Expert Systems With Applications 169 (2021) 114417

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

A review of deep learning methods for semantic segmentation of remote sensing imagery

Xiaohui Yuan [b,*], Jianfang Shi [a,b], Lichuan Gu [b,c]

[a] Taiyuan University of Technology, Taiyuan 030024, China
[b] University of North Texas, Denton, TX 76203, USA
[c] Anhui Agricultural University, Hefei 230036, China

ARTICLE INFO

Keywords:
Semantic image segmentation
Deep neural networks
Remote sensing imagery

ABSTRACT

Semantic segmentation of remote sensing imagery has been employed in many applications and is a key research topic for decades. With the success of deep learning methods in the field of computer vision, researchers have made a great effort to transfer their superior performance to the field of remote sensing image analysis. This paper starts with a summary of the fundamental deep neural network architectures and reviews the most recent developments of deep learning methods for semantic segmentation of remote sensing imagery including non-conventional data such as hyperspectral images and point clouds. In our review of the literature, we identified three major challenges faced by researchers and summarize the innovative development to address them. As tremendous efforts have been devoted to advancing pixel-level accuracy, the emerged deep learning methods demonstrated much-improved performance on several public data sets. As to handling the non-conventional, unstructured point cloud and rich spectral imagery, the performance of the state-of-the-art methods is, on average, inferior to that of the satellite imagery. Such a performance gap also exists in learning from small data sets. In particular, the limited non-conventional remote sensing data sets with labels is an obstacle to developing and evaluating new deep learning methods.

1. Introduction

Semantic image segmentation is a fundamental task in computer vision that assigns a label to each pixel, a.k.a. pixel-level classification. It serves as a vital component in computer vision-based applications including lane analysis for autonomous vehicles (Fischer, Azimi, Roschlaub, & Krauß, 2018) and geolocalization for Unmanned Aerial Vehicles (Nassar, Amer, ElHakim, & ElHelw, 2018). In contrast to making a prediction for an image, semantic segmentation generates a fine-grained delineation of objects that embeds their spatial information. As the semantic segmentation techniques continuously advance, they have been employed to address remote sensing problems that are diverse and data-rich in nature (Ball, Anderson, & Chan, 2017). Examples of semantic segmentation of remote sensing imagery include environmental monitoring (Blaschke, Lang, Lorup, Strobl, & Zeil, 2000; Yuan & Sarma, 2011), crop cover and type analysis (Yang, Chen, Yuan, & Liu, 2016; Kussul, Lavreniuk, Skakun, & Shelestov, 2017; Jadhav & Singh, 2018), tree species in forests (Dechesne, Mallet, Le Bris, & Gouet-Brunet, 2017), building classification and land use analysis in urban spaces (Rottensteiner et al., 2012; Volpi & Ferrari, 2015; Fang, Yuan, Wang, Liu, & Luo, 2018).

In the past decade, deep learning methods have demonstrated much superior performance in many traditional computer vision applications including object classification (Liu, Deng, & Yang, 2018; Shi, Yuan, Elhoseny, & Yuan, 2020), detection (Yuan, Xie, & Abouelenien, 2018), and semantic segmentation (Long, Shelhamer, & Darrell, 2015; Noh, Hong, & Han, 2015; Badrinarayanan, Kendall, & Cipolla, 2017; Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2018). Deep learning methods automatically derive features that are tailored for the targeted classification tasks, which makes such methods better choices for handling complicated scenarios. The great success in other domains excited the adoption and extension of deep learning methods for the problems in the field of remote sensing. Despite decades of efforts, assigning meaningful labels to the elements of a remote sensing image is still very challenging. Considering the enormous quantity and a large number of modalities of the remote sensing data, the in-the-loop feature extraction and learning methods facilitated by deep neural networks could be of great benefit to researchers and practitioners that are knowledgeable in geosciences and

* Corresponding author.
E-mail address: xiaohui.yuan@unt.edu (X. Yuan).

https://doi.org/10.1016/j.eswa.2020.114417
Received 3 June 2020; Received in revised form 8 November 2020; Accepted 29 November 2020
Available online 14 December 2020

need less programming-intensive tools for high-level data analysis.

Reviews of deep learning methods for remote sensing problems have been conducted in the past years. Zhang, Zhang, and Du (2016) reviewed the fundamental deep learning techniques including convolutional neural networks (CNNs), autoencoders, and restricted Boltzmann machines, and gave a technical tutorial on the design of deep neural network-based methods for target recognition and scene understanding from satellite imagery. A follow-up survey by Zhu et al. (2017) reviewed additional deep neural network architectures such as recurrent neural networks and covered a number of applications in the field of remote sensing. Ball et al. (2017) presented a review of the deep learning methods applied to applications such as anomaly detection, disaster analysis, and assessment, land cover classification, and segmentation.

Most recently, there are reviews on semantic image segmentation using deep learning methods (Liu et al., 2018; Guo, Liu, Georgiou, & Lew, 2018; Ajmal et al., 2018; Yu et al., 2018; Garcia-Garcia et al., 2018; Hoeser & Kuenzer, 2020). Guo et al. (2018) reviewed deep learning architectures for semantic segmentation of optical images and provided a categorical view of the existing methods. Liu et al. (2018) summarized methods based on deep neural networks from the architectural aspects such as means of up-sampling, convolution approaches, weakly-supervised and unsupervised methods, etc. In addition to five deep learning architectures, Garcia-Garcia et al. (2018) reviewed the methods for semantic segmentation and data sets used for evaluation. This study presented error metrics adopted for performance analysis and gave a quantitative comparison in terms of time, memory footprint, and accuracy of the existing methods for image and video semantic segmentation (Ajmal et al., 2018; Yu et al., 2018). Ma et al. (2019) reviewed deep learning methods for remote sensing in terms of study targets, deep learning models, image resolution, type of study area, and level of classification accuracy. The review provides a summary of publications and applications with respect to venues and years. Hoeser and Kuenzer (2020) gave a detailed review of deep learning methods for Earth observation data. The focus is on object detection with a gentle coverage of image segmentation.

The aforementioned reviews focus on recent development on deep neural networks and the applications to remote sensing imagery in general. To the best of our knowledge, there is no extensive survey that covers the deep learning methods for semantic segmentation of remote sensing imagery. As new deep learning methods emerge quickly in recent years, it is necessary to summarize the development and provide scholars and practitioners a comprehensive review as well as identify open challenges in the semantic segmentation of remote sensing imagery.

The rest of this paper is organized as follows. Section 2 presents the variants of convolutional neural network architectures designed for semantic image segmentation and the fundamental ideas of these architectures. Section 3 discusses the open challenges and the developments of deep learning methods to address these challenges for semantic segmentation of remote sensing imagery. Section 4 concludes this review with a summary.

2. CNN Architectures for semantic segmentation

There have been a number of deep network architectures devised for image classification and segmentation. This section first gives a brief introduction of the fundamental ideas of CNNs and then we focus on variants of CNN designed toward semantic segmentation and present their network structures and key ideas. This section also includes deep learning architectures that are not directly applicable to semantic segmentation problems, but their ideas have been adopted in a few methods in the field of remote sensing.

2.1. AlexNet, VGGNet, and GoogLeNet

AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), VGGNet (Simonyan & Zisserman, 2015), and GoogLeNet (Szegedy et al., 2015) are three deep neural networks designed for image classification. Many of the later developments are built upon them, and the fundamental ideas in these architectures support the development of the deep network architectures for semantic segmentation. Hence, we briefly review these networks and the key components to lay the ground for further discussions.

AlexNet (Krizhevsky et al., 2012) consists of five convolutional layers and three fully-connected layers. Fig. 1(a) depicts the network architecture of AlexNet. The convolutional layers are also known as feature extraction layers. There is a pooling layer in between two adjacent convolutional layers, which aims at reducing dimensionality and hence reducing the computational complexity. Common pooling schemes include max pooling and average pooling. In AlexNet, max pooling is used, which computes the largest value of the image covered by the filter and discards the noisy components in the filter window.

In AlexNet, the first and the second convolutional layers apply filters of size 11×11 and 5×5 for feature extraction, and the other three convolutional layers use a filter of a smaller size 3×3. The idea of various filter sizes is to accommodate an object on different scales. The fully-connected layers take flattened feature vectors as input and learn a classification function.

AlexNet pioneers of the evolution of CNNs in three aspects:

(1) it applies the non-saturating Rectified Linear Unit (ReLU):

$$f(x) = \max(x, 0).$$

This activation function is computationally efficient because only a comparison operation is performed.

(2) it applies the overlapping max pooling, i.e., the stride (or step size) of each filtering operation is less than the size of the filter.

(3) it introduces the dropout technique in the fully-connected layers to reduce overfitting, which randomly assigns zero to the output following a probability of 0.5.

VGGNet (Simonyan & Zisserman, 2015) consists of a number of convolutional layers and three fully connected layers. Fig. 1(b) illustrates a VGG-16 network structure. By varying the number of convolutional layers, a suite of VGGNets can be created, e.g., VGG-11 and VGG-16. A significant difference from AlexNet is that VGGNet employs filters of size 3×3 in the convolutional layers (Lin, Chen, & Yan, 2014). In addition, the stride of convolution is one pixel. Spatial padding is used to preserve the spatial resolution, i.e. the padding is one pixel for the 3×3 convolutional layer. The max-pooling is performed over a 2×2 window with a stride of two pixels. For each neuron in the hidden layers, the ReLU activation function is used.

The employment of small filters by VGGNets reduces the number of weights of the network and hence the training complexity. The multiple convolutional layers followed by a max-pooling result in a very similar effect in comparison to the employment of a large filter as used in AlexNet. The simplification of the convolutional layers allows a greater number of network depth and enables improved accuracy. The extracted features from the convolutional layers form a hierarchy of scales and the network performs well in many tasks such as semantic segmentation and target detection (Gatys, Ecker, & Bethge, 2016). The features can be used by other classifiers such as support vector machines without fine-tuning (Penatti, Nogueira, & dos Santos, 2015).

GoogLeNet (Szegedy et al., 2015) differs from other variants of CNN in three aspects: the employment of an inception module, auxiliary classifiers at the training stage, and usage of one fully connected layer. Fig. 1(c) illustrates a GoogLeNet structure. The inception model applies filters of three sizes: $1 \times 1, 3 \times 3, 5 \times 5$ to the input and concatenates the filtering results with the max-pooling result. A naive version of the inception model is shown in Fig. 1(d). Max pooling is used between inception modules, and an average pooling (Lin et al., 2014) that

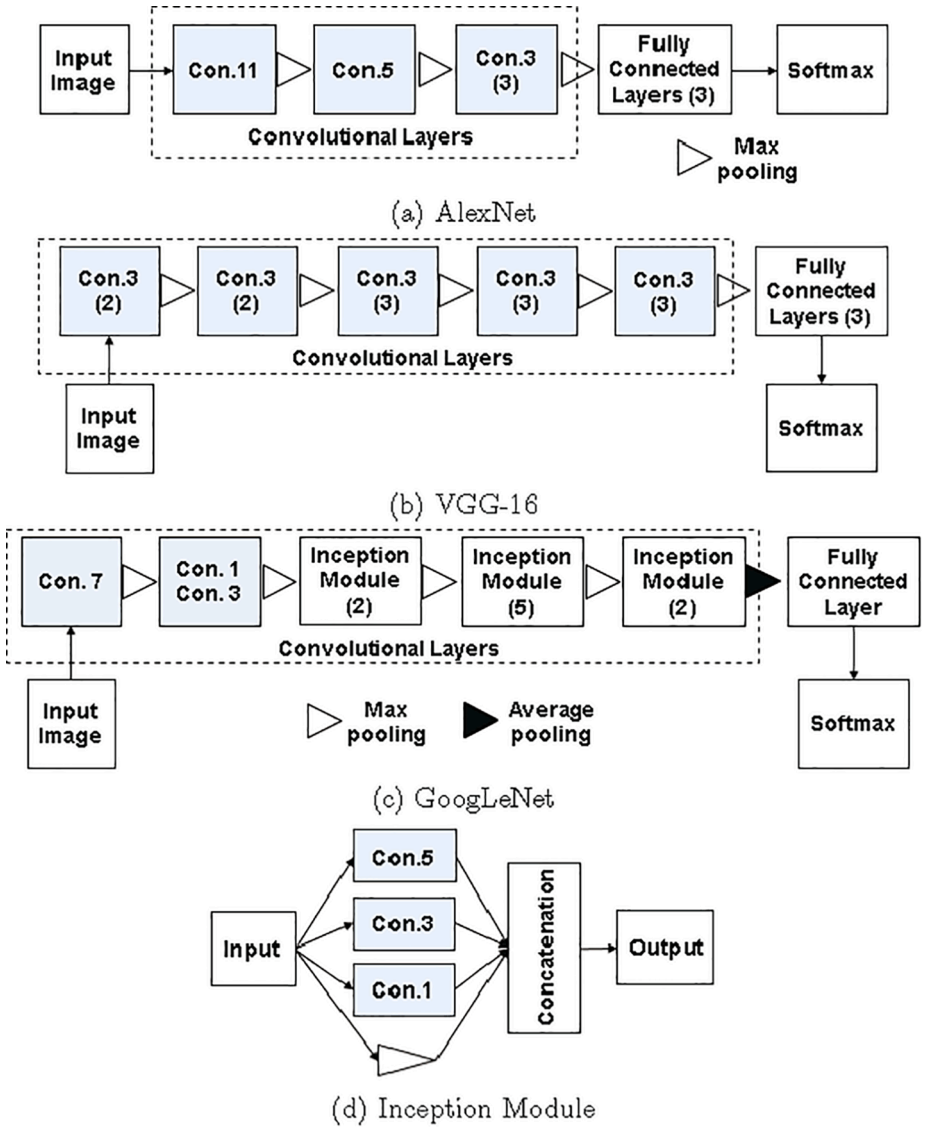

Fig. 1. The schematic network architectures of AlexNet, VGGNet, and GoogLeNet. To highlight the most significant ideas, we omit the details of each layer and adopt the following abbreviations for the building blocks of the network: *Con.N* denotes a convolutional layer using N × N filters; the number in parenthesis indicates the number of consecutive layers; and triangles and solid triangles represent max and average pooling processes, respectively.

employs dropout is used after the last inception module.

The nine inception modules and three convolutional layers make GoogLeNet fairly deep. Given such a relatively large depth of the network, the effectiveness of propagating gradients through layers becomes a concern. To address this issue, GoogLeNet adds auxiliary classifiers to the intermediate layers. These auxiliary classifiers take the form of small convolutional networks and take the output of the Inception modules. During training, the loss from these classifiers is added to the total loss of the network. In the prediction phase, the auxiliary classifiers are excluded from making decisions.

2.2. Fully Convolutional Network

Long et al. (Long et al., 2015; Shelhamer, Long, & Darrell, 2017) extended AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan & Zisserman, 2015) and GoogLeNet (Szegedy et al., 2015) and developed Fully Convolutional Network (FCN) for image semantic segmentation. The general idea of FCN consists of three steps: multi-layer convolution, deconvolution, and fusion. FCN replaces the fully connected layers with convolutional layers. Specifically, a 1 × 1 convolution (a.k.a. pixel-wise convolution) is used to compute a score for each class in an image. Because of the pooling operations that follow the convolutional layers, the output has a smaller size than the input image.

To recover the size of the original image, which is a key requirement of the segmentation process, deconvolution is used to bilinearly upsample these coarse outputs. The deconvolution process follows the same mechanism of the convolution process but operates to 'enlarge' the input by padding the matrix and integrating the elements within a deconvolution filter. The stride (or the step size) of the deconvolution is inverse proportional to the upsampling factor. Hence, the outcome from deconvolution consists of a label matrix of an improved scale.

Despite the recovery of the size of the original image using deconvolution, the class scores are diluted and details are lost. To recuperate the spatial details, a skip architecture is used to combine semantic information extracted from a deep layer with location details from its previous layers to produce the final segmentation. The upsampled deep layer is fused with the output of a shallow layer by element-wise addition. Fig. 2 gives an illustration of the fusion process.

2.3. U-Net

U-Net (Ronneberger, Fischer, & Brox, 2015) aims at image

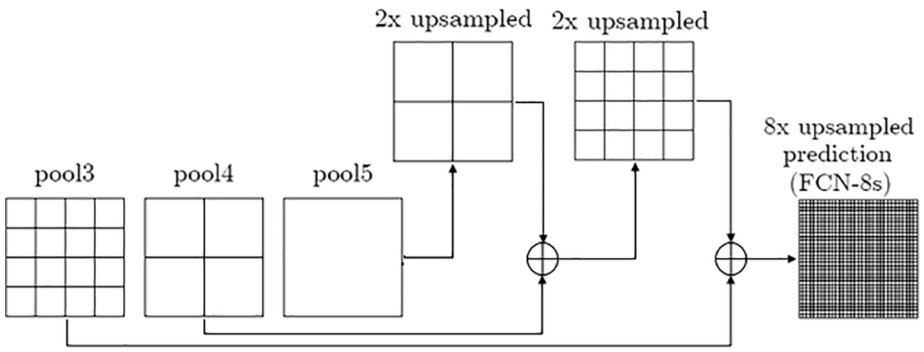

Fig. 2. The fusion process of FCN to recover the fine details of the segmentation results.

segmentation using a small training data set. Fig. 3 depicts the network architecture of the U-Net. The network consists of convolution and deconvolution layers. The convolution layers apply two 3×3 filters and the outputs are processed with ReLU followed by max pooling. The stride of this max pooling is two, which generates a downsampled outputs. However, the number of feature channels doubles at every step in the convolutional layers. In the deconvolution layers, the feature map is upsampled and a 2×2 convolution is applied to reduce the number of features. The feature maps generated by convolution layers are cropped to the size of the input. The cropping operation deals with the loss of border pixels in the convolution process and makes the dimension of the convolution results consistent with that of the deconvolution results. The cropped feature maps are stacked with the deconvolution results via shortcut connections. The network applies a 1×1 convolution to the feature map to label pixels and generates the segmentation result.

2.4. SegNet

SegNet (Badrinarayanan et al., 2017) consists of two sub-networks: an encoder and a decoder network. The encoder network in SegNet is structured contains a number of convolution and max pooling operations to extract features, which follows the architecture of FCNs. The deeper layer of this network extracts features of greater semantic meanings. Yet, the spatial information in the output of deeper layers becomes ambiguous. To address this issue, SegNet stores the element index (i.e., location of an element within the filter window) and uses it in the upsampling process of the decoder network.

The decoder network follows a symmetric structure to the encoder network. It maps low-resolution features to the higher resolution versions via convolutions and guided upsampling processes using the pooling index from the encoder network. For instance, a 2×2 low-resolution feature map is scaled up to a 4×4 matrix filled with zeros. The contents of the 2×2 map are placed to the location where they are pooled from the 4×4 matrix in the corresponding encoder layer. Such reuse of the pooling index helps to recover the spatial information and improve the boundary accuracy. It shares a similar architecture to U-Net but differs in that U-Net transfers the extracted features to the corresponding decoders, which are then concatenated into upsampled feature maps. The schematic network architecture of SegNet is shown in Fig. 4.

2.5. DeepLab

DeepLab (Chen et al., 2018) extends FCN and employs atrous convolution that enlarges the scope of filters to incorporate image context in a larger neighborhood and, hence, enables explicit control of the resolution of feature responses. Fig. 5 illustrate the idea of atrous convolution. The atrous convolution takes the form of

$$y(a) = \sum_{p=1}^{K} w(p)x(a + rp)$$

where a is the index of the element of the input/output path, p is the offset of the index of the element of the filter window, $w(p)$ is the weight to the element p, and r is the sampling rate to the input x. When $r = 1$, the atrous convolution degenerates to the conventional convolution. Atrous convolution is used to replace convolutional layers with $r > 1$. An advantage of atrous convolution is the capability of using a larger filter but maintaining the count of network parameters. Atrous Spatial Pyramid Pooling (ASPP) applies several atrous convolutions using the same kernel but with different sampling rates. The outputs from all convolutions are combined with addition.

The employment of downsampling and max pooling operations makes the segmentation results prone to losing fine details. To improve the spatial localization of segmentation, especially boundary details, DeepLab applies Conditional Random Field (CRF). A fully connected CRF is applied to the network output after bi-linear interpolation. The

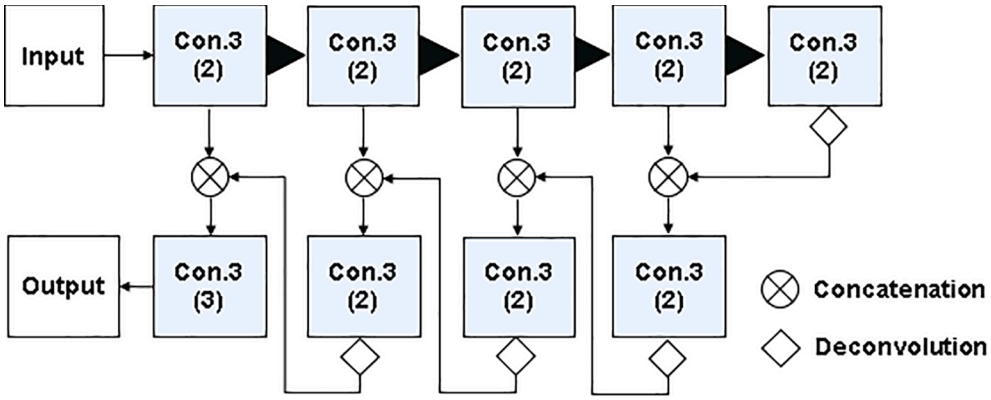

Fig. 3. The schematic network architecture of U-Net.

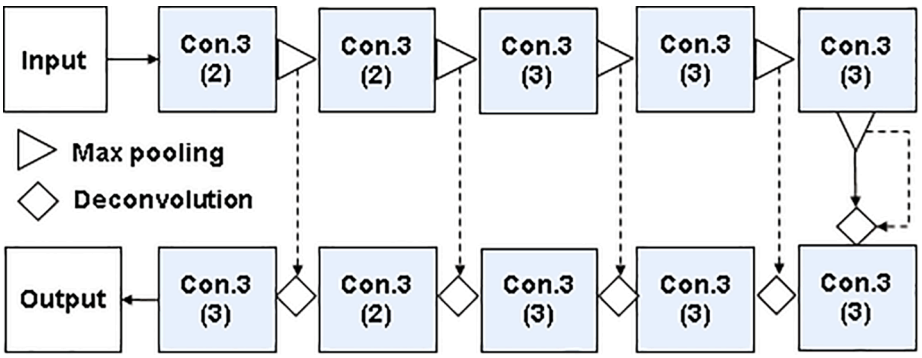

Fig. 4. The schematic network architecture of SegNet.

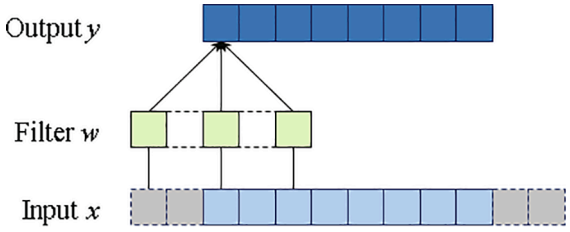

Fig. 5. A one-dimensional illustration of atrous convolution. The light blue cells depict the input and the dark blue cells depict the output. Two elements are padded on both ends of the input as shown with gray cells. The filter size is three (as shown in green). The sampling rate is two and hence a zero is inserted into every adjacent element in the filter kernel. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CRF employs the energy function that consists of two components: a pairwise potential and a unary potential. The pairwise potential follows the form presented in (Krähenbühl & Koltun, 2011) that penalizes the nodes with distinct labels. This potential integrates a Gaussian kernel for both color and pixel coordinates and another Gaussian kernel for pixel positions via a weighted summation. The unary potential encodes the label assignment probability for each pixel. The fully connected CRF is trained separately from the rest of the network.

Besides DeepLab, DeepLab V3 (Chen, Papandreou, Schroff, & Adam, 2017) and DeepLab V3+ (Chen, Zhu, Papandreou, Schroff, & Adam, 2018) are developed. Different from DeepLab, DeepLab V3 (Chen et al., 2017) uses multiple consecutive atrous convolutions with different sampling rates. In ASPP, it applies global average pooling on the last feature map of the model, feeds the features to a 1×1 convolution with 256 filters, and performs bilinearly upsample to recover the desired spatial size. In addition, DeepLab V3 abandons CRF. Instead, the features are concatenated and aggregated with a 1×1 convolution before computing the prediction scores. DeepLab V3+ (Chen et al., 2018) adds a decoder module to the DeepLabV3 network to refine the boundary details. The method applies a separable, depth-wise convolution in the decoder modules and the ASPP pooling. For each input channel, a depth-wise, spatial convolution is performed. In addition, the outputs from depth-wise convolutions are combined with a 1×1 convolution operation.

2.6. Other deep learning methods

Besides the aforementioned CNN architectures, there are other developments that have been adopted or extended for remote sensing image segmentation to address issues such as computational complexity. Specifically, we have seen ideas from ResNet (He, Zhang, Ren, & Sun, 2016), Densely Connected Convolutional Network (DenseNet) (Huang, Liu, van der Maaten, & Weinberger, 2017), and ShuffleNet (Zhang, Zhou, Lin, & Sun, 2018) used by researchers in their designs of the network for semantic segmentation of remote sensing data (Liu et al., 2018; Chen et al., 2018; Chen et al., 2018; Pan, Gao, Zhang, Yang, & Liao, 2018; Fang, Li, Zhang, & Chan, 2019). We briefly summarize these architectures in this section.

Similar to VGGNet, ResNet (He et al., 2016) has a suite of network structures with a different number of layers. Fig. 6(a) depicts the network diagram of a ResNet34. Besides one convolutional layer that uses a 7×7 filter and a single fully-connected layer, the rest of a ResNet is constructed with residual modules, which consists of two 3×3 filters and a shortcut connection to add the input to the convolution results. The diagram of a residual module is shown in Fig. 6(b). Similar to GoogLeNet, one average pooling is used after the last residual module. The classification is achieved with a softmax layer. ResNet differs from the other network architecture in that it learns a residual mapping with shortcut connections, which allows the construction of very deep networks.

DenseNet extends ResNet by introducing connections from one layer to its subsequent layers, which increases information flow and feature reusing (Huang et al., 2017). A flowchart of the DenseNet is shown in Fig. 6(d). The building block of a DenseNet, namely dense block, consists of layers of stacked two filters (a 3×3 filter followed by a 1×1 filter). Each layer receives input from every previous layer including the input layer. Four dense blocks are connected with a 1×1 convolutional layer for feature reduction. The design of DenseNet aims to address the vanishing gradient problem and enable feature reuse. This leads to a fewer number of features per layer, and hence fewer parameters to learn.

ShuffleNet (Zhang et al., 2018) improves the computational efficiency by leveraging group convolutions (Krizhevsky et al., 2012) to reduce the computation complexity of 1×1 convolutions and uses channel shuffle to help the information flow across feature channels. Fig. 6(c) depicts the network structure. The group convolution divides the computation into multiple independent shares to be processed in parallel, e.g., using GPU processing. The outputs from the group convolutions are reorganized into a matrix, where the number of rows in this matrix equals the group count and the number of columns equals the channel count. In a ShuffleNet Unit, the 3×3 convolution is replaced with depth-wise convolution (Chollet, 2017) (separable convolution on each channel). The second group convolution restores channel dimension to match the residual for concatenation.

Table 1 summarizes a list of methods that extend the aforementioned CNN architectures for semantic segmentation of remote sensing images. There are many other CNN-based methods that integrate several ideas or inherit FCN architecture but aim at processing non-traditional imagery data are not included in this table and will be discussed later.

3. Deep Learning Methods for Semantic Segmentation of Remote Sensing Imagery

Inspired by the superior performance and explosion of new

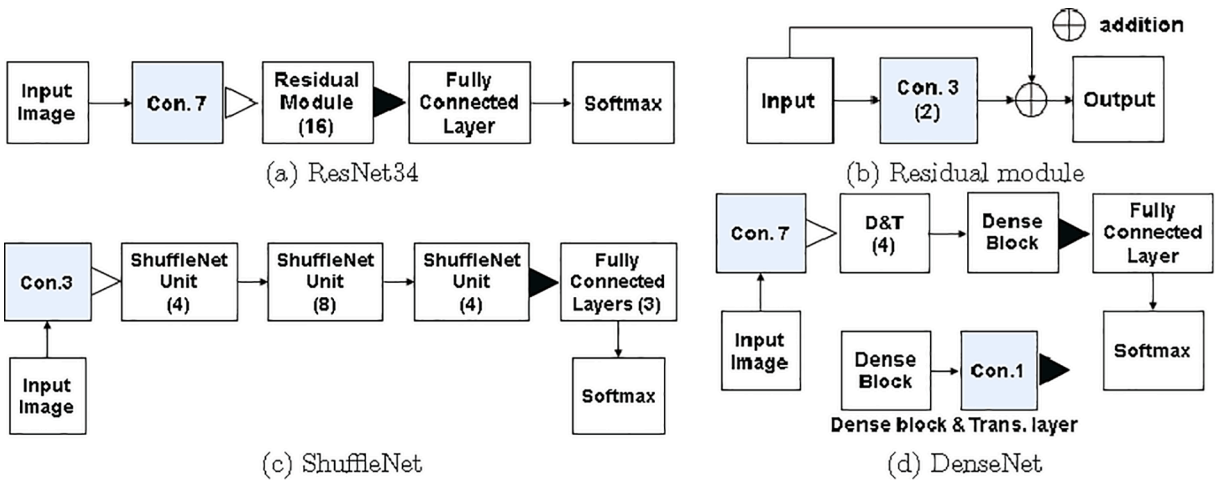

Fig. 6. The architecture of ResNet34, residual module, DenseNet, and ShuffleNet.

Table 1
Methods that extend CNN architectures for semantic segmentation of remote sensing imagery.

CNN	Method
FCN	Kampffmeyer et al. (2016), Maggiori et al. (2017), Fu et al. (2017), Henry et al. (2018), Marmanis et al. (2018), Sun and Wang (2018),† Pan et al. (2018)*‡
SegNet	Audebert et al. (2016), Liu et al. (2018), Cheng et al. (2017), Marmanis et al. (2016), Audebert et al. (2018)†
U-Net	Zhang et al. (2018), Henry et al. (2018), Xu et al. (2018), Wang et al. (2017), Wu et al. (2018), Li et al. (2018)
DeepLab	Arief et al. (2018), Liu et al. (2018), Henry et al. (2018), Chen et al. (2018),* Liu et al. (2018)
DenseNet	Fang et al. (2019)

‡ DenseNet idea is used.
* ShuffleNet idea is used.
† ResNet is used.

architectures of CNNs, great efforts have been devoted to transfer the success of deep learning methods to the segmentation of remote sensing data (Liu et al., 2018; Guo et al., 2018; Ajmal et al., 2018; Garcia-Garcia et al., 2018). Our survey of the literature reveals the following major challenges that require investigation and development of novel methods:

(1) demand for pixel-level accuracy,
(2) analysis of non-conventional data, and
(3) lack of training examples.

Every pixel in an image has a semantic meaning, which makes remote sensing imagery differ significantly from scenic and portrait images available in most public image databases such as PASCAL VOC (Zhao & Du, 2016). Besides the conventional "objects" of interest, e.g., buildings and bridges, remote sensing images contain semantically meaningful "background" such as water bodies, roads, and open fields. Such objects and background need accurate delineation to facilitate further extraction of geometric properties such as width and perimeter (Dechesne et al., 2017; Rottensteiner et al., 2012; Volpi & Ferrari, 2015). Hence, pixel-level spatial accuracy, especially at the boundary of different objects, is the utmost goal for semantic segmentation of remote sensing imagery (Liu et al., 2018; Marmanis et al., 2018).

Besides the conventional three-channel raster images, point clouds, and data with a large number of channels (e.g., hyperspectral images) are common modalities of remote sensing data. Designed for the convolution of the raster matrix, automated learning from disordered 3D points using the deep neural network is non-trivial. Point clouds are unevenly distributed in the space. Applying convolution operation is not straight forward and classifying every point is also difficult, especially in urban scenes with a variety of objects, scales, and occlusions. Another non-conventional data modality is Hyperspectral Images (HSI), which usually have dozens, if not hundreds, of channels that capture rich spectral information. A large number of channels make it difficult if not impossible to apply the existing deep learning frameworks for semantic segmentation (Ball et al., 2017; Signoroni, Savardi, Baronio, & Benini, 2019).

Lack of training examples despite a large volume of imagery is a common issue (Ma, Wang, & Wang, 2016; Kemker, Luu, & Kanan, 2018). Training a high quality deep neural network model requires an enormous number of examples. In addition, generating such a training set is extraordinarily tedious and labor-intensive (Gao et al., 2019). Strictly speaking, this is a challenge faced in many real-world applications that leverage deep learning methods. However, unlike scenic images, remote sensing imagery usually require professionals with extensive training to achieve satisfactory accuracy in the delineation (i.e., labeling) of various objects, and the crowdsourcing strategy used to successfully label massive scenic images is not always applicable to the processing of remote sensing imagery.

In the rest of this section, we organize our discussions around the three challenges we are facing in the semantic segmentation of remote sensing imagery and the deep learning methods proposed to address them.

3.1. *Methods towards pixel-level accuracy*

To avoid loss of spatial details from convolution and hence achieve pixel-level accuracy, three strategies have been explored: combining multiscale features, fusing data of different modalities, and enhancing the resulted segmentation with post-processing techniques. In the rest of this section, we review the methods relevant to each strategy and explain the key ideas.

3.1.1. *Multiscale strategy*

To achieve pixel-level accuracy, dilated convolutions (Yu & Koltun, 2016), a.k.a., atrous convolution, are often used, in which the elements at noncontiguous positions in a kernel are integrated to increase the amount of spatial context. Zhao and Du (2016) proposed a multiscale convolutional neural network (MCNN) to learn deep features of spatial relationships. MCNN constructs a pyramid structure from the image, which presents spatial features at different scales. The high-level spatial features are concatenated with spectral features to form a data set for training a logistic regression. The final results are produced with majority voting. Längkvist, Kiselev, Alirezaie, and Loutfi (2016) employed

four convolutional neural networks in parallel, each of which focuses on a contextual size. The mid-level features with semantic meanings (i.e., objects) are derived from low-level features. The fully-connected layers are fine-tuned using backpropagation. The filters in CNN are pre-trained. Audebert, Le Saux, and Lefèvre (2016) extended SegNet by introducing three parallel convolution layers that apply filters with kernel sizes at 3×3, 5×5, and 7×7. The predictions from these layers are aggregated via averaging the receptive cells of different scales. To handle both large manmade and fine-structured objects simultaneously, Liu et al. (2018) proposed a self-cascaded network that uses dilated convolutions (Yu & Koltun, 2016) for multi-scale representation on the last layer of the encoder. Besides including a larger scope of contextual information, multi-scale representation integrates hierarchical dependencies of the context. A number of dilation rates are used to extract scaled features, the results of which are aggregated in a coarse-to-fine manner with skip connections from the encoder to achieve a refined target object. Chen et al. (2018) extended shuffling CNNs for the segmentation of aerial images. A shuffling operator is used to convert low-resolution feature maps from downsampling to a higher resolution version. To make a prediction for pixels in an overlapped region, decisions of the nearest patches are used and score maps are aggregated via average for the final segmentation. Wu et al. (2018) proposed a multi-constraint, fully convolutional network for building segmentation, which adopts U-Net (Ronneberger et al., 2015) architecture with scale constraints for the intermediate layers. These constraints are computed based on the prediction at different layers in the decoder and the corresponding ground truth. Zhang et al. (2018) constructed a convolutional encoder neural networks consisting of two layers: the first layer has two sets of convolutional kernels for extracting the features of farmland and woodland and the second layer consists of two encoders that use nonlinear functions to encode the learned features and map the encoding results to the corresponding category.

Alternatively, deconvolution and skip strategy are employed to interpolate and upsample the output to achieve the spatial details. Chen et al. (2018) proposed DeepLab semantic segmentation network to mitigate the loss of details from downsampling using atrous spatial pyramid pooling, which delineates objects at various scales by probing the previous convolutional layer with filters at several sampling rates. Li et al. (2018) extended U-Net using DownBlocks (two convolution layers concatenated through a ReLU layer) in the contracting path and using UpBlocks (two convolutional layers followed by an upsampling layer) in the expansion path. These two units are similar to the residual units in the deep Residual U-Net (Zhang, Liu, & Wang, 2018), but contain two convolution layers with a fixed number of channels. High-resolution feature maps are generated from the upsampled outputs and a successive convolution layer learns to combine the outputs. In the DownBlock layers, features fed into a convolution layer are integrated with the outputs of this convolution layer using addition. The same integration strategy is applied to the UpBlock layers. Chen et al. (2018) developed a symmetrical encoder-decoder network from fully convolutional networks with shortcut blocks. The decoders generate boundary locations and labeling results from features extracted by the encoders. The shortcut block is a convolutional layer that uses a point-wise filter and a batch normalization layer (He et al., 2016). Henry, Azimi, and Merkle (2018) applied fully convolutional neural networks including FCN-8s based on VGG-19, Deep Residual U-Net, and DeepLabv3+ (Chen et al., 2018) for road segmentation.

Another idea for improving accuracy is integrating edge maps into the segmentation process. Cheng, Meng, Xiang, and Pan (2017) extended SegNet and proposed an edge-aware convolutional network by constructing an edge detection network and a segmentation network. Semantic features in different scales are extracted with the segmentation network, which is used in the training of the edge detection network. The edge map from the edge detection network is used to fine-tune the network. Similarly, Marmanis et al. (2018) extended the SegNet encoder-decoder architecture by adding boundary detection, which makes object boundaries explicit in the form of pixel-wise contour likelihood in the model. Xu, Wu, Xie, and Chen (2018) applied ResNet and used a 'guide' filter (He, Sun, & Tang, 2013) for building extraction from images. The guided filter is an edge-preserving smoothing technique that involves the original image as the guidance and a filtering image. The guidance image is used to optimize object boundaries. The guided filter is employed at the layer of the network to improve the segmentation accuracy by learning linear transformation from the input of the filter and the desired output. Audebert, Boulch, Le Saux, and Lefevre (2019) leveraged multi-task learning (Ruder, 2017) and combined negative log-likelihood loss and L1 loss to achieve a better structured semantic map. The proposed method computes the distance transform on the label masks and trains an FCN in a multi-task setting of learning classification and distance regression. Diakogiannis, Waldner, Caccetta, and Wu (2020) proposed ResUNet-a that uses a UNet encoder/decoder backbone that combines residual connections, atrous convolutions, pyramid scene parsing pooling, and multi-tasking inference. The proposed method infers object boundaries, the distance transform of the segmentation mask, the segmentation mask, and a reconstruction of the input.

3.1.2. Fusion based strategies

The idea of integrating geometry and spectral information to improve segmentation accuracy is increasingly popular given the availability of drastically different, geo-registered data in the remote sensing field. When the input data are of similar structure or representation, feature level fusion is usually performed. Marmanis et al. (2016) proposed an ensemble of FCNs following the VGG-16 architecture. Each FCN is trained with a different initialization from ImageNet (Russakovsky et al., 2015), Pascal VOC (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010), and Places (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014) data sets. The predictions are combined with average. The method employs a skip connection from the early layer to retain fine boundary details. Maggiori, Tarabalka, Charpiat, and Alliez (2017) extracted intermediate features from the network at various resolutions. The method leveraged neural networks to learn an optimal way of combining features and make the final classification. The proposed method consists of a feature extraction step to get a subset of features of intermediate resolution from the network, from which a higher resolution feature map is generated via an upsampling process.

When the input data are of different structures, a separate network is usually used to handle each data type and fusion happens at the classification stage. In the ensemble of FCNs (Marmanis et al., 2016), the incorporation of height information from the Digital Elevation Model (DEM) using the same FCN architecture is performed by merging the outputs at the very last layer. Marmanis et al. (2018) used two SegNets in parallel: one for color channels and the other for DEM. The network for color channels is initialized by training a SegNet with the Pascal data set. The network for elevation processes DSM and normalized DSM. The initialization of this network is performed with randomization but it maintains the magnitude of gradients roughly the same across layers (Glorot & Bengio, 2010). The outputs from these two networks are concatenated and processed with a point-wise convolution layer to combine the feature responses and generate a classification score for every class. Audebert, Le Saux, and Lefèvre (2018) applied FCNs to extract semantic feature maps at multiple resolutions and extend the method to integrate DSM generated from Light Detection and Ranging (LiDAR) point clouds. Two fusion strategies were developed: early fusion with FuseNet at the encoder stages and late fusion using residual correction. The late fusion strategy aims at addressing topological inconsistency among the input data such as 3D point clouds and 2D images. Two fully convolutional networks with residual connections for 2D images and 3D point clouds are trained on the respective source of data. The classification outputs are fused with average to obtain a smooth map. The correction modules are re-trained following a residual strategy. Yousefhussien, Kelbe, Ientilucci, and Salvaggio (2018)

presented a modified FCN that processes normalized point clouds and the corresponding spectral data to make classification of each point. Features are extracted from geo-referenced point clouds of LiDAR data or multiple-view imagery.

Alternatively, features from one type of data are used as auxiliary information to assist the segmentation of images. Wang, Wang, Zhang, Xiang, and Pan (2017) proposed a method that uses a gate mechanism to integrate the features. The gate mechanism is implemented with entropy maps to compute weights to feature maps in the integration. The application of gates allows the network to focus on the pixels that receive confusing decisions across layers as well as to integrate the details from the earlier network layers to improve the classification accuracy of these pixels. Sun and Wang (2018) introduced the maximum fusion to generate an initial segmentation from the color images and used a DSM backend to correct erroneous segmentation produced by an FCN. The DSM backend computes the elevation difference of each pixel to the surrounding ground. It removes false top-hats and false ground pixels.

3.1.3. Post-processing techniques

Post-processing techniques are usually used to refine the segmentation results for improved classification accuracy and smooth object boundary (Kemker, Salvaggio, & Kanan, 2018). The intuition is that the labels of the adjacent pixels are strongly correlated because of the spatial continuity of objects. If nearby segments are classified into the same class, these segments are likely to be merged as one object. Following this idea, Simple Linear Iterative Clustering (SLIC) superpixel segmentation (Achanta et al., 2012) has been used. Längkvist et al., 2016 smoothed the segmentation map using the pre-generated superpixels from SLIC. The class of each pixel is corrected using the average classifications of pixels in each superpixel of the RGB channels. Alshehhi, Marpu, Woon, and Dalla Mura (2017) used color similarity and spatial proximity based on pixel coordinates and applied SLIC to generate an initial image segmentation result. The shape features including compactness, elongation, asymmetry, and density of adjacent superpixels are used to merge discontinued segments of the same objects.

Another widely adopted post-processing technique is Conditional Random Fields (CRFs) (Chen et al., 2018; Zheng et al., 2015; Fu, Liu, Zhou, Sun, & Zhang, 2017; Henry et al., 2018; Liu et al., 2018; Pan et al., 2018). In this technique, the unary potential is the class labels produced by the network and the pairwise potential takes the input image that includes both location and spectral information (Fu et al., 2017). The mean-field approximation method is employed to reach a solution for CRF models, and the class label of each pixel is refined following the position-color constraints. Henry et al. (2018) applied fully connected CRFs that optimizes an energy function that consists of spatial potential and a correlation potential based on color information to suppress the inconsistency in the road segmentation. Liu et al. (2018) studied the impact of CRF on remote sensing image segmentation. Objects with a relatively narrow shape such as roads appeared to have a degraded accuracy, whereas large, open areas received a boost with reduced misclassification and smoothed boundary. Pan et al. (2018) adopted fully connected CRFs as a post-processing method. The output of the softmax layer from the network is encoded as the unary potential of CRF. The color-infrared image provides the pair-wise potential to compute the “distance” between pixels.

Besides SLIC and CRF methods, Sun, Zhang, Xin, and Huang (2018) combines CNN with multi-resolution segmentation (MRS). Point clouds are integrated with high-resolution images via a data fusion process for improved semantic labeling. The method reduces the salt-and-pepper distortions and hence improves the delineation of object boundaries. Liu, Minh Nguyen, Deligiannis, Ding, and Munteanu (2017) developed a weighted belief-propagation module based on Markov random field to refine the coarse border between objects. Jiang (2019) applied wavelet packet to filter the distortions such as shadows for road classification from high-resolution remote sensing images. Geng et al. (2015) applied morphological smoothing to remove the isolated misclassification from the segmentation outputs.

Table 2 summarizes the overall accuracy of the aforementioned methods as well as the data sets used in the evaluation. The methods are grouped according to the data sets used and ordered based on the accuracy if the same data set is used in the evaluation. This ordering does not mean to be a ranking because the training and testing were often conducted in different ways. In addition to the common practice of using image tiles as training examples, researchers also used a sliding window technique to create training images of smaller size but with redundancy (Audebert et al., 2018; Xu et al., 2018). Also, in some studies, precision and F-score are reported in (Chen et al., 2018; Cheng et al., 2017; Henry et al., 2018) instead of the accuracy. The results are included in this table but they are not put into order. The number of classes in the data sets is

Table 2
Performance and data sets used in learning. The ones with ∗ indicate precision and the ones with † indicate F-score. Dash – indicates no clear statement of the training data size. ‘pts’ stands for points.

Method	Accuracy	Data set	# of Class	Training size
Marmanis et al. (2018)	86.2%	ISPRS Potsdam	5	6 tiles
Chen et al. (2018)	86.92%	ISPRS Potsdam	5	18 tiles
Maggiori et al. (2017)	87.02%	ISPRS Potsdam	5	17 tiles
Audebert et al. (2018)	90.6%	ISPRS Potsdam	5	–
Sun et al. (2018)	90.62%	ISPRS Potsdam	5	21 tiles
Liu et al. (2018)	91.1%	ISPRS Potsdam	5	14 tiles
Liu et al. (2020)	92.8%/ 90.9%†	ISPRS Potsdam	5	–
Xu et al. (2018)	96.91%	ISPRS Potsdam	5	–
Chen et al. (2018)	85.78%*	ISPRS Potsdam	5	17 tiles
Kampffmeyer et al. (2016)	87.03%	ISPRS Vaihingen	5	11 tiles
Chen et al. (2018)	87.79%	ISPRS Vaihingen	5	12 tiles
Marmanis et al. (2016)	88.5%	ISPRS Vaihingen	5	12 tiles
Maggiori et al. (2017)	88.92%	ISPRS Vaihingen	5	11 tiles
Audebert et al. (2016)	89.8%	ISPRS Vaihingen	5	16 tiles
Audebert et al. (2018)	90.0%	ISPRS Vaihingen	5	–
Wang et al. (2017)	90.03%	ISPRS Vaihingen	5	12 tiles
Marmanis et al. (2018)	90.03%	ISPRS Vaihingen	5	12 tiles
Sun and Wang (2018)	90.06%	ISPRS Vaihingen	5	15 tiles
Liu et al. (2018)	91.1%	ISPRS Vaihingen	5	8 tiles
Xu et al. (2018)	97.71%	ISPRS Vaihingen	5	–
Chen et al. (2018)	86.23%*	ISPRS Vaihingen	5	11 tiles
Yousefhussien et al. (2018)	81.6%	ISPRS Vaihingen 3D	9	753,859 pts
Kemker et al. (2018)	94.19%	Pavia Univ.	9	450 pixels
Zhao and Du (2016)	96.78%	Pavia Univ.	9	3,921 pixels
Zhao and Du (2016)	99.65%	Pavia Center	9	7,456 pixels
Alshehhi et al. (2017)	91.7%/ 94.6%	Massachusetts	2	137 images
Liu et al. (2020)	94.3%/ 92.9%†	Massachusetts	2	137 images
Wu et al. (2018)	97.6%	Land Info.	2	70% data
Längkvist et al. (2016)	94.49%	Satellite images	5	70,000 pixels
Li et al. (2018)	99.04%	See-Land	2	122 images
Cheng et al. (2017)	99.36%†	See-Land	2	140 images
Geng et al. (2015)	90.68%	TerraSAR-X	5	4M pixels
Henry et al. (2018)	71.69%*	TerraSAR-X	2	201.3M pixels
Zhou and Gong (2018)	93.7%*	NOAA LiDAR	2	10K pts

also reported in this table, which is a factor the method accuracy.[1]

3.1.4. Loss function

Loss function plays an important role in deep learning methods. A majority of deep learning methods for segmenting remote sensing images inherit cross-entropy as the loss function (Alshehhi et al., 2017; Cheng et al., 2017; Fu et al., 2017; Kussul et al., 2017; Li, Zhang, & Shen, 2017; Wang et al., 2017; Yu, Jia, & Xu, 2017; Chen et al., 2018; Chen et al., 2018; Chen et al., 2018; Kestur et al., 2018; Liu et al., 2018; Liu, Yu, Yu, & Wan, 2018; Pan et al., 2018; Paoletti, Haut, Plaza, & Plaza, 2018; Sun et al., 2018; Xu et al., 2018; Zhang, Li, Li, & Wang, 2018; Zhou & Gong, 2018; Feng, Wang, Yu, Jiao, & Zhang, 2019; Li, Wang, Liu, Yu, & Lan, 2019; Sellami, Farah, Farah, & Solaiman, 2019) as these methods rely on the existing backbone network architecture. The loss function is the summation of pixel-wise cross-entropy between the predicted label-probability and true ground-truth patches. For a two-class problem, the cross-entropy is expressed as follows:

$$-y\log(p) - (1-y)\log(1-p) \tag{1}$$

where y is a binary indicator of the correct classification of an instance and p is the probability of an instance belongs to the target class. In case when there are more than two classes, cross-entropy is expressed as a weighted summation as follows:

$$-\sum_{c=1}^{C} y_c \log(p_c) \tag{2}$$

where c denotes the class label. In practice, the outputs from the softmax function are used to provide the a posteriori probability for each class, i.e., p_c. Since remote sensing images are often large, patches are generated as inputs to the deep networks. Hence, normalized cross-entropy over the entire patch is used (Audebert et al., 2016; Audebert et al., 2018).

Besides cross-entropy, the mean-squared error (MSE) between the input data and the reconstructed output has been used as the learning loss (Henry et al., 2018; Kemker et al., 2018; Zhang et al., 2018). Henry et al. (2018) adapted CNN by replacing the softmax with a sigmoid function and employing a class-weighted MSE as the loss function:

$$\frac{1}{N}\sum_{i=1}^{N} w_i (y_i - \widehat{y}_i)^2 \tag{3}$$

where y_i denotes the ground truth, $\widehat{y}_i$ denotes the sigmoid value of the prediction, and N is the count of pixels in the image.

To improve accuracy, semantic segmentation of remote sensing images is formulated as a multi-task learning problem (Yang, Luo, & Urtasun, 2018; Volpi & Tuia, 2018). Yang et al. (2018) used cross-entropy loss on the classification output and a smooth L1 loss on the regression output. The loss function is obtained by adding these two functions together. Similarly, the loss function in (Volpi & Tuia, 2018) is a linear combination of the independent losses for the semantic segmentation and the semantic boundary detection tasks:

$$\sum_{i\in\{S,B\}}^{N} w_i L^i(y, \widehat{y}) \tag{4}$$

where w_i is a weight to balance the contribution of each loss L^i. S and B denote segmentation and boundary detection tasks, respectively. The loss function of each task takes the form of cross-entropy. The probability is obtained by normalizing the activation through a sigmoid function. Alternatively, Geng, Wang, Fan, and Ma (2018) included an intraclass compactness term based on Fisher discriminant analysis to fine-tune the network. The loss function is hence a combination of the total error of the encoding network and the Fisher constraint. Gao et al. (2019) employed the GAN model for segmenting HSI images. Because the output of the discriminator is no longer the probability, the semi-supervised GAN loss function has two parts: one is supervised learning loss function, the other is unsupervised loss function. Both losses take the form of cross-entropy.

To handle imbalanced datasets, Kampffmeyer, Salberg, and Jenssen (2016) and Kemker and Kanan (2017) modified the cross-entropy loss function by introducing a different weight mechanism. Kampffmeyer et al. (2016) trained two FCN models: one using the standard cross-entropy loss, and one where the loss of the classes is weighted using median frequency balancing. Median frequency balancing weights the class loss by the ratio of the median class frequency in the training set and the actual class frequency. Kemker and Kanan (2017) introduced the inverse class frequency as weights to compute the cross-entropy. Besides modifying cross-entropy, Arief, Strand, Tveite, and Indahl (2018) used Intersection-over-Union (IoU) as the loss function to handle highly imbalanced datasets.

3.2. Methods for non-conventional data

Besides visible bands, remote sensing technology uses passive or active imaging sensors with a wide range of spectrum or much finely divided spectral bands, which resulted in image-like data sets, e.g., Synthetic Aperture Radar (SAR) images, Hyper Spectral Images, and scatter point clouds, e.g., Light Detection and Ranging. Data in these modalities differ in structure and representation from visible band images processed in computer vision applications, which poses great challenges when deep learning methods are deployed (Yang, Liu, Yuan, Chen, & Tong, 2020).

Geng et al. (2015) developed a deep convolutional autoencoder that consists of a convolutional layer, a scale transformation layer, four layers based on sparse autoencoders, and two post-processing layers. Filters used in convolution layers are devised to include gray-level co-occurrence matrix and Gabor features. A scale transformation integrates the correlated neighbor pixels to suppress the impacts of noise. Geng et al. (2018) extended long short term memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) to extract contextual dependency. A SAR image is divided into patches and each patch is converted into a vector to accommodate the input structure requirement of LSTM, which learns the latent spatial correlations. A nonnegative and Fisher constrained autoencoders were proposed, in which constraints are implemented in each autoencoder to regularize the network training. Ren, Hou, Wen, and Jiao (2018) proposed a sorting idea to the deep neural network for unsupervised feature learning. The method randomly selects image patches and put them into order according to the distance to the prototype. The patches close to the prototypes are the representative ones and the ones that are far from the prototype are the confusing ones. A dual-sparse autoencoder is used to learn the weights and a CNN is used to extract both spatial and structural features for classification.

For a hyperspectral image with b bands, a convolutional layer requires a number of filters, the number of which is proportional to b. Hence, the larger number of bands translates to an increased number of network parameters. An intuitive method to apply deep networks to HSI is reducing the number of spectral bands to bridge the gap between HSI and tri-color images (Ghamisi, Chen, & Zhu, 2016). Zhao and Du (2016) handled spectral and spatial features separately using two CNNs. A 1D CNN is used to extract the spectral features; whereas a 2D CNN is used to extract the spatial features. The features from these two networks are concatenated and used as the input to a classifier. Yu et al. (2017) used 1 $\times$ 1 convolutional kernel to extract features from different bands and employed an average pooling layer and larger dropout rates in the CNN. Li et al. (2017) handles the hyperspectral cube with 3D CNNs, which processes the multiple bands with 3D convolutions. Paoletti et al. (2018) also developed a 3D CNN network that is capable of processing both spectral and spatial features of the hyperspectral image simultaneously.

[1] https://data.linz.govt.nz/layer/53413-nz-building-outlines-pilot/.

The input layer accepts a volume of data of size $n \times n \times b$, where n denotes the width and height of a band and b denotes the number of bands. Fang et al. (2019) extended the DenseNet (Huang et al., 2017) with spectral attention mechanism for HSI image classification. The method uses 3D dilated convolutions to extract features at different spatial scales and spectral bands. In addition, the method adopts the squeeze-and-excitation block to model the inter-dependency between spectral features to emphasize informative spectral features.

Alternatively, Kemker et al. (2018) presented a stacked convolutional auto-encoder model to extract features for HSI classification. The method used unsupervised learning to create a pool of spatial-spectral feature extractors. Xu, Du, Zhang, and Zhang (2018) proposed a Random Patches Network that takes image patches as the convolution kernels, which requires no training. The spectral bands within a patch are flattened and used as inputs for the network. Feng et al. (2019) devised a divide-and-conquer, dual-architecture CNN. The method separates homogeneous regions from heterogeneous regions and processes them differently. To process homogeneous regions, a multi-scale CNN architecture is constructed to learn joint spatial-spectral features, which takes large image patches as input. To process heterogeneous regions, a fine-grained CNN architecture is constructed to learn hierarchical spectral features, which takes small image patches as inputs.

LiDAR point clouds have been used in many semantic segmentation applications to provide an extra dimension of information (Marmanis et al., 2018; Audebert et al., 2018; Sun & Wang, 2018; Zhou & Gong, 2018). To process LiDAR point cloud with CNNs, points are usually converted into raster images via a gridding process (Arief et al., 2018; Zhou & Gong, 2018; Sun et al., 2018). Arief et al. (2018) extracted the normalized elevation from the LiDAR point cloud and created a two-dimensional matrix of the elevation. The proposed a network architecture that integrates an atrous network with the stochastic depth method and imposes a regularization. Zhou and Gong (2018) converted point clouds into gray-scale images, each pixel of which represents the quantized elevation of the corresponding points relevant to the ground. CNN is trained to derive features for differentiating buildings from other objects. Sun et al. (2018) also gridded normalized point clouds into images and employed multi-filter CNN to aggregate point clouds and images for semantic labeling. A bottom-up merging method is used to combine segments. Homogeneous regions adjacent to each other are merged according to their scale, shape, and compactness. Yousefhussien et al. (2018) modified PointNet (Qi, Su, Mo, & Guibas, 2017) to operate directly on point clouds. The spectral features are extracted from georeferenced images. The network takes normalized point clouds with respect to the ground and the corresponding spectral features to generate labels for each point.

Table 3 summarizes the average accuracy of the methods for learning from non-conventional remote sensing data and the data sets used in learning and evaluation. Note that for methods evaluated with multiple data sets, overall average accuracy is reported, i.e., average across all evaluation cases. For the methods that used the same or similar data sets, the accuracy is arranged in ascending order.

Table 3
Performance of methods that learn from non-conventional data sets and the data sets used in learning. The ones with $*$ indicate precision.

Methods	Accuracy	Data set	Modality
Geng et al. (2015)	88.11%	TerraSAR-X	SAR
Geng et al. (2018)	86.61%	TerraSAR-X/Radarsat-2/ALOS-2	
Ren et al. (2018)	83.22%	Traunstein/PoDelta	
Ghamisi et al. (2016)	82.17%	Indian Pines/Pavia Univ.	HSI
Paoletti et al. (2018)	98.09%	Indian Pines/Pavia Univ.	
Yu et al. (2017)	66.02%	Indian Pines/Pavia Univ./ Salinas Valley	
Kemker et al. (2018)	97.3%	Indian Pines/Pavia Univ./ Salinas Valley	
Feng et al. (2019)	97.85%	Indian Pines/Pavia Univ./ Salinas Valley	
Li et al. (2017)	99.23%	Indian Pines/Pavia Univ./ Botswana	
Fang et al. (2019)	87.47%	Indian Pines/Pavia Univ./Univ. of Houston	
Xu et al. (2018)	98.78%	Indian Pines/Pavia Univ./KSC	
Zhao and Du (2016)	97.88%	Pavia Center/Pavia Univ.	
Sun et al. (2018)	89.27%	Guangzhou	LiDAR
Arief et al. (2018)	93.64%	Follo 2014	
Zhou and Gong (2018)	94.5%*	LiDAR	
Yousefhussien et al. (2018)	81.6%	ISPRS 3D Vaihingen	

3.3. Methods to learn from small training set

Deep neural networks for semantic segmentation of high-resolution imagery usually have millions of parameters, which requires a huge amount of labeled examples to train. In many computer vision applications, large data sets have been created and made available for public access, for example, ImageNet (Russakovsky et al., 2015) has 1.28M training images. A common strategy is to construct a deep network using a rich data set. The trained model is then fine-tuned using a smaller data set of the target problem (Pan, Shi, & Xu, 2018; Penatti et al., 2015). However, for remote sensing imagery, there are very few large labeled data sets, to begin with. Table 4 lists popular public data sets used in the existing studies. Properties of the data sets are also reported including the number of channels (if available), number of classes, spatial resolution, and modality. Unlike scenic image data sets available for object classification and recognition, remote sensing data sets are usually available in the form of one image or a number of tiles that can be pieced together into an image. Hence, the size of the image is reported.[2]

To deal with the problem of scarce training examples, an intuitive way is to create synthetic images to increase the training set size. Kemker et al. (2018) used the Digital Imaging and Remote Sensing Image Generation software (free to use but limit to qualified users) to create synthetic multi-spectral images and the labels to the images. Each labeled image is divided into patches of 160 by 160 pixels and used as training examples. The authors adapted Sharpmask (Pinheiro, Lin, Collobert, & Dollár, 2016) and RefineNet (Lin, Milan, Shen, & Reid, 2017) to initialize with the synthetic data. Zhou and Gong (2018) included a data augmentation technique to create additional labeled examples by randomly rotating and flipping an annotated image patch. Ma et al., 2016 proposed a method that integrates a local decision from the weighted samples within a neighborhood and a global decision from the trained deep network. The unlabeled samples with high confidence are used for training a deep network.

The main advantage of creating synthetic images is that such images are normally cheaper and easier to obtain and the labeling takes no time (you know what to create). Yet, the synthetic gap (the difference in feature-space distributions) makes it difficult to transform features of synthetic examples to that of real images. To close this gap, Feng et al. (2019) selected unlabeled samples according to the spectral similarity under adaptively spatial constraint and the ones with high similarity are included in the training set. Gao et al. (2019) employed the Generative Adversarial Network (GAN) model (Radford, Metz, & Chintala, 2016; Bashmal et al., 2018) by adding a softmax layer. The discriminator of the GAN generates label categories, which allows the network to classify labeled examples and unlabeled samples. Ghamisi et al. (2016) proposed a self-improving convolutional neural network that employs particle swarm optimization for band selection from HSI images. The optimization method fractional-order Darwinian idea and the selected bands are used for network training.

Another way to handle example scarcity is semi-supervised learning for classification incorporated with unsupervised feature learning.

[2] the road image set has a resolution of 609×914.

Table 4
Public remote sensing data sets for semantic segmentation and the key properties of them. Dash – denotes that the property is inapplicable.

Data set	Image/Data Size (Channel)	# of Classes	Spatial Resolution	# of Tiles	Modality
Vaihingen (2019)	2000 × 2500 (3)	6	9 cm	33	RGB/
Potsdam (2019)	6000 × 6000 (4)	6	5 cm	38	IR-RGB
Inria Aerial Image (Maggiori et al., 2017)	1500 × 1500(3)	2	30 cm	360	
Building & Road (Mnih, 2013)	1500 × 1500(3)	2	1 m	1352	
Indian Pines (Baumgardner et al., 2015)	145 × 145 (200)	16	20 m	–	
Salinas Valley (Salinas, 2019)	512 × 217 (204)	16	3.7 m	–	
Pavia Univ. (Comp. Intelligence Group, 2019)	610 × 610 (103)	9	1.3 m	–	
Pavia Center (Comp. Intelligence Group, 2019)	1096 × 715 (102)	9	1.3 m	–	HSI
Botswana Scene (Botswana, 2019)	1476 × 256 (145)	14	30 m	–	
Kennedy Space Center (2019)	512 × 614 (176)	13	18 m	–	
University of Houston (2019)	349 × 1905 (144)	15	2.5 m	–	
Quebec and Canada (2019)	795 × 564 (84)	7	1 m	-	
SpaceNet (2018)	666 × 666 (8)	2	30/50 cm	–	
TerraSAR-X (2019)	3580 × 2250	5	0.38 m	-	
RADARSAT-2 (2019)	2000 × 1600	6	1.5 m	-	SAR
ALOS-2 (2019)	900 × 1600	6	3 m	-	
F-SAR (Microwaves, 2019)	6187 × 4278	4	1 m × 0.67 m	–	
ISPRS Vaihingen 3D (2019)	753,859 pts	9	4 pts/m^2	–	LiDAR
Follo 2014 (Kartverket, 2019)	2.4 M pts/set	11	5 pts/m^2	1877	

Kemker and Kanan (2017) proposed a method that employs three stacked convolutional autoencoders trained separately with unlabeled images to derive high-level features, which follows an unsupervised learning strategy. The features extracted by autoencoders are concatenated and processed via average pooling and dimension reduction prior to classification using a support vector machine classifier. A later work of the authors extended the stacked autoencoder framework and developed an unsupervised feature learning model to reconstruct inputs (Kemker et al., 2018). This module consists of an encoder to learn the representations and a decoder to reconstruct the input and includes multi-reconstruction and classification errors in the loss function. A supervised multilayer perception module is used to classify the extracted features. Yu et al. (2017) used two dropout layers with a high dropout rate at 0.6 to help the optimization of the CNN parameters from training with a small data set. The dropout randomly suppresses the output of each neuron to zero and the network is forced to learn more robust features. Pan et al. (2018) proposed a multi-grained network. The multi-grained scanning technique extracts spectral and spatial information and combines spectral information among different bands and spatial correlation within neighboring pixels at different scales. Sellami et al. (2019) extended the semi-supervised 3D CNN that performs adaptive dimensionality reduction for the classification of HSI images. It circumvents the challenge of limited training samples by selecting multiple sets of spectral bands that are most representative of the ground objects to enrich the training set.

The transfer learning has also been explored to deal with example scarcity. In general, by training a CNN with a large number of examples other than the actual ones to be processed, the model parameters are learned. Such a trained model is then refined with a small set of examples of the target problem to get the parameters fine-tuned. Following this idea, Liu et al. (2018) proposed a Siamese network that integrates two convolutional neural networks. The Siamese network is trained with a large volume of image data, the model of which is refined with a small number of labeled HSI images and features extracted from the spectral bands. Li et al. (2019) trained a deep learning network with a large data set that has high similarity to the target data set. Common features of the data in the source and target domains are identified to facilitate the transfer of the learned model.

In summary, we report the performance of the aforementioned methods and the data sets used in the evaluation in Table 5. In many of these studies, more than one data set is used in the evaluation. The table reports the one with the best overall accuracy as well as the size of the data set used in the training phase of the networks.

4. Conclusion

CNNs and the variants have demonstrated great success in numerous computer vision applications. The great success in other domains excited the adoption and extension of deep learning methods for the problems in the field of remote sensing and researchers are transferring such superior performance of deep learning methods to the field of remote sensing image analysis. This paper reviews recent developments of deep learning methods for semantic segmentation of remote sensing imagery including non-conventional data such as HSI and LiDAR point clouds to provide scholars and practitioners a comprehensive review as well as identify open challenges in the semantic segmentation of remote sensing imagery. Specifically, we review the fundamental and advanced CNN architectures that enable the transformation of deep learning methods to accommodate a variety of data modalities and structures in remote sensing. In our review of the literature, we identified three major challenges:

1. Demand for pixel-level accuracy. Every object in the scope of a remote sensing image carries meaningful information and needs to be accurately separated from the adjacent ones. Great efforts have

Table 5
The average accuracy of methods that learn from small data sets. This table also reports the data sets and the training data size in the number of pixels, if not specified, used in training. The ones with * indicate precision.

Methods	Accuracy	Data set	Training Size (pixels)
Ma et al. (Ma et al., 2016)	92.11%	Indian Pines	160
Pan et al. (Pan et al., 2018)	95.48%	Indian Pines	350
Liu et al. (Liu et al., 2018)	98.72%	Indian Pines	1,620
Kemker et al. (Kemker & Kanan, 2017)	98.06%	Salinas	800
Gao et al. (Gao et al., 2019)	96.19%	Salinas	16,238
Feng et al. (Feng et al., 2019)	98.8%	Salinas	543
Yu et al. (Yu et al., 2017)	85.24%	Salinas Valley	48
Ghamisi et al. (Ghamisi et al., 2016)	82.67 %	Pavia University	3,912
Li et al. (Li et al., 2019)	90.12%	Pavia University	1,608
Kemker et al. (Kemker et al., 2018)	98.18%	Pavia University	450
Sellami et al. (Sellami et al., 2019)	98.45	Pavia University	851
Zhou and Gong (Zhou & Gong, 2018)	94.5%*	NOAA LiDAR	10,000 pts

been devoted to addressing this issue and methods have been developed that extend FCN and regularization such as object boundary information. The emerged deep learning methods demonstrated much-improved performance on several public data sets. The success is more prominent for color and infrared satellite images, which are most similar to the image sets used in the scenic and portrait computer vision tasks. Throughout the review, we found that it is of great importance to have public data sets such as ISPRS data sets and a number of HSI image sets, which enable the development of deep learning methods and facilitate comparison study. On the other hand, the diverse data modality and evaluation metrics make comparison difficult.

2. Non-conventional data. Besides RGB and infrared images, point clouds and HSI images with a large number of bands are common modalities of remote sensing applications. Handling such unstructured point cloud and rich channel data requires a redesign of the network structure or a conversion of the non-conventional data to a similar format to RGB images. The accuracy for processing non-conventional data is, on average, inferior to that of the satellite imagery. Even with an extension to a network for point data, the average accuracy is in the range of the lower eighties.
3. Lack of training examples despite a large volume of data. This challenge is not unique to remote sensing applications, but it is much more pressing, especially for non-conventional data sources such as SAR, HSI, and LiDAR. Our review of the literature identifies that researchers strive for learning from small examples with HSI imagery. This is mostly because the HSI data set usually much less number of pixels to acquire rich spectral information. It is clear that limited, non-conventional remote sensing data sets with labels make developing and evaluating new deep learning methods a great challenge. Semi-supervised methods that leverage unlabeled data have demonstrated potentials.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The work of this paper is partially supported by Ministry of Science and Technology (G20200204007). We thank Chengyuan Zhuang at the University of North Texas for his assistance of searching and selection of relevant works and anonymous reviewers for their constructive comments and suggestions.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*, 2274–2282.

Ajmal, H., Rehman, S., Farooq, U., Ain, Q. U., Riaz, F., & Hassan, A. (2018). Convolutional neural network based image segmentation: a review. In *Pattern recognition and tracking XXIX* (p. 106490N). International Society for Optics and Photonics volume 10649.

ALOS-2 (2019). https://www.eorc.jaxa.jp/ALOS-2/en/about/palsar2.htm access in Sept. 2019.

Alshehhi, R., Marpu, P. R., Woon, W. L., & Dalla Mura, M. (2017). Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing, 130*, 139–149.

Arief, H. A., Strand, G.-H., Tveite, H., & Indahl, U. G. (2018). Land cover segmentation of airborne lidar data using stochastic atrous network. *Remote Sensing, 10*(973) 1–22.

Audebert, N., Boulch, A., Le Saux, B., & Lefevre, S. (2019). Distance transform regression for spatially-aware deep semantic segmentation. *Computer Vision and Image Understanding, 189*, 102809.

Audebert, N., Le Saux, B., & Lefèvre, S. (2016). Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In *Asian conference on computer vision* (pp. 180–196). Springer.

Audebert, N., Le Saux, B., & Lefèvre, S. (2018). Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing, 140*, 20–32.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* (pp. 2481–2495).

Ball, J. E., Anderson, D. T., & Chan, C. S. (2017). Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of Applied Remote Sensing, 11*, 42609.

Bashmal, L., Bazi, Y., AlHichri, H., AlRahhal, M., Ammour, N., & Alajlan, N. (2018). Siamese-GAN: Learning invariant representations for aerial vehicle image categorization. *Remote Sensing, 10*, 351.

Baumgardner, M. F., Biehl, L. L., & Landgrebe, D. A. (2015). 220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3. URL: https://purr.purdue.edu/publications/1947/1.

Blaschke, T., Lang, S., Lorup, E., Strobl, J., & Zeil, P. (2000). Object-oriented image processing in an integrated GIS/remote sensing environment and perspectives for environmental applications. *Environmental Information for Planning, Politics and the Public, 2*, 555–570.

Botswana (2019). http://aviris.jpl.nasa.gov/data/free_data.html access in Sept. 2019.

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. Technical Report Google Inc.

Chen, K., Fu, K., Yan, M., Gao, X., Sun, X., & Wei, X. (2018). Semantic segmentation of aerial images with shuffling convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters, 15*, 173–177.

Cheng, D., Meng, G., Xiang, S., & Pan, C. (2017). FusionNet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*, 5769–5783.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*, 834–848.

Chen, G., Zhang, X., Wang, Q., Dai, F., Gong, Y., & Zhu, K. (2018). 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*, 1633–1644.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision* (pp. 801–818).

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).

Comp. Intelligence Group (2019). Hyperspectral remote sensing scenes. http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes#Pavia_Centre_and_University access in Sept. 2019.

Dechesne, C., Mallet, C., Le Bris, A., & Gouet-Brunet, V. (2017). Semantic segmentation of forest stands of pure species as a global optimization problem. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, (pp. 141–148).

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

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision, 88*, 303–338.

Fang, B., Li, Y., Zhang, H., & Chan, J. C.-W. (2019). Hyperspectral images classification based on dense convolutional networks with spectral-wise attention mechanism. *Remote Sensing, 11*(159) 1–18.

Fang, F., Yuan, X., Wang, L., Liu, Y., & Luo, Z. (2018). Urban land-use classification from photographs. *IEEE Geoscience and Remote Sensing Letters, 15*, 1927–1931.

Feng, J., Wang, L., Yu, H., Jiao, L., & Zhang, X. (2019). Divide-and-conquer dual-architecture convolutional neural network for classification of hyperspectral images. *Remote Sensing, 11*(484) 1–26.

Fischer, P., Azimi, S. M., Roschlaub, R., & Krauß, T. (2018). Towards HD maps from aerial imagery: Robust lane marking segmentation using country-scale imagery. *ISPRS International Journal of Geo-Information, 7*, 458.

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*(498) 1–21.

Gao, H., Yao, D., Wang, M., Li, C., Liu, H., Hua, Z., & Wang, J. (2019). A hyperspectral image classification method based on multi-discriminator generative adversarial networks. *Sensors, 19*, 3269.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., & Garcia-Rodriguez, J. (2018). A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing, 70*, 41–65.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2414–2423).

Geng, J., Fan, J., Wang, H., Ma, X., Li, B., & Chen, F. (2015). High-resolution sar image classification via deep convolutional autoencoders. *IEEE Geoscience and Remote Sensing Letters, 12*, 2351–2355.

Geng, J., Wang, H., Fan, J., & Ma, X. (2018). Sar image classification via deep recurrent encoding neural networks. *IEEE Transactions on Geoscience and Remote Sensing, 56*, 2255–2269.

Ghamisi, P., Chen, Y., & Zhu, X. X. (2016). A self-improving convolution neural network for the classification of hyperspectral data. *IEEE Geoscience and Remote Sensing Letters, 13*, 1537–1541.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256). Chia Laguna Resort, Sardinia, Italy volume 9.

Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval, 7*, 87–93.

Henry, C., Azimi, S. M., & Merkle, N. (2018). Road segmentation in SAR satellite images with deep fully-convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 1–5.

He, K., Sun, J., & Tang, X. (2013). Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*, 1397–1409.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In *European conference on computer vision* (pp. 630–645). Springer.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.

Hoeser, T., & Kuenzer, C. (2020). Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends. *Remote Sensing, 12*, 1667.

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

ISPRS Vaihingen 3D (2019). http://www2.isprs.org/commissions/comm3/wg4/3d-semantic-labeling.html access in Sept. 2019.

Jadhav, J. K., & Singh, R. P. (2018). Automatic semantic segmentation and classification of remote sensing data for agriculture. *Mathematical Models in Engineering, 4*, 112–137.

Jiang, Y. (2019). Research on road extraction of remote sensing image based on convolutional neural network. *EURASIP Journal on Image and Video Processing, 2019*, 31.

Kampffmeyer, M., Salberg, A.-B., & Jenssen, R. (2016). Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 1–9).

Kartverket (2019). Follo 2014 LiDAR data set. https://hoydedata.no/LaserInnsyn/ access in Sept. 2019.

Kemker, R., & Kanan, C. (2017). Self-taught feature learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 55*, 2693–2705.

Kemker, R., Luu, R., & Kanan, C. (2018). Low-shot learning for the semantic segmentation of remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing, 56*(6214), 6223.

Kemker, R., Salvaggio, C., & Kanan, C. (2018). Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing, 145*, 60–77.

Kennedy Space Center (2019). http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes#Kennedy_Space_Center_.28KSC.29 access in Sept. 2019.

Kestur, R., Farooq, S., Abdal, R., Mehraj, E., Narasipura, O., & Mudigere, M. (2018). UFCN: a fully convolutional neural network for road extraction in RGB imagery acquired by remote sensing from an unmanned aerial vehicle. *Journal of Applied Remote Sensing, 12*(016020) 1–15.

Krähenbühl, P., & Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems 24* (pp. 109–117).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

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*, 778–782.

Längkvist, M., Kiselev, A., Alirezaie, M., & Loutfi, A. (2016). Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sensing, 8*, 329.

Li, Y., Zhang, H., & Shen, Q. (2017). Spectral-spatial classification of hyperspectral imagery with 3d convolutional neural network. *Remote Sensing, 9*(67) 1–17.

Li, K., Wang, M., Liu, Y., Yu, N., & Lan, W. (2019). A novel method of hyperspectral data classification based on transfer learning and deep belief network. *Applied Sciences*, (pp. 1379 (1–17)).

Li, R., Liu, W., Yang, L., Sun, S., Hu, W., Zhang, F., & Li, W. (2018). DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1–9.

Lin, M., Chen, Q., & Yan, S. (2014). Network in network. In *Proceedings of the international conference on machine learning*.

Lin, G., Milan, A., Shen, C., & Reid, I. (2017). 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).

Liu, X., Deng, Z., & Yang, Y. (2018). Recent progress in semantic image segmentation. *Artificial Intelligence Review*, 1–18.

Liu, B., Yu, X., Yu, A., & Wan, G. (2018). Deep convolutional recurrent neural network with transfer learning for hyperspectral image classification. *Journal of Applied Remote Sensing, 12*(026028) 1–18.

Liu, Y., Fan, B., Wang, L., Bai, J., Xiang, S., & Pan, C. (2018). Efficient patch-wise semantic segmentation for large-scale remote sensing images. *Sensors, 18*(3232) 1–16.

Liu, Y., Fan, B., Wang, L., Bai, J., Xiang, S., & Pan, C. (2018). Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS Journal of Photogrammetry and Remote Sensing, 145*, 78–95.

Liu, Y., Minh Nguyen, D., Deligiannis, N., Ding, W., & Munteanu, A. (2017). Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery. *Remote Sensing, 9*, 522.

Liu, J., Wang, S., Hou, X., & Song, W. (2020). A deep residual learning serial segmentation network for extracting buildings from remote sensing imagery. *International Journal of Remote Sensing, 41*, 5573–5587.

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

Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In *IEEE international geoscience and remote sensing symposium (IGARSS)*.

Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). High-resolution aerial image labeling with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing, 55*, 7092–7103.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing, 152*, 166–177.

Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M., & Stilla, U. (2018). Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing, 135*, 158–172.

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

Ma, X., Wang, H., & Wang, J. (2016). Semisupervised classification for hyperspectral image based on multi-decision labeling and deep feature learning. *ISPRS Journal of Photogrammetry and Remote Sensing, 120*, 99–107.

Microwaves and Radar Inst. (2019). F-sar at traunstein in bavaria, germany. https://www.dlr.de/hr/en/DesktopDefault.aspx/tabid-4698/7782_read-12248/gallery-1/gallery_read-Image. 32.6097/ access in Sept. 2019.

Mnih, V. (2013). Machine learning for aerial image labeling. Ph.D. thesis University of Toronto. URL https://www.cs.toronto.edu/ vmnih/data/.

Nassar, A., Amer, K., ElHakim, R., & ElHelw, M. (2018). A deep CNN-based framework for enhanced aerial imagery registration with applications to UAV geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 1513–1523).

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1520–1528).

Pan, X., Gao, L., Zhang, B., Yang, F., & Liao, W. (2018). High-resolution aerial imagery semantic labeling with dense pyramid network. *Sensors, 18*, 3774.

Pan, B., Shi, Z., & Xu, X. (2018). Mugnet: Deep learning for hyperspectral image classification using limited samples. *ISPRS Journal of Photogrammetry and Remote Sensing, 145*, 108–119.

Paoletti, M., Haut, J., Plaza, J., & Plaza, A. (2018). A new deep convolutional neural network for fast hyperspectral image classification. *ISPRS Journal of Photogrammetry and Remote Sensing, 145*, 120–147.

Penatti, O. A. B., Nogueira, K., & dos Santos, J. A. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 44–51).

Pinheiro, P. O., Lin, T.-Y., Collobert, R., & Dollár, P. (2016). Learning to refine object segments. In *European conference on computer vision* (pp. 75–91). Springer.

ISPRS Potsdam (2019). http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html access in Sept. 2019.

Qi, C., Su, H., Mo, K., & Guibas, L. (2017). Pointnet: Deep learning on point sets for 3D classification and segmentation. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Quebec, Canada (2019). Ieee grss 2014 data fusion contest. http://www.grss-ieee.org/community/technical-committees/data-fusion/2014-ieee-grss-data-fusion-contest/ access in Sept. 2019.

RADARSAT-2 (2019). https://mdacorporation.com/geospatial/international/satellites/RADARSAT-2 access in Sept. 2019.

Radford, A., Metz, L., & Chintala, S. (2016). *Unsupervised representation learning with deep convolutional generative adversarial networks*. URL https://arxiv.org/abs/1511.06434.

Ren, Z., Hou, B., Wen, Z., & Jiao, L. (2018). Patch-sorted deep feature learning for high resolution sar image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11*, 3113–3126.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., & Breitkopf, U. (2012). The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci, 1*, 293–298.

Ruder, S. (2017). *An overview of multi-task learning in deep neural networks*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*, 211–252.

Salinas (2019). http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes#Salinas access in Sept. 2019.

Sellami, A., Farah, M., Farah, I. R., & Solaiman, B. (2019). Hyperspectral imagery classification based on semi-supervised 3-d deep neural network and adaptive band selection. *Expert Systems with Applications, 129*, 246–259.
Shelhamer, E., Long, J., & Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*, 640–651.
Shi, J., Yuan, X., Elhoseny, M., & Yuan, X. (2020). Weakly supervised deep learning for objects detection from images. In *Urban intelligence and applications proceedings of ICUIA 2019* (pp. 231–242). Springer International Publishing.
Signoroni, A., Savardi, M., Baronio, A., & Benini, S. (2019). Deep learning meets hyperspectral image analysis: A multidisciplinary review. *Journal of Imaging, 5*(52) 1–32.
Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations*.
SpaceNet (2018). Spacenet on amazon web services (aws). https://spacenetchallenge.github.io/datasets/datasetHomePage.html last modified April 30, 2018, access in Nov. 2020.
Sun, W., & Wang, R. (2018). Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined With DSM. *IEEE Geoscience and Remote Sensing Letters, 15*, 474–478.
Sun, Y., Zhang, X., Xin, Q., & Huang, J. (2018). Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data. *ISPRS Journal of Photogrammetry and Remote Sensing, 143*, 3–14.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).
TerraSAR-X (2019). https://spacedata.copernicus.eu/web/cscda/missions/terrasar-x access in Sept. 2019.
University of Houston (2019). Ieee grss 2013 data fusion contest. http://www.grss-ieee.org/community/technical-committees/data-fusion/2013-ieee-grss-data-fusion-contest/ access in Sept. 2019.
ISPRS Vaihingen (2019). http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html access in Sept. 2019.
Volpi, M., & Ferrari, V. (2015). Semantic segmentation of urban scenes by learning local class interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 1–9).
Volpi, M., & Tuia, D. (2018). Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing, 144*, 48–60.
Wang, H., Wang, Y., Zhang, Q., Xiang, S., & Pan, C. (2017). Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sensing, 9*, 446.
Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., Xu, Y., & Shibasaki, R. (2018). Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sensing, 10*, 407.
Xu, Y., Du, B., Zhang, F., & Zhang, L. (2018). Hyperspectral image classification via a random patches network. *ISPRS Journal of Photogrammetry and Remote Sensing, 142*, 344–357.
Xu, Y., Wu, L., Xie, Z., & Chen, Z. (2018). Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sensing, 10*, 144.
Yang, S., Chen, Q., Yuan, X., & Liu, X. (2016). Adaptive coherency matrix estimation for polarimetric sar imagery based on local heterogeneity coefficients. *IEEE Transactions on Geoscience and Remote Sensing, 54*, 6732–6745.
Yang, S., Liu, X., Yuan, X., Chen, Q., & Tong, S. (2020). A Unified Coherent-Incoherent Target Decomposition Method for Polarimetric SAR. In *Urban intelligence and applications proceedings of ICUIA 2019* (pp. 69–79). Springer International Publishing.
Yang, B., Luo, W., & Urtasun, R. (2018). PIXOR: Real-time 3D object detection from point clouds. In *IEEE conference on computer vision and pattern recognition* (pp. 7652–7660).
Yousefhussien, M., Kelbe, D. J., Ientilucci, E. J., & Salvaggio, C. (2018). A multi-scale fully convolutional network for semantic labeling of 3D point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing, 143*, 191–204.
Yuan, X., & Sarma, V. (2011). Automatic urban water-body detection and segmentation from sparse alsm data via spatially constrained model-driven clustering. *IEEE Geoscience and Remote Sensing Letters, 8*, 73–77.
Yuan, X., Xie, L., & Abouelenien, M. (2018). A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. *Pattern Recognition, 77*, 160–172.
Yu, S., Jia, S., & Xu, C. (2017). Convolutional neural networks for hyperspectral image classification. *Neurocomputing, 219*, 88–98.
Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *International conference on learning*. representations.
Yu, H., Yang, Z., Tan, L., Wang, Y., Sun, W., Sun, M., & Tang, Y. (2018). Methods and datasets on semantic segmentation: A review. *Neurocomputing, 304*, 82–103.
Zhang, R., Li, G., Li, M., & Wang, L. (2018). Fusion of images and point clouds for the semantic segmentation of large-scale 3D scenes based on deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*.
Zhang, Z., Liu, Q., & Wang, Y. (2018). Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters, 15*, 749–753.
Zhang, C., Liu, J., Yu, F., Wan, S., Han, Y., Wang, J., & Wang, G. (2018). Segmentation model based on convolutional neural networks for extracting vegetation from gaofen-2 images. *Journal of Applied Remote Sensing, 12*, 1–18.
Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine, 4*, 22–40.
Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.
Zhao, W., & Du, S. (2016). Learning multiscale and deep representations for classifying remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing, 113*, 155–165.
Zhao, W., & Du, S. (2016). Spectral-spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Transactions on Geoscience and Remote Sensing, 54*, 4544–4554.
Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., & Torr, P. H. S. (2015). Conditional random fields as recurrent neural networks. In *2015 IEEE international conference on computer vision (ICCV)* (pp. 1529–1537).
Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems* (pp. 487–495).
Zhou, Z., & Gong, J. (2018). Automated residential building detection from airborne LiDAR data with deep neural networks. *Advanced Engineering Informatics, 36*, 229–241.
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*, 8–36.