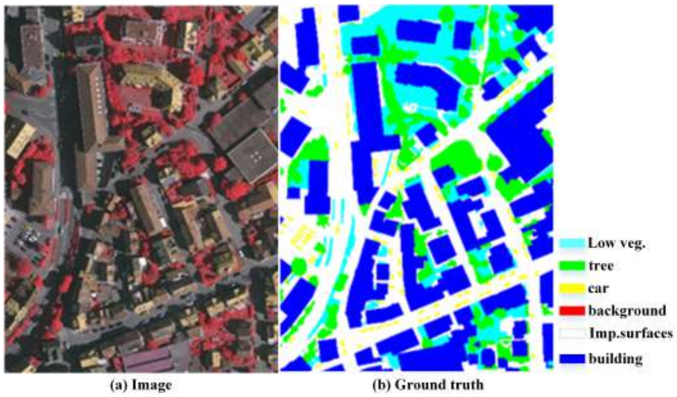
Land-use Mapping for High Spatial Resolution Remote Sensing Image via Deep Learning: A Review

Ning Zang, Yun Cao, Yuebin Wang, *Member, IEEE*, Bo Huang, Liqiang Zhang, *Member, IEEE*, and P. Takis Mathiopoulos, *Senior Member, IEEE*

***Abstract*—Land-use mapping (LUM) using high spatial resolution remote sensing images (HSR-RSIs) is a challenging and crucial technology. However, due to the characteristics of HSR-RSIs, such as different image acquisition conditions and massive, detailed information, performing LUM faces unique scientific challenges. With the emergence of new deep learning (DL) algorithms in recent years, methods to LUM with DL have achieved huge breakthroughs, which offers novel opportunities for the development of LUM for HSR-RSIs. This paper aims to provide a thorough review of recent achievements in this field. Existing high spatial resolution datasets in the research of semantic segmentation and single object segmentation are presented firstly. Next, we introduce several basic DL approaches that are frequently adopted for LUM. After reviewing DL-based LUM methods comprehensively, which highlights the contributions of researchers in the field of LUM for HSR-RSIs, we summarize these DL-based approaches based on two LUM criteria. Individually, the first one has supervised learning, semi-supervised learning, or unsupervised learning, while another one is pixel-based or object-based. We then briefly review the fundamentals and the developments of the development of semantic segmentation and single object segmentation. At last, quantitative results that experiment on the dataset of ISPRS Vaihingen and ISPRS Potsdam are given for several representative models such as FCN and U-Net, following up with a comparison and discussion of the results.**

***Index Terms*—HSR-RSIs, deep learning, land-use mapping, semantic segmentation**

I. Introduction



(a) Image

(b) Ground truth

Fig. 1. Examples of remote sensing images (left) and corresponding land use labels (right) from the ISPRS Vaihingen dataset.

In recent years, HSR-RSIs, including satellite (e.g., IKONOS, Quickbird, SPOT, GaoFen) and airborne (e.g., unmanned aerial vehicle) remote sensing imagery [1], are steadily becoming widespread and available [2]. This paper mainly considering optical images. Accurate and timely LUM for HSR-RSIs plays a significant part in a variety of fields, such as precision agriculture, land use retrieval, and land management [3]–[8]. The essence of LUM for HSR-RSIs is semantic segmentation (or scene segmentation), which is directed to correctly labeling each pixel of the entire image with the corresponding semantic category of what is being represented, as shown in Fig. 1. The land-cover maps are critical products that present the forms of land use and practical use, which have an indispensable referential value for the aggregate plans of land-cover [9].

The complexity of HSR-RSIs increases swiftly as the observation scale turns finer [10] and the details of the objects get richer. This leads to intra-class variability increased while decreasing the inter-class disparity, bringing more challenges to the LUM of HSR-RSIs [11]. On the one hand, diverse imaging conditions usually reduce the separability among different classes [12]. On the other hand, each land parcel used for one purpose often includes multiple categories of land-use with distinct characteristics [13]. Traditionally, on the basis of the spatial unit of representation, artificially designed feature extractor methods that are popular in the past few decades have experienced three stages: pixel-level, object-based, and per-field [14]–[17]. Nevertheless, traditional

Manuscript received February 18, 2021; revised April 5, 2021 and April 23, 2021; accepted May 2, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 41801241, in part by the Fundamental Research Funds for the Central Universities under Grant 292018029 and Grant 375201906, in part by the Key Research and Development Projects of Shanxi Province under Grant 201903D121142, in part by the Open Fund of the State Key Laboratory of Remote Sensing Science under Grant OFSLRSS201923, and in part by the Guizhou Science and Technology Plan Project under Grant Qiankehezhicheng [2020] 4Y022. *(Ning Zang and Yun Cao contributed equally to this work.) (Corresponding author: Yuebin Wang.)*

N. Zang, Y. Cao, and Y. Wang are with the School of Land Science and Technology, China University of Geosciences, Beijing 100083, China (e-mail: zangning97@126.com, cy12160019@163.com, xxgcdxwyb@163.com).

B. Huang is with the Chinese University of Hong Kong, Hong Kong (e-mail: bohuang@cuhk.edu.hk).

L. Zhang is with the Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China (e-mail: zhanglq@bnu.edu.cn).

P. Takis Mathiopoulos is with the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, 15784 Athens, Greece (e-mail: mathio@hol.gr).

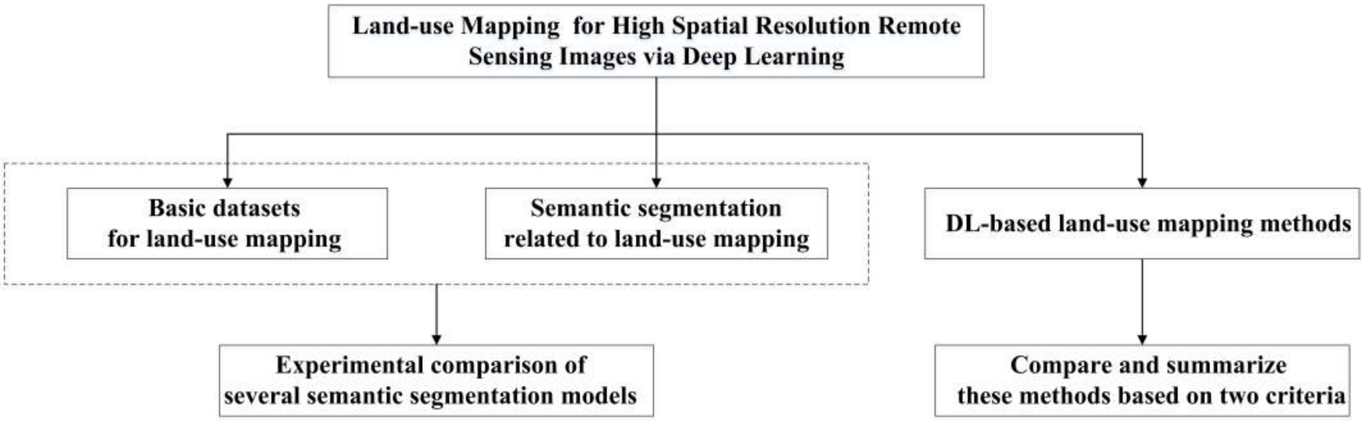

Land-use Mapping for High Spatial Resolution Remote Sensing Images via Deep Learning
Basic datasets for land-use mapping
Semantic segmentation related to land-use mapping
DL-based land-use mapping methods
Experimental comparison of several semantic segmentation models
Compare and summarize these methods based on two criteria

Fig. 2. The framework of this paper.

approaches that utilize hand-crafted features lack the ability to precisely describe features of complex ground objects [4], [18].

And the shallow classifiers lack discrimination because of the small parameter scale [19]. DL is based on deep architectures that are comprised of multiple nonlinear transformations [4], [20]. It emphasizes automatic feature learning from a huge dataset and tries to resolve the problems of feature extraction and classifier design. Recently, deep architectures, such as Convolutional Neural Network (CNN), which have its superiorities in high-level semantic features representation, have indicated tremendous potential in semantic segmentation [21]. In the aspect of segmentation accuracy and even efficiency, CNN greatly surpasses other approaches mentioned previously.

Though new DL techniques have made great contributions in LUM for HSR-RSIs in recent years, to the best of our knowledge, there is still lacking a relatively general and systematic survey that covers the existing methods of this field. This paper, therefore, aims to summarize the development of DL-based LUM methods for HSR-RSIs. Most recently, there also have been a series of reviews related to DL [22]–[25] in remote sensing. These papers and our review all present basic DL models of current-state-of-the-art DL methods and classifiers for remote sensing data. However, the aforementioned reviews [22]–[24] focus on reviewing remotely sensed hyperspectral image classification. The review of [25] mainly concentrates on providing a general framework of DL for RS data analysis, including image processing, high-level semantic feature extraction, scene understanding, and etc. Moreover, our paper mainly focuses on considering optical HSR-RSIs and providing an updated review about widely used DL models for LUM. We further compare several semantic segmentation models that are related to LUM based on two largely used datasets. The main contributions of our paper are summarized as follows:

1) A detailed and in-depth review of the DL-based LUM methods is provided. We also summarize the DL-based methods that are mainly described in this paper from two LUM criteria.

2) We provide an extensive survey of existing datasets, which may be useful for the LUM of HSR-RSIs.

3) Performance evaluation of representative semantic segmentation models is given. The overall performance of LUM has gradually improved, and the U-Net performs best both on the Vaihingen and Potsdam datasets.

This paper (as shown in Fig. 2) is organized as follows: Section 2 introduces high spatial resolution datasets commonly used in the literature for semantic segmentation. In section 3, The related basic DL models for computer vision are given. We then exhaustively review the DL-based LUM methods for HSR-RSIs. In section 4, we summarize these aforementioned methods based on two criteria. The developments of semantic segmentation and single object segmentation related to LUM are described in Section 5. In section 6, the performances of several current-state-of-the-art DL models are compared and discussed on two widely used semantic segmentation benchmarks. In section 7, we conclude this paper and the emerging research trend.

II. Datasets for DL-based LUM

The increasing number of HSR-RSIs enable building large-scale segmentation datasets that play an indispensable part in advance of semantic segmentation. In the past few years, several publicly available HSR-RSIs benchmark datasets have been proposed by different research groups for LUM of remote sensing images [26]–[35].

As a matter of fact, in the following, we firstly illustrate publicly available and the most popular semantic segmentation datasets currently for LUM of HSR-RSIs. Then we describe several single object segmentation datasets for road and building detection. Single object segmentation, a branch of semantic segmentation, extracts a certain kind of object from HSR-RSIs. In order to review datasets comprehensively and clearly, we list these kinds of datasets separately. All datasets pointed here to provide proper pixel-based ground truths. Table I lists 8 publicly available semantic segmentation datasets for LUM of HSR-RSIs. Table II describes single object segmentation datasets for road and building detection especially. Some examples of semantic segmentation and single object segmentation datasets can be found in Fig. 3.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3078631, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

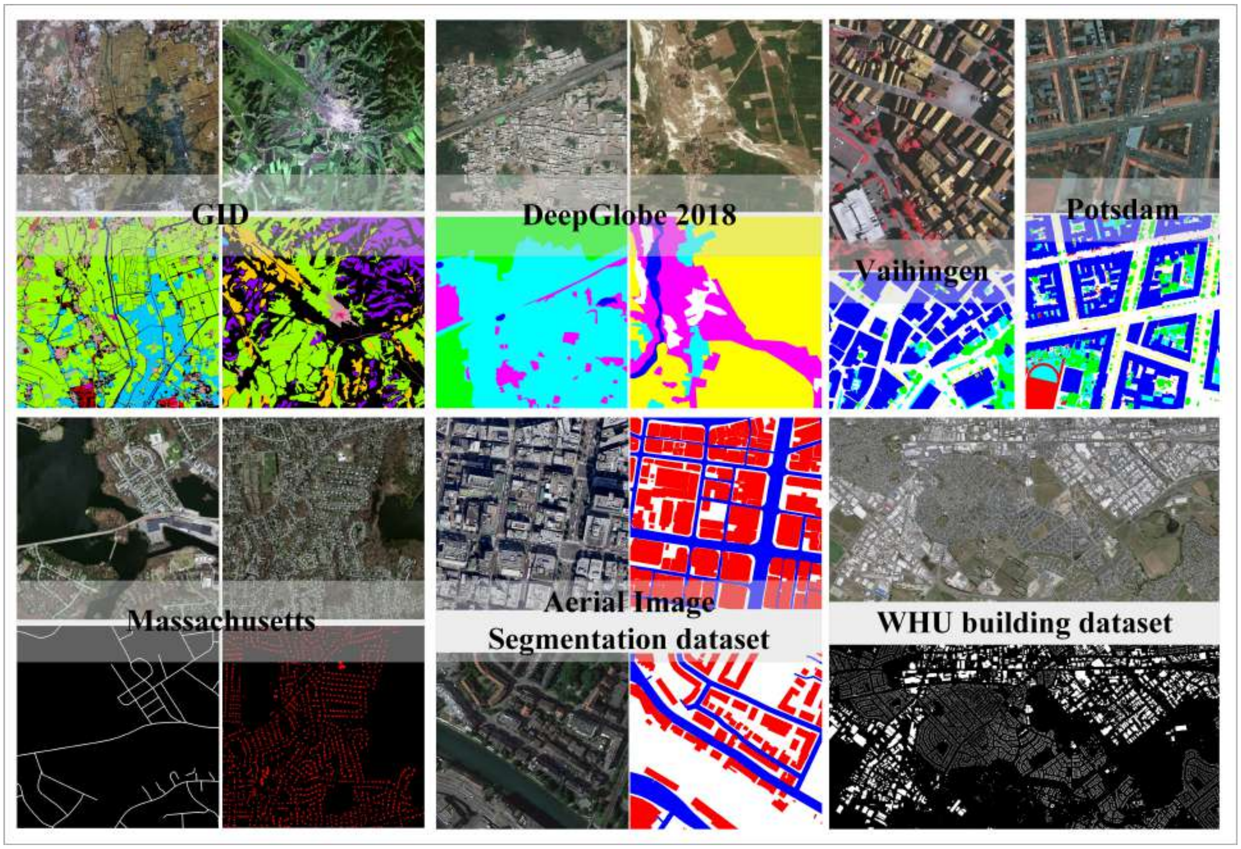


Fig. 3. Some examples of semantic segmentation and single object segmentation datasets.

Zurich Summer dataset [32] is taken from the Quickbird images of Zurich city in 2002. It contains 20 multispectral images at a high resolution of 0.62 meters and is classified into eight classes. The size of the images is 1000 × 1150 approximate.

EvLab-SS dataset [31] contains 35 satellite images and 25 aerial images with a different resolution from 0.1 meters to 2 meters. The average size of the images is 4500 × 4500 approximate. It is classified into 11 major classes.

DeepGlobe Land Cover Classification dataset [36] is a public dataset that focuses on rural areas. It comprises a total of 1146 satellite images, the size of 2448 × 2448. It is divided into training/validation/test sets (803/171/172).

RIT-18 dataset [37] is taken by an unmanned aircraft system (UAS) in Hamlin Beach State Park, New York. The average size of the images is 10000 × 7000 approximately. Each image (0.047 meters of resolution) comprises six bands: near-infrared, red, green, blue, and two other infrared bands.

Gaofen Image Dataset (GID) [38] contains 150 Gaofen-2 images (4 meters) acquired from China. The size of the images is 6800 × 7200 approximately.

2018 IEEE GRSS Data Fusion Contest dataset [36] is obtained from the National Center for Airborne Laser Mapping at a resolution of 0.05 meters. It is classified into 20 classes.

Massachusetts dataset [39], aiming to detect roads and buildings, utilizes images released by the state of Massachusetts state. All images were 3-channel at a resolution of 1 meter and 2250 Km^2 of coverage. The input images and the target maps generated from OpenStreetMap are publicly available. The Massachusetts Building dataset is composed of 151 aerial images, and the roads datasets contain 1171 aerial images.

Buffalo dataset [39] is composed of 30 aerial images of Buffalo city at a resolution of 1 meter and all with a size of 609 × 914.

Inria Aerial Image Labeling dataset [40] covers 810 Km^2. The ground truth is labeled into two semantic categories: buildings or not buildings. It contains 180 aerial images (0.3 meters). The size of the images is 5000 × 5000 .

SpaceNet dataset [41] comes from five SpaceNet regions, and the image size is 650 × 650. These five areas are Rio (0.5 meters), Las Vegas (0.3 meters), Paris (0.3 meters), Shanghai (0.3 meters), and Khartoum (0.3 meters).

AIRS dataset [42] is taken from the city of Christchurch with a very high ground resolution of 0.075 meters. It is composed of 226342 labeled buildings. The ground truth is labeled into two semantic classes: roof and not the roof.

WHU dataset [43] comprises an aerial dataset, satellite dataset I, and satellite II, all with a size 512 × 512 at a spatial resolution of 0.075 meters. It contains about 22000 independent buildings. The satellite dataset I contains 204 images with a different resolution from 0.3 meters to 2.5 meters. The satellite dataset II is cropped into 29085 buildings with a 2.7 meters ground resolution.

TABLE I
8 PUBLICLY AVAILABLE SEMANTIC SEGMENTATION DATASETS FOR LUM OF HSR-RSIs.

Datasets	Image size (pixel)	Number of classes	Resolution (m)	Total image number
ISPRS Vaihingen [35]	About 2500 × 2500	6	0.09	33
ISPRS Potsdam [35]	6000 × 6000	6	0.05	38
Zurich Summer [32]	About 1000 × 1150	8	0.62	20
EvLab-SS [31]	About 4500 × 4500	11	0.1-2	60
RIT-18 [34]	About 10000 × 7000	18	0.047	3
GID [38]	6800 × 7200	5	0.8-10	150
DeepGlobe Land Cover Classification [36]	2448 × 2448	7	0.5	1146
2018 IEEE GRSS Data Fusion Contest [29]	11920 × 12020	20	0.05	14

TABLE II
SINGLE OBJECT SEGMENTATION DATASETS FOR ROAD AND BUILDING DETECTION.

-	Datasets	Image size (pixel)	Resolution (m)	Total image or titles number
Road detection	Massachusetts [39]	1500×1500	1.0	1171
	Buffalo [39]	609×914	1.0	30
	DeepGlobe 2018 [36]	1024×1024	0.5	8570
Building detection	Massachusetts [39]	1500×1500	1.0	151
	Inria Aerial Image Labeling [40]	5000×5000	0.3	180
	SpaceNet (Rio) [41]	650×650	0.5	-
	SpaceNet (Las Vegas)	650×650	0.3	-
	SpaceNet (Paris)	650×650	0.3	-
	SpaceNet (Shanghai)	650×650	0.3	-
	SpaceNet (Khartoum)	650×650	0.3	-
	WHU Aerial imagery dataset [43]	512×512	0.075	8189
	WHU Satellite dataset I (global cities)	512×512	0.3-2.5	204
	WHU Satellite datasetII (East Asia)	512×512	2.7	17388
	AIRS [42]	10000×10000	0.075	1047

III. REVIEW ON BASIC DL METHODS AND LUM METHODS

In this section, we firstly discuss three basic DL methods (i.e., DBN, stacked autoencoder, and CNN) that have been used for LUM. We next review DL-based LUM methods comprehensively. As we all know, there are two criteria to determine which type the LUM method belongs to. As a result, we then compare and summarize the mentioned DL-based LUM methods according to the two criteria.

A. Basic DL methods

For achieving better segmentation results of remote sensing, it cannot lack the related leading approaches to provide solutions. Recently, DL algorithms can offer basic tools for solving this problem. DL has a wide application in varieties of computing vision challenges now, such as semantic segmentation [44], [45], image classification [46], [47], image retrieval [48], [49], and object detection [50], [51]. To distinguish the land-use category of each pixel of the HSR-RSIs (e.g., building, road, and water), LUM is regarded as multi-level semantic segmentation [52]. Three related DL approaches for computer vision are listed below, namely Deep Belief Network (DBN) [53], Stacked (Denoising) autoencoder [54], and CNN [55], all of which have made major contributions to the LUM of HSR-RSIs.

1) DBN: DBN [53] proposed by Hinton et al. has demonstrated robust unsupervised characteristic learning capability in the field of computer vision. Classic DBN structure contains multi-layer restricted Boltzmann machines (RBMs) and a back-propagation (BP) network. Fig. 4 presents the graph architecture of an RBM. RBM is a two-layer neural network that including visible layer and hidden layer. Vector V, H represents the value of the neurons in the visible layer and hidden layer separately. The visible layer and the hidden layer are fully connected, which is similar to deep CNN. Fig. 5

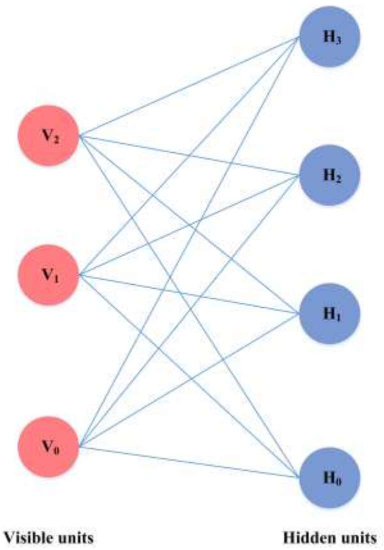


Fig. 4. The graph architecture of RBM.

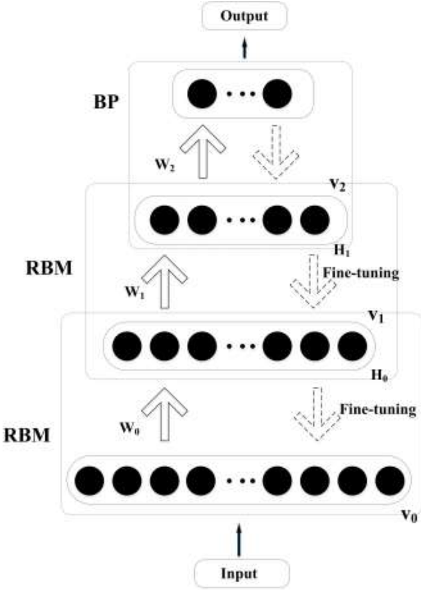


Fig. 5. The graph architecture of classic DBN.

illustrates a classic DBN comprised of stacking multi-layer RBMs and a BP network. W represents the RBM weight matrix. The training process of DBN contains pre-training and fine-tuning. Pre-training is carried out through unlabeled samples in an unsupervised manner. The greedy algorithm is used to optimize each layer during training. The parameters of each layer of RBM are adjusted separately. After training one layer, the output of this layer is regarded as the input of the next layer to continue training the next RBM. After pre-training, supervised learning is used to train the last layer of the BP network. The error is propagated back layer by layer. Finally, the weight of the entire DBN network is fine-tuned through the back-propagation method. It overcomes two problems of long training time and easily falls into local optimal.

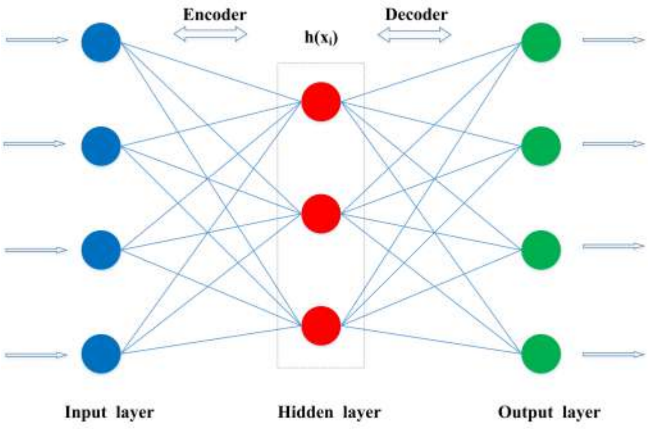


Fig. 6. The graph architecture of Autoencoder.

2) Stacked (Denoising) autoencoder: Stacked autoencoder, the main idea is to train the input of every level of the encoder to learn more powerful feature expression [56]. It contains a multi-layer unsupervised autoencoder presented in Fig. 6, similarly to the way that DBN utilizes a BP network and RBMs. Autoencoder consists of two parts: an encoder and a decoder. The encoder creates a hidden layer (i.e., h(xi)) containing a low-dimensional vector of the information of the input data. The decoder reconstructs the input data from the low-dimensional vector of the hidden layer. The theory of training stacked autoencoder is equal to that previously illustrated for DBN, except that autoencoders are used instead of RBM as the main building block. Unlike RBM, one obvious advantage of the autoencoder is to allow nearly any layers to be parameterized [57]. Some variants of autoencoder include Sparse Autoencoder (SAE) [58], De-noising autoencoder (DAE) [59], [60], and Contractive autoencoder (CAE) [61].

3) CNN: CNN that imitates the biological visual perception mechanism is a kind of neural network model with a deep architecture [20]. It has shown a strong feature learning ability in the computing vision domain. After [62] proposed the AlexNet that outperforms previously proposed models and makes a breakthrough in the contest, there has emerged a series of superior CNN models, such as VGGNet [63], GoogleNet [64], ResNet [65], MoblieNet [66], DenseNet [67], SENet [68], and SKNet [69].

A CNN model usually contains various layers of different functions (see Fig. 7), where conv, pool, and F denote convolutional, pooling, and fully-connected layer. Convolutional layers play a significant role in feature extraction from HSR-RSIs. The operation of the pooling layer can down-sampling the feature map spatially. The fully connected layer aims to perform global features extraction and classification.

In recent years, inspired by the successful breakthrough of DL and the development of computer vision, deep CNN, among the related computer vision methods, has gradually become the leading model in semantic segmentation field [70] and has a significant impact on HSR-RSIs for LUM. As a result, we mainly focus on reviewing CNN-based approaches of LUM for HSR-RSIs in the following content of this paper.

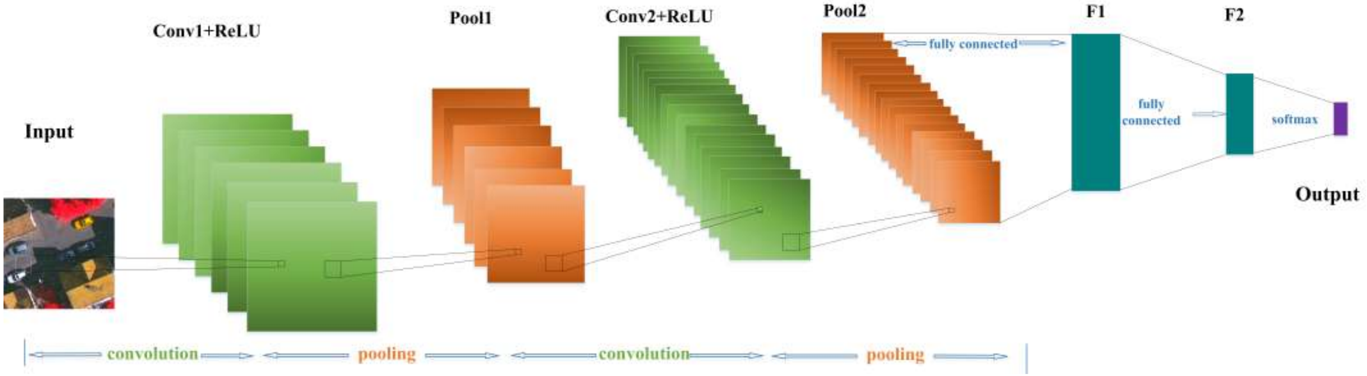


Fig. 7. The graph architecture of a conventional CNN, which contains two convolutional layers (conv1, conv2) followed respectively by a Rectified Linear Units (ReLU) to process the output, two pooling layers (pool1, pool2), two fully connected layers, and a SoftMax layer.

B. DL-based LUM methods

Recently, many researchers have made significant contributions to LUM for HSR-RSIs, driven by its wide range of applications. From DBN to stacked autoencoder and then to CNN, DL models constantly update LUM records. These deep architectures that extract image information in terms of computer vision can greatly improve the accuracy of LUM for HSR-RSIs with a large amount of unknown information, thereby have achieved outstanding performance [4]. Compared with the previously mentioned LUM methods, CNN has the superiority of end-to-end feature learning and the potentiality of learning relevant contextual features automatically. Just being given the input data and output, the end-to-end network can automatically learn “hand-crafted” features that traditional methods have to get from the input data. Meanwhile, CNN explores the complex and high-level visual features hidden in the image, which cannot be extracted by hand-craft features based approaches. After successfully applying CNN with a strong generalization and transferability to large-scale computer vision classification tasks, by 2015, the use of CNN finally stands out in the remote sensing data analysis domain [71], [72]. A great number of CNNs that are considered the most successful DL models have produced the best LUM performance.

Currently, most of the CNN-based LUM approaches can generally be assigned to two main categories: 1) extracting high-level image features through deep CNN; 2) semantic segmentation through the deep end-to-end model. The second method uses the category probability output by the deep model to predict the category of land cover. Specifically, this section is devoted to reviewing LUM approaches for HSR-RSIs based on five kinds of architectures, including DBN or stacked autoencoder, combining CNN and shallow classifier such as Support Vector Machine (SVM) and Logistic Regression (LR), category probability generated by CNN, FCN, also including recently popular transfer learning. Table III gives a summary of LUM methods based on DL.

1) LUM based on DBN or stacked autoencoder: Due to their own limitations, DBN and stacked autoencoder are not widely applied to LUM for HSR-RSIs, and the architectures of these two models are similar, so we review them together. In practical applications, the depth of the DBN network has a significant impact on the classification effect. Higher network depth can discover more abstract feature representations and improve classification performance [73]. However, too many layers may increase training time, reduce network generalization performance and training efficiency. The appropriate network depth is often related to specific applications and datasets. There also is a phenomenon of overfitting [74].

We can also find some works proposed to tempt to tackle these problems. Mnih et al. [75] used DBN detecting roads in high-resolution aerial images, which initially applies the DL model to remote sensing. A method in [76] based on DBN that combines the advantage of supervised learning and unsupervised learning achieved better homogenous mapping results than SVM, neural networks, and stochastic Expectation-Maximization (SEM) in polarimetric synthetic aperture radar data. However, this based on the DBN method cannot directly extract high-dimensional image features, and the learning process is slow [77]. The original stacked autoencoder [78] focused on extracting one-dimensional spectral features that are widely used in hyperspectral images and not enough to support HSR-RSIs classification. Chen et al. applied a stacked autoencoder to hyperspectral remote sensing image LUM [79]. Applied stacked autoencoder to African LUM and got the conclusion that stacked autoencoder has obvious superiority in classification accuracy, predicted time, and LUM performance was done by [80].

2) LUM based on combining CNN and shallow classifier: This kind of approach uses CNN as an image feature exactor to extract high-level semantic information of HSR-RSIs and combines a shallow structure classifier, such as RF and Multilayer Perceptron (MLP) for feature classification [81]–[86]. Razavian et al. [87] showed that training linear SVM classifier on CNN deep feature representation performs better than highly tuned most advanced algorithms in all classification tasks of computer vision on all kinds of datasets. Zhao et al. [81] utilized multi-scale CNN (MCNN) to train LR classifier for initial LUM. MCNN can learn spatial-related deep features combined with spectral features. The ability to learn new spatial features performs better than existing methods such as a multi-index learning approach [88]. Paisitkriangkrai et al. [84] introduced an effecting semantic pixel labeling approach. They used multi-resolution CNN deep features, handcrafted features extracted by RF classifier, and a pixel-level Conditional Random Fields (CRFs) that is applied to the label probabilities for pixel classification in HRS-RSIs. However, these methods are suitable for the situation where the number of samples is limited. Because they do not need to adjust the parameters of the model, they only need to train the shallow classifier that cannot discriminate complex information well.

3) LUM based on category probability generated by CNN: The CNN-based methods use pixels or grids as the unit to assign category labels of HSR-RSIs and combine segmentation or CRF to extract accurate feature edge information. They can make up for the inaccuracy of traditional classifiers in classifying complex HSR-RSIs [89]–[93]. Maggiori et al. [89] used input images to produce classification maps. They train CNN directly and conclude that CNN can be utilized end-to-end to process a great number of satellite images. A multi-layer DL architecture for multi-source multitemporal image classification presented by [90] is based on a pixel level. Volpi et al. [91] introduced a CNN-based approach labeling each pixel on the initial resolution of HSR-RSIs. The novel architecture proposed by [93] follows an hourglass-shaped network (HSN) designed for the per-pixel semantic segmentation of HRS-RSIs. HSN uses down-sampling and up-sampling separately, predicting LUM results. Maggiori et al. [92] designed a novel semantic labeling network architecture called MLP (after multi-layer perceptron). Their experiment shows that such appropriate architecture leads to a win-win situation. An attentive spatial temporal graph convolutional neural network (GCNN) proposed by [94] utilizes spatio-temporal information. It is the first spatial temporal GCNN strategy specifically designed to deal with specific features characterizing HSR-RSIs. However, these CNN-based approaches lack in extracting boundary features accurately [89], [92].

4) LUM based on FCN: As an alternative to the above method, end-to-end CNN models [92], [95] such as FCN [89], [96] do not need to use other classifiers to label the land use of HSR-RSIs. FCN that consists of an encoder-decoder architecture and removes a fully connected layer can predict the correct label maps of the entire input image directly. It also can restrain the fine structure of spatial information without segmentation post-processing. Thus, such a model is more suitable for the LUM of submeter-level ultra-high remote sensing images [97], [98]. Recently, there have emerged a variety of FCN-based LUM methods by exploiting different strategies of FCN. Fu et al. [99] introduced an HSR-RSIs LUM approach based on an improved FCN model. In order to reduce the noise generated by pixel-based LUM, the region boundaries were refined utilizing fully connected CRFs according to the approaches of [100], [101] and so on. Guo et al. [102] coupled a supervised LUM method that relied on an ASPP network with post-processing. This method outperforms the basic FCN and FCN-8s methods of [103], the MLP approach presented in [104], and the ASPP approach introduced by [105] that all performed HRS-RSIs land-use classification and image segmentation successfully. Persello et al. [98] delved into a deep FCN that outperforms state-of-the-art CNNs. They use dilated convolutions of increasing spatial support to detect informal settlements in HRS-RSIs. Sherrah et al. [97] used FCN with no down-sampling to predict aerial imagery labels. To make better use of imagery features, they experiment with fine-tuning a pre-trained CNN. In [106], a non-overlapping grid-based method is proposed to train FCN-8s, which develops a novel framework for better boundary segmentation. A time and memory efficient LUM approach named FastFCN was designed by [107], which also has not a great loss of accuracy than other existing methods when experiments on GID.

However, FCN techniques usually depend on deep multi-scale CNN frameworks, which need numerous trainable parameters [27] and cause a loss of fine resolution details. There also exists lots of redundancy that often lead to vanish gradients in BP and diminish features reutilize in forwarding propagation [108]. In addition, FCN-based LUM approaches do not consider the relationship among pixels and spatial regularity [109]. To resolve these problems, an effective approach is to retain the structure of detailed spatial information [98], [110] obtained through a complementary classification framework instead of the down-sampling process. A hybrid classification method that uses a rule-based fusion scheme, which combines CNN and MLP, was devised by [111]. The integrated classifier MLP-CNN respectively compensates for the restrictions of CNN and MLP. To improve the segmentation accuracy, an object-based CNN (OCNN) combined with small and large windows that mapping on very fine spatial resolution images was introduced by [112]. In practical LUM applications, the number of parameters in the deep CNN (DCNN) increases with layers, which causes ground-truth samples insufficient to train high-quality classifiers.

5) LUM based on transfer learning: Since per-pixel labeled HSR-RSIs are not publicly accessible, they are difficult to get. To decrease the number of images need for training, transfer learning is recognized as a potential method [113]. It can make trained models resolve specific tasks and adapt them to new but related tasks. Therefore, a method based on the semi-shifted DCNN (STDCNN) was devised for multispectral image classification [4]. An unsupervised restricted deconvolution neural network (URDNN) framework that uses an FCN and few labeled pixels was designed by [114]. This model learns pixel-to-pixel LUM for HSR-RSIs. However, there are still some problems in applying deep models to multi-source HSR-RSIs, such as the lack of transferability of the model. To solve this problem, Tong et al. [115] introduced the approach of pseudo-labeling and sample selection. They formulate a hybrid mapping scheme by combining hierarchical segmentation and patch-based mapping.

IV. Summary of DL-based LUM methods

In what follows, an in-depth summary of the above-mentioned LUM approaches, according to two criteria, is carried out. The first criterion is whether the mentioned method belongs to supervised learning, semi-supervised learning, or unsupervised learning. The other one is whether the method is pixel-based or object-based.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3078631, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

TABLE III
SUMMARY OF LUM METHODS BASED ON DL

Method	Reference	Architecture	Contribution(s)
DBN or stacked autoencoder	[116]	RBMs	Layer-wise greedy learning strategy
	[56]	Autoencoder and BP neural network	SAE-based
Combine CNN and shallow classifier	[117]	Multi-scale CNN and LR classifier	Extend traditional CNN to the MCNN
	[83]	CNN and MLP	Combine deep architecture and shallow structure
Category probability generated by CNN	[91]	downsampling-then-upsampling	Learn CNN directly for segmentation
	[93]	Hourglass-Shaped CNN	A weighted belief-propagation post-processing module
FCN	[118]	FCN	FCN-based classification, fully connected CRFs
	[96]	ASPP network	Supervised and pixel-wise classification
	[14]	CNN	Two CNNs with different architectures and windows
Transfer learning	[4]	AlexNet	Transferred DCNN and Small DCNN
	[18]	CNN + ResNet	Patch-wise classification and hierarchical segmentation

A. Supervised Learning, Semi-Supervised Learning, or Unsupervised Learning

There are three related primary types of LUM algorithms applied to HSR-RSIs. A supervised DL model [10], [92], [98], [119] generally requires a massive number of labeled images input for training. Although the CNN model [83], [118] that relied on supervised learning has greatly improved the LUM performance of HSR-RSI. An unsupervised learning approach [120]–[122] that utilizes small amounts of images with no labels is still aroused attention continuously as labeled training samples are not largely available until now. It is used for pre-training that can initialize the parameters to the local minimum. In the domain of remote sensing, the supervised learning approaches used for semantic segmentation are costly in labeling images, while a small amount of labels leads to a decline in the performance of the trained network. Semi-supervised learning techniques [123]–[125], the combination of supervised learning and unsupervised learning, can solve this problem.

1) Supervised learning: In [56], Ding et al. proposed a stacked encoder-based LUM approach. Their experiments show that stacked encoder outperforms artificial neural networks, SVM, decision trees, and a series of nonparametric classifiers [126], [127] verified by the image of GF-1. Paisitkriangkrai et al. [84] used massively available training data to train CNN for learning features. However, pixel-wise labeled HSR-RSIs are not publicly accessible. They are cost-intensive and time-consuming. To overcome these problems, some augmentation techniques such as transfer learning [128], [129] and active learning [130], [131] have been developed. Two main transfer learning methods have been studied: supervised learning and semi-supervised learning methods. In the former method, the training dataset can be used both in the source and target domains is presumed.

In comparison, if only use unlabeled data in the target domain, these methods are defined as semi-supervised. However, semi-supervised approaches do not need strict and standard matching between the domains of source and target [132], [133] but largely rely on the ability of the classifier to learn the structural information of the target domain. Moreover, a lot of complex models contain a huge number of parameters, which easily lead to overfitting and bring greater challenges to train a high-performance classifier.

2) Unsupervised learning: Some unsupervised learning approaches, such as DBN and stacked autoencoder, have been successfully applied to LUM. However, the characteristics of unsupervised learning, training samples with no label bring lots of challenges and limits for the LUM of HSR-RSIs. Fortunately, domain adaptation [134]–[137], a particularly representative approach of transfer learning that tries to harness information from the dataset of other areas where have available labels, aims to compensate for the mismatch between the training images and testing image distributions [138]. Its purpose is to use informative source and fully labeled domain samples to improve performance on an unlabeled target domain [139].

Domain adaption approaches have been applied to unsupervised classification problems [140]. In [141], a domain adaptation algorithm was designed for the LUM of remote sensing images. This framework is based on class centroid and covariance alignment (CCCA) that incorporates spatial knowledge of images. Liu et al. [142] proposed a novel domain adaptation for unsupervised transfer learning, named multikernel jointly domain matching (MKJDM). They perform their experiments on HSR-RSIs and multimodal remote sensing datasets, which shows the performance of LUM is improved than other state-of-the-art domain results. In addition, more recently explored methods [143]–[145] have adopted the adversarial training framework, where the feature network generates domain-invariant features to fool the discriminator that works on image-level. Another method of semantic segmentation based on unsupervised domain adaptation is pseudo-label retraining [146], which finetunes the trained model on the source images by taking high-confident predictions as pseudo ground truth for the unlabeled images. Other approaches for working in an unsupervised manner, such as the framework based on robust manifold matrix factorization and its out-of-sample extension, which achieves competitive clustering accuracy and running time for hyperspectral image classification [147].

3) Semi-supervised learning: These techniques based on semi-supervised learning require few labeled data and plenty of unlabeled data to train the classifier. The core idea of semi-supervised is to try to find a more precise classification criterion than utilizing merely labeled samples [148]–[150]. Laine et al. [151] presented a simple CNN-based approach for training CNN in a semi-supervised way that effectively reduces the classification error. Self-training and co-training are widely used semi-supervised learning techniques [152]. In general, semi-supervised based learning approaches are well-suited for LUM as a great number of unlabeled HSR-RSIs exist.

At present, the research on unsupervised LUM of HSR-RSIs based on DL is still in the development stage. In addition to the difficulty of the problem, its research progress is not as good as other computer vision directions. At the same time, experiments are only performed on some simple datasets and have not been applied to actual scenes on a large scale. Table IV summarizes the training process and classification process of several typical approaches for LUM of HSR-RSIs.

B. Pixel-based or object-based

The pixel-based LUM approaches act on a single pixel. In contrast, the object-based LUM approaches split an HSR-RSI into segmented objects or separated regions as its functional units. As for HSR-RSIs that contain complex details and small objects, it is using the pixel-based LUM approaches may cause poorer interpretation effects owing to the "salt and pepper effect" and lack of semantic meaning of the objects. As a result, pixel-based semantic classification unable to meet the increasing demand for HSR-RSIs. It makes more sense to identify ground objects to efficiently classify HSR-RSIs rather than pixels.

1) Pixel-based: Recently, CNN has been adjusted to perform pixel-based LUM (i.e., semantic segmentation) of HSR-RSIs. In practical applications, there are two methods that use CNN for HSR-RSIs segmentation, as shown in Fig. 8. The first one is a patch-based approach [84], [154], [155] that trains CNN to infer the central pixel of patches segmented from the original input image by looking over the surrounding area.

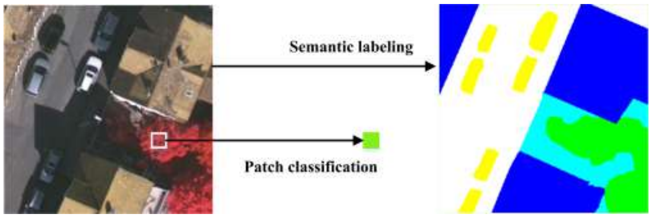


Fig. 8. Comparing patch-based LUM and pixel-to-pixel semantic labeling.

This usually trains small HSR-RSIs patches and then to classify every pixel by utilizing a sliding window way. So obviously, this kind of approach does not apply the whole image as input, which leads to redundant processing and decrease efficiency when predicting labels of large scale HSR-RSIs. The second segmentation method is based on pixel-to-pixel and end-to-end [33], [98], [118], [156]. It can directly infer pixel-based labels of the whole patch or image. Deconvolution, the inverse process of convolution, realizes directly generates the results of per-pixel classification [89], [103], [157]. It takes advantage of the idea of FCN and converts the feature map from a convolutional layer to the original size [114]. Table V illustrates the comparison of patch-based and pixel-to-pixel based LUM for HSR-RSIs.

2) Object-based: Compared with patch-based LUM methods, the object-based approaches can use the segmented image with precise boundaries to classify image objects efficiently. Segmented images have more useful information (e.g., objects' shape and topologies) than pixels and patches of image [10]. An object-based LUM method [4], [10], [14], [84], [96], [158]–[161] consists of two steps [162]: (i) HSR-RSIs segmentation to generate a segmented image (i.e., objects) and (ii) semantic segmentation of the segmented image. The results of semantic segmentation are therefore influenced by the performance of the HSR-RSIs segmentation process. Table VI presents two steps of the object-based classification of several representative papers.

V. Semantic segmentation related to DL-based LUM

Semantic segmentation (or scene segmentation), the essence of LUM for HSR-RSIs, is dedicated to split an input scene or image into its various object components associated with semantic categories that including discrete objects (e.g., car, tree) and stuff (e.g., forest, grass, water, and so on) in computer vision research [164], [165]. The performance of semantic segmentation in natural datasets has been continuously improving, and research outcomes have gradually been applied to the field of remote sensing, especially LUM for HSR-RSIs. It is one of the long-standing and challenging problems. Recently, it has been dramatically improved over the past years thanks to huge breakthroughs of DL models [166]–[169].

The most advanced end-to-end semantic segmentation models have been encouraged mainly by FCN [170], in which the convolutional layer replaces the fully connected layer in standard CNN. They have achieved perfect results on lots of natural datasets, for example, Cityscapes [171]. To mitigate the

TABLE IV
THE TRAINING AND CLASSIFICATION PROCESS OF SEVERAL TYPICAL APPROACHES FOR LUM OF HSR-RSIS.

Sub-class	Reference	Training process	Classification process
Supervised learning	[84]	Train RF classifier on hand-crafted features to complement the features learned by three CNNs with different input image spatial resolutions (i.e., multiresolution CNN).	Vectorizing extracted features and then applying logistic regression weights for classifying objects to different categories.
	[118]	Use images and ground truth to train FCN, and the parameters of FCN are updated based on the error between predict labels and corresponding ground truth.	Perform the trained FCN on test images to roughly predict categories, then use CRF to refine labeling results.
	[14]	Train a large input window CNN and a series of small window CNN models.	Characterize images into functional units, and combine two CNN models and rule-based decision fusion to perform LUM.
	[4]	Use natural images and HSR multispectral images training transferred DCNN and small DCNN, respectively.	Use trained STDCNN classifying processing units.
	[96]	Use the original images and augmented patches to train the ASPP network.	Put test images into trained ASPP network for pixel-wise LUM, then use CRF to refine labeling results.
Unsupervised learning	[153]	Use an unlabeled dataset to train UDCNN (for local structure extraction) and UDFNN (for global semantic feature abstraction).	UDCNN and UDFNN are used to recognize the sliding window.
Semi-supervised learning	[149]	Utilize labeled and unlabeled input images to train ResNet, and a discriminator is also used as an auxiliary network for training.	The residual blocks are utilized to segment images.

TABLE V
THE COMPARISON OF PATCH-BASED AND PIXEL-TO-PIXEL BASED SEMANTIC LABELING FOR HSR-RSIS.

Sub-class	Characteristics	Advantages	Disadvantages
Patch-based	(1) The network ends up with added fully connected layers. (2) Infer the central of patches by looking over the surrounding pixel and then utilize a sliding window to classify every pixel.	-	(1) Do not apply the whole image as input. (2) Cause distortion on the boundary of the classification patch. (3) Not suitable for large scale HSR-RSIs.
Pixel-to-pixel based	(1) The convolutional network removes fully connected layers. (2)End-to-end and pixel-to-pixel classifying whole patch or image directly.	(1) Improve the segmentation performance due to lacking fully-connected layers. (2) Models applied to LUM with different sizes. (3) Avoid the problem of repeated operations of pixels.	-

TABLE VI
TWO STEPS OF OBJECT-BASED CLASSIFICATION.

Reference	Year	(i) HSR-RSIs segmentation	(ii) Semantic segmentation rules
[160]	2014	**A street blocks.**	Using the very fast **multi-threshold segmentation** within eCognition.
[10]	2017	Take image segments as **building blocks**.	For each image object, the optimal statistical method is applied to determine the land cover classes.
[161]	2017	Use **graph-based minimal-spanning-tree** approach to segment images into objects.	Use trained **stacked autoencoders** and **stacked denoising autoencoders** network to classify objects.
[4]	2018	A **skeleton-based decomposition** method splits every street block into processing units with regular shapes.	Utilize trained **STDCNN** classifying processing units into different land-cover categories.
[14]	2018	Characterize the landscape into **linear shape objects** and other **general objects**.	Combine **rule-based decision fusion** and **two CNNs** for LUM.
[96]	2018	Use **Graph-based segmentation** segmenting samples into small patches.	Use FCN to perform **pixel-based** LUM.
[163]	2018	Employ **multiresolution segmentation** method generating highly irregular objects.	Utilizing **CNN** to classify objects.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3078631, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

issue of spatial information loss caused by FCN, Ronneberger et al. [172] presented U-Net that adopts skip connections between each encoder and decoder module. There also are some model variants introduced to enhance contextual aggregation. Noh et al. [44] introduced a deconvolution network to predict segmentation masks. Chen et al. [173] exploited Atrous Spatial Pyramid Pooling (ASPP) that captures object and image context to segment object at multiscale. Based on DeepLabv3 [174], another recent DeepLabv3+ [175] was added a functional decoder module to optimize the segmentation performance. Jegou et al. [176] extended DenseNet [177], namely Fully Convolutional DenseNet, to handle the problem of semantic segmentation issues, which do not need any pre-training and further post-processing module. To decrease the number of training parameters and computational time, Badrinarayanan et al. [178] designed a deep FCN architecture termed SegNet. The SegNet provides better performance in inference memory-wise than the well-known FCN [170], DeepLab-LargeFOV [179], and DeconvNet [44]. The pyramid scene parsing network (PSPNet) provided by [180] achieves the most advanced results on the scene parsing task. To address the problem of lacking pixel-level annotated data, Souly et al. [181] introduced a semi-supervised architecture that contained a generator network to give additional training images. This architecture relies on Generative Adversarial Network (GAN). To increase feature similarity of the same object, [182] explored to spread information in the entire image under the control of the object boundary and proposed unidirectional acyclic graphs (UAGs). Du et al. [183] firstly incorporated DeepLabv3+ and object-based image analysis (OBIA) strategy to label HSR-RSIs, which achieves completive accuracy. Table VII shows the extending CNN architectures for semantic segmentation applied to remote sensing data.

TABLE VII
THE EXTENDING CNN ARCHITECTURES FOR SEMANTIC SEGMENTATION APPLIED TO REMOTE SENSING DATA.

Architecture	Method
FCN	[40], [99], [156], [157], [184], [185]*
SegNet	[186], [187]
U-Net	[188], [189]
DeepLab	[190]

* The idea of DenseNet is used.

Single object segmentation is a branch of semantic segmentation. It extracts a certain kind of object (e.g., building, road, vehicle, and car) from HSR-RSIs based on the given specific features and rules. Building segmentation [188], [191], [192] uses specific criteria of building characteristics such as the shadow that they cast [193], the uniform spectral reflectance values [194], full resolution binary building mask [195], and so on. HSR-RSIs also provide a possibility for segmenting linear features such as road [196]–[198]. There have been proposed a series of CNN-based road segmentation models such as StixelNet [199], FCN [190], and MAP [200]. In the past several years, numerous models for segmenting road, building, and vehicle have been used for practical applications. Amit et al. detected building changes via semantic segmentation to update maps [201]. Mnih et al. [75] proposed to detect the road by using RBMs to initialize the feature detectors. In [202], a novel algorithm that classifies on-board images was presented. This method trains a general dataset to generate training labels and segments road areas in an individual image. Buslaev et al. [203] proposed an FCN that consists of ResNet-34 and the decoder to automatic extract road. Nicolas et al. [204] applied SegNet to vehicle detection and segmentation in remote sensing images. In [205], a global context based dilated CNN (GC-DCNN) that is similar to the structure of U-Net was proposed, which aims to address the challenges of complex backgrounds and view occlusions of buildings and trees around a road when segmenting road.

The essential difference between semantic segmentation and single object segmentation is that single object segmentation belongs to binary classification. Their data input and network architecture are the same. The activation function of the last layer of semantic segmentation is softmax, while single object segmentation is sigmoid. The loss functions of these two kinds of segmentation are also different. The loss function of semantic segmentation is categorical_crossentropy, while the other loss function is binary_crossentropy. A comparison of semantic segmentation and single object segmentation can be concisely illustrated in Table VIII.

VI. Performance Comparison and Discussion

In recent years, a variety of semantic segmentation models have been proposed. We select four state-of-the-art architectures, including SegNet [178], U-Net [172], FCN-32s [103], and FCN-8s [103] to compare segmentation performance. They all take VGG-16 as the backbone. We evaluate them on two widely used datasets of ISPRS Vaihingen and Potsdam. Because DL-based semantic segmentation models rely on large-scale data, we augment training samples.

A. Datasets

The ISPRS 2D semantic labeling contest dataset contains aerial images of Vaihingen and Potsdam cities in Germany. Each dataset is labeled into six categories: impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. The background category contains water bodies and other objects (such as containers) that are different from other defined categories. These objects usually belong to uninteresting semantic objects in urban scenes.

The Vaihingen dataset: It comprises 33 orthophoto titles at a ground resolution of 9 cm. 16 of them are labeled. 17 of them are used as the test image. The size of the images is 2100×2100 approximate. The single image comprises three bands: near-infrared, red, and green (IRRG).

The Potsdam dataset: It contains 38 orthophoto titles (5 cm). 24 of them are labeled. 14 of them are used as the test images. The size of the tiles is 6000×6000. The single image comprises four bands: red, green, blue, and near-infrared bands (RGB-IR).

TABLE VIII
COMPARISON OF SEMANTIC SEGMENTATION AND SINGLE OBJECT SEGMENTATION.

Methods	Characteristics	Network architecture	The activation function of the last layer	Loss function
semantic segmentation	Classify each pixel into a corresponding category	Encoder-decoder	SoftMax	Categorical_crossentropy
Single object segmentation	Binary classification	Encoder-decoder	Sigmoid	Binary_crossentropy

B. Evaluation metrics

we use three confusion metrics, including Kappa coefficient, overall accuracy (OA), and user's accuracy [206]. Let k denote the number of categories, and let N be the total number of pixels, let N_{ij} denote the number of pixels that should be of class i but are predicted to be of class j, let N_{i+} be the total number of pixels of class i in the test images, let N_{+j} be the number of pixels predicted to class j. The metrics are defined as follow:

Kappa coefficient: It evaluates the inter-rater consistency and reliability for the segmentation result.

$$Kappa = \frac{N\sum_{i=1}^{k} N_{ii} - \sum_{i=1}^{k} N_{i+} * N_{+j}}{N^2 - \sum_{i=1}^{k} N_{i+} * N_{+j}} \tag{1}$$

where

$$N_{i+} = \sum_{j=1}^{k} N_{ij} \tag{2}$$

$$N_{+j} = \sum_{\mathrm{i}=1}^{k} N_{ij} \tag{3}$$

Overall accuracy: It is a metric that measures the number of truly classified pixels divided by the total pixels of the whole test image.

$$OA = \frac{\sum_{i=1}^{k} N_{ii}}{N} \tag{4}$$

User's accuracy: It refers to the possibility that the corresponding ground truth category is i when the classifier is assumed to classify pixels into category i.

$$\text{user's accuracy} = \frac{N_{ii}}{N_{i+}} \tag{5}$$

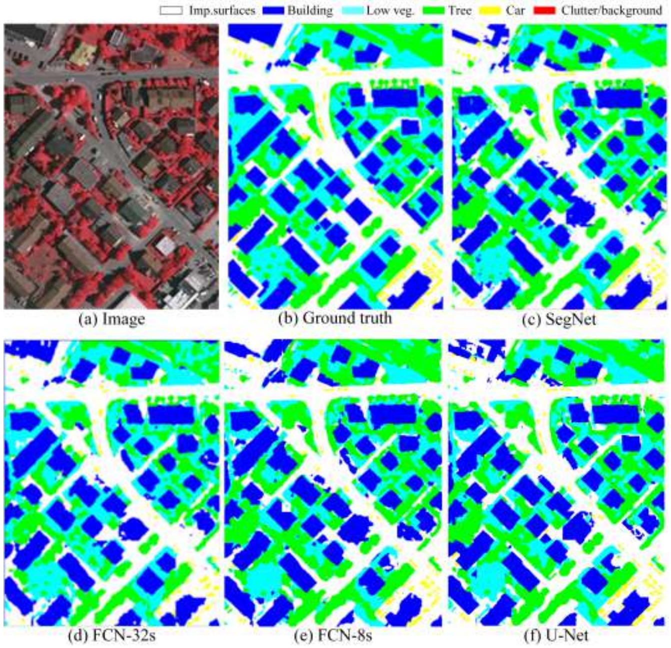
C. Implementation details

We split the labeled images of the Vaihingen dataset into a training dataset (12 images of ID 1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28) and a test dataset (4 images of ID 30, 32, 34, 37). We randomly crop the training images into a size of 256 × 256 and flip and rotation images for data augmentation. Thus, we can obtain 12000 patches for the process of training. We train these four models with a batch size of 16 and other same hyper-parameters setting, except that we use different learning rates for different models. For the Potsdam dataset, we also divided the labeled images into a test dataset (6 images of ID 2_10, 3_10, 4_10, 5_10, 6_10, 7_10) and a training dataset (the remaining 18 images). Then, we can obtain 14000 256 × 256 patches for training when performs the same process as the Vaihingen dataset. Pixels of clutter/background occupy a tiny percentage. Therefore, we report the accuracy of the remaining five classes merely. We are training from scratch of models without bells and whistles. All performed experiments are conducted in the TensorFlow framework with the platform of an NVIDIA 2080Ti GPU.

D. Experimental results

1) Vaihingen dataset: The accuracy results of the semantic segmentation of the four models are listed in Table IX. We also visualize the results of ID 30, as displayed in Fig. 9, to more easily compare the semantic segmentation performance of different models. As can we see from Table IX, the overall performance of LUM of the ISPRS Vaihingen dataset has gradually improved, though small objects such as cars show a relatively low accuracy. We observe that the U-Net achieved the best OA of 86.08% and Kappa of 0.740, but the segmentation maps of the building are jagged at the edge. As for FCN, the boundary of the object is blurred, and the result is reduced visually, so it is usually impossible to detect objects that are small or with many boundaries. But the performance of the car category in FCN-8s (user's accuracy: 50.55%) is higher than in FCN-32s (user's accuracy: 13.05%), which demonstrates it is significant for segmenting small areas of low-level features. The category of impervious (user's accuracy: 85.93%) also illustrates a definite response that probably interferes with the car region in the SegNet network (OA:79.73%, Kappa: 0.727).

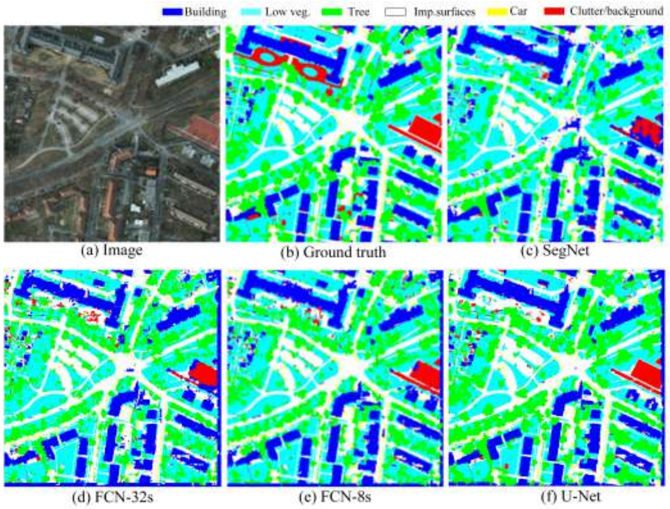
2) Potsdam dataset: In order to compare these four models comprehensively, we also experiment on the ISPRS Potsdam dataset, and quantitative maps are listed in Table X. We also visualize the results of ID 3_10, as displayed in Fig. 10. The LUM results of the Potsdam dataset are slightly worse than those of Vaihingen in general. The difference between these four models is mainly reflected in the performance of segmenting small objects such as car, which is the case for both datasets, as shown in Fig. 11. The U-Net (OA:80.86%, Kappa: 0.524) model is superior to the other three models, but it cannot completely recognize the boundaries of the car category with various appearances of vision. The FCN-8s (OA:76.54%, Kappa: 0.702), the best accurate model in the FCN series [207], is a bit more accurate than SegNet (OA:75.32%, Kappa: 0.693) in this experiment. However, it



(a) Image (b) Ground truth (c) SegNet

(d) FCN-32s (e) FCN-8s (f) U-Net

Fig. 9. Visualization results of the ID 30 of the Vaihingen dataset. (a) is a raw image. (b) is corresponding ground truth. (c) (d) (e) (f) are the semantic segmentation results of SegNet, FCN-32s, FCN-8s, and U-Net, respectively.



(a) Image (b) Ground truth (c) SegNet

(d) FCN-32s (e) FCN-8s (f) U-Net

Fig. 10. Visualization results of the ID 3_10 of the Potsdam dataset. (a) is raw image. (b) is the corresponding ground truth. (c) (d) (e) (f) are the semantic segmentation results of SegNet, FCN-32s, FCN-8s, and U-Net, respectively.

mistakes certain impervious surface regions for buildings. The predicted mapping results outputted by FCN-32s (OA:75.24%, Kappa: 0.665) are rough and easy to lose relatively small objects (car: 15.40% and tree: 65.61% user's accuracy).

E. Discussion

As we have observed from Table IX, Table X, and Fig.11, the performance of LUM for HSR-RSIs has been successfully advanced as the continuous breakthrough of semantic segmentation models. DL-based LUM methods were mainly based on FCN during the early stages, and researchers often utilize the ISPRS Vaihingen and Potsdam datasets [97], [189], [208]–[211] and the DeepGlobe land cover classification dataset [212]–[217] to perform FCN extended algorithms evaluation. Until now, FCN-based approaches are still promising on the semantic segmentation datasets of HSR-RSIs for directly predict semantic labels of input images, which shows end-to-end networks have got exceeding success under the BP. Nevertheless, if the amounts of test images are much smaller or larger than the training images, the mapping results are worse because the fusion strategy adds pool features of the previous layer, which results in high-level features not being used well. Blurry object boundary is also a usual problem in the mapping results of FCN owing to downsampling operations ignore local information. Moreover, though FCN-8s performs better than FCN-16s and FCN-32s, which indicates that the shallow predicted results contain more detailed features, the labeling maps of FCN-8s and FCN-32s are relatively rough than SegNet and U-Net, as shown in Fig. 9 and Fig. 10. Therefore, some researchers illustrated a series of approaches to improve the LUM results, such as combining with DSM [156], [218].

Fortunately, other modified and extended variants based on FCN, such as encoder-decoder structures (notably SegNet and U-Net), are remarkable, aiming to make transforms more suitable for semantic segmentation. Furthermore, researchers demonstrate that encoder-decoder structures trained with ImageNet weights are more easily transferred to the remote sensing domains [186]. To restore the feature map to the original input image size, SegNet utilizes max-pooling indices for nonlinear upsampling, while FCN uses deconvolution. SegNet that adopts dilated convolution to decrease local information loss can also balance performance and computational cost. As a result, SegNet takes up less memory and provides a competitive inference time than FCN. However, the edge problem is serious when the sliding window is too large in segmentation. Specifically, the splicing edges of the predicted labels are obvious. When SegNet performs semantic segmentation, a CRF module is usually used to refine the output results.

As for U-Net, skip connection can concatenate high-level semantic and low-level fine-grained information, which meets the requirements of semantic segmentation. To increases output resolution, it also uses unsampled operators to replace pooling operators. The biggest advantage of U-Net is that it can be trained well with small scale datasets, which is more suitable for the current lacking sufficient training labeled HSR-RSIs of semantic segmentation. But it still needs data augmentation, just like the comparative experiment in this paper. U-Net also cannot expand the difference between classes because it does not fully explore all level semantic information. This results in limiting its applications in semantic segmentation for HSR-RSIs. Hence, novel end-to-end models such as CSE-UNet [219] are proposed to resolve these inter-class homogeneity challenges.

With several publicly available datasets (e.g., Zurich Summer dataset, EvLab-SS, and GID) releasing, it becomes easier to compare semantic segmentation models comprehensively. And the performance of DL-based approaches also highly lies in the amounts of training images. Therefore, sample-driven

TABLE IX
Semantic mapping results of the ISPRS Vaihingen dataset.

Models	User's Accuracy (%)						
	OA (%)	Kappa	Imp. Surfaces	Building	Low Veg.	Tree	Car
FCN-32s [103]	74.57	0.662	75.61	86.53	58.85	80.62	13.05
SegNet [178]	79.73	0.727	85.93	88.11	63.32	79.42	48.57
FCN-8s [103]	81.68	0.761	82.45	90.92	79.38	76.24	50.55
U-Net [172]	86.08	0.740	86.47	91.63	80.58	72.81	52.71

TABLE X
Semantic mapping results of the ISPRS Potsdam dataset.

Models	User's Accuracy (%)						
	OA (%)	Kappa	Imp. Surfaces	Building	Low Veg.	Tree	Car
FCN-32s [103]	75.24	0.665	76.72	85.52	67.44	65.61	25.40
SegNet [178]	75.32	0.693	84.53	93.41	70.11	70.14	77.78
FCN-8s [103]	76.54	0.702	85.10	93.25	62.71	69.96	82.28
U-Net [172]	80.86	0.524	85.69	90.41	74.50	74.66	51.20

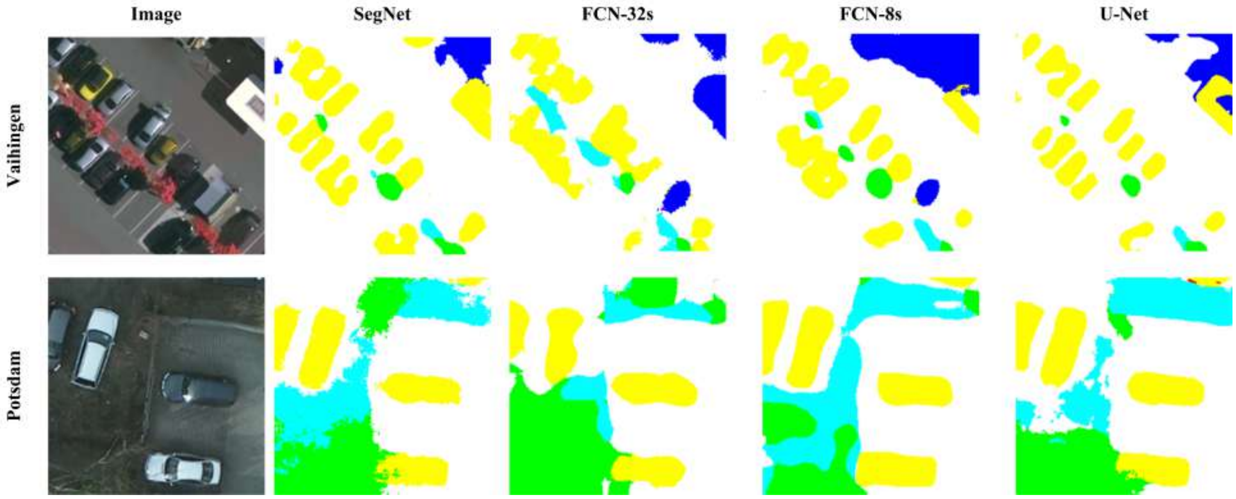


Fig. 11. Visualization analysis results of car category. The first row is in the Vaihingen dataset, while the second row is the results of the Potsdam dataset.

semantic segmentation schemes can be further promoted by constructing large scale and challenging HSR-RSIs datasets. Generally, transferring the successful experience of semantic segmentation models from computer vision to the remote sensing domain is also an urgent and challenging task for improving the performance of LUM for HSR-RSIs.

VII. Conclusions

LUM of HSR-RSIs has obtained significant achievements through several decades of rapid development. To our knowledge, the number of papers on LUM of HSR-RSIs, especially about DL-based methods, is breathtaking. This paper is the first one that focuses on exhaustively reviewing LUM approaches based on the rising topic of DL, covering the current work in this field. We have also compared and discussed the quantitative performance of such representative models. The performance of these models proves their effectiveness in resolving practical issues, though it has not yet reflected the full potential of DL.

Due to the increased availability of the HSR-RSIs dataset and computational resources of DL, it is expected that DL rapidly develops in the LUM of HSR-RSIs in the next few years. Nevertheless, the research in DL-based LUM for HSR-RSIs is still immature and remains many unanswered questions. It is quite a long way to reach its full potential when addressing numerous unsolved challenges. Currently, the difficulties and key points of LUM focus on the lack of labeled training samples, small object segmentation, and accurate edge segmentation. Thus, the following are several potentially interesting topics in the LUM for HSR-RSIs.

1) The complexity of HSR-RSIs: Unlike natural scene images, each land parcel used for one purpose of HSR-RSIs often includes multiple categories of land-use with distinct characteristics. The complexity of HSR-RSIs increases, leading to the difficulties of learning discriminative features from image scenes with DL algorithms.

2) The number of labeled training images: The existing datasets for LUM mostly cover a small area and concentrated locations, which cannot fully reflect the true distribution of ground truth. It also limits the discriminative ability of the CNN model that relies heavily on the quality and quantity of the training images. In addition, the labeled training HSR-RSIs are not largely available until now. In this case, maintaining the representation learning performance of the DL-based approaches with fewer labeled training samples is still a huge challenge. Based on this problem, semi-supervised learning, weakly supervised learning, and unsupervised learning methods have great potential.

3) Small object segmentation and edge segmentation: The present work to improve accuracy is close to saturation, and as a result, researches mainly focus on obtaining accurate small object segmentation performance and high-quality boundaries [220]–[222].

References

[1] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "A review of supervised object-based land-cover image classification," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 130, no. aug., pp. 277–293, 2017.

[2] A. S. Belward and J. O. Skoien, "Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 103, no. may, pp. 115–128, 2015.

[3] C. Yao, Y. Zhang, and H. Liu, "Application of convolutional neural network in classification of high resolution agricultural remote sensing images," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-2/W7, pp. 989–992, 2017.

[4] H. Bo, Z. Bei, and S. Yimeng, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sensing of Environment*, vol. 214, pp. 73–86, 2018.

[5] J. P. Ardila, V. A. Tolpekin, W. Bijker, and A. Stein, "Markov-random-field-based super-resolution mapping for identification of urban trees in vhr images," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 6, pp. 762–775, 2011.

[6] O. O. Asli, O. Ali, and S. Konrad, "Mapping of agricultural crops from single high-resolution multispectral images—data-driven smoothing vs. parcel-based smoothing," *Open Remote Sensing Journal*, vol. 7, no. 5, pp. 5611–5638, 2015.

[7] C. Zhang and J. M. Kovacs, "The application of small unmanned aerial systems for precision agriculture: A review," *Precision Agriculture*, vol. 13, no. 6, 2012.

[8] H. Shi, L. Chen, F. K. Bi, H. Chen, and Y. Yu, "Accurate urban area detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 9, pp. 1948–1952, 2015.

[9] X. Yang, Z. Chen, B. Li, D. Peng, and B. Zhang, "A fast and precise method for large-scale land-use mapping based on deep learning," 2019.

[10] W. Zhao, S. Du, and W. J. Emery, "Object-based convolutional neural network for high-resolution imagery classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 7, pp. 3386–3396, 2017.

[11] L. Bruzzone and L. Carlin, "A multilevel context-based system for classification of very high spatial resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 9, pp. 2587–2600, 2006.

[12] T. Devis, C. V. Gustau, and D. Zhaohong, "Kernel manifold alignment for domain adaptation," *Plos One*, vol. 11, no. 2, p. e0148655, 2016.

[13] B. Zhao, Y. Zhong, and L. Zhang, "A spectral–structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 116, no. jun., pp. 73–85, 2016.

[14] Kelvin, Chew, Preetha, and Thulasiramany, "An object-based convolutional neural network (ocnn) for urban land use classification," 2017.

[15] Q. Zhu, Y. Zhong, B. Zhao, and G. S. Xia, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, 2016.

[16] X. Huang, H. Liu, and L. Zhang, "Spatiotemporal detection and analysis of urban villages in mega city regions of china using high-resolution remotely sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3639–3657, 2015.

[17] M. Voltersen, C. Berger, S. Hese, and C. Schmullius, "Object-based land cover mapping and comprehensive feature calculation for an automated derivation of urban structure types at block level," *Remote Sensing of Environment*, vol. 154, pp. 192–201, 2014.

[18] X. Y. Tong, G. S. Xia, Q. Lu, H. Shen, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.

[19] F. Hu, G. S. Xia, and L. Zhang, "Deep sparse representations for land-use scene classification in remote sensing images," in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, 2016.

[20] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[21] W. Zhao, S. Du, Q. Wang, and W. J. Emery, "Contextually guided very-high-resolution imagery classification with semantic segments," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 132, no. oct., pp. 48–60, 2017.

[22] S. Li, W. Song, L. Fang, Y. Chen, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–20, 2019.

[23] N. Audebert, B. L. Saux, and S. Lefèvre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, 2019.

[24] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, no. Dec., pp. 279–317, 2019.

[25] L. Zhang, L. Zhang, and D. Bo, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, 2016.

[26] H. Li, K. Qiu, L. Chen, X. Mei, and C. Tao, "Scattnet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. PP, no. 99, pp. 1–5, 2020.

[27] Q. Liu, M. Kampffmeyer, R. Jessen, and A. B. Salberg, "Dense dilated convolutions merging network for land cover classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 9, pp. 6309–6320, 2020.

[28] N. Yokoya, P. Ghamisi, J. Xia, S. Sukhanov, R. Heremans, I. Tankoyeu, B. Bechtel, B. Le Saux, G. Moser, and D. Tuia, "Open data for global multimodal land use classification: Outcome of the 2017 ieee grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, no. 5, pp. 1–15, 2018.

[29] Y. Xu, B. Du, L. Zhang, D. Cerra, M. Pato, E. Carmona, S. Prasad, N. Yokoya, R. Hänsch, and B. Le Saux, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 ieee grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 6, pp. 1709–1724, 2019.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3078631, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

[30] J. Gong, X. Hu, S. Pang, and K. Li, "Patch matching and dense crf-based co-refinement for building change detection from bi-temporal aerial images," *Sensors*, vol. 19, no. 7, 2019.

[31] M. Zhang, X. Hu, L. Zhao, Y. Lv, M. Luo, and S. Pang, "Learning dual multi-scale manifold ranking for semantic segmentation of high-resolution images," *Remote Sensing*, vol. 9, no. 9, 2017.

[32] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Computer Vision and Pattern Recognition Workshops*, 2015.

[33] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6054–6068, 2017.

[34] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, no. NOV., pp. 60–77, 2018.

[35] M. Gerke, F. Rottensteiner, J. D. Wegner, and G. Sohn, "Isprs semantic labeling contest," in *PCV - Photogrammetric Computer Vision*, 2014.

[36] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," 2018.

[37] B. Ksenia, A. Fathalrahman, C. Shiyong, K. Marco, and R. Peter, "Building footprint extraction from vhr remote sensing images combined with normalized dsms using fused fully convolutional networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2615–2629, 2018.

[38] X. Tong, G. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Learning transferable deep models for land-use classification with high-resolution remote sensing images," *CoRR*, vol. abs/1807.05713, 2018. [Online]. Available: http://arxiv.org/abs/1807.05713

[39] Mnih and Volodymyr., "Machine learning for aerial image labeling." Ph.D. dissertation.

[40] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *Igarss IEEE International Geoscience and Remote Sensing Symposium*, 2017.

[41] N. Weir, D. Lindenbaum, A. Bastidas, A. V. Etten, and H. Tang, "Spacenet mvoi: a multi-view overhead imagery dataset," 2020.

[42] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "Temporary removal: Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 147, no. JAN., pp. 42–55, 2019.

[43] Shunping, Ji, Shiqing, Wei, Meng, and Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2019.

[44] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1520–1528.

[45] S. Song and J. Xiao, "Sliding shapes for 3d object detection in depth images," in *European Conference on Computer Vision*, 2014.

[46] Y. Zhu and S. Newsam, "Land use classification using convolutional neural networks applied to ground-level images," pp. 1–4, 2016.

[47] M. Jianwen and H. Bagan, "Land-use classification using aster data and self-organized neutral networks," *International Journal of Applied Earth Observation and Geoinformation*, vol. 7, no. 3, pp. 183–188, 2005.

[48] Cao, Zhang, Zhu, Li, Li, Liu, and Qiu, "Enhancing remote sensing image retrieval using a triplet deep metric learning network," *International Journal of Remote Sensing*, 2020.

[49] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," 2014.

[50] W. Ouyang, X. Zeng, X. Wang, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, and H. a. Li, "Deepid-net: Object detection with deformable part based convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 7, pp. 1320–1334, 2017.

[51] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. V. Gool, "Weakly supervised cascaded convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[52] C. Tian, C. Li, and J. Shi, "Dense fusion classmate network for land cover classification," 2019.

[53] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets." *Neural Computation*, 2006.

[54] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 1096–1103. [Online]. Available: https://doi.org/10.1145/1390156.1390294

[55] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 38, no. 2, pp. 295–307, 2016.

[56] A. Ding and X. Zhou, "Land-use classification with remote sensing image based on stacked autoencoder," in *International Conference on Industrial Informatics-computing Technology*, 2017.

[57] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–13, 2018.

[58] X. Jiang, Y. Zhang, W. Zhang, and X. Xiao, "A novel sparse auto-encoder for deep unsupervised learning," in *Sixth International Conference on Advanced Computational Intelligence*, 2013.

[59] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, 2008.

[60] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. 12, pp. 3371–3408, 2010.

[61] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *ICML*, 2011.

[62] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[64] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, and A. Rabinovich, "Going deeper with convolutions," 2014.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[66] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.

[67] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[68] H. Jie, S. Li, and S. Gang, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[69] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 510–519.

[70] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[71] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, 2016.

[72] X. X. Zhu, D. Tuia, L. Mou, G. S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2018.

[73] N. L. Roux and Y. Bengio, "Representational power of restricted boltzmann machines and deep belief networks," *Neural Computation*, vol. 20, no. 6, pp. 1631–1649, 2008.

[74] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 1, no. 10, pp. 1–40, 2009.

[75] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI*, 2010.

[76] L. Qi, D. Yong, N. Xin, X. Jiaqing, X. Jinbo, and X. Fei, "Urban land use and land cover classification using remotely sensed sar data through deep belief networks," *Journal of Sensors*, vol. 2015, pp. 1–10, 2015.

[77] L. Dawei, H. Ling, and H. Xiaoyong, "High spatial resolution remote sensing image classification based on deep learning," *Acta Optica Sinica*, vol. 36, p. 4, 2016.

[78] W. Li, H. Fu, L. Yu, P. Gong, D. Feng, C. Li, and N. Clinton, "Stacked autoencoder-based deep learning for remote-sensing image classification: a case study of african land-cover mapping," *International Journal of Remote Sensing*, vol. 37, no. 23-24, pp. 5632–5646, 2016.

[79] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2094–2107, 2014.

[80] H. Gao, J. Guo, P. Guo, and X. Che, "Classification of very-high-spatial-resolution aerial images based on multiscale features with limited semantic information," *Remote Sensing*, vol. 13, no. 3, p. 364, 2021.

[81] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 113, no. Mar., pp. 155–165, 2016.

[82] Wenzhi, Zhao, Zhou, Guo, Jun, Yue, Xiuyuan, Zhang, Liqun, and Luo, "On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery: International journal of remote sensing: Vol 36, no 13," *International Journal of Remote Sensing*, 2015.

[83] C. Zhang, X. Pan, H. Li, A. Gardiner, I. Sargent, J. Hare, and P. M. Atkinson, "A hybrid mlp-cnn classifier for very fine resolution remotely sensed image classification," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 140, no. JUN., pp. 133–144, 2018.

[84] S. Paisitkriangkrai, J. Sherrah, P. Janney, and V. D. Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Computer Vision and Pattern Recognition Workshops*, 2015.

[85] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. V. D. Hengel, "Semantic labeling of aerial and satellite imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 7, pp. 2868–2881, 2017.

[86] N. Audebert, B. L. Saux, and S. Lefèvre, "How useful is region-based classification of remote sensing images in a deep learning framework?" in *Geoscience and Remote Sensing Symposium*, 2016.

[87] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *2014 IEEE conference on computer vision and pattern recognition workshops*, 2014.

[88] X. Huang, Q. Lu, and L. Zhang, "A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 90, no. APR., pp. 36–48, 2014.

[89] Emmanuel, Maggiori, Yuliya, Tarabalka, Guillaume, Charpiat, Pierre, and Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 645–657, 2017.

[90] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. PP, no. 99, pp. 1–5, 2017.

[91] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, 2017.

[92] X. Sun, S. Shen, X. Lin, and Z. Hu, "Semantic labeling of high-resolution aerial images using an ensemble of fully convolutional networks," *Journal of Applied Remote Sensing*, vol. 11, no. 4, 2017.

[93] L. Yu, M. N. Duc, D. Nikos, D. Wenrui, and M. Adrian, "Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery," *Remote Sensing*, vol. 9, no. 6, p. 522, 2017.

[94] A. M. Censi, D. Ienco, Y. Gbodjo, R. G. Pensa, and R. Gaetano, "Attentive spatial temporal graph cnn for land cover mapping from multi temporal remote sensing data," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2021.

[95] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. PP, no. 99, pp. 1–5, 2017.

[96] G. Rui, L. Jianbo, L. Na, L. Shibin, C. Fu, C. Bo, D. Jianbo, L. Xinpeng, and M. Caihong, "Pixel-wise classification method for high resolution remote sensing imagery using deep neural networks," *International Journal of Geo-Information*, vol. 7, no. 3, p. 110, 2018.

[97] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," *CoRR*, vol. abs/1606.02585, 2016. [Online]. Available: http://arxiv.org/abs/1606.02585

[98] C. Persello and A. Stein, "Deep fully convolutional networks for the detection of informal settlements in vhr images," *IEEE Geoscience and Remote Sensing Letters*, 2017.

[99] G. Fu, C. Liu, R. Zhou, T. Sun, and Q. Zhang, "Classification for high resolution remote sensing imagery using a fully convolutional network," *Remote Sensing*, vol. 9, no. 6, p. 498, 2017.

[100] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015.

[101] S. Paisitkriangkrai, J. Sherrah, P. Janney, and V. D. Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Computer Vision and Pattern Recognition Workshops*, 2015.

[102] G. Rui, L. Jianbo, L. Na, L. Shibin, C. Fu, C. Bo, D. Jianbo, L. Xinpeng, and M. Caihong, "Pixel-wise classification method for high resolution remote sensing imagery using deep neural networks," *International Journal of Geo-Information*, vol. 7, no. 3, p. 110, 2018.

[103] Long, Jonathan, Shelhamer, Evan, Darrell, and Trevor, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[104] X. Yao, H. Yang, Y. Wu, P. Wu, and S. Wang, "Land use classification of the deep convolutional neural network method reducing the loss of spatial features," *Sensors*, vol. 19, no. 12, pp. 2792–, 2019.

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

[106] A. Nayem, A. Sarker, O. Paul, A. Ali, M. A. Amin, and A. M. Rahman, "Lulc segmentation of rgb satellite image using fcn-8," 2020.

[107] M. Onim, A. R. Ehtesham, A. Anbar, A. Islam, and A. Rahman, "Lulc classification by semantic segmentation of satellite images using fastfcn," in *2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)*, 2020.

[108] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, "Deep networks with stochastic depth," 2016.

[109] G. Rui, L. Jianbo, L. Na, L. Shibin, C. Fu, C. Bo, D. Jianbo, L. Xinpeng, and M. Caihong, "Pixel-wise classification method for high resolution remote sensing imagery using deep neural networks," *International Journal of Geo-Information*, vol. 7, no. 3, p. 110, 2018.

[110] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–13, 2020.

[111] C. Zhang, X. Pan, H. Li, A. Gardiner, I. Sargent, J. Hare, and P. M. Atkinson, "A hybrid mlp-cnn classifier for very fine resolution remotely sensed image classification," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 140, no. JUN., pp. 133–144, 2018.

[112] Kelvin, Chew, Preetha, and Thulasiramany, "An object-based convolutional neural network (ocnn) for urban land use classification."

[113] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 41–57, 2016.

[114] Y. Tao, M. Xu, F. Zhang, B. Du, and L. Zhang, "Unsupervised-restricted deconvolutional neural network for very high resolution remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–19, 2017.

[115] X. Y. Tong, G. S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," 2018.

[116] Q. Lv, Y. Dou, X. Niu, J. Xu, and B. Li, "Classification of land cover based on deep belief networks using polarimetric radarsat-2 data," in *IGARSS 2014 - 2014 IEEE International Geoscience and Remote Sensing Symposium*, 2014.

[117] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 113, no. Mar., pp. 155–165, 2016.

[118] Y. Tan, S. Xiong, Z. Li, J. Tian, and Y. Li, "Accurate detection of built-up areas from high-resolution remote sensing imagery using a fully convolutional network," *Photogrammetric Engineering and Remote Sensing*, 2019.

[119] X. Yang, Z. Chen, B. Li, D. Peng, and B. Zhang, "A fast and precise method for large-scale land-use mapping based on deep learning," 2019.

[120] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical

representations," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, 2009.

[121] A. Romero, P. Radeva, and C. Gatta, "Meta-parameter free unsupervised sparse feature learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1716–22, 2015.

[122] Y. Li, C. Tao, Y. Tan, K. Shang, and J. Tian, "Unsupervised multilayer feature learning for satellite image scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 2, pp. 157–161, 2016.

[123] S. Kiyasu, Y. Uraguchi, K. Sonoda, and T. Sakai, *Semi-supervised method for land cover classification of remotely sensed image considering spatial arrangement of the pixels*, 2011.

[124] G. Camps-Valls, T. V. B. Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3044–3054, 2007.

[125] S. Kiyasu, Y. Yamada, and S. Miyahara, "Semi-supervised land cover classification of remotely sensed data using two different types of classifiers," in *Iccas-sice*, 2009.

[126] P. K. S. A, D. H. A, R. R. A, M. B. B, and T. I. A, "Selection of classification techniques for land use/land cover change investigation," *Advances in Space Research*, vol. 50, no. 9, pp. 1250–1265, 2012.

[127] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[128] L. Xue, Z. Liangpei, D. Bo, Z. Lefei, and S. Qian, "Iterative reweighting heterogeneous transfer learning framework for supervised remote sensing image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 5, pp. 2022–2035, 2017.

[129] B. Zhao, B. Huang, and Y. Zhong, "Transfer learning with fully pretrained deep convolution networks for land-use classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 99, pp. 1436–1440, 2017.

[130] C. Persello and L. Bruzzone, "Active learning for domain adaptation in the supervised classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 11, pp. 4468–4483, 2012.

[131] Q. Lu, Y. Ma, and G. S. Xia, "Active learning for training sample selection in remote sensing image classification using spatial information," *Remote Sensing Letters*, vol. 8, no. 10-12, pp. 1211–1220, 2017.

[132] G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia, "Semisupervised transfer component analysis for domain adaptation in remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3550–3564, 2015.

[133] L. Gomez-Chova, G. Camps-Valls, J. Munoz-Mari, and J. Calpe, "Semisupervised image classification with laplacian support vector machines," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 3, pp. 336–340, 2008.

[134] P. Pandey, A. K. Tyagi, S. Ambekar, and A. P. Prathosh, "Unsupervised domain adaptation for semantic segmentation of nir images through generative latent search," in *European Conference on Computer Vision*, 2020.

[135] S. Paul, Y. H. Tsai, S. Schulter, A. K. Roy-Chowdhury, and M. Chandraker, *Domain Adaptive Semantic Segmentation Using Weak Labels*, 2020.

[136] G. Li, G. Kang, W. Liu, Y. Wei, and Y. Yang, "Content-consistent matching for domain adaptive semantic segmentation," *Springer, Cham*, 2020.

[137] J. Yang, W. An, S. Wang, X. Zhu, C. Yan, and J. Huang, "Label-driven reconstruction for domain adaptation in semantic segmentation," 2020.

[138] Y. Sun, E. Tzeng, T. Darrell, and A. A. Efros, "Unsupervised domain adaptation through self-supervision," 2019.

[139] S. Ozan, H. OHsong, A. Saxena, and S. Savarese, "Learning transferrable representations for unsupervised domain adaptation," *in neural information processing systems*, pp. 2110–2118, 2016.

[140] Saeid, Niazmardi, Begüm, Demir, Lorenzo, Bruzzone, Abdolreza, Safari, Saeid, and Homayouni, "Multiple kernel learning for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 3, pp. 1425–1443, 2017.

[141] L. Ma, M. M. Crawford, L. Zhu, and Y. Liu, "Centroid and covariance alignment-based domain adaptation for unsupervised classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 2305–2323, 2019.

[142] W. Liu and R. Qin, "A MultiKernel Domain Adaptation Method for Unsupervised Transfer Learning on Cross-Source and Cross-Region Remote Sensing Data Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 4279–4289, Jun. 2020.

[143] Z. Wang, M. Yu, Y. Wei, R. Feris, J. Xiong, W. M. Hwu, T. S. Huang, and H. Shi, "Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation," *IEEE*, 2020.

[144] Y. Chen, L. Wen, and L. V. Gool, "Road: Reality oriented adaptation for semantic segmentation of urban scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[145] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," 2017.

[146] Q. Lian, F. Lv, L. Duan, and B. Gong, "Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[147] L. Zhang, L. Zhang, B. Du, J. You, and D. Tao, "Hyperspectral image unsupervised classification by robust manifold matrix factorization," *Information Sciences*, 2019.

[148] C. Persello and L. Bruzzone, "Active and semisupervised learning for the classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 11, pp. 6937–6956, 2014.

[149] X. Sun, A. Shi, H. Huang, and H. Mayer, "\\mathrm{BAS}^4$net:boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, no. 99, pp. 1–1, 2020.

[150] M. S. Ibrahim, A. Vahdat, M. Ranjbar, and W. G. Macready, "Semi-supervised semantic image segmentation with self-correcting networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 712–12 722.

[151] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," 2016.

[152] Biplab, Banerjee, Krishna, Mohan, and Buddhiraju, "A novel semi-supervised land cover classification technique of remotely sensed images," *Journal of the Indian Society of Remote Sensing*, vol. 43, no. 4, pp. 719–728, 2015.

[153] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1349–1362, 2015.

[154] X. Y. Tong, G. S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," 2018.

[155] K. Nogueira, M. D. Mura, J. Chanussot, W. R. Schwartz, and J. A. D. Santos, "Learning to semantically segment high-resolution remote sensing images," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016.

[156] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with dsm," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2018.

[157] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2016.

[158] M. Li, L. Ma, T. Blaschke, L. Cheng, and D. Tiede, "A systematic comparison of different object-based classification techniques using high spatial resolution imagery in agricultural environments," *International Journal of Applied Earth Observation and Geoinformation*, vol. 49, pp. 87–98, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243416300125

[159] T. Fu, L. Ma, M. Li, and B. A. Johnson, "Using convolutional neural network to identify irregular segmentation objects from very high-resolution remote sensing imagery," *Journal of Applied Remote Sensing*, vol. 12, no. 2, pp. 1–, 2018.

[160] M. Voltersen, C. Berger, S. Hese, and C. Schmullius, "Object-based land cover mapping and comprehensive feature calculation for an automated derivation of urban structure types at block level," *Remote Sensing of Environment*, vol. 154, pp. 192–201, 2014.

[161] X. Zhang, G. Chen, W. Wang, Q. Wang, and F. Dai, "Object-based land-cover supervised classification for very-high-resolution uav images using stacked denoising autoencoders," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2017.

[162] N. Aggarwal, M. Srivastava, and M. Dutta, "Comparative analysis of pixel-based and object-based classification of high resolution remote

sensing images – a review," *International Journal of Engineering Trends and Technology*, vol. 38, no. 1, pp. 5–11, 2016.

[163] T. Fu, L. Ma, M. Li, and B. A. Johnson, "Using convolutional neural network to identify irregular segmentation objects from very high-resolution remote sensing imagery," *Journal of Applied Remote Sensing*, vol. 12, no. 2, pp. 1–, 2018.

[164] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *International Journal of Computer Vision*, vol. 81, no. 1, pp. 2–23, 2009.

[165] J. Krapac and I. K. S. Segvic, "Ladder-style densenets for semantic segmentation of large natural images," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2018.

[166] S. R. Bulo, L. Porzi, and P. Kontschieder, "In-place activated batchnorm for memory-optimized training of dnns," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[167] T. Wang, Y. Piao, H. Lu, X. Li, and L. Zhang, "Deep learning for light field saliency detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2020.

[168] L. Zhang, Z. Lin, J. Zhang, H. Lu, and Y. He, "Fast video object segmentation via dynamic targeting network," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[169] Y. Zeng, Z. Yunzhi, L. Huchuan, and L. Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," *in international conference on computer vision*, pp. 7223–7233, 2019.

[170] Long, Jonathan, Shelhamer, Evan, Darrell, and Trevor, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[171] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[172] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[173] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," 2016.

[174] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[175] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision*, 2018.

[176] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016.

[177] B. Fang, Y. Li, H. Zhang, and J. Chan, "Hyperspectral images classification based on dense convolutional networks with spectral-wise attention mechanism," *Remote Sensing*, vol. 11, no. 2, 2019.

[178] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2017.

[179] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015.

[180] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[181] N. Souly, C. Spampinato, and M. Shah, "Semi and weakly supervised semantic segmentation using generative adversarial network," pp. 5689–5697, 2017.

[182] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2020.

[183] S. Du, S. Du, B. Liu, and X. Zhang, "Incorporating deeplabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images," *International Journal of Digital Earth*, 2020.

[184] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 135, 2017.

[185] X. Pan, L. Gao, B. Zhang, F. Yang, and W. Liao, "High-resolution aerial imagery semantic labeling with dense pyramid network," *Sensors*, vol. 18, no. 11, 2018.

[186] N. Audebert, B. L. Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," 2016.

[187] D. Cheng, G. Meng, S. Xiang, and C. Pan, "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*, vol. PP, no. 99, pp. 1–15, 2017.

[188] Yuchu, Qin, Yunchao, Bin, Shuai, Gao, Miao, Liu, Yulin, and Zhan, "Semantic segmentation of building roof in dense urban environment with deep convolutional neural network: A case study using gf2 vhr imagery in china." *Sensors*, 2019.

[189] H. Wang, Y. Wang, Q. Zhang, S. Xiang, and C. Pan, "Gated convolutional neural network for semantic segmentation in high-resolution images," *Remote Sensing*, vol. 9, no. 5, 2017.

[190] C. Henry, S. M. Azimi, and N. Merkle, "Road segmentation in sar satellite images with deep fully convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, 2018.

[191] X. Yongyang, W. Liang, X. Zhong, and C. Zhanlong, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sensing*, vol. 10, no. 1, p. 144, 2018.

[192] Y. Shi, Q. Li, and X. X. Zhu, "Building segmentation through a gated graph convolutional neural network with deep structured feature embedding," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 184–197, 2020.

[193] A. O. Ok, "Automated detection of buildings from single vhr multispectral images using shadow information and graph cuts," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 86, no. 12, p. 21–40, 2013.

[194] Zhang, L., Huang, and X., "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE journal of selected topics in applied earth observations and remote sensing*, 2012.

[195] B. Ksenia, A. Fathalrahman, S. Cui, K. Marco, and R. Peter, "Building footprint extraction from vhr remote sensing images combined with normalized dsms using fused fully convolutional networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2615–2629, 2018.

[196] Weixing, Wang, Nan, Yang, Yi, Zhang, Fengping, Wang, Ting, and Cao, "A review of road extraction from remote sensing images," *Journal of Traffic and Transportation Engineering*, 2016.

[197] E. Y. Lam, K. S. Niel, S. Saito, and Y. Aoki, "Building and road detection from large aerial imagery," in *Image Processing: Machine Vision Applications VIII*, 2015.

[198] M. Yuan, Z. Liu, F. Wang, and F. Jin, "Rethinking labelling in road segmentation," *International Journal of Remote Sensing*, pp. 1–20, 2019.

[199] L. Dan, N. Garnett, and E. Fetaya, "Stixelnet: A deep convolutional network for obstacle detection and road segmentation." in *British Machine Vision Conference*, 2015.

[200] A. Laddha, M. K. Kocamaz, L. E. Navarro-Serment, and M. Hebert, "Map-supervised road detection," in *Intelligent Vehicles Symposium*, 2016.

[201] S. Amit, S. Saito, Y. Aoki, and Y. Kiyoki, "Building change detection via semantic segmentation and difference extraction method," in *Information Modelling and Knowledge Bases XXVIII*, ser. Frontiers in Artificial Intelligence and Applications, vol. 292. IOS Press, 2017, pp. 249–257.

[202] J. M. Alvarez, T. Gevers, Y. Lecun, and A. M. Lopez, "Road scene segmentation from a single image," in *European Conference on Computer Vision*, 2012.

[203] A. V. Buslaev, S. S. Seferbekov, V. I. Iglovikov, and A. A. Shvets, "Fully convolutional network for automatic road extraction from satellite imagery," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[204] A. Nicolas, L. S. Bertrand, and L. Sébastien, "Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images," *Remote Sensing*, vol. 9, no. 4, pp. 368–, 2017.

[205] M. Lan, Y. Zhang, L. Zhang, and B. Du, "Global context based automatic road segmentation via dilated convolutional neural network," *Information Sciences*, vol. 535, 2020.

[206] P. Olofsson, G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder, "Good practices for estimating area and assessing

accuracy of land change," *Remote Sensing of Environment*, vol. 148, pp. 42–57, 2014.

[207] G. Fu, C. Liu, R. Zhou, T. Sun, and Q. Zhang, "Classification for high resolution remote sensing imagery using a fully convolutional network," *Remote Sensing*, vol. 9, no. 6, p. 498, 2017.

[208] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 135, 2017.

[209] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Semantic segmentation of multisensor remote sensing imagery with deep convnets and higher-order conditional random fields," *Journal of Applied Remote Sensing*, vol. 13, no. 1, p. 1, 2019.

[210] M. Kampffmeyer, A. B. Salberg, and R. Jenssen, "Urban land cover classification with missing data modalities using deep convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 6, pp. 1758–1768, 2018.

[211] L. Mou and X. X. Zhu, "Rifcn: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images," *CoRR*, vol. abs/1805.02091, 2018. [Online]. Available: http://arxiv.org/abs/1805.02091

[212] M. Samy, K. Amer, K. Eissa, M. Shaker, and M. Elhelw, "Nu-net: Deep residual wide field of view convolutional neural network for semantic segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[213] A. Davydow and S. Nikolenko, "Land cover classification with superpixels and jaccard index post-optimization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[214] C. Bian, C. Yuan, J. Wang, M. Li, and Y. Zheng, "Uncertainty-aware domain alignment for anatomical structure segmentation," *Medical Image Analysis*, vol. 64, p. 101732, 2020.

[215] S. Seferbekov, V. Iglovikov, A. Buslaev, and A. Shvets, "Feature pyramid network for multi-class land segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 272–2723.

[216] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," 2018.

[217] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation," 2018.

[218] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. a. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[219] F. Wang and J. Xie, "A context and semantic enhanced unet for semantic segmentation of high-resolution aerial imagery," *Journal of Physics: Conference Series*, vol. 1607, no. 1, p. 012083 (6pp), 2020.

[220] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2020.

[221] Y. Lu, Y. Chen, D. Zhao, and J. Chen, "Graph-fcn for image semantic segmentation," 2020.

[222] J. Chen, H. Wang, Y. Guo, G. Sun, and M. Deng, "Strengthen the feature distinguishability of geo-object details in the semantic segmentation of high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, no. 99, pp. 1–1, 2021.