Deep learning models with multiscale architectures are extensively used in semantic segmentation of remote sensing images. Li et al. (2019) enhanced the DeepLab algorithm in water extraction by assigning various weights to the output features at each scale and managing the effect of each scale feature on the water extraction outcomes using a multi-scale feature perception technique. Wang et al. (2020a, 2020b, 2020c) proposed a multi-scale lake water extraction network (MSLWNET), which obtained multi-scale information through different expansion rates and well extracted the water bodies of small lakes. Guo et al. (2020) proposed a multi-scale water extraction convolutional neural network (MWEN) that obtains multiscale information by combining expansion convolution with different expansion rates and automating the extraction of various water bodies of various sizes from GF-1 remote sensing images. In terms of sea-land segmentation, Pan et al. (2018) proposed a MIFNET, a CNN-based multiinformation fusion network that took into account multiscale edge and multi-scale segmentation information, as well as global context information, demonstrated advanced performance in sea-land segmentation of Google maps natural imagery. Cui et al. (2021) proposed a SANet is a scale-adaptive semantic segmentation network (SANet) that replaces serial convolution with an adaptive multiscale feature learning module (AML). SANet can adaptively fuse feature maps of different scales while achieving multi-scale detail information and contextual semantic capture, and the segmentation results of SANet for various natural and artificial coastlines are more accurate and c Building extraction is one of the applications of the deep RSISS model, which is based on multi-scale feature fusion. Our feature-based segmentation approach is essentially a clustering procedure that may consider a variety of variables such as color, motion, disparities, object location, and gradient (Zhu et al., 2020), land cover segmentation (Wang et al. 2020a, 2020b, 2020c), and satellite image cloud and cloud shadow segmentation (Xia et al., 2021a, 2021b).

Attention-Based Mechanism Network

Chen et al. (2016) proposed a weighted multiscale feature extraction method that teaches itself to assign weight values to each pixel. The attention strategy improves average and maximum pooling while also allowing the model to assess the importance of items in various placements and sizes. Huang et al. (2017) suggested a semantic segmentation strategy based on reverse attention. Li et al. (2018a) proposed a pyramidal attention network for semantic segmentation that takes advantage of the influence of global appropriate semantic data in semantic separation by combining attention mechanisms as well as spatial pyramids towards excerpt accurate compressed features for pixel labeling rather than using complex convolution and manually designed decoder networks.

Remote sensing images, in contrast to ordinary natural photos, are large and complex, posing significant issues including spatial target distribution diversity besides spectral information removal. Semantic segmentation methods based on attention mechanism can help RSISS to achieve a balance between feature representation capability and spatial localization accuracy. A number of researchers have already adopted attention mechanism when performing RSISS. Ni et al. (2019) introduced the attention mechanism into DeepLab v3+, they used attention information to extract image semantic information and richer image features to achieve finer segmentation of remote sensing image target regions. Chen et al. (2020) developed a channel attention module that collects multidimensional global context and improves class-specific feature representation, as well as a decoding stage that captures multi-scale spatial information with a lightweight global feature attention module. Shang et al. (2020) also introduced channel attention mechanism in MANet to fuse multi-scale contextual semantic features to generate global features and collect adaptive weight information for each channel, utilize global characteristics as channel weights, so as to achieve effective remote sensing image semantic fusion. Dong et al. (2020) also used the channel attention mechanism to improve the segmentation effect when surveying and mapping oil palm plantations with remote sensing images. Xu et al. (2020) proposed various network modules for fusing attention mechanisms, and combined with the attention module, proposed a new segmentation framework for high-resolution remote sensing images, the heavy-weight spatial feature fusion pyramid network (FFPNet), covers a wide range of target geometries in large scale remote sensing images. based on regional attention, with a regional pyramidal attention mechanism By incorporating the dual attention mechanism into densely linked convolutional networks, Hu et al. (2020) constructed a densely connected global entropy network (DGEN) for semantic segmentation of remote sensing images (DenseNets). Methods for remote sensing image processing based on the attention mechanism have also been applied to several practical application areas, such as land cover segmentation (Wang et al. 2020a, 2020b, 2020c), water body information extraction (Xia et al., 2021a), cloud segmentation (Xia et al., 2021b), and tropical forest monitoring (Yu et al., 2021), etc.

GAN-Based Network

GANs are a new type of deep learning model that essentially consists of a generator and a discriminator. GANs (Generative adversarial networks) are computational frameworks that pit two neural networks against each other to produce new, synthetic data instances that seem real. They're frequently utilized in the creation of images and videos, as well as in the creation of audio. Luc et al. (2016) proposed a semantic segmentation adversarial training method in which they used another adversarial network to discriminate true segmentation labels from the segmentation network's segmentation results after training a convolutional semantic segmentation network as a generative network. On the PASCAL VOC 2012 dataset, the approach performed well in terms of segmentation accuracy. Pascal VOC is a dataset collection for object detection. For benchmarking, the most frequent combination is to use 2007 trainval and 2012 trainval for training and 2007 test for validation. For semi-supervised semantic segmentation, Hung et al. (2018) proposed an adversarial network-based paradigm. The logistic regression of this may include cross-entropy results in the destruction on segmented labeling, aggressive reduction of the fully convolutional system, semi-supervised reduction depending on the confident mapping, as well as the discriminator's outputs.

A number of researchers in the field of remote sensing image processing have also tried to implement RSISS using the network architecture of GAN. As the training image data of remote sensing images is limited, GAN has insufficient confrontation information to explain the problem of the inverse process of segmentation, and there is also a lack of proper objective loss function to overwhelm the vanishing gradient issue. Zhang and Hu (2017) proposed conditional least squares generative adversarial network (CLS-GAN) for semantic segmentation, using a special f-divergence class as the optimal objective function, the network achieved high accuracy segmentation results in a limited number of high-resolution remote sensing images. In remote sensing images, the same target may be observed differently at different times, and for such dynamic object extraction, Kniaz proposed a semi-supervised GAN model in 2018 (Kniaz, 2018), using the Pix2Pix model as the starting point of the study. Later Kniaz (2019) used GeoGAN network for water body extraction, which is able to densely label water bodies in different seasons. Bona et al. (2019) also used a GAN model to refine the segmentation results when performing high turbidity seawater extraction. Xiong et al. (2020) proposed an end-to-end Bayesian RSISS network based on GANs, which is more stable than previous GAN-based networks, by using FCNs and GANs to realize the likelihood derivation of pre probability and posterior probability in Bayesian theory.

R-CNN Based Network

Regional convolutional networks (R-CNN) (Girshick et al., 2014) and its extensions fast R-CNN (Girshick, 2015), faster R-CNN (Ren et al., 2017) are well-known target detection networks that have been widely used in the problem of instance segmentation, that is, performing semantic segmentation while performing target detection. The mask R-CNN network proposed by He et al. (2020) is a well-known instance segmentation network that has achieved excellent results in many computer vision challenges. The model based on faster R-CNN can effectively detect objects in images and perform high-quality segmentation of each extracted instance object using regression branching.

One of the significant advantages of instance segmentation is that it can distinguish the difference between different individuals of the same class and achieve the extraction of individuals. Wu et al. (2020a, 2020b) applied instance segmentation to orchard crop data acquisition. They used faster R-CNN to detect each apple tree and then segmented each tree with U-Net so that apple trees could be detected and counted. Mask R-CNN was used by Zhang et al. (2018) to recover Arctic ice wedge polygons from high-resolution remote sensing images, with a classification accuracy of 79%. Wu et al. (2020a, 2020b) used the mask R-CNN network to extract clouds from remote sensing images and enhanced it by integrating population training and boundary optimization. Zhang and Chi (2020) proposed a mask R-FCN network that uses the R-CNN network as a complementary network for the FCN network. It helps the FCN network to strike a balance between background semantics and edge details, and enable small target objects to be extracted accurately. Soloy et al. (2020) proposed an instance segmentation network based on mask R-CNN that can be used to measure the size of coarse sediment debris on the surface, thus allowing monitor the spatial variability of particle size before and after storms. Song et al. (2020) proposed an adaptive mask R-CNN network for the extraction of surface water bodies.

Other Methods

Several strategies have been used to semantically segment remote sensing images in addition to the primary kinds of semantic segmentation methods listed above. For better segmentation, Guo et al. (2019) suggested a learnable gated network (L-GCNN) for global and local contextual spatial connection analysis in remote sensing images. The learnable gated-deep convolutional neural network (L-GCNN) was created to solve issues encountered by a range of artificial objects with significant visual appearance besides size changes via multiscale information fusion. Panboonyuen et al. (2019) used migration learning to separate remote sensing images into semantic categories. Sun et al. (2020) created a boundary-aware semi-supervised semantic segmentation network (BAS4Net) that improves segmentation accuracy while reducing annotation time. Most deep learning-based remote sensing image processing methods use a variety of deep learning model architectures to maximize the benefits of several models and improve semantic segmentation accuracy.

Remote Sensing Datasets and Data Processing

Remote Sensing Image Datasets

Most of the research papers will use publicly available remote sensing image datasets when verifying the validity of the model, and some of them will also use home-made remote sensing image datasets for experiments. Figure 1 provides a statistical analysis of the datasets used by the referenced papers on semantic segmentation algorithms for remote sensing images.

By analyzing the researches of remote sensing semantic segmentation, it is found that the most commonly used datasets for semantic segmentation of remote sensing images are the Vaihingen and Potsdam datasets of ISPRS,

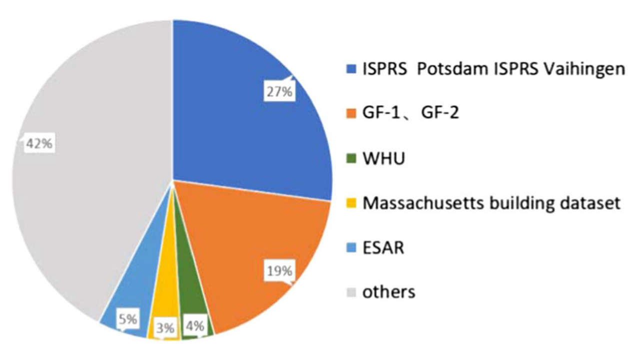

Fig. 1 The proportion of remote sensing image data sets used in papers

and among the references of remote sensing semantic segmentation algorithms, almost 30% of them use these two datasets, followed by the Gaofen-1 and Gaofen-2 satellite images, with nearly 20% of the literature used such datasets, and some other datasets such as the German ESAR and GID (Tong et al., 2020), the WHU (Ji et al., 2019) dataset for building extraction, and the Massachusetts roads and buildings dataset (Mnih, 2013) for road and building extraction were also frequently used. For the extraction of natural objects, such as water bodies, vegetation, crops, ice wedges, shorelines, and surface water resources, nearly half of the research teams have adopted home-grown datasets, some using Sentinel (Xia et al., 2021a, 2021b), Worldview (Song et al., 2020), QuickBird (Zhang & Chi, 2020), Landsat series (Panboonyuen et al., 2019; Wu et al., 2020a, 2020b), RADARSAT (Venugopal, 2020) and other satellite sensor images to produce datasets, some are downloaded directly from map websites such as Google Maps satellite images (Pan et al., 2018; Zhang & Chi, 2020), data published by satellite data and resource application research centers in various countries (Xia et al., 2021a, 2021b), etc., and others use unmanned aerial photography to obtain remote sensing images (Chen et al., 2020; Kniaz, 2019; Li et al. 2018b; Song et al., 2020; Zhao et al., 2018) and perform real-time kinematic (RTK) measurements (Huang et al., 2018) to obtain tagging data. Real-time kinematic (RTK) is the technique, which works for the carrier-based ranging to produce ranges the orders of magnitude more exact than code-based positioning, but the RTK methods are difficult for master. It is mainly used to minimize and remove the mistakes that happened between a rover pair and base station. It is utilized in the applications for demanding high accuracies.

Table 1 lists the datasets often used for semantic segmentation of remote sensing images in order to enable future research on semantic segmentation of remote sensing images more accessible.

Data Preparation Method

Data preparation methods include image preprocessing, label preparation, and image amplification. Image preprocessing changes the features of the image at the pixel level or spectral level, and these methods include atmospheric correction (Bona et al, 2019; Peng et al., 2019; Yu et al., 2021), radiometric correction (Peng et al., 2019; Yu et al., 2021) geometric correction (Yang et al., 2019), and image cropping and stitching (He et al., 2020), and some researchers use histogram specification algorithms to correct for visible exposure problems (Song et al., 2020). Histogram specification is a broader variant of histogram equalization and a common image processing technique. At all brightness levels, an equalized image has the same amount of pixels, resulting in a straight horizontal line on the histogram graph. When histogram is delivered to images, the histogram intended is defined. Meantime, the nonlinear stretch operation is used to cause the image histogram to take that shape. It is mainly used for condensing an image's dynamic range and removing pixel values with minimal information to make image easily displayed in the monitor. Histogram equalization is a computer image processing technique, which is mainly used to maximize the contrast of the image. The contrast of the image is improved to spread the common intensity values effectively. This strategy commonly improves the global contrast of images when the important data is represented by near contrast values. This allows regions with low local contrast to benefit from a contrast increase. Since remote sensing images from high altitude are often obscured by clouds, some cases require land masking and

Table 1 Commonly used RSISS data source

Datasets	Year	Scene classes	Total image	Image sizes	Spatial resolution
ISPRS Vaihingen	–	6	33	2494 × 2064	9 cm
ISPRS Potsdam	–	6	38	6000 × 6000	5 cm
GID (Tong et al., 2020)	2014	5	150	6800 × 7200	1 m/4 m
WHU-RS19 (Ji et al., 2019)	2012	19	950	600 × 600	> 0.5 m
Massachusetts building dataset (Mnih, 2013)	2013	2	151	1500 × 1500	0.5 m
UC Merced LandUse (Zou et al., 2015)	2010	21	2100	256 × 256	0.3 m
SIRI-WHU (Zhu et al., 2015)	2016	19	3800	600 × 600	0.5 m
RSSCN7 (Cheng et al., 2014)	2015	7	2800	400 × 400	–
RSC11 (Zhao et al., 2016)	2016	11	1100	512 × 512	0.2 m
EuroSAT	2017	10	27,000	64 × 64	–
PatternN (Zhou et al., 2018)	2017	38	30,400	256 × 256	0.062–4.6 m
BigEarthNet (Sumbul et al., 2019)	2019	–	590,326	Non-fixed	–

cloud masking processes (Bona et al., 2019). In general, papers that use standard datasets such as ISPRS do not require further image preprocessing because these datasets are already prepared with standard image and labeled data and provide methods on how to use datasets for training, testing, and validation. ISPRS datasets also provide matching DSM data, etc. To assist in the analysis, in special circumstances, some papers adjust and reclassify the ISPRS dataset (Song & Kim, 2020). Some datasets require manual labeling (Huang et al., 2018), some use software generation (Song et al., 2020), and some use traditional image segmentation methods to create labels (Bona et al., 2019), more labeled data help to improve the accuracy of image segmentation. In order to provide an auxiliary or complementary analysis for remote sensing image data, some papers also calculate DEM (Liu et al., 2020), DSM and NDSM (Guo et al., 2019; Liu et al., 2019; Osco et al., 2021), and ground metric data such as NDWI (Song et al., 2020).

Most papers perform data augment before conducting algorithm experiments because remote sensing data labels are expensive to obtain, the number of experimental data that can be obtained is limited, and in comparison to natural images, remote sensing images are more complicated. Common data augment means include cropping, flipping, rotating, adding noise, upsampling, and interpolating. Because of the limitations of computer GPU computational speed, the big high-resolution remote sensing images are required to reduce their size to 256*256 or 512*512 pixels, and the methods of image reduction include downsampling and image cropping, etc. The downsampling method often leads to the loss of detail information, so image cropping is generally used when reducing the image size. Image cropping is divided into cropping with overlap (30–50% overlap) (Hu et al., 2020; Kniaz, 2018; Li et al., 2018b; Liu et al., 2019, 2020; Xia et al., 2021a, 2021b) and cropping without overlap (He et al. 2019a, 2019b; Xiong et al., 2020). Some use sliding window (Hu et al., 2020; Li et al., 2020; Zhang & Chi, 2020) method to crop patches from the image and corresponding labels, and some articles use random extraction method (Huang et al., 2018) to extract fixed size from the original images. Because the remote sensing image itself has the feature of multi-directionality, image flipping and rotation are also commonly used in order to enhance the training diversity (Panboonyuen et al., 2019; Abdollahi et al., 2020; Liu et al., 2020; Xia et al., 2021a, 2021b). Amplifying the data set by adding noise (Liu et al., 2019; Song et al., 2020) can improve the robustness of the algorithm. The up–downsampling of the image (Cui et al., 2019; Venugopal, 2020) can obtain multi-scale remote sensing image data with different resolutions, which is also conducive to improving the segmentation accuracy of the algorithm; most of the papers use a mixture of data augment methods to achieve data augmentation. In data analysis, the data augmentation technique is used to improve the quantity of the data. Meantime, the data quantity is improved by adding the copies of current data or by creating new synthetic data from existing data. The data augmentation technique function is used as a regularizer to minimize the overfitting in machine learning model training.

Discussion

Challenges

Deep learning-based semantic segmentation methods for remote sensing images have significantly improved segmentation effects compared with traditional methods. They solve the problem of accurately locating object boundaries that most traditional pixel-level segmentation methods completely ignored, and they are also robust against salt-and-pepper noise (Pan et al., 2020a, 2020b). However, deep learning-based methods still have problems and drawbacks, as listed below.

1. Any deep learning-based semantic segmentation method requires a large amount of training data, but the cost of remote sensing data collection is high, so there will be a lack of training data for the model training. Usually, researchers augment the dataset by performing data augmentation. ISPRS-labeled dataset tries to solve the problem of data shortage by collecting images with higher accuracy such as 5 cm resolution. Migration learning also solves the problem of data shortage to some extent by training the model on public dataset and migrating it to semantic segmentation of remotely sensed images.
2. Deep learning method also needs a large amount of label data. Usually, the labels of remote sensing images need to be manually labeled, and the public dataset will provide part of the labels, but if you want to make your own dataset, it is very difficult to obtain the labels. Some papers try to generate labels by traditional methods, and some choose to obtain label data from moving maps, but the accuracy of such label data is difficult to guarantee.
3. High-resolution remote sensing images have more complex spatial structure compared to natural images, high spectral heterogeneity, and complex situations such as image occlusion and artifacts, which require higher segmentation performance of the algorithm. The current processing method is to mix multiple network architectures to improve the algorithm performance, but the complexity of the algorithm will

increase, and the cost for model training will be larger, so it is also difficult to achieve RSISS in real time.

4. Deep learning algorithms require high performance of computer GPU in terms of computation and storage, especially in training complex algorithms. Currently, there are some cloud computing services such as Google Colaboratory providing free usage time, and there are also some researchers attempt to propose lightweight networks to reduce the training parameters to improve the training efficiency.

Future Directions

Even though deep learning has shown promising results in the field of semantic segmentation of remote sensing images, there is still room for development. The following are the main research directions for the future: (1) Because deep learning model training still needs a large amount of data and collecting individual datasets is difficult, new public datasets equivalent to ISPRS should be made available. Currently, public remote sensing datasets specialized in buildings, roads, and cities are abundant, while public remote sensing datasets for various natural research objects, such as hydrology, glaciers, crops, etc., are still very few. (2) More auxiliary data such as ground indicators (NDVI, NWVI), ground surface models (DEM, DSM), edge information, spectral features, etc., will be further integrated into the deep learning algorithm to alleviate the problem of lacking image labels and improve the accuracy of the algorithm, and the use of integrating multiple data sources of remote sensing image is also conducive to the algorithm to obtain more semantic information, so as to achieve a good segmentation effect. (3) The efficient learning of small samples and the optimization of network architecture are still the main research directions at present. How to generate a good network from the training data of small samples, and how to solve the problems of network overfitting and network layering are problems that need to be solved urgently. Lightweight networks, migration learning, GAN, and other network architectures are still hot research topics in the future.

Conclusion

Semantic segmentation is crucial in remote sensing image analysis. Deep learning-based semantic segmentation approaches for remote sensing images were discovered to outperform standard methods and provide better performance, and they have been successfully applied to urban planning, crop classification, forest and water extraction, coastline segmentation, cloud extraction, and other application fields. Deep learning-based segmentation algorithms are data-driven, as opposed to standard model-driven segmentation algorithms, which opens up new potential and problems for remote sensing image segmentation.

In this paper, we looked at deep learning-based semantic segmentation approaches for remote sensing pictures. Our response to the survey is as follows: First, we presented our findings in semantic segmentation of remote sensing photographs using both traditional model-driven techniques and more current deep learning-based approaches. Second, we showed datasets that deep learning-based algorithms and data preparation methods frequently use. Third, we discussed current difficulties and future advancements in deep learning-based semantic segmentation. This review study is expected to widen knowledge limits and give important material for researchers interested in continuing their research on this issue.

Acknowledgements This work was partially supported by China Scholarship Council, State Key Laboratory of Geo-Information Engineering (No. SKLGIE2019-Z-4-1), Key Laboratory of Geological Survey and Evaluation of Ministry of Education (No. GLAB2020ZR11), the Fundamental Research Funds for the Central Universities, and National Natural Science Foundation of China (41871305).

Authors' Contributions BJ contributed to methodology, project administration, and manuscript editing; XA contributed to software and validation; SX performed visualization, and manuscript review and editing; ZC helped with design framework, resources, and validation.

Funding No funding support was received.

Availability of Data and Material Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest The authors report no conflict of interest.

References

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder–decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Bona, D. S., Murni, A., & Mursanto, P. (2019). Semantic segmentation and segmentation refinement using machine learning case study: Water turbidity segmentation. In *Proceedings of the 2019 IEEE international conference on aerospace electronics and remote sensing technology, ICARES 2019*. https://doi.org/10.1109/ICARES.2019.8914551.

Chen, G., Zhang, X., Wang, Q., Dai, F., Gong, Y., & Zhu, K. (2018a). Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11*(5), 1633–1644. https://doi.org/10.1109/JSTARS.2018.2810320

Chen, J., Chen, G., Wang, L., Fang, B., Zhou, P., & Zhu, M. (2020). Coastal land cover classification of high-resolution remote sensing images using attention-driven context encoding network. *Sensors (switzerland), 20*(24), 1–22. https://doi.org/10.3390/s20247032

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected CRFs. *Complex Variables, Theory and Application: An International Journal, 7*(4), 357–361. https://doi.org/10.1080/17476938708814211

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018b). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv. Available at: http://arxiv.org/abs/1706.05587.

Chen, L. C., Yang, Y., Wang, J., Xu, W., & Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition, 2016-Decem* (pp. 3640–3649). https://doi.org/10.1109/CVPR.2016.396.

Cheng, B., Liang, C., Liu, X., Liu, Y., Ma, X., & Wang, G. (2020). Research on a novel extraction method using deep learning based on GF-2 images for aquaculture areas. *International Journal of Remote Sensing, 41*(9), 3575–3591. https://doi.org/10.1080/01431161.2019.1706009

Cheng, G., Han, J., Zhou, P., & Guo, L. (2014). Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing, 98*, 119–132. https://doi.org/10.1016/j.isprsjprs.2014.10.002

Cui, B., Jing, W., Huang, L., Li, Z., & Lu, Y. (2021). SANet: A sea-land segmentation network via adaptive multiscale feature learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14*, 116–126. https://doi.org/10.1109/JSTARS.2020.3040176

Cui, W., Wang, F., He, X., Zhang, D., Xu, X., Yao, M., Wang, Z., & Huang, J. (2019). Multi-scale semantic segmentation and spatial relationship recognition of remote sensing images based on an attention model. *Remote Sensing, 11*(9), 1044. https://doi.org/10.3390/rs11091044

Dong, R., Li, W., Fu, H., Gan, L., Yu, L., Zheng, J., & Xia, M. (2020). Oil palm plantation mapping from high-resolution remote sensing images using deep learning. *International Journal of Remote Sensing, 41*(5), 2022–2046. https://doi.org/10.1080/01431161.2019.1681604

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision, 88*(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

Fu, G., Liu, C., Zhou, R., Sun, T., & Zhang, Q. (2017). Classification for high resolution remote sensing imagery using a fully convolutional network. *Remote Sensing, 9*(5), 1–21. https://doi.org/10.3390/rs9050498

Gao, J., Wang, H., & Shen, H. (2020). Task failure prediction in cloud data centers using deep learning. *IEEE Transactions on Services Computing*. https://doi.org/10.1109/tsc.2020.2993728

Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S. R., Tiede, D., & Aryal, J. (2019). Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sensing, 11*(2), 196. https://doi.org/10.3390/rs11020196

Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision, 2015 Inter* (pp. 1440–1448). https://doi.org/10.1109/ICCV.2015.169.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). 'Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 580–587). https://doi.org/10.1109/CVPR.2014.81.

Guo, H., He, G., Jiang, W., Yin, R., Yan, L., & Leng, W. (2020). A multi-scale water extraction convolutional neural network (MWEN) method for GaoFen-1 remote sensing images. *ISPRS International Journal of Geo-Information, 9*(4), 1–20. https://doi.org/10.3390/ijgi90401899

Guo, S., Jin, Q., Wang, H., Wang, X., Wang, Y., & Xiang, S. (2019). Learnable gated convolutional neural network for semantic segmentation in remote-sensing images. *Remote Sensing, 11*(16), 1–22. https://doi.org/10.3390/rs11161922

Han, Z., Dian, Y., Xia, H., Zhou, J., Jian, Y., Yao, C., Wang, X., & Li, Y. (2020). Comparing fully deep convolutional neural networks for land cover classification with high-spatial-resolution gaofen-2 images. *ISPRS International Journal of Geo-Information, 9*(8), 478. https://doi.org/10.3390/ijgi9080478

He, C., Fang, P., Zhang, Z., Xiong, D., & Liao, M. (2019a). An end-to-end conditional random fields and skip-connected generative adversarial segmentation network for remote sensing images. *Remote Sensing, 11*(13), 1–22. https://doi.org/10.3390/rs11131604

He, C., Li, S., Xiong, D., Fang, P., & Liao, M. (2020). Remote sensing image semantic segmentation based on edge information guidance. *Remote Sensing, 12*(9), 1–20. https://doi.org/10.3390/RS12091501

He, H., Yang, D., Wang, S., Wang, S., & Li, Y. (2019b). Road extraction by using atrous spatial pyramid pooling integrated encoder–decoder network and structural similarity loss. *Remote Sensing, 11*(9), 1–16. https://doi.org/10.3390/rs11091015

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90.

Hu, H., Li, Z., Li, L., Yang, H., & Zhu, H. (2020). Classification of very high-resolution remote sensing imagery using a fully convolutional network with global and local context information enhancements. *IEEE Access, 8*, 14606–14619. https://doi.org/10.1109/ACCESS.2020.2964760

Huang, Q., Xia, C., Wu, C., Li, S., Wang, Y., Song, Y., & Kuo, C. C. J. (2017). Semantic segmentation with reverse attention. In *British machine vision conference 2017*.

Hung, W. C., Tsai, Y. H., Liou, Y. T., Lin, Y. Y., & Yang, M. H. (2018). Adversarial learning for semi-supervised semantic segmentation. https://arxiv.org/1802.07934v2.

Ji, S., Wei, S., & Lu, M. (2019). Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing, 57*(1), 574–586. https://doi.org/10.1109/TGRS.2018.2858817

Kattenborn, T., Eichel, J., & Fassnacht, F. E. (2019). Convolutional neural networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. *Scientific Reports, 9*(1), 1–9. https://doi.org/10.1038/s41598-019-53797-9

Kniaz, V. V. (2018). Conditional GANs for semantic segmentation of multispectral satellite images. In L. Bruzzone, F. Bovolo, & J.

A. Benediktsson (Eds.), *Image and signal processing for remote sensing XXIV* (p. 28). SPIE. https://doi.org/10.1117/12.2325601
Kniaz, V. V. (2019). Deep learning for dense labeling of hydrographic regions in very high resolution imagery. In L. Bruzzone, F. Bovolo, & J. A. Benediktsson (Eds.), *Image and signal processing for remote sensing XXV* (p. 63). SPIE. https://doi.org/10.1117/12.2533161
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84–90. https://doi.org/10.1145/3065386
Krogh Mortensen, A., Dyrmann, M., Karstoft, H., Jørgensen, R. N., & Gislum, R. (2016). Semantic segmentation of mixed crops using deep convolutional neural network. In *CIGR-AgEng conference* (pp. 1–6). Available at: www.elementar.de.
Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters, 14*(5), 778–782. https://doi.org/10.1109/LGRS.2017.2681128
Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 34*, 436–444. https://doi.org/10.1038/nature14539
Li, H., Xiong, P., An, J., & Wang, L. (2018a). Pyramid attention network for semantic segmentation. https://arxiv.org/1805.10180v3.
Li, R., Liu, W., Yang, L., Sun, S., Hu, W., Zhang, F., & Li, W. (2018b). DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11*(11), 3954–3962. https://doi.org/10.1109/JSTARS.2018.2833382
Li, Q., Shi, Y., Huang, X., & Zhu, X. X. (2020). Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF). *IEEE Transactions on Geoscience and Remote Sensing, 58*(11), 7502–7519. https://doi.org/10.1109/TGRS.2020.2973720
Li, Z., Wang, R., Zhang, W., Hu, F., & Meng, L. (2019). Multiscale features supported DeepLabV3+ optimization scheme for accurate water semantic segmentation. *IEEE Access, 7*, 155787–155804. https://doi.org/10.1109/ACCESS.2019.2949635
Lin, G., Milan, A., Shen, C., & Reid, I. (2017a). Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1925–1934). https://doi.org/10.1109/CVPR.2017.549.
Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017b) Feature pyramid networks for object detection. In *IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE Computer Society.
Liu, S., Ding, W., Li, H., & Liu, C. (2019) PLFCN: Pyramid loss reinforced fully convolutional network. In *ACM international conference proceeding series*. https://doi.org/10.1145/3349801.3349819.
Liu, W., Zhang, Y., Fan, H., Zou, Y., & Cui, Z. (2020). A new multichannel deep convolutional neural network for semantic segmentation of remote sensing image. *IEEE Access, 8*, 131814–131825. https://doi.org/10.1109/ACCESS.2020.3009976
Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3431–3440). IEEE. https://doi.org/10.1109/CVPR.2015.7298965.
Lopez, J., Santos, S., Atzberger, C., & Torres, D. (2019). Convolutional neural networks for semantic segmentation of multispectral remote sensing images. In *Proceedings—2018 10th IEEE Latin–American conference on communications, LATINCOM 2018* (pp. 1–5). https://doi.org/10.1109/LATINCOM.2018.8613216.
Luc, P., Couprie, C., Chintala, S., & Verbeek, J. (2016). Semantic segmentation using adversarial networks. Available at: http://arxiv.org/abs/1611.08408.
Mangalraj, P., Sivakumar, V., Karthick, S., Haribaabu, V., Ramraj, S., & Samuel, D. J. (2019). A review of multi-resolution analysis (MRA) and multi-geometric analysis (MGA) tools used in the fusion of remote sensing images. *Circuits, Systems, and Signal Processing, 39*(6), 3145–3172. https://doi.org/10.1007/s00034-019-01316-6
Mnih, V. (2013). *Machine learning for aerial image labeling*. PhD Thesis.
Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science, 7*(September), 1–10. https://doi.org/10.3389/fpls.2016.01419
Ni, X., Cheng, Y., & Wang, Z. (2019). Remote sensing semantic segmentation with convolution neural network using attention mechanism. In *2019 14th IEEE international conference on electronic measurement and instruments, ICEMI 2019* (pp. 608–613).
Osco, L. P., Nogueira, K., Ramos, A. P., Pinheiro, M. M., Furuya, D. E., Gonçalves, W. N., de Castro Jorge, L. A., Junior, J. M., & dos Santos, J. A. (2021). Semantic segmentation of citrus-orchard using deep neural networks and multispectral UAV-based imagery. *Precision Agriculture*. https://doi.org/10.1007/s11119-020-09777-5
Pan, X., Zhao, J., & Xu, J. (2020a). An end-to-end and localized post-processing method for correcting high-resolution remote sensing classification result images. *Remote Sensing, 12*(5), 852. https://doi.org/10.3390/rs12050852
Pan, Z., Dou, H., Mao, J., Dai, M., & Tian, J. (2018). MIFNet: Multi-information fusion network for sea-land segmentation. In *ACM international conference proceeding series* (pp. 24–29). ACM Press. https://doi.org/10.1145/3239576.3239578.
Pan, Z., Xu, J., Guo, Y., Hu, Y., & Wang, G. (2020b). Deep learning segmentation and classification for urban village using a worldview satellite image based on U-net. *Remote Sensing, 12*(10), 1574. https://doi.org/10.3390/rs12101574
Panboonyuen, T., Jitkajornwanich, K., Lawawirojwong, S., Srestasathiern, P., & Vateekul, P. (2019). Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning. *Remote Sensing, 11*(1), 1–22. https://doi.org/10.3390/rs11010083
Peng, C., Li, Y., Jiao, L., Chen, Y., & Shang, R. (2019). Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12*(8), 2612–2626. https://doi.org/10.1109/JSTARS.2019.2906387
Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *IEEE Access* (pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28.
Shang, R., Zhang, J., Jiao, L., Li, Y., Marturi, N., & Stolkin, R. (2020). Multi-scale adaptive feature fusion network for semantic segmentation in remote sensing images. *Remote Sensing, 12*(5), 1–20. https://doi.org/10.3390/rs12050872
Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition* (pp. 1–13). Available at: https://arxiv.org/abs/1409.1556.

Soloy, A., Turki, I., Fournier, M., Costa, S., Peuziat, B., & Lecoq, N. (2020). A deep learning-based method for quantifying and mapping the grain size on pebble beaches. *Remote Sensing, 12*(21), 3659.

Song, A., & Kim, Y. (2020). Semantic segmentation of remote-sensing imagery using heterogeneous big data: International society for photogrammetry and remote sensing potsdam and cityscape datasets. *ISPRS International Journal of Geo-Information, 9*(10), 601. https://doi.org/10.3390/ijgi9100601

Song, S., Liu, J., Liu, Y., Feng, G., Han, H., Yao, Y., & Du, M. (2020). Intelligent object recognition of urban water bodies based on deep learning for multi-source and multi-temporal high spatial resolution remote sensing imagery. *Sensors (switzerland), 20*(2), 397. https://doi.org/10.3390/s20020397

Sumbul, G., Charfuelan, M., Demir, B., & Markl, V. (2019). Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *arXiv* (pp. 2–5). https://doi.org/10.1109/igarss.2019.8900532.

Sun, X., Shi, A., Huang, H., & Mayer, H. (2020). BAS Net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13*, 5398–5413. https://doi.org/10.1109/JSTARS.2020.3021098

Tong, X. Y., Xia, G. S., Lu, Q., Shen, H., Li, S., You, S., & Zhang, L. (2020). Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment, 237*, 111322. https://doi.org/10.1016/j.rse.2019.111322

Venugopal, N. (2020). Automatic semantic segmentation with DeepLab dilated learning network for change detection in remote sensing images. *Neural Processing Letters, 51*(3), 2355–2377. https://doi.org/10.1007/s11063-019-10174-x

Wang, S., Mu, X., Yang, D., He, H., & Zhao, P. (2020a). Attention guided encoder–decoder network with multi-scale context aggregation for land cover segmentation. *IEEE Access, 8*, 215299–215309. https://doi.org/10.1109/ACCESS.2020.3040862

Wang, W., Samuel, R. D., & Hsu, C. (2020b). Prediction architecture of deep learning assisted short long term neural network for advanced traffic critical prediction system using remote sensing data. *European Journal of Remote Sensing, 54*(Sup2), 65–76. https://doi.org/10.1080/22797254.2020.1755998

Wang, Z., Gao, X., Zhang, Y., & Zhao, G. (2020c). Mslwenet: A novel deep learning network for lake water body extraction of google remote sensing images. *Remote Sensing, 12*(24), 1–19. https://doi.org/10.3390/rs12244140

Weng, L., Xu, Y., Xia, M., Zhang, Y., Liu, J., & Xu, Y. (2020). Water areas segmentation from remote sensing images using a separable residual SegNet network. *ISPRS International Journal of Geo-Information, 9*(4), 256. https://doi.org/10.3390/ijgi9040256

Wu, J., Yang, G., Yang, H., Zhu, Y., Li, Z., Lei, L., & Zhao, C. (2020a). Extracting apple tree crown information from remote imagery using deep learning. *Computers and Electronics in Agriculture, 174*, 105504. https://doi.org/10.1016/j.compag.2020.105504

Wu, W., Gao, X., Fan, J., Xia, L., Luo, J., & Zhou, Y. N. (2020b). Improved mask R-CNN-based cloud masking method for remote sensing images. *International Journal of Remote Sensing, 41*(23), 8908–8931. https://doi.org/10.1080/01431161.2020.1792576

Xia, M., Cui, Y., Zhang, Y., Xu, Y., Liu, J., & Xu, Y. (2021a). DAU-Net: A novel water areas segmentation structure for remote sensing image. *International Journal of Remote Sensing, 42*(7), 2594–2621. https://doi.org/10.1080/01431161.2020.1856964

Xia, M., Wang, T., Zhang, Y., Liu, J., & Xu, Y. (2021b). Cloud/shadow segmentation based on global attention feature fusion residual network for remote sensing imagery. *International Journal of Remote Sensing, 42*(6), 2022–2045. https://doi.org/10.1080/01431161.2020.1849852

Xiong, D., He, C., Liu, X., & Liao, M. (2020). An end-to-end Bayesian segmentation network based on a generative adversarial network for remote sensing images. *Remote Sensing, 12*(2), 216. https://doi.org/10.3390/rs12020216

Xu, Q., Yuan, X., Ouyang, C., & Zeng, Y. (2020). Attention-based pyramid network for segmentation and classification of high-resolution and hyperspectral remote sensing images. *Remote Sensing, 12*(21), 1–34. https://doi.org/10.3390/rs12213501

Yang, Q., Liu, M., Zhang, Z., Yang, S., Ning, J., & Han, W. (2019). Mapping plastic mulched farmland for high resolution images of unmanned aerial vehicle using deep semantic segmentation. *Remote Sensing*. https://doi.org/10.3390/rs11172008

Yu, L., Ma, F., Jayasuriya, A., Sigelle, M., & Perreau, S. (2007). A new contour detection approach in mammogram using rational wavelet filtering and MRF smoothing. In *Proceedings—Digital image computing techniques and applications: 9th biennial conference of the Australian pattern recognition society, DICTA 2007* (pp. 106–111). https://doi.org/10.1109/DICTA.2007.4426783.

Yu, T., Wu, W., Gong, C., & Li, X. (2021). Residual multi-attention classification network for a forest dominated tropical landscape using high-resolution remote sensing imagery. *ISPRS International Journal of Geo-Information, 10*(1), 22. https://doi.org/10.3390/ijgi10010022

Zhan, Z., Zhang, X., Liu, Y., Sun, X., Pang, C., & Zhao, C. (2020). Vegetation land use/land cover extraction from high-resolution satellite images based on adaptive context inference. *IEEE Access, 8*, 21036–21051. https://doi.org/10.1109/ACCESS.2020.2969812

Zhang, M., & Hu, X. (2017). Translation-aware semantic segmentation via conditional least-square generative adversarial networks. *Journal of Applied Remote Sensing, 11*(04), 042622. https://doi.org/10.1117/1.jrs.11.042622

Zhang, W., Witharana, C., Liljedahl, A. K., & Kanevskiy, M. (2018). Deep convolutional neural networks for automated characterization of arctic ice-wedge polygons in very high spatial resolution aerial imagery. *Remote Sensing, 10*(9), 1487. https://doi.org/10.3390/rs10091487

Zhang, X., Xiao, Z., Li, D., Fan, M., & Zhao, L. (2019). Semantic segmentation of remote sensing images using multiscale decoding network. *IEEE Geoscience and Remote Sensing Letters, 16*(9), 1492–1496. https://doi.org/10.1109/lgrs.2019.2901592

Zhang, Y., & Chi, M. (2020). Mask-R-FCN: A deep fusion network for semantic segmentation. *IEEE Access, 8*, 155753–155765. https://doi.org/10.1109/ACCESS.2020.3012701

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017) Pyramid scene parsing network. In *Proceedings—30th IEEE conference on computer vision and pattern recognition, CVPR 2017* (pp. 6230–6239). https://doi.org/10.1109/CVPR.2017.660.

Zhao, L., Tang, P., & Huo, L. (2016). Feature significance-based multibag-of-visual-words model for remote sensing image scene classification. *Journal of Applied Remote Sensing, 10*(3), 035002. https://doi.org/10.1117/1.jrs.10.035004

Zhao, T., Yang, Y., Niu, H., Wang, D., & Chen, Y. (2018). Comparing U-Net convolutional networks with fully convolutional networks in the performances of pomegranate tree canopy segmentation. In A. M. Larar, M. Suzuki, & J. Wang (Eds.), *Multispectral, hyperspectral, and ultraspectral remote sensing technology, techniques and applications VII* (p. 64). SPIE. https://doi.org/10.1117/12.2325570

Zhou, W., Newsam, S., Li, C., & Shao, Z. (2018). PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote*

Sensing, 145, 197–209. https://doi.org/10.1016/j.isprsjprs.2018.01.004

Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., & Jiao, J. (2015). Orientation robust object detection in aerial images using deep convolutional neural network. In *Proceedings—International conference on image processing, ICIP* (pp. 3735–3739). https://doi.org/10.1109/ICIP.2015.7351502.

Zhu, Q., Li, Z., Zhang, Y., & Guan, Q. (2020). Building extraction from high spatial resolution remote sensing images via multiscale-aware and segmentation-prior conditional random fields. *Remote Sensing, 12*(23), 1–18. https://doi.org/10.3390/rs12233983

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine, 5*(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

Zou, J., Li, W., & Du, Q. (2015). deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters, 12*(11), 2321–2325. https://doi.org/10.1109/LGRS.2015.2475299