XIAO XIANG ZHU, DEVIS TUIA, LICHAO MOU, GUI-SONG XIA,
LIANGPEI ZHANG, FENG XU, AND FRIEDRICH FRAUNDORFER

Deep Learning in Remote Sensing

A comprehensive review and list of resources

Central to the looming paradigm shift toward data-intensive science, machine-learning techniques are becoming increasingly important. In particular, deep learning has proven to be both a major breakthrough and an extremely powerful tool in many fields. Shall we embrace deep learning as the key to everything? Or should we resist a black-box solution? These are controversial issues within the remote-sensing community. In this article, we analyze the challenges of using deep learning for remote-sensing data analysis, review recent advances, and provide resources we hope will make deep learning in remote sensing seem ridiculously simple. More importantly, we encourage remote-sensing scientists to bring their expertise into deep learning and use it as an implicit general model to tackle unprecedented, large-scale, influential challenges, such as climate change and urbanization.

MOTIVATION

Deep learning is the fastest-growing trend in big data analysis and was deemed one of the ten breakthrough technologies of 2013 [1]. It is characterized by neural networks (NNs) involving usually more than two hidden layers (for this reason, they are called *deep*). Like shallow NNs, deep NNs exploit feature representations learned exclusively from data, instead of handcrafting features that are designed based mainly on domain-specific knowledge. Deep learning research has been extensively pushed by Internet companies, such as Google, Baidu, Microsoft, and Facebook, for several image analysis tasks, including image indexing, segmentation, and object detection.

Based on recent advances, deep learning is proving to be a very successful set of tools, sometimes able to surpass

Digital Object Identifier 10.1109/MGRS.2017.2762307
Date of publication: 27 December 2017

even humans in solving highly computational tasks (consider, e.g., the widely reported Go match between Google's AlphaGo artificial intelligence program and the world Go champion Lee Sedol). Based on such exciting successes, deep learning is increasingly the model of choice in many application fields.

For instance, convolutional NNs (CNNs) have proven to be good at extracting mid- and high-level abstract features from raw images by interleaving convolutional and pooling layers (i.e., by spatially shrinking the feature maps layer by layer). Recent studies indicate that the feature representations learned by CNNs are highly effective in large-scale image recognition [2]–[4], object detection [5], [6], and semantic segmentation [7], [8]. Furthermore, recurrent NNs (RNNs), an important branch of the deep learning family, have demonstrated significant achievement on a variety of tasks involved in sequential data analysis, such as action recognition [9], [10] and image captioning [11].

In the wake of this success and thanks to the increased availability of data and computational resources, the use of deep learning is finally taking off in remote sensing as well. Remote-sensing data present some new challenges for deep learning, because satellite image analysis raises unique issues that pose difficult new scientific questions.

BRAIN IMAGE WITH MAP—IMAGE LICENSED BY INGRAM PUBLISHING, ELECTRONIC CIRCUIT BOARD—©ISTOCKPHOTO.COM/HENRIK5000

- Remote-sensing data are often multimodal, e.g., from optical (multi- and hyperspectral), Lidar, and synthetic aperture radar (SAR) sensors, where the imaging geometries and content are completely different. Data and information fusion uses these complementary data sources in a synergistic way. Already, prior to a joint information extraction, a crucial step involves developing novel architectures to match images taken from different perspectives and even different imaging modalities, preferably without requiring an existing three-dimensional (3-D) model. Also, in addition to conventional decision fusion, an alternative is to investigate the transferability of trained networks to other imaging modalities.
- Remote-sensing data are geolocated, i.e., they are naturally located in the geographical space. Each pixel corresponds to a spatial coordinate, which facilitates the fusion of pixel information with other sources of data, such as geographic information system layers, geotagged images from social media, or simply other sensors (as just discussed). On the one hand, this allows tackling data fusion with nontraditional data modalities. On the other hand, it opens the field to new applications, such as picture localization, location-based services, and reality augmentation.
- Remote-sensing data are geodetic measurements in which quality is controlled. This enables us to retrieve geoparameters with confidence estimates. However, unlike purely data-driven approaches, the role of prior knowledge concerning the sensors' adequacy and data quality becomes especially crucial. To retrieve topographic information, e.g., even at the same spatial resolution, interferograms acquired using a single-pass SAR system are considered to be more reliable than the ones acquired in a repeat-pass manner.
- The time variable is becoming increasingly important in the field. The Copernicus program guarantees continuous data acquisition for decades; e.g., Sentinel-1 images the entire Earth every six days. This capability is triggering a shift from individual image analysis to time-series processing. Novel network architectures must be developed to optimally exploit the temporal information jointly with the spatial and spectral information of these data.
- Remote sensing also faces the "big data" challenge. In the Copernicus era, we are dealing with very large and ever-growing data volumes, often on a global scale. Even if they were launched in 2014, e.g., Sentinel satellites have already acquired about 25 PB of data. The Copernicus concept calls for global applications, i.e., algorithms must be fast enough and sufficiently transferrable to be applied for the whole Earth surface. However, these data are well annotated and contain plenty of metadata. Hence, in some cases, large training data sets might be generated (semi)automatically.
- In many cases, remote sensing aims at retrieving geophysical or biochemical quantities rather than detecting or classifying objects. These quantities include mass movement rates, mineral composition of soils, water constituents, atmospheric trace gas concentrations, and terrain elevation of biomass. Often, process models and expert knowledge exist and are traditionally used as priors for the estimates. This suggests, in particular, that the dogma of expert-free, fully automated deep learning should be questioned for remote sensing and that physical models should be reintroduced into the concept, as, e.g., in the concept of emulators [12].

Remote-sensing scientists have exploited the power of deep learning to tackle these different challenges and instigated a new wave of promising research. In this article, we review these advances.

FROM PERCEPTRON TO DEEP LEARNING

The perceptron is the basis of the earliest NNs [13]. It is a bioinspired model for binary classification that aims to mathematically formalize how a biological neuron works. In contrast, deep learning has provided more sophisticated methodologies to train deep NN architectures. In this section, we recall the classic deep learning architectures used in visual data processing.

AUTOENCODER MODELS

AUTOENCODER AND STACKED AUTOENCODER

An autoencoder (AE) [14] takes an input $\boldsymbol{x} \in \mathbb{R}^D$ and, first, maps it to a latent representation $\boldsymbol{h} \in \mathbb{R}^M$ via a nonlinear mapping:

$$\boldsymbol{h} = f(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{\beta}), \tag{1}$$

where $\boldsymbol{W}$ is a weight matrix to be estimated during training, $\boldsymbol{\beta}$ is a bias vector, and f stands for a nonlinear function, such as the logistic sigmoid function or a hyperbolic tangent function. The encoded feature representation $\boldsymbol{h}$ is then used to reconstruct the input $\boldsymbol{x}$ by a reverse mapping, leading to the reconstructed input $\boldsymbol{y}$:

$$\boldsymbol{y} = f(\boldsymbol{W}'\boldsymbol{h} + \boldsymbol{\beta}'), \tag{2}$$

where $\boldsymbol{W}'$ is usually constrained to be the form of $\boldsymbol{W}' = \boldsymbol{W}^T$, i.e., the same weight is used for encoding the input and decoding the latent representation. The reconstruction error is defined as the Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ that is constrained to approximate the input data $\boldsymbol{x}$ (i.e., minimizing $\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$). The parameters of the AE are generally optimized by stochastic gradient descent (SGD).

A stacked AE (SAE) is an NN consisting of multiple layers of AEs in which the outputs of each layer are wired to the inputs of the following one.

SPARSE AUTOENCODER

The conventional AE relies on the dimension of the latent representation $\boldsymbol{h}$ being smaller than that of input $\boldsymbol{x}$, i.e., $M < D$, which means that it tends to learn a low-dimensional, compressed representation. However, when $M > D$, one can still discover interesting structures by enforcing a sparsity constraint on the hidden units. Formally, given a set of unlabeled data $X = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N\}$, training a sparse

AE [15] boils down to finding the optimal parameters by minimizing the following loss function:

$$L = \frac{1}{N}\sum_{i=1}^{N}\left(J(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{\Theta}, \boldsymbol{\beta}) + \lambda \sum_{j=1}^{M} \mathrm{KL}(\rho \,\|\hat{\rho}_j)\right), \tag{3}$$

where $J(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{\Theta}, \boldsymbol{\beta})$ is an average sum-of-squares error term, which represents the reconstruction error between the input $\boldsymbol{x}^i$ and its reconstruction $\boldsymbol{y}^i$. $\mathrm{KL}(\rho \| \hat{\rho}_j)$ is the Kullback–Leibler (KL) divergence between a Bernoulli random variable with mean ρ and a Bernoulli random variable with mean $\hat{\rho}_j$:

$$\mathrm{KL}(\rho \,\|\hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho)\log \frac{1-\rho}{1-\hat{\rho}_j}. \tag{4}$$

KL divergence is a standard function for measuring the similarity between two distributions. In the sparse AE model, the KL divergence is a sparsity penalty term, and λ controls its importance. ρ is a free parameter corresponding to a desired average activation value, and $\hat{\rho}$ indicates the actual average activation value of hidden neuron $\boldsymbol{h}_j$ over the training samples. An activation corresponds to how often a region of the image reacts when convolved with a filter. In the first layer, e.g., each location in the image receives a value that corresponds to a linear combination of the original input and the filter applied. The higher such value, the more activated this filter is on that region. When convolved over the whole image, a filter produces an activation map, which is the activation at each location where the filter has been applied. Similar to the AE, the optimization of a sparse AE can be achieved via SGD.

RESTRICTED BOLTZMANN MACHINE AND DEEP BELIEF NETWORK

Unlike the deterministic network architectures, such as AEs or sparse AEs, a restricted Boltzmann machine (RBM) (see Figure 1) is a stochastic undirected graphical model consisting of a visible layer and a hidden layer. No connections exist within the hidden layer or the input layer. The energy function of an RBM can be defined as follows:

$$\mathbb{E}(\boldsymbol{x}, \boldsymbol{h}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{x} - (\boldsymbol{h}^T\boldsymbol{W}\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{h}), \tag{5}$$

where $\boldsymbol{W}$, $\boldsymbol{c}$, and $\boldsymbol{b}$ are learnable weights. Here, the input x is also named as the visible random variable, which is denoted as $\boldsymbol{v}$ in the original RBM paper [16]. The joint probability distribution of the RBM is defined as

$$p(\boldsymbol{x}, \boldsymbol{h}) = \frac{1}{Z}\exp(-\mathbb{E}(\boldsymbol{x}, \boldsymbol{h})), \tag{6}$$

where Z is a normalization constant. The form of the RBM makes the conditional probability distribution computationally feasible when $\boldsymbol{x}$ or $\boldsymbol{h}$ is fixed.

The feature representation ability of a single RBM is limited. However, its real power emerges when two or more RBMs are stacked, forming a deep belief network (DBN) [16]. Hinton et al. [16] proposed a greedy approach that trains RBMs in each layer to efficiently train the whole DBN.

CONVOLUTIONAL NEURAL NETWORKS

Supervised deep NNs have come under the spotlight in recent years. The leading model is the CNN, which studies the filters performing convolutions in the image domain. Here, we briefly review some successful CNN architectures recently offered for computer vision. For a comprehensive introduction to CNNs, we refer readers to the excellent book by Goodfellow and colleagues [17].

ALEXNET

In 2012, Krizhevsky et al. [2] created AlexNet, a "large, deep convolutional neural network" that won the 2012 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). The year 2012 is marked as the first year that a CNN was used to achieve a top-five test error rate of 15.4%.

AlexNet (see Figure 2) scaled the insights of LeNet [18] into a deeper and much larger network that could be used to learn the appearance of more numerous and complicated objects. The contributions of AlexNet include the following:

- using rectified linear units (ReLUs) as nonlinearity functions capable of decreasing training time because a ReLU is several times faster than the conventional hyperbolic tangent function
- implementing dropout layers to avoid the problem of overfitting
- using data augmentation techniques to artificially increase the size of the training set (and observe a more diverse set of situations); from this, the training patches are translated and reflected on the horizontal and vertical axes.

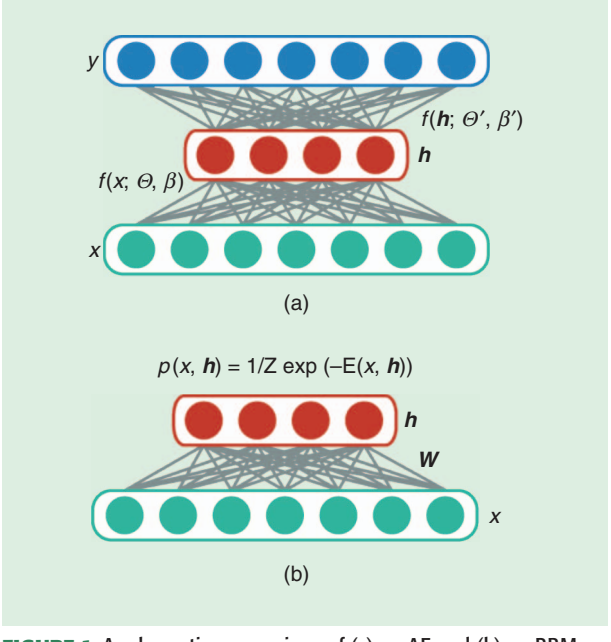

FIGURE 1. A schematic comparison of (a) an AE and (b) an RBM.

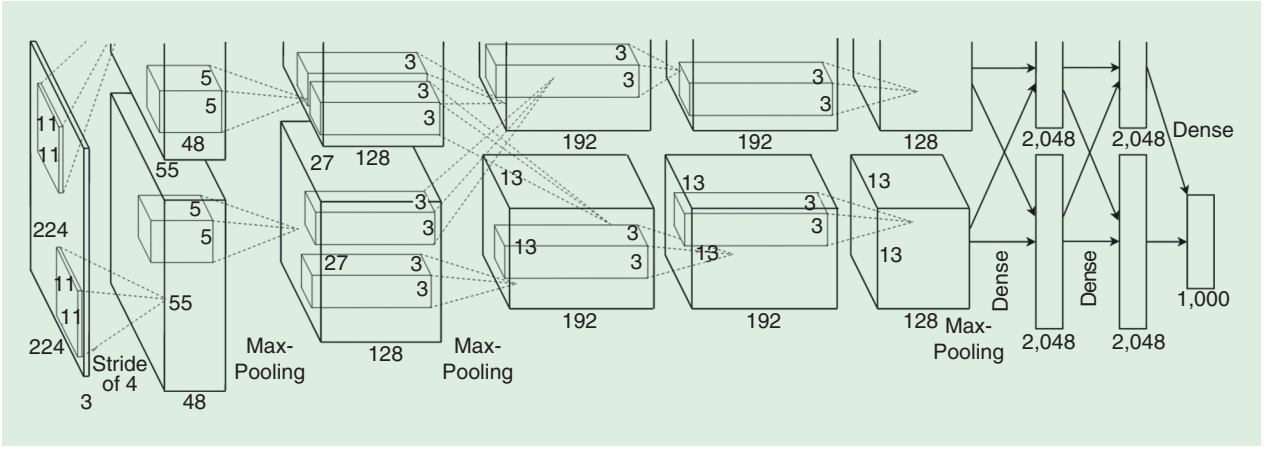

FIGURE 2. The architecture of AlexNet, as shown in [2].

One of the keys of the success of AlexNet is that the model was trained on graphics processing units (GPUs). The fact that GPUs can offer a much larger number of cores than central processing units allows much faster training, which, in turn, allows the use of larger data sets and bigger images.

VGG NETWORKS

The design philosophy of VGG networks (named for Oxford University's Visual Geometry Group) [3] is simplicity and depth. In 2014, Simonyan and Zisserman created VGG networks that make use strictly of 3 × 3 filters with stride and padding of 1, along with 2 × 2 max-pooling layers with stride of 2. The main points of VGG networks are that they

- use filters with a small receptive field of 3 × 3, rather than larger ones (5 × 5 or 7 × 7, as in Alexnet)
- have the same feature map size and number of filters in each convolutional layer of the same block
- increase the size of the features in the deeper layers, roughly doubling after each max-pooling layer
- use scale jittering as one data augmentation technique during training.

VGG networks are one of the most influential CNN models, as they reinforce the notion that CNNs with deeper architectures can promote hierarchical feature representations of visual data, which, in turn, improves classification accuracy. A drawback is that training such a model from scratch requires large computational power and a very large labeled training set.

RESNET

He et al. [4] pushed the idea of very deep networks even further by proposing the 152-layer ResNet, which won ILSVRC 2015 with a top-5 error rate of 3.6% and set new records in classification, detection, and localization through a single network architecture. In [4], the authors provide an in-depth analysis of the degradation problem, i.e., that simply increasing the number of layers in plain networks results in higher training and test errors, and they claim that it is easier to optimize the residual mapping in the ResNet than to optimize the original, unreferenced mapping in conventional CNNs. The core idea of ResNet is to add shortcut connections that bypass two or more stacked convolutional layers by performing identity mapping. The connections are then added together with the output of stacked convolutions.

FULLY CONVOLUTIONAL NETWORK

The fully convolutional network (FCN) [7] is the most important work in deep learning for semantic segmentation, i.e., the task of assigning a semantic label to every pixel in the image. To perform this task, the output of the CNN must be of the same pixel size as the input (contrary to the single class per image of the aforementioned models). FCN introduces many significant ideas, such as

- end-to-end learning of the upsampling algorithm via an encoder/decoder structure that first downsamples the activation's size and then upsamples it again
- using a fully convolutional architecture, which allows the network to take images of arbitrary size as input because there is no fully connected layer at the end that requires a specific size of the activations
- introducing skip connections as a way of fusing information from different depths in the network for the multiscale inference.

Figure 3 shows the FCN architecture.

REMOTE SENSING MEETS DEEP LEARNING

Deep learning is taking off in remote sensing, as shown in Figure 4, which illustrates the number of papers published on the topic since 2014. The exponential increase confirms the rapid surge of interest in deep learning for remote sensing. In this section, we focus on a variety of remote-sensing applications that are achieved by deep learning and provide an in-depth investigation from the perspectives of hyperspectral image analysis, interpretation of SAR images, interpretation of high-resolution satellite images, multimodal data fusion, and 3-D reconstruction.

HYPERSPECTRAL IMAGE ANALYSIS

Hyperspectral sensors are characterized by hundreds of narrow spectral bands. This very high spectral resolution enables us to identify the materials contained in the pixel via spectroscopic analysis. Analysis of hyperspectral data is of great importance in many practical applications, such as land cover/use classification or change and object detection. Because high-quality hyperspectral satellite data are becoming available (e.g., via the launch of EnMAP, planned for 2020, and the DESIS on the International Space Station, planned for 2018), hyperspectral image analysis has been one of the most active research areas within the remote-sensing community over the last decade.

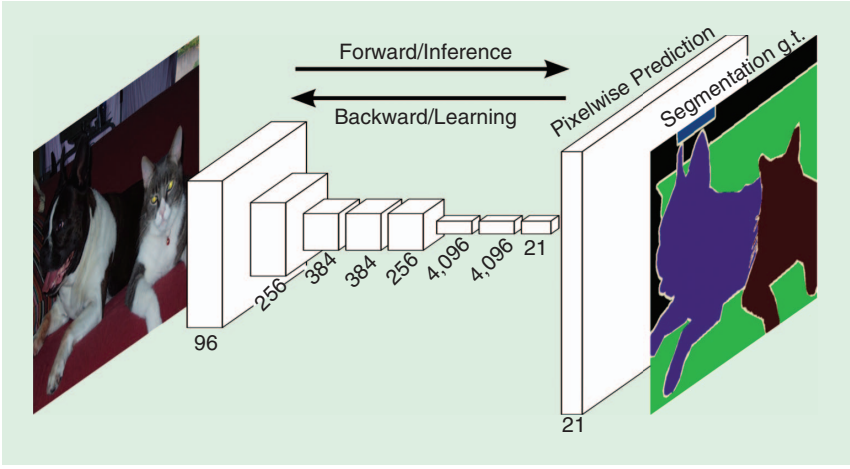

FIGURE 3. The FCN architecture [7]. g.t.: ground truth.

Inspired by the success of deep learning in computer vision, preliminary studies have been carried out on deep learning in hyperspectral data analysis, which brings new momentum to this field. In the following, we review two application cases, land cover/use classification and anomaly detection.

HYPERSPECTRAL IMAGE CLASSIFICATION

Supervised classification is probably the most active research area in hyperspectral data analysis. There is a vast literature on this topic using conventional supervised machine-learning models, such as decision trees, random forests, and support vector machines (SVMs) [20]. With the investigation of hyperspectral image classification [21], a major finding was that various atmospheric scattering conditions, complicated light-scattering mechanisms, interclass similarity, and intraclass variability result in the hyperspectral imaging procedure being inherently nonlinear. It is believed that, compared to the previously mentioned shallow models, deep learning architectures are able to extract high-level, hierarchical, and abstract features, which are generally more robust to the nonlinear processing.

The following sections discuss research on hyperspectral image classification.

SAE FOR HYPERSPECTRAL DATA CLASSIFICATION

A first attempt in this direction can be found in [22], where the authors make use of an SAE to extract hierarchical features in the spectral domain. Subsequently, in [23], the authors employ DBN. Similarly, Tao et al. [24] use sparse SAEs to learn an effective feature representation from input data; then, the learned features are fed into a linear SVM for hyperspectral data classification.

SUPERVISED CNNs

In [25], the authors train a simple one-dimensional (1-D) CNN that contains five layers—i.e., an input layer, a convolutional layer, a max-pooling layer, a fully connected layer, and an output layer—and directly classify hyperspectral images in the spectral domain.

Makantasis et al. [26] exploited a two-dimensional (2-D) CNN to encode spectral and spatial information, followed by a multilayer perceptron performing the actual classification. In [27], the authors attempted to carry out the classification of crop types using 1-D CNN and 2-D CNN. They concluded that the 2-D CNNs can outperform the 1-D CNNs, but some small objects in the final classification map provided by 2-D CNNs are smoothed and misclassified. To avoid overfitting, Zhao and Du [28] suggest a spectral-spatial-feature-based classification framework, which jointly makes use of a local-discriminant embedding-based dimension-reduction algorithm and a 2-D CNN. In [21], the authors propose a self-improving CNN model that combines a 2-D CNN with a fractional-order Darwinian particle swarm optimization algorithm to iteratively select the most informative bands suitable for training the designed CNN. Santara et al. [29] discuss

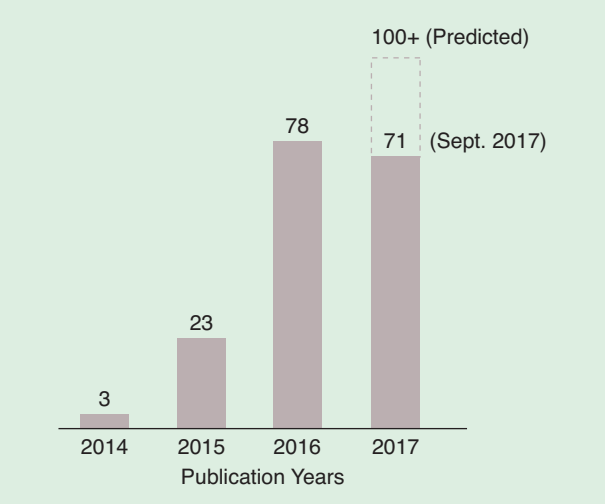

FIGURE 4. The statistics for published papers related to deep learning in remote sensing [187].

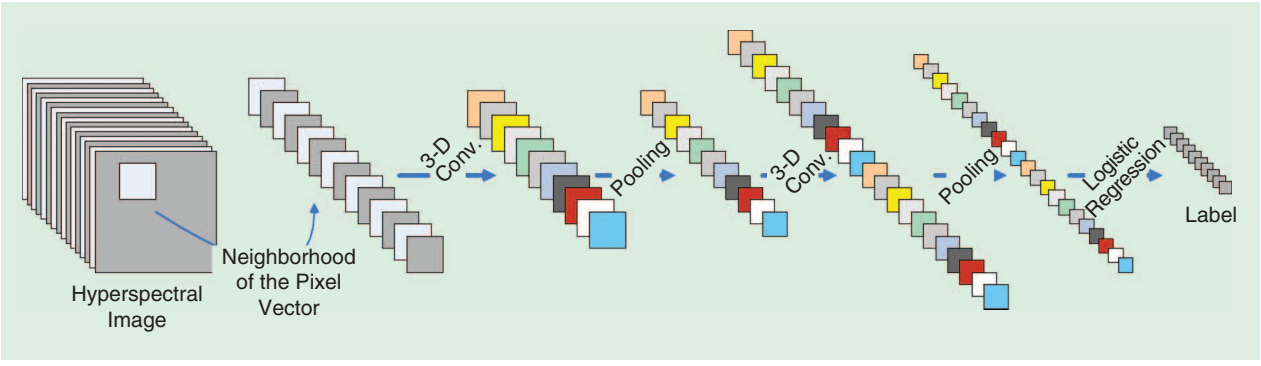

FIGURE 5. A flowchart of the 3-D CNN architecture proposed in [19] for spectral-spatial hyperspectral image classification. Conv.: convolution.

FIGURE 6. The object-detection maps using learned filters of the first residual block in the unsupervised residual conv-deconv network [33], [34], where some neurons own good description power for semantic visual patterns at the object level. The feature maps activated by the convolutional filter numbers 52 and 03, e.g., in the first residual block can be used to precisely capture (a) metal sheets and (b) vegetative covers, respectively.

an end-to-end, band-adaptive spectral-spatial-feature-learning network to address the problems of the curse of dimensionality. In [30], to allow a CNN to be appropriately trained using limited labeled data, the authors present a novel pixel-pair CNN to significantly augment the number of training samples.

Following recent vision developments in 3-D CNNs [31], in which the third dimension usually refers to the time axis, such architecture has also been employed in hyperspectral classification. In other words, in a 3-D CNN, convolution operations are performed spatial spectrally, while in 2-D CNNs, they are done only spatially. The authors in [19] introduce a supervised, ℓ_2-regularized 3-D CNN-based model (see Figure 5), while the authors of [32] follow a similar idea for spatial-spectral classification.

UNSUPERVISED DEEP LEARNING

To allow less dependence on the existence of large annotated collections of labeled data, unsupervised feature extraction is of great interest. The authors of [35] propose an unsupervised convolutional network for learning spectral-spatial features using sparse learning to estimate the network weights in a greedy layerwise fashion instead of end-to-end learning. Mou et al. [33], [34] present a network architecture called a *fully residual conv-deconv network* for unsupervised spectral-spatial feature learning of hyperspectral images. They report an extensive study of the filters learned (see Figure 6).

RECURRENT NEURAL NETWORKS FOR HYPERSPECTRAL IMAGE CLASSIFICATION

In [36], the authors propose an RNN model with a new activation function and modified gated recurrent unit for hyperspectral image classification that can effectively analyze hyperspectral pixels as sequential data and then determine information categories via network reasoning (see Figure 7).

ANOMALY DETECTION

In a hyperspectral image, the pixels whose spectral signatures are significantly different from the global background pixels are considered anomalies. Because the prior knowledge of the anomalous spectrum is difficult to obtain in practice, anomaly detection is usually solved by background modeling or statistical characterization for hyperspectral data. So far, the only attempt to address this problem via deep learning can be found in [37], where Li et al. propose an anomaly detection framework in which a multilayer CNN is trained using the differences in values between neighboring pixel pairs in the reference image as input data. Then, in the test phase, anomalies are detected by evaluating differences between neighboring pixel pairs using the trained CNN.

In summary, deep learning has been widely applied to multi/hyperspectral image classification, and some promising results have been achieved. In contrast, for other hyperspectral data analysis tasks, such as change and anomaly detections, deep learning is just beginning to make its mark [37], [38]. Some potential problems to

be further explored include nonlinear spectral unmixing, hyperspectral image enhancement, and hyperspectral time-series analysis.

INTERPRETATION OF SYNTHETIC APERTURE RADAR IMAGES

Over the past several years, many studies related to deep learning for SAR image analysis have been published. Among these, deep learning techniques have been used most in typical applications, including automatic target recognition (ATR), terrain surface classification, and parameter inversion. This section reviews some of the relevant studies in this area.

AUTOMATIC TARGET RECOGNITION

SAR ATR is an important application, in particular, for military surveillance [39]. A standard architecture for efficient ATR consists of three stages: detection, discrimination, and classification. Each stage tends to perform a more complicated and refined processing than its predecessor and selects the candidate objects for the next-stage processing. However, all three stages can be treated as a classification problem and, for this reason, deep learning has made its mark.

Chen and Wang [40] introduced CNNs into SAR ATR and tested them on the standard ATR data set MSTAR [41]. They found the major issue to be the lack of sufficient training samples as compared to optical images. This might cause severe overfitting and, therefore, greatly limit the capability of generalizing the model, so data augmentation is employed to counteract overfitting. Chen et al. [42] propose to further remove all fully connected layers from conventional CNNs, which are accountable for most trainable parameters. The final performance is demonstrated as superior compared to conventional CNNs on the MSTAR data set (i.e., a state-of-the-art accuracy of 99.1% in standard operating condition). Extensive experiments have been conducted to test the generalization capability of the so-called AConvNets, and they have proved to be quite robust in several extended operating conditions. The removal of the fully connected layers, originally designed to be trainable classifiers, might be justifiable in this case because the limited number of target types can be seen as the feature templates that the AConvNets are extracting.

Many authors have applied CNNs to SAR ATR and tested the results on the MSTAR data set, e.g., [43]–[46]. Among these studies, the one common finding is that data augmentation is necessary and the most critical step for SAR ATR using CNNs. Various augmentation strategies have been offered, including translation, rotation, and interpolation. Cui et al. [47] introduce DBN to SAR ATR, where stacked RBMs are used to extract features that are then fed to a trainable classifier.

SAR ATR IS AN IMPORTANT APPLICATION, IN PARTICULAR, FOR MILITARY SURVEILLANCE.

Wagner [48] suggests using a CNN to first extract feature vectors and then feed them to an SVM for classification. The CNN is trained with a fully connected layer, but only the previous activations are used. A systematic data augmentation approach is employed, which includes elastic distortions and affine transformations. It is intended to mimic typical imaging errors, such as a changing range (which is scale dependent on the depression angle) or an incorrectly estimated aspect angle.

Additional studies applying CNNs to the ATR problem are also of interest. Bentes et al. [49] applied a CNN to ship–iceberg discrimination, tested on TerraSAR-X StripMap images. Schwegmann et al. [50] applied a specific type of deep NNs, highway networks, to the discrimination of ships in SAR imagery and achieved promising results. Ødegaard et al.

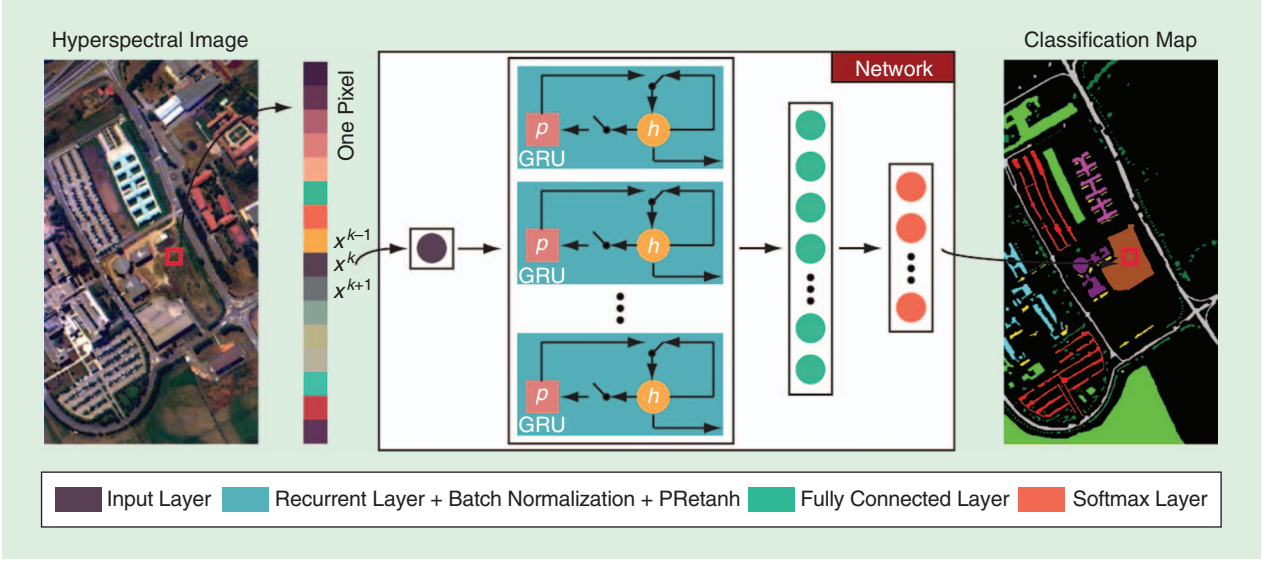

FIGURE 7. The RNN proposed for the hyperspectral image classification task in [36]. GRU: gated recurrent unit; PRetanh: parametric rectified tanh.

[51] applied a CNN to detect ships in a harbor background in SAR images; to address the issue of a lack of training samples, they employed a simulation software to generate simulated data for training. Song et al. [52] follow this idea, introducing a deep generative NN for SAR ATR. A generative deconvolutional NN is first trained to generate a simulated SAR image from a given target label, while a feature space is simultaneously constructed in the intermediate layer. A CNN is then trained to map an input SAR image to the feature space. The goal is to develop an extended ATR system capable of interpreting a previously unseen target in the context of all known targets.

WHEN TERRAIN SURFACE CLASSIFICATION USES SAR, IN PARTICULAR POLARIMETRIC SAR, DATA MEET ANOTHER IMPORTANT APPLICATION IN RADAR REMOTE SENSING.

TERRAIN SURFACE CLASSIFICATION

When terrain surface classification uses SAR, in particular polarimetric SAR (PolSAR), data meet another important application in radar remote sensing. This is very similar to the task of image segmentation in computer vision. Conventional approaches are based mostly on pixelwise polarimetric target decomposition parameters [54]. They hardly consider the spatial patterns, which convey rich information in high-resolution SAR images [55]. Deep learning provides a tool for automatically extracting features that represent spatial patterns as well as polarimetric characteristics.

One large stream of studies employs at least one type of unsupervised generative graphical models, such as the DBN, SAE, or RBM. Xie et al. [56] first introduced multilayer feature learning for PolSAR classification; here, an SAE is employed to extract useful features from a channel PolSAR image.

Geng et al. [57] proposed a deep convolutional AE (DCAE) to extract features and conduct classification automatically. The DCAE consists of a handcrafted first layer of convolution that contains kernels, such as gray-level co-occurrence matrix and Gabor filters, and a handcrafted second layer of scale transformation that integrates correlated neighbor pixels. The remaining layers are trained SAEs. This approach was tested on high-resolution, single-polarization TerraSAR-X images. Geng et al. [58] later presented a similar framework, deep supervised and contractive NN, for SAR image classification; this framework additionally includes the histogram of oriented gradient descriptors as handcrafted kernels. The trainable AE layers employ a supervised penalty that captures the relevant information between features and labels, as well as a contractive restriction that enhances local invariance. An interesting finding of Geng et al. [58] is that speckle reduction yields the worst performance, and the authors suspect that speckle reduction might smooth out some useful information.

Lv et al. [59] tested DBN on urban land use and land cover classification using PolSAR data. Hou et al. [60] proposed an SAE combined with superpixels for PolSAR image classification. Here, multiple AE layers are trained on a pixel-by-pixel basis, and superpixels are formed based on Pauli-decomposed pseudocolor images. The output of the SAE is used as a feature in the final step for k-nearest neighbor clustering of superpixels. Zhang et al. [61] applied a stacked sparse SAE to PolSAR image classification, while Qin et al. [62] applied adaptive boosting of RBMs to PolSAR image classification. Zhao et al. [63] proposed discriminant DBN for SAR image classification; here, the discriminant features are learned by combining ensemble learning with a DBN in an unsupervised manner.

Jiao and Liu [64] presented a deep stacking network for PolSAR image classification, which mainly takes advantage of fast Wishart distance calculation through linear projection. The proposed network aims to perform a k-means clustering/classification task where Wishart distance is used as the similarity metric.

The other stream of studies involves CNNs. Zhou et al. [65] applied CNNs to PolSAR image classification; here, a covariance matrix is extracted as six real-channel data input. Duan et al. [66] suggested replacing the conventional pooling layer in CNNs by a wavelet-constrained pooling layer. The so-called convolutional-wavelet NN is then used in conjunction with superpixels and a Markov random field (MRF) to produce the final segmentation map.

Zhang et al. [53] described a complex-valued (CV) CNN (see Figure 8) specifically designed to process complex values in PolSAR data, i.e., the off-diagonal elements of coherency or covariance matrix. The CV CNN not only takes complex numbers as input but also employs complex weights and

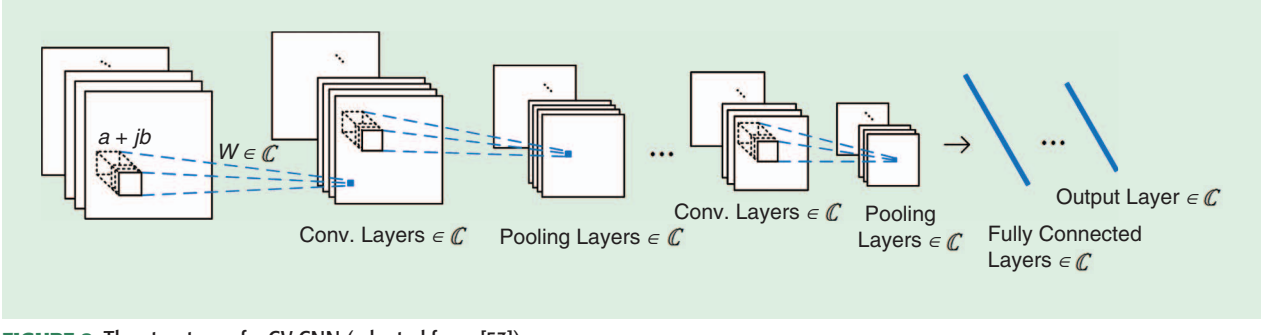

FIGURE 8. The structure of a CV CNN (adapted from [53]).

complex operations throughout different layers. A CV backpropagation algorithm is also developed to train it. Figure 9 shows an example of PolSAR classification using a CV CNN.

PARAMETER INVERSION

The authors of [67] applied CNNs to estimate ice concentration using SAR images during melt season. The labels were produced by visual interpretation by ice experts and tested on dual-polarized RadarSat-2 data. Because the problem considered is regression of a continuous value, the loss function is selected as mean-squared error. The final results suggest that CNNs can produce a more detailed result than operational products.

INTERPRETATION OF HIGH-RESOLUTION SATELLITE IMAGES

SCENE CLASSIFICATION

Scene classification, which aims to automatically assign a semantic label to each scene image, has been an active research topic in the field of high-resolution satellite images in past decades [68]–[74]. As a key problem in the interpretation of satellite images, it has widespread applications, including object detection [75], [76], change detection [77], urban planning, and land resource management. However, due to the high spatial resolutions, different scene images may contain the same kinds of objects or share similar spatial arrangement, e.g., both residential areas and commercial areas may contain buildings, roads, and trees, but they are two different scene types. Therefore, the great variations in the spatial arrangements and structural patterns make scene classification a considerably challenging task.

Generally, scene classification can be divided into two steps: feature extraction and classification. With the growing number of images, training a complicated nonlinear classifier is time consuming. Hence, extracting a holistic and discriminative feature representation is the most significant step for scene classification. Traditional approaches are most often based on the bag-of-visual-words (BoVW) model [78], [79], but their potential for improvement has been limited by the ability of experts to design the feature extractor and the expressive power encoded.

The deep arhitectures discussed in the "Convolutional Neural Networks" section have been applied to the scene classification problem of high-resolution satellite images and led to state-of-the-art performance [71], [74], [80]–[87]. As deep learning is a multilayer feature-learning architecture, it can learn more abstract and discriminative semantic features as the depth grows and achieve far better classification performance compared to midlevel approaches. In this section, we summarize the existing deep learning-based methods according to the following three categories:

- *Using pretrained networks*: The deep CNN pretrained on a natural image data set, e.g., OverFeat [88] and GoogLeNet [89], has led to impressive results on the scene classification of high-resolution satellite images by directly extracting the features from the intermediate layers to form global feature representations [81]–[83], [87]; e.g., [74], [81], and [82] directly use the features from the fully connected layers as the input of the classifier, while [83] takes the CNN as a local feature extractor and combines it with feature coding techniques, such as BoVW [78] and vector of locally aggregated descriptors (VLAD), to generate the final image representation.
- *Making a pretrained model adapt*: Making a pretrained model adapt to the specific conditions observed in a data set under study, one can decide to fine-tune it on a smaller labeled data set of satellite images. The authors of [82] and [86], e.g., fine-tune some high-level layers of the GoogLeNet [89] using the University of California–Merced

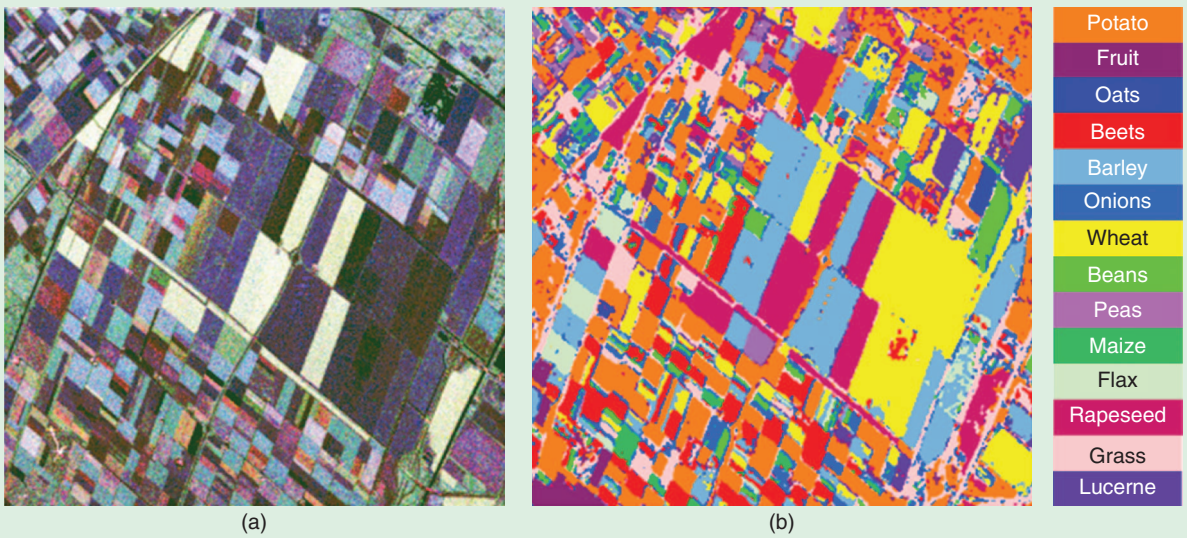

FIGURE 9. The Flevoland data set. (a) The Pauli RGB of the PolSAR data set. (b) The classification result from [53].

(UC Merced) data set [90] (see the "Remote-Sensing Data for Training Deep Learning Models" section), thus obtaining better results than when directly using only the pretrained CNNs. This can be explained, because the features learned are more oriented to the satellite images after fine-tuning, which can help exploit the intrinsic characteristic of satellite images. Nonetheless, compared with the natural image data set consisting of more than 10 million samples, the scales of public satellite image data sets (i.e., UC Merced data set [90], RSSCN7 [80], and WHU-RS19 [91]) are fairly small, e.g., up to several thousands, for which we cannot fine-tune whole CNNs to make them more adaptive to satellite images.

- *Training new networks*: In addition to the previous two ways to use deep learning methods for classifying satellite images, some researchers train the network from scratch using satellite images. The authors of [82] and [86], e.g., train the networks by using only the existing satellite image data set, which suffers a drop in classification accuracy compared with using the pretrained networks as global feature extractors or fine-tuning the pretrained networks. The reason lies in the fact that large-scale networks usually contain millions of parameters to be learned. Thus, training them using small-scale satellite image data sets will easily cause overfitting and local minimum problems. Consequently, some construct a new smaller network and train it from scratch using satellite images to better fit the satellite data [80], [84], [85], [92]. However, such small-scale networks are often easily oriented to the training images, and the generalization ability decreases. For each satellite data set, the network needs to be retrained.

OBJECT DETECTION

Object detection is another important task in the interpretation of high-resolution satellite images [93]: one wishes to localize one or more specific ground objects of interest (such as a building, vehicle, or aircraft) within a satellite image and predict the corresponding categories, as shown in Figure 10. Due to the powerful ability of learning high-level (more abstract and semantically meaningful) feature representations, deep CNNs are being explored in object-detection systems in contrast to the more traditional methods followed by a classifier based on handcrafted features [94], [95]. Here, we review most existing works using CNNs for both specific and generic object detection.

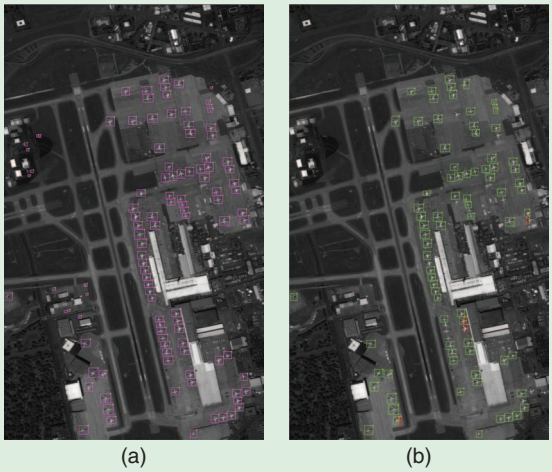

FIGURE 10. An illustration of a typical object-detection result within a high-resolution satellite image. (a) The annotated ground truth of targets of interests (airplanes). (b) The airplanes detected by a CNN-based detector.

Jin and Davis [96] proposed a vector-guided vehicle detection approach for IKONOS satellite imagery using a morphological shared-weight NN that learns the implicit vehicle model and incorporates both spatial and spectral characteristics and classifies pixels into vehicles and nonvehicles. To address the problem of large-scale variance of objects, Chen et al. [97] suggested a hybrid deep CNN model for vehicle detection in satellite images; this model divides all feature maps of the last convolutional and max-pooling layer of the CNN into multiple blocks of variable-receptive-field size or pooling size to extract multiscale features. Jiang et al. [98] proposed a CNN-based vehicle detection approach, wherein a graph-based superpixel segmentation is used to extract image patches and a CNN model is trained to predict whether a patch contains a vehicle.

A few detection methods transfer the pretrained CNNs for object detection. Zhou et al. [99] presented a weakly supervised learning framework to train an object detector; here, a pretrained CNN model is transferred to extract high-level features of objects, and the negative bootstrapping scheme is incorporated into the detector training process to provide faster convergence of the detector. Zhang et al. [100] advanced a hierarchical oil tank detector, which combines deep surrounding features extracted from the pretrained CNN model with local features (histogram of oriented gradients [101]). The candidate regions are selected by an ellipse and line segment detector. Salberg [102] proposed extracting features from the pretrained AlexNet model and applying the deep CNN features for automatic detection of seals in aerial images. Ševo and Avramović [103] suggested a two-stage approach for CNN training and developed an automatic object-detection method based on a pretrained CNN, where the GoogLeNet is first fine-tuned twice on the UC Merced data set using different fine-tuning options and then the fine-tuned model is used for sliding-window object detection. To address the problem of orientation variations of objects, Zhu et al. [104] have employed the pretrained CNN features that are extracted from combined layers and implemented orientation-robust object detection in a coarse localization framework.

Zhang et al. [105] proposed a weakly supervised learning approach using coupled CNNs for aircraft detection. The authors employed an iterative, weakly supervised framework that simply requires image-level training data to automatically mine and augment the training data set

from the original image, which can dramatically decrease human labor. A coupled CNN model, composed of a candidate region proposal network, and a localization network were developed to generate region proposals and locate aircraft simultaneously, which is suitable and effective for large-scale, high-resolution satellite images.

For enhancing the performance of generic object detection, Cheng et al. [76] proposed an effective approach to learning a rotation-invariant CNN (RICNN) to improve invariance to object rotation. In their paper, they add a new rotation-invariant layer to the off-the-shelf AlexNet model. The RICNN is learned by optimizing a new object function, including an additional regularization constraint that enforces the training samples before and after being rotated to share similar features and guarantee the rotation-invariant ability of the RICNN model.

Finally, several papers considering methods other than CNNs exist. Tang et al. [106] offered a compressed-domain ship detection framework combined with SDA and an extreme learning machine (ELM) [107] for optical spaceborne images. Two SDA models are employed for hierarchical ship feature extraction in the wavelet domain, which can yield more robust features under changing conditions. The ELM was introduced for efficient feature pooling and classification, making the ship detection accurate and fast. Han et al. [108] advanced an effective object-detection framework, exploiting weakly supervised learning and DBNs. The system requires only weak labels to identify the presence of an object in the whole image and significantly reduces the labor of manually annotating training data.

IMAGE RETRIEVAL

Remote-sensing image retrieval aims at retrieving images having a similar visual content with respect to a query image from a database [109]. A common image-retrieval system needs to compute image similarity based on image feature representations, and thus the performance of a retrieval depends to a large degree on the descriptive capability of image features. Building image representation via feature coding methods (e.g., BoVW and VLAD) using low-level handcrafted features has been proven to be very effective in aerial image retrieval [109], [110]. Nevertheless, the discriminative ability of low-level features is very limited, and thus it is difficult to achieve substantial performance gain. Recently, a few works have investigated extracting deep feature representations from CNNs. Napoletano [111] extracts deep features from the fully connected layers of the pretrained CNN models, and the deep features prove to perform better than low-level features regardless of the retrieval system. Zhou et al. [112] proposed a CNN architecture followed by a three-layer perceptron, which is trained on a large remote-sensing data set and able to achieve remarkable performance even with low-dimensional deep features. Jiang et al. [113] present a sketch-based satellite-image-retrieval method that involves learning deep cross-domain features, which enables the retrieval of satellite images with hand-free sketches only.

Although there is still a lack of sufficient study in terms of exploiting deep learning approaches for remote-sensing image retrieval at present, considering the great potential for learning high-level features with deep learning methods, we believe that more deep learning-based image-retrieval systems will be developed in the near future. It is also worth noticing how feedback from users is integrated into the deep learning retrieval scheme.

MULTIMODAL DATA FUSION

Data fusion is one of the fast-moving areas of remote sensing [114]–[116]. Due to recent increases in the availability of sensor data, using big and heterogeneous data to study environmental processes has become more tangible. Of course, when data are big and relations to be unveiled are complex, one would favor high-capacity models. In this respect, deep NNs are natural candidates to tackle the challenges of modern data fusion in remote sensing. In this section, we review three areas of remote-sensing image analysis where data fusion tasks have been approached with deep learning: pansharpening, feature and decision-level fusion, and fusion of heterogeneous sources.

WE BELIEVE THAT MORE DEEP LEARNING-BASED IMAGE-RETRIEVAL SYSTEMS WILL BE DEVELOPED IN THE NEAR FUTURE.

PANSHARPENING AND SUPERRESOLUTION

Pansharpening is the task of improving the spatial resolution of multispectral data by fusing these with data characterized by sharper spatial information. It is a special instance of the more general problem of superresolution. Traditionally, the field was dominated by works fusing multispectral data with panchromatic bands [117], but more recently it has been extended to thermal [118] or hyperspectral images [119]. Most techniques rely either on projective methods, sparse models, or pyramidal decompositions. Using deep NNs for pansharpening multispectral images is certainly an interesting concept, because most images acquired by satellite such as the WorldView series or Landsat come with a panchromatic band. In this respect, training data are abundant, which is in line with the requirements of modern CNNs.

A first attempt in this direction can be found in [120], where the authors use a shallow network to upsample the intensity component obtained after the intensity, hue, and saturation of color images [red, green, blue (RGB)]. Once the multispectral bands have been upsampled with the CNN, a traditional Gram–Schmidt transform is used to perform the pansharpening. The authors use a data set of QuickBird images for their analysis. Even though this is interesting, in [120], the authors simply replace one operation (the nearest neighbor or bicubic convolution) with a CNN.

In [121], the authors propose using a CNN to learn the pansharpening transform end to end, i.e., letting the CNN perform the whole pansharpening process. In their CNN, they stack upsampled spectral bands with the panchromatic band and then learn, for each patch, the high resolution values of the central pixel.

In [122], the authors use a superresolution CNN trained on natural images [123] as a pretrained model and fine-tune it on a data set of hyperspectral images. By doing so, they make an attempt at transfer learning [124] between the domains of color (three bands, large bandwidths) and hyperspectral images (many bands, narrow bandwidths). Fine-tuning existing architectures that have been trained on massive data sets with very large models is often a relevant solution, because one makes use of discriminative strong features and injects only task-specific knowledge.

In [125], the authors present an upsampling of the panchromatic band via a stack of AEs: the model is trained to predict the full-resolution panchromatic image from a downsampled version of itself (at the resolution of the multispectral bands). Once the model is trained, the multispectral bands are fed into the model one by one and thereby upsampled using the data relationships learned from the panchromatic images.

FEATURE- AND DECISION-LEVEL FUSION FOR IMAGE CLASSIFICATION

Most of the current remote-sensing literature dealing with deep NNs studies the problem of image classification, i.e., the task of assigning each pixel in the image to a given semantic class (land use, land cover, damage level, and so forth). In the following, we review recent approaches dealing with image classification problems, mostly at very high resolution, using two strategies: feature-level fusion and decision-level fusion. In the last part of this section, we review works using different data sources to tackle separate but related predictive tasks, or multitask problems.

FEATURE-LEVEL FUSION

Feature-level fusion uses multiple sources simultaneously in a network. Like most image-processing techniques, deep NNs use d-dimensional inputs. A very simple way of using multiple data sources in a deep network is to stack them, i.e., to concatenate the image sources into a single data cube to be processed. The filters learned by the first layer of the network will, therefore, depend on a stack of different sources. Studies considering this straightforward extension of NNs are numerous and, in [126], the authors compared networks trained on RGB data (fine-tuned from existing architectures) with networks including a digital surface model (DSM) channel, using the 2015 Data Fusion Contest data set over the city of Zeebruges [127] (data are available from [188]; also see the "Remote-Sensing Data for Training Deep Learning Models" section). They use the CNN as a feature extractor and then use the features to train an SVM, predicting a single semantic class for the entire patch. They then apply the classifier in a sliding-window manner.

Parallel research has considered spatial structures in the network by training architectures predicting all labels in the patch instead of a single label to be attributed to the central pixel. By doing so, spatial structures are inherently included in the filters. Fully convolutional and deconvolutional approaches are natural candidates for such a task: in the first, the last fully connected layer is replaced with a convolutional layer (see [88]) to have a downsized patch prediction that then needs to be upsampled. In the second, a series of deconvolutions (transposed convolutions [7], [8]) are learned to upsample the convolutional fully connected layer. Both approaches have been compared in [92] using the International Society for Photogrammetry and Remote Sensing (ISPRS) Vaihingen and Potsdam benchmark data sets (available in [189]; also see the "Remote-Sensing Data for Training Deep Learning Models" section), stacking color infrared (CIR) and normalized digital elevation models. The architectures are compared and some zoomed results are reported in Figures 11 and 12, respectively. Other strategies for spatial upsampling have been proposed in recent literature, including the direct use of upsampled activation maps as features to train the final classifier [128]. In [129], the authors studied the possibility of visualizing uncertainty of predictions (applying the model of [130]). They stacked CIR, DSM, and normalized DSM data as inputs to the CNN.

In addition to dense predictions, other strategies have been presented to include spatial information in deep NNs. For example, the authors of [58] extract different types of spatial filters and stack them in a single tensor, which is then used to learn a supervised stack of AEs. They apply their models on the classification of SAR images, so

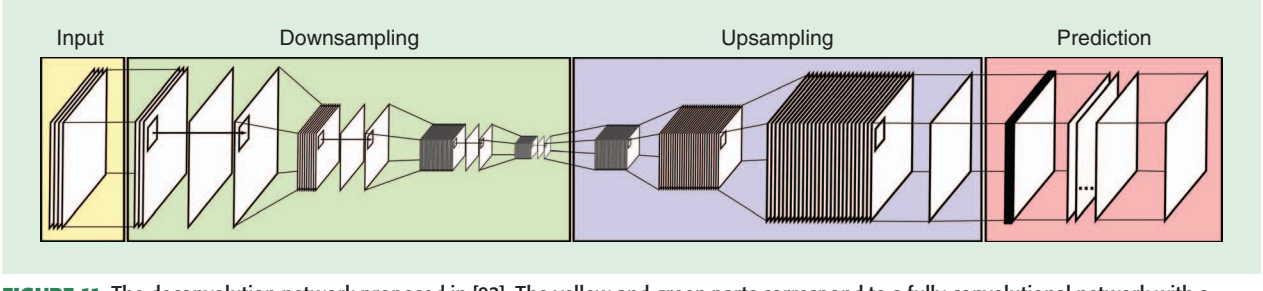

FIGURE 11. The deconvolution network proposed in [92]. The yellow and green parts correspond to a fully convolutional network with a 9 × 9-pixels bottleneck; then, a deconvolutional block (purple) leads to predictions of the same size as the input image (in [92], 65 × 65 pixels).

the fusion here is to be considered between different types of spatial information. The NN is then followed by a conditional random field (CRF) to decrease the effect of speckle noise inherent in SAR images. In [131], the authors present a model that learns combinations of spatial filters extracted from hyperspectral image bands and DSMs. Even if the model is not a traditional deep network, it learns a sequence of recombinations of filters, thereby extracting higher-level information in an automatic way as deep NNs do, i.e., it learns the right filter parameters (along with their combinations) instead of learning the filter coefficients themselves.

Data fusion is also a key component in change detection, where one wishes to extract joint features from a bitemporal sequence. The aim is to learn a joint representation in which both (coregistered) images can be compared. This area is especially interesting when methods can align data from multiple sensors (see [132] and [133]). Three studies employ deep learning to this end:

- In [134], the authors present a model that learns a joint representation of two images with DBNs. Feature vectors issued from the two image acquisitions are stacked and used to learn a representation, where changes stand out more clearly. Using such representation, changes are more easily detected by image differencing. This approach is applied on optical images from the Chinese GaoFen-1 satellite and WorldView-2.
- In [135], the joint representation is learned via a stack of AEs using the single temporal acquisitions at each end of the encoder–decoder system. By doing so, they learn a representation useful for change detection at the bottleneck of the system (i.e., in the middle). The authors show the versatility of their approach by applying it to several data sets, including pairs of optical and SAR images, and an example performing change detection between optical and SAR images.
- More recent work addresses the transferability of deep learning for change detection, while analyzing data of long time series for large-scale problems. In [38], e.g., the authors make use of an end-to-end RNN to solve the multi/hyperspectral change detection task, because RNN is well known to be good at processing sequential data. In their framework, an RNN based on long short-term memory is employed to learn joint spectral feature representations from a bitemporal image sequence. In addition, the authors show that their network can detect multiclass changes and has a good transferability for change detection in a new scene without fine-tuning. The authors of [136] introduce an RNN-based transfer-learning approach to detect annual urban dynamics of four cities

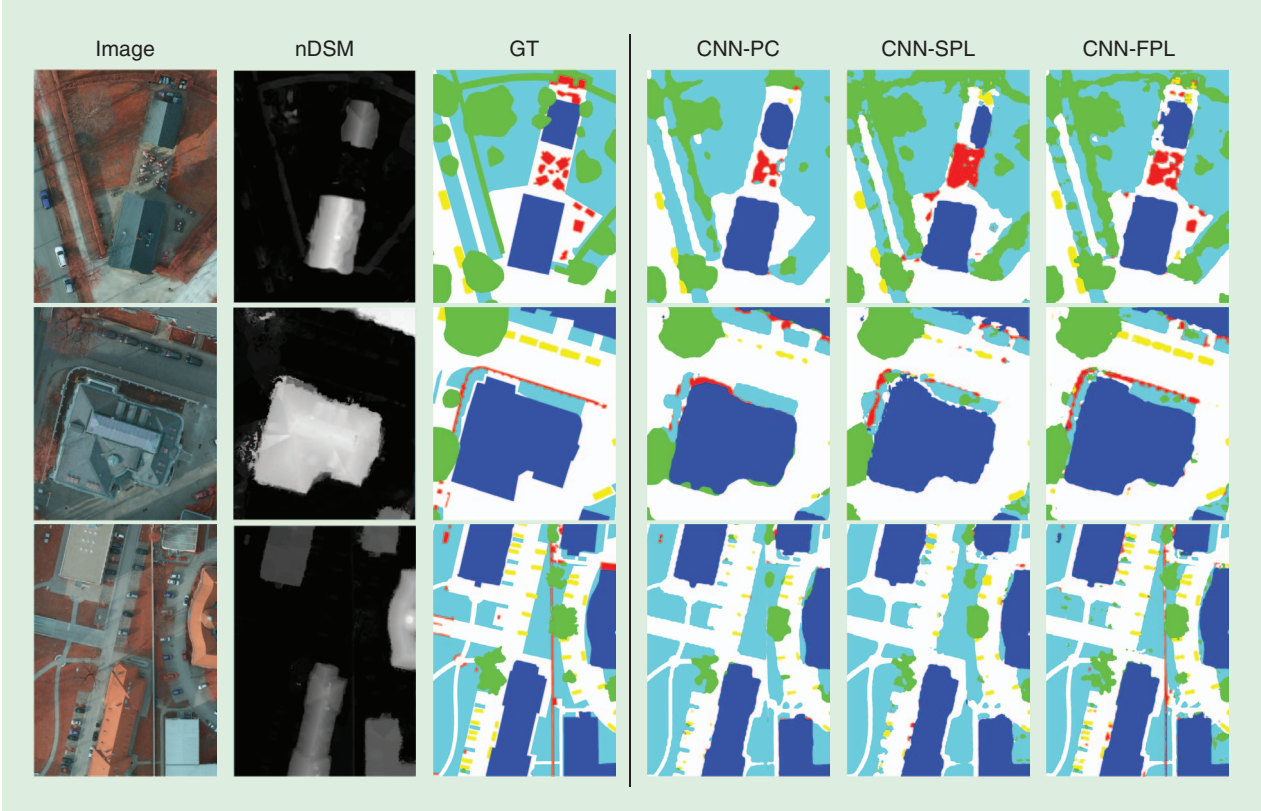

FIGURE 12. The image classification results on the Potsdam data sets, considering 65 × 65-pixels patches (from [92]). CNN-PC: patch-based CNN, predicting single labels per patch and using a sliding-window approach; CNN-SPL: fully convolutional CNN, predicting a 9 × 9 output, then upsampled to the original size via interpolation; CNN-FPL: deconvolutional network predicting the 65 × 65 output at full resolution; nDSM: normalized digital surface model; GT: ground truth.

(Beijing, New York, Melbourne, and Munich) from 1984 to 2016, using Landsat data. The main challenge here is that training data in such a large-scale and long-term image sequence are very scarce. By combining RNN and transfer learning, the authors are able to transfer the feature representations learned from a few training samples to new target scenes directly. Some zoomed results are reported in Figure 13.

Another view of feature fusion involves NNs fusing features obtained from different inputs: two (or more) networks are trained in parallel, and their activations are then fused at a later stage, e.g., by feature concatenation. The author of [137] studies a solution in this direction that fuses two CNNs: the first considers CIR images of the Vaihingen data set and passes them through the pretrained VGG network to learn color features, while the second considers the DSM and learns a fully connected network from scratch. Both models' features are then concatenated, and two randomly initialized, fully connected layers are learned from this concatenation. A similar logic is also followed in [138], where the authors present a model that learns a fully connected layer performing the fusion between networks learned at different spatial scales. They apply their model to the tasks of buildings and road detection. In [139], the authors train a two-stream CNN with two separate yet identical convolutional streams that process the PolSAR and hyperspectral data in parallel and only fuse the resulting information at a later convolutional layer for the purpose of land cover classification. With a similar network architecture and contrastive loss function, the authors of [140] present a model that learns a network for the identification of corresponding patches in SAR and optical imagery of urban scenes.

DECISION-LEVEL FUSION

Decision-level fusion fuses CNN (and other) outputs. While the works reviewed previously use a single network to learn the semantics of interest all at once (either by extracting relevant features or by learning the model end to end), another series of works has studied ways of performing decision fusion with deep learning. Even though the distinction between these and the models reviewed previously might seem artificial, here we review approaches including an explicit fusion layer between land cover maps. We distinguish between two families of approaches, depending on whether the decision fusion is performed as a postprocessing step or learned.

- *Fusing semantic maps obtained with CNNs*: In this case, different models predict the classes, and their predictions are then fused. Two works are particularly notable in this respect. On the one hand [141], the authors fuse a classification map obtained by a CNN with another obtained by a random forest classifier trained using handcrafted features. Both models use CIR, DSM, and normalized DSM inputs from the Vaihingen data set. The two maps are fused by multiplication of the posterior probabilities, and an edge-sensitive CRF is also learned on top to improve the quality of the final labeling. On the other hand, the authors in [133] consider the learning of an ensemble of CNNs and then averaging their predictions: their proposed pipeline has two main streams,

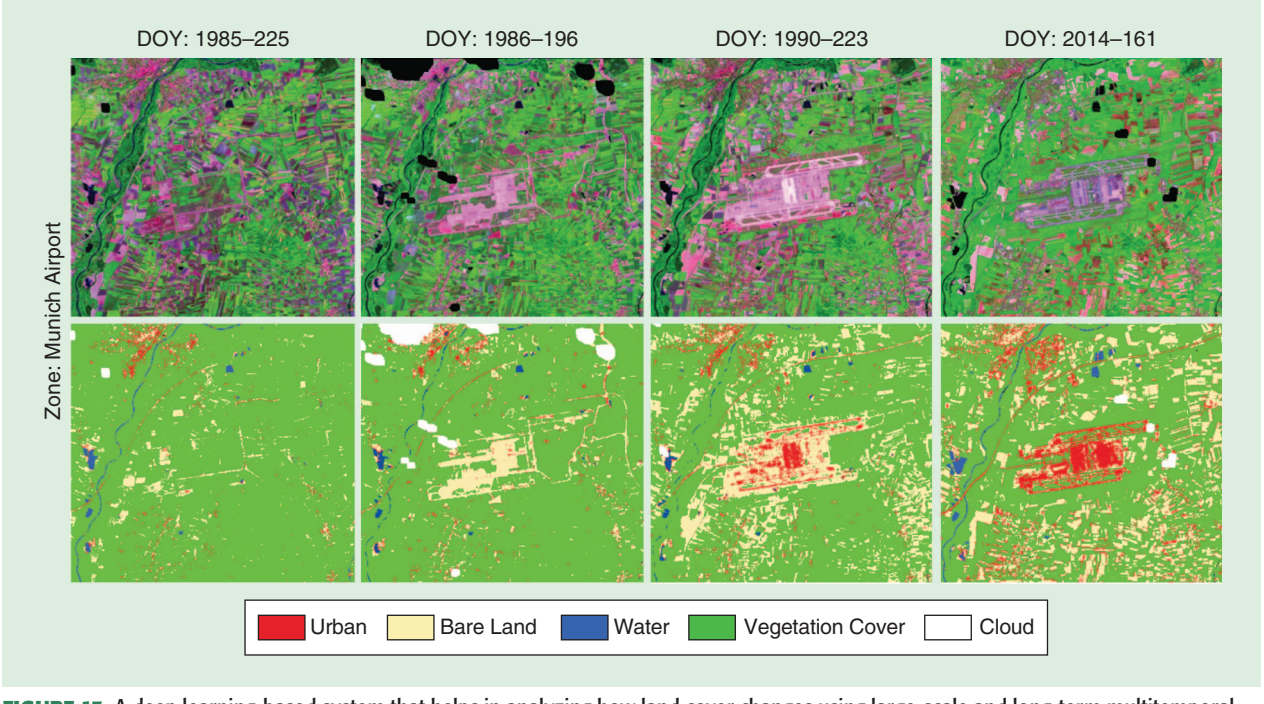

FIGURE 13. A deep learning-based system that helps in analyzing how land cover changes using large-scale and long-term multitemporal image sequences. This example shows how Munich airport was built out over the past 30 years. DOY: day of year.

one processing the CIR data and another processing the DSM. They train several CNNs, using them as inputs for the activation maps of each layer of the main model as well as one fusing the CIR and DSM main streams, as in [137]. By doing so, they obtain a series of land cover maps to nourish the ensemble, which improves performances by considering classifiers issued from different data sources and levels of abstraction. Compared to the previous model discussed in this section, this model has the advantage of being entirely learned in an end-to-end fashion, but it also incurs an extreme computational load and a complex architecture involving many skip connections and fusion layers.

- *Decision fusion learned in the network*: This is an alternative to an ad hoc fusion (multiplication or averaging of the posterior maps), in which one may learn the optimal fusion. In [142], the authors perform the fusion between two maps obtained by pretrained models by learning a fusion network based on residual learning [4] logic. In their architecture, they learn how to correct the average fusion result by learning extra coefficients favoring one or the other map. Their results show that such a learned fusion outperforms the more intuitive, simple averaging of the posterior probabilities.

USING CNNs FOR SOLVING DIFFERENT TASKS

So far, only literature dealing with a single task (image classification) has been reviewed. But, besides this, one might want to predict other quantities or use the image-classification results to improve the quality of related tasks such as image registration. In this case, predicting different outputs jointly allows one to tighten feature representations with different meanings, thereby leading to another type of data fusion with respect to the ones described earlier (which were concerned mainly with fusing different inputs). Here, we discuss fusing outputs and describe three examples from recent literature wherein alternative tasks are learned together with image classification.

- *Edges*: In the previous section, we discussed the work of Marcos et al. [133], in which the authors produced and fused an ensemble of land cover maps. In [143], that work was extended by including the idea of predicting object boundaries jointly with the land cover. The intuition behind this is that predicting boundaries helps to achieve sharper (and therefore more useful) classification maps. In [143], the authors present a model that learns a CNN to separately output edge likelihoods at multiple scales from CIR and height data. Then, the boundaries detected with each source are added as an extra channel to each source, and an image classification network, similar to the one in [133], is trained. The predictions of such a model are very accurate, but the computational load involved becomes very high: the authors report models involving up to 800 million parameters to be learned.
- *Depth*: Some approaches discussed previously include the DSM as an input to the network. But, often, such information is not available (and it is certainly not when working on historical data). A system predicting a height map from image data would indeed be very valuable, because it could generate reasonably accurate DSM for color image acquisitions. This is known in vision as the *problem of estimating depth maps* [144] and has been considered in [145] for monocular subdecimeter images. In their models, the authors use a joint-loss function, which is a linear combination of a dense-image-classification loss and a regression loss minimizing DSM predictions errors. The model can be trained by traditional backpropagation by alternating over the two losses. Note that, in this case, the DSM is used as an output (contrary to most approaches discussed previously) and is, therefore, not needed at prediction time.

SERVICES LIKE GOOGLE STREET VIEW AND FLICKR PROVIDE ENDLESS SOURCES OF GROUND IMAGES DESCRIBING CITIES FROM THE HUMAN PERSPECTIVE.

- *Registration*: When performing change detection, one expects perfect coregistration of the sources. But, especially when working at very high resolution, this is difficult to achieve. Think of urban areas, e.g., where buildings are tilted by the viewing angle. In their entry to the IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest 2016 (data are available from [188]), the authors of [146] present a model that learns jointly the registration between the images, the land use classification of each input, and a change detection map with a CRF model. The land use classifier used is a two-layer CNN trained from scratch; the model is applied successfully either to pairs of very high-resolution (VHR) images or to data sets composed of VHR images and video frames from the International Space Station.

FUSING HETEROGENEOUS SOURCES

Data fusion is not only about fusing image data with the same viewpoint. Multimodal remote-sensing data that exceed these restrictive boundaries and approaches to tackle new, exciting problems with remote sensing are beginning to appear in the literature. An excellent example is the joint use of ground-based and aerial images [147]: services like Google Street View and Flickr provide endless sources of ground images describing cities from the human perspective. These data can be fused to overhead views to provide better object detection, localization, or re-creation of virtual environments. In the following, we review a series of applications in this area.

In [148], the authors consider the task of detecting and classifying urban trees. To this end, they exploit Faster R-CNN [149], an object detector developed for general-purpose object detection in vision. After detecting the trees in

the aerial view and the Google Street View panoramas, they minimize an energy function to detect trees jointly in all sources but also avoid multiple and illogical detections (e.g., trees in the middle of a street). They use a trees inventory from the city of Pasadena to validate their detection model and train a fine-grained CNN based on GoogLeNet [89] to perform a fine-grained classification of the tree species on the detections, with impressive results. The authors of [150] take advantage of an approach that combines a CNN and an MRF and can estimate fine-grained categories (e.g., road, sidewalk, background, building, and parking) by performing joint inference over both monocular aerial imagery and ground images taken from a stereo camera on top of a car.

THE PROCESSING OF IMAGE DATA FROM AIRBORNE SENSORS OR SATELLITE SYSTEMS IS A LONG-STANDING TRADITION.

Many papers in geospatial computer vision work toward cross-view image localization: when presented to a ground picture, it would be relevant to be able to locate images in space. This is very important in photo-sharing platforms, for which only a fraction of the uploaded photos comes with geolocation. The authors of [151] and [152] worked toward this aim, by training a cross-view Siamese network [153] to match ground images and aerial views. Siamese networks have also been recently applied [147] to detect changes between matched ground panoramas and aerial images. Returning to more traditional CNNs, the authors of [154] and [155] study the specificity of images to refer to a given city: they study how closely images of Charleston, South Carolina, resemble those of San Francisco, and the other way around, by using the fully connected layers of Places CNN [156] and then translating this into differences in the respective aerial images. Moreover, in [154], they also present applications on image localization similar to those mentioned previously, where the likelihood of localization is given by a similarity score between the features of the fully connected layer of Places CNN.

3-D RECONSTRUCTION

The 3-D data generation from image data plays an important role for remote sensing. The 3-D data (e.g., in the form of a DSM or digital terrain model is a basic data layer for further processing or analysis steps. The processing of image data from airborne sensors or satellite systems is a long-standing tradition. In a typical 3-D data-generation workflow, two main steps must be performed. First is camera orientation, which refers to computing the position and orientation of the cameras that produced the image. This can be computed from the image data themselves, by identifying and matching tie points and then performing camera resectioning. The second step is triangulation, which calculates the 3-D measurements for point correspondences established through stereo matching. The fundamental algorithms in this pipeline are geometric in nature, and the implementations are based on analytical calculations. So far, machine learning has not played a major role in this pipeline. However, there are steps in this pipeline that could be improved significantly by using machine-learning techniques.

TIE-POINTS IDENTIFICATION AND MATCHING

During camera orientation, the identification and matching of tie points have long been accomplished manually by operators. The task of the operator was to identify corresponding locations in two or more images. This process has been automated by clever engineering of computer algorithms to detect point locations in images that will be easy to redetect in other images (e.g., corners) as well as algorithms for computing similarities of image patches for finding a tie-point correspondence. Many different detectors and similarity measures have been engineered so far—famous examples are the SIFT [157] or SURF [158] features. However, all these engineered methods fall short (i.e., they are still less accurate than humans). This is a domain in which machine learning and, in particular, CNNs are employed to learn, based on an enormous number of correct tie-point matches and point locations, the similarity metrics between image patches.

In the area of tie-point detection and matching, Fischer et al. [159] used a CNN to learn a descriptor for image patch matching from training examples, similar to the well-known SIFT descriptor. In this article, the authors trained a CNN with five convolutional layers and two fully connected layers. The trained network computes a descriptor for a given image patch. In the experiments on standard data sets, the authors could show that the trained descriptors outperform engineered descriptors (i.e., SIFT) significantly in a tie-point matching task. Similar successes are described in other works such as those by Handa et al. [160], Lenc and Vedaldi [161], and Han et al. [162] The work of Yi et al. [163] takes this idea one step further: the authors propose a deep CNN to detect tie-point locations in an image and output a descriptor vector for each tie point.

STEREO PROCESSING USING CNNs

The second important step in this workflow is stereo matching, i.e., the search for corresponding pixels in two or more images. In this step, a corresponding pixel is sought for every pixel in the image. In most cases, this search can be restricted to a line in the corresponding image. However, current methods still make mistakes in this process. The semiglobal matching (SGM) approach by Hirschmueller [164] served as the gold-standard method for a considerable time.

Since 2002, progress on stereo processing is tracked by the Middlebury stereo evaluation benchmark (http://vision.middlebury.edu/stereo/). The benchmark allows comparison of results of stereo-processing algorithms to a carefully maintained ground truth. The performance of the different algorithms can be viewed as a ranked list. This

ranking reveals that, today, the top performing method is based on CNNs.

Most stereo methods in this ranking proceed along the following main steps. First, a stereo correspondence search is performed by computing a similarity measure between image locations. This is typically carried out exhaustively for all possible depth values. Next, the optimal depth values are searched by optimization on the cost value. Different optimization schemes—convex optimization, local-optimization strategies (e.g., SGM), and probabilities methods (e.g., MRF inference)—are used. Finally, some heuristic filtering is typically applied to remove gross outliers (e.g., left–right check).

The pioneering work of Zbontar and LeCun [165] utilized a CNN in the first step of the typical stereo pipeline. In their work, the authors suggested training a CNN to compute the similarity measure between image patches (instead of using normalized cross correlation or the census transform). This change proved to be significant. Compared to SGM, which is often considered a baseline method, the proposed method achieved a significantly lower error rate. For SGM, the error rate was still 18.4%, whereas for the matching-cost(MC)-CNN method, the error rate was only 8%. After that, other variants of CNN-based stereo methods have been offered, and the best ranking method today has an error rate of only 5.9%. Table 1 lists the error rates of the top-ranking CNN-based methods.

TABLE 1. THE TOP-RANKED STEREO METHODS FROM THE MIDDLEBURY STEREO EVALUATION BENCHMARK AS OF MAY 2017.

METHOD	BAD PIXEL ERROR RATE %
3DMST [166]	5.92
MC-CNN + TDSR [167]	6.35
LW-CNN [168]	7.04
MC-CNN-acrt [165]	8.08
SGM [164]	18.4

LW-CNN: look wider CNN; TDSR: top-down segmented regression.

In addition to similarity measures, a typical stereo-processing pipeline contains other engineered decisions as well. After creating a so-called cost volume from the similarity measures, most methods use specifically engineered algorithms to find the depths (e.g., based on neighborhood constraints) and heuristics to filter out wrong matches. New proposals, however, suggest that these other steps can also be replaced solely by a CNN. Mayer et al. [169] offered such a paradigm-shifting design for stereo processing. In their proposal, the stereo-processing problem is modeled solely as a CNN. The proposed CNN takes two images of a stereo pair as an input and directly outputs the final disparity map. A single CNN architecture replaces all the individual algorithm steps utilized so far. The CNN of Mayer is based on an encoder–decoder architecture with a total of 26 layers. In addition, it includes crosslinks between contracting and expanding network parts. To train the CNN architecture, end-to-end training using ground truth image-depth map pairs is performed. The fascinating aspect of the proposed method is that the stereo algorithm itself can be learned from data only. The network architecture does not define the algorithm, but the data and the end-to-end training define what type of processing the network should perform.

LARGE-SCALE SEMANTIC 3-D CITY RECONSTRUCTION

The availability of semantics (e.g., the knowledge of what type of object a pixel in the image represents) through CNN-based classification is also changing the way that 3-D information is generated from image data. The traditional 3-D generation process neglected object information: the 3-D data were generated from geometric constraints only, and image data were treated as pure intensity values without any semantic meaning.

The availability of semantic information from CNN-based classification now makes it possible to utilize this information in the 3-D generation process. CNN-based classification allows one to assign class labels to aerial imagery with unprecedented accuracy [170]. Pixels in the images are then assigned labels like vegetation, road, building, and so on. This semantic information can then be used to steer the 3-D data generation process. Class label-specific parameters can be chosen for the 3-D data generation process.

The latest proposal in this area, however, is a joint reconstruction of 3-D and semantic information (Häne et al. [171]), where 3-D reconstruction is performed with a volumetric method. The area to be reconstructed is partitioned into small cells, the size of which define the resolution of the 3-D reconstruction. The reconstruction algorithm now finds the optimal partitioning of this voxel grid into occupied and nonoccupied voxels that fit to the image data. The result is a 3-D reconstruction of the scene. The work of Häne et al. also jointly assigns the 3-D reconstruction to a class label for each voxel, e.g., vegetation, building, road, and sky. Each generated 3-D data point now also has a class label. The 3-D reconstruction is semantically interpretable. This process is a joint process, with the computation of the occupied and nonoccupied voxels taking into account the class labels in the original images. If a voxel corresponds to a building pixel in the image, it is set to "occupied" with high probability. If a voxel corresponds to a sky pixel in the image, it has a high probability of being "unoccupied." If a set of voxels is stacked on top of one another, it is likely that these belong

CNN-BASED CLASSIFICATION ALLOWS ONE TO ASSIGN CLASS LABELS TO AERIAL IMAGERY WITH UNPRECEDENTED ACCURACY.

to some building, i.e., the probability for assigning the label "class of building" is increased for this structure.

This semantic 3-D reconstruction method has been successfully applied to 3-D reconstruction from aerial imagery by Bláha et al. [172], [173]. In their work, they achieved a semantic 3-D reconstruction of cities on large scales. The 3-D model contains not only 3-D data but also class labels, e.g., a 3-D structure that represents buildings gets the class label "building." Even more, every building has its roof structures labeled as "roof." Figure 14 shows an image of a semantic 3-D reconstruction produced by the method described in [172].

In summary, we can say that CNNs quickly took on a significant role in 3-D data generation. Utilizing CNNs for stereo processing significantly boosted the accuracy and precision of depth estimation. The availability of reliable class labels extracted from CNN classifiers opened the possibility of creating semantic 3-D reconstructions, a research area that is poised to grow significantly.

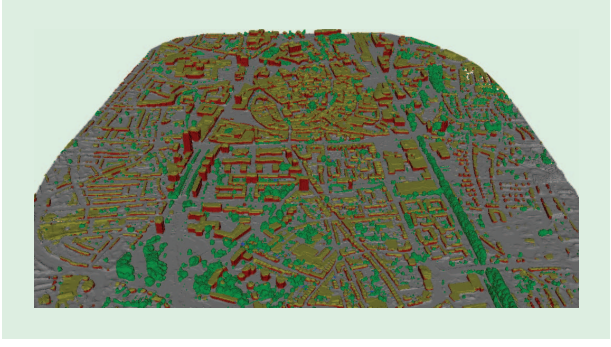

FIGURE 14. A semantic 3-D reconstruction from the Enschede aerial image data set computed with the method described in [172]. The different colors represent different class labels: "ground" (gray), "building" (red), "roof" (yellow), "vegetation" (green), and "clutter" (blue). (Image courtesy of the authors of [172].)

DEEP LEARNING IN REMOTE SENSING MADE RIDICULOUSLY SIMPLE

To provide an easy starting point for researchers attempting to work on deep learning in remote sensing, we list some available resources, including tutorials and open-source deep learning frameworks. In addition, we provide a selected list of open remote-sensing data for training deep learning models as well as some showcasing examples with source codes developed using different deep learning frameworks.

TUTORIALS

Some valuable tutorials for those new to deep learning, including books, survey papers, code tutorials, and videos, can be found at http://deeplearning.net/reading-list/tutorials/. In addition, we list two references [174], [175] that provide some general recommendations for the choice of the parameters.

OPEN-SOURCE DEEP LEARNING FRAMEWORKS

When diving deep into deep learning, the choice of an open-source framework is of great importance. Figure 15 shows the most popular open-source deep learning frameworks, such as Caffe, Torch, Theano, TensorFlow, and Microsoft-CNTK. Because the field and surrounding technologies are relatively new and have been developing rapidly, the most common concerns among people who would like to work on deep learning are how these frameworks differ, where they fall short, and which ones are worth investing in. A detailed discussion of popular deep learning frameworks can be found in [190].

REMOTE-SENSING DATA FOR TRAINING DEEP LEARNING MODELS

To train deep learning methods with good generalization abilities, one needs large data sets. This is true for both fine-tuning models and training small networks from scratch (although if we consider training large architectures, one should preferably resort to pretrained methods) [176]. In

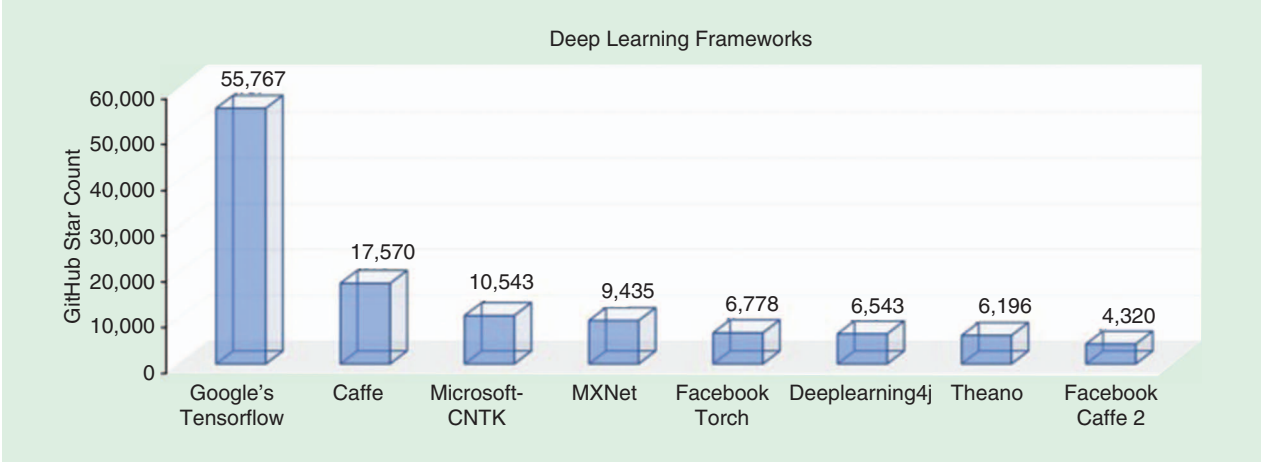

FIGURE 15. The most popular open-source deep learning frameworks. The ranking is based on the number of stars awarded by developers in GitHub. (Image courtesy of [190].)

recent years, several data sets have been made public that can be used to train deep NNs. The following is a nonexhaustive list.

SCENE CLASSIFICATION (ONE IMAGE IS CLASSIFIED INTO A SINGLE LABEL)

- *UC Merced data set* [177]: This data set is a collection of aerial images (256 × 256 pixels in RGB space) depicting 21 land use classes. Each class is made up of 100 images. Because every image comes with a single label, the data set can be used only for image-classification purposes, i.e., to classify the whole image into a single land use class. The data set can be downloaded from [191].
- *Aerial Image data set (AID)* [74]: This data set is a collection of 10,000 annotated aerial images distributed in 30 land use scene classes and can be used for image-classification purposes. In comparison with the UC Merced data set, the AID contains many more images and covers a wider range of scene categories. Thus, it is in line with the data requirements of modern deep learning. The data set can be downloaded from [192].
- *Northwestern Polytechnical University–Remote Sensing Image Scene Classification 45 data set* [178]: This data set contains 31,500 aerial images spread over 45 scene classes. So far, it is the largest data set for land use scene classification in terms of both total number of images and number of scene classes. The data set can be obtained from [193].

IMAGE CLASSIFICATION (EACH PIXEL OF AN IMAGE IS CLASSIFIED INTO A LABEL)

- *Zurich Summer data set* [179]: This data set is a collection of 20 image chips from a single large QuickBird image acquired over Zurich, Switzerland, in 2002. Each image chip is pansharpened to 0.6-m resolution, and eight land use classes are presented. All images are released, along with their ground truths. The data set can be obtained from [194].
- *Zeebruges, or the Data Fusion Contest 2015 data set* [127]: In 2015, the Image Analysis and Data Fusion Technical Committee of the IEEE GRSS organized a data processing competition aimed at 5-cm resolution land mapping. To do so, the organizers provided both an RGB aerial image and a dense (65-points/m^2) lidar point cloud over the harbor of Zeebruges (Belgium). The data are organized on seven 10,000 × 10,000 pixels tiles. All the tiles have been labeled densely in eight land classes, including land use (building, roads) and objects (vehicles, boats) [126]. The data can be obtained from the Data and Algorithm Standard Evaluation (DASE) website, http://dase.ticinumaerospace.com/. On DASE, users can download the seven tiles and labels for five tiles. To assess models on the two remaining tiles, users can upload the classified maps on the DASE server.
- *ISPRS 2-D semantic labeling challenge*: The working group II/4 of the ISPRS 3-D Scene Reconstruction and Analysis provided a subdecimeter resolution data set over the two cities of Vaihingen and Potsdam (Germany). The data are similar to those of the Zeebruges data, with the difference that the height information is provided as a DSM at the same resolution of the image data. Moreover, images are provided with an infrared channel. The data set is also fully labeled into six classes, including land classes (roads, meadows) and objects (cars). It also comes with a clutter class gathering all unknown objects. The Vaihingen data set comes with 33 tiles having an average size of 2,000 × 3,000 pixels. Half the tiles come with labels. The other 17 tiles come with no labels, and participants must upload classification maps for evaluation. The Potsdam data set comes with 24 labeled tiles (6,000 × 6,000 pixels) and 14 unlabeled ones. Both data sets can be obtained from [189].

WITH THE GROWING ATTENTION ON VHR SAR DATA, THE FUSION OF OPTICAL AND SAR IMAGES IN DENSE URBAN AREAS HAS BECOME AN EMERGING AND TIMELY TOPIC.

REGISTRATION/MATCHING

- *SARptical data set* [180]. With the growing attention on VHR SAR data, the fusion of optical and SAR images in dense urban areas has become an emerging and timely topic. At the core of such a fusion topic is the challenging task of coregistering SAR and optical images. Two such images are acquired with intrinsically different imaging geometries and thus are nearly impossible to be coregistered without a precise 3-D model of the imaged scene. SARptical is a unique data set for SAR and optical image matching in dense urban areas. It consists of 10,000 pairs of corresponding SAR and optical image patches in central Berlin, with the center pixels of each patch pair precisely coregistered. They are generated based on coregistered 3-D interferometric SAR point clouds (which are reconstructed by SAR tomography using tens of TerraSAR-X high-resolution spotlight images) and 3-D optical point clouds (which are reconstructed by structure from motion, followed by dense stereo matching using several UltraCam images with a ground spacing of 20 cm). This data set can be downloaded from https://www.sipeo.bgu.tum.de/downloads.

SHOWCASING

Starting to work with CNNs from scratch might seem a titanic task. The number of models available is large, and setting up an architecture from zero is challenging. In this section, we point to three showcasing example that have been recently provided by remote-sensing researchers. All these examples are offered with open licenses, and the corresponding papers must be acknowledged when using those codes. The rules on the respective websites apply. Please

read the specific terms and conditions carefully. Each example, uses a different deep learning library (and programming language).

- *Deconvolution network in MatConvNet*: The first example is released by the authors of [92] and corresponds to the architecture in Figure 11. It exploits the MatConvNet library for MATLAB (http://www.vlfeat.org/matconvnet/) and provides a pretrained network for both the Vaihingen and Potsdam data sets described previously. The initial models are specific to remote-sensing data and have been trained on each data set separately. This example is primarily meant to show how to fine-tune an existing model in MatConvNet by training a few extra iterations to improve the model weights. It can, of course, be trained from scratch by reinitializing the weights randomly. A function to test the additional images of the data sets is also provided. Overall, it allows one to reproduce the results in [92], which are similar to the right-hand column in Figure 12. By removing the deconvolutional part of the network and adding a fully connected layer at the bottleneck, one can reproduce the CNN-PC model. If, instead, one adds a spatial upsampling layer (e.g., a spatial interpolation of the bottleneck), one can also reproduce the results of the CNN-SPL model of Figure 12. In both cases, the models must be retrained (or, at least, heavily fine-tuned). The code can be downloaded from [195].
- *Fully convolutional (SegNet) architecture in Caffe*: This second example is released by the authors of [142] and exploits the Caffe library (http://caffe.berkeleyvision.org/). The model uses the SegNet architecture from Kendall et al. [181]. The authors released the pretrained model to reproduce the results of [142] on the Vaihingen data set. The network configuration, database generation, and training files are given in Python. The code can be downloaded from https://github.com/nshaud/DeepNetsForEO.
- *AConvNet for SAR ATR in Caffe*: The third example is released by the authors of [42]. It implements a CNN-based SAR target recognition demonstrated via the MSTAR data set. It includes the model configuration file and the source code for training and testing as well as a successfully trained CNN model. The code can be downloaded from https://github.com/fudanxu/MSTAR-AConvNet.
- *Residual conv-deconv network in TensorFlow*: This final example is released by the authors of [33] and [34] and shows how to build up a residual conv-deconv network for unsupervised spectral-spatial feature learning of hyperspectral data. It exploits the TensorFlow (https://www.tensorflow.org/) and Keras (https://keras.io/) libraries. The trained network can be transferred for the user's own classification purpose by fine-tuning the target data sets; alternatively, free object detection can be obtained using the learned filters in the first residual block of the residual conv-deconv network. The code can be downloaded from https://www.sipeo.bgu.tum.de/downloads.

THANKS TO THE ENORMOUS SUCCESS ENCOUNTERED IN SEVERAL AREAS OF RESEARCH, REMOTE SENSING IS SURFING THE WAVE OF DEEP NNS.

CONCLUSIONS AND FUTURE TRENDS

In this article, we have reviewed the current state of the art in deep learning for remote sensing. Thanks to the enormous success encountered in several areas of research, remote sensing is surfing the wave of deep NNs, following a trend similarly being pursued in other fields: deep networks are solid models that tend to improve over classical approaches using handcrafted features. Yet, this field is still relatively young and, in the upcoming years, rapid advancement of deep learning in remote sensing is expected. Technical challenges obviously remain, however.

- What further applications in remote sensing might potentially benefit from deep learning? In general, deep networks are particularly beneficial for remote-sensing problems whose physical models are complicated, e.g., nonlinear, or even not yet well understood and/or cannot be generalized. Yet, so far, in various remote-sensing fields, most deep learning-related research has focused on classification- and detection-related tasks using a number of benchmark data sets.
- Is the transferability of deep networks sufficient to extract geoinformation on a global scale? Complex light-scattering mechanisms in natural objects, various atmospheric scattering conditions, intraclass variability, culture-dependent features, and limited training samples make the use of deep learning for global tasks challenging [182]. To meet the need of large-scale applications, possible solutions are never-ending learning [183] and self-taught learning [184].
- How should problems raised by very limited annotated data in remote sensing be tackled? Is it possible to learn deep hierarchical models for remote-sensing image understanding in a weakly supervised, semisupervised, or even unsupervised way? A few inspiring works in machine learning and computer vision are [34], [185], and [186]. How do we benchmark the fast-growing deep-learning algorithms in remote-sensing applications? Some recent initiatives include the 2017 IEEE GRSS Data Fusion Contest data set [196] and the Functional Map of the World Challenge data set [197].

The fusion of physics-based modeling and deep NN is a promising direction. Remote-sensing imagery is a direct product of physical processes, such as light reflection and microwave scattering. It must resort to a synergy of the physics-based models that describe the a priori knowledge of the process behind the imagery and newly developed artificial intelligence technologies.

Besides focusing on technical challenges, deep learning in remote sensing opens up opportunities for new

applications, such as monitoring global changes or evaluating strategies for the reduction of resources consumption. In this context, deep learning offers an incredible tool box that allows researchers in remote sensing to exceed the boundaries of the field, to move beyond traditional small-scale benchmarking tasks and tackle large-scale, real-life problems with implicit models that generalize well. The data are now here, the hardware is ready, and deep learning frameworks are openly available, so it is now time to design models that are tailored to big remote-sensing data and the multimodal, geolocated, and multitemporal aspects we raised in the introduction.

Commercial players are on the march toward remote sensing and Earth observation. Planet, e.g., has launched approximately 140 small satellites that map the whole Earth daily. Standing on the paradigm shift from computational science to data-driven science, we, as remote-sensing experts, must appropriately position ourselves among other data scientists also trying to use deep learning for innovative remote-sensing applications. This requires us, in turn, to bring our domain expertise into deep learning to provide prior knowledge that is tailored to specific remote-sensing problems.

Last but not least, we encourage efforts within the community to share data and architectures and so be able to answer the challenges of the years to come.

ACKNOWLEDGMENTS

The work of Xao Xiang Zhu and Lichao Mou is supported by the European Research Council under the European Unions Horizon 2020 research and innovation program (grant agreement no. ERC-2016-StG-714087, So2Sat), the Helmholtz Association under the framework of the Young Investigators Group SiPEO (VH-NG-1018, www.sipeo.bgu.tum.de), and the China Scholarship Council. The work of Devis Tuia is supported by the Swiss National Science Foundation under project no. PP0P2 150593. The work of Gui-Song Xia and Liangpei Zhang is supported by the National Natural Science Foundation of China (NSFC) projects with grant no. 41501462 and no. 41431175. The work of Feng Xu is supported by the NSFC projects with grant no. 61571134.

AUTHOR INFORMATION

Xiao Xiang Zhu (xiao.zhu@dlr.de) received her M.Sc. degree, Dr.-Ing. degree, and habilitation in the field of signal processing from the Technical University of Munich (TUM), Germany, in 2008, 2011, and 2013, respectively. She has been a professor of signal processing in Earth observation since 2015 at TUM and the German Aerospace Center (DLR); head of Team Signal Analysis (since 2011) at DLR with the Remote Sensing Technology Institute; and head of the Helmholtz Young Investigator Group SiPEO (since 2013), DLR and TUM. Her main research interests are advanced interferometric synthetic aperture radar techniques, computer vision in remote sensing including object reconstruction and multidimensional data visualization, big data analysis in remote sensing, and modern signal processing. She is an associate editor of *IEEE Transactions on Geoscience and Remote Sensing*.

Devis Tuia (devis.tuia@wur.nl) received his M.Sc. and Ph.D. degrees from the University of Lausanne, Switzerland, in 2005 and 2009, respectively. He was a postdoctoral scholar at the University of Valéncia, Spain, the University of Colorado, Boulder, and École Polytechnique Fédérale de Lausanne, Switzerland. He is currently an associate professor with the GeoInformation Science and Remote Sensing Laboratory at Wageningen University, The Netherlands. Between 2014 and 2017, he was an assistant professor at the University of Zurich, Switzerland. His research interests include algorithms for information extraction and geospatial data fusion (including remote sensing) using machine learning and computer vision.

Lichao Mou (lichao.mou@dlr.de) received his bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, China, in 2012 and his master's degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, in 2015. In 2015, he spent six months with the Computer Vision Group at the University of Freiburg in Germany. He is currently working toward his Ph.D. degree at the German Aerospace Center (DLR) and the Technical University of Munich, Germany. His research interests include remote sensing, computer vision, and machine learning, especially remote-sensing video analysis and applications for deep neural networks in remote sensing. He was first place in the 2016 IEEE Geoscience and Remote Sensing Society's Data Fusion Contest and a finalist for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event.

Gui-Song Xia (guisong.xia@whu.edu.cn) received his B.S. degree in electronic engineering and M.S. degree in signal processing from Wuhan University, China, in 2005 and 2007, respectively, and a Ph.D. degree in image processing and computer vision from the Centre National de la Recherche Scientifique (CNRS), Laboratoire de Traitement et Communication de l' Information, Telecom ParisTech, France, in 2011. He is currently a professor with the State Key Laboratory of Information Engineering, Surveying, Mapping, and Remote Sensing, Wuhan University. He was a postdoctoral researcher with the Centre de Recherche en Mathmatiques de la Decision, CNRS, Paris-Dauphine University, France, for one-and-a-half years beginning in 2011. His current research interests include mathematical image modeling, texture synthesis, image indexing and content-based retrieval, structure from motion, perceptual grouping, and remote-sensing imaging.

Liangpei Zhang (zlp62@whu.edu.cn) received his B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, his M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences in 1988, and his Ph.D. degree in photogrammetry and remote sensing from Wuhan University, China, in 1998. He is currently a Chang-Jiang Scholar Chair

Professor at Wuhan University, appointed by the Ministry of Education of China. He has published more than 500 research papers and five books, and he holds 15 patents. He is a fellow of the Institution of Engineering and Technology, an executive member (Board of Governors) of the China National Committee of the International Geosphere-Biosphere Programme, and an executive member of the China Society of Image and Graphics. He was a recipient of the 2010 Best Paper Boeing Award and the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing. He is as an associate editor of *IEEE Transactions on Geoscience and Remote Sensing*.

Feng Xu (fengxu@fudan.edu.cn) received his B.E. degree in information engineering from Southeast University, Nanjing, China, and his Ph.D. degree in electronic engineering from Fudan University, Shanghai, China, in 2003 and 2008, respectively. In 2012, he was accepted into China's Global Experts Recruitment Program and returned to Fudan University in June 2013, where he is currently a professor in the School of Information Science and Technology and the vice director of the Key Lab for Information Science of Electromagnetic Waves. He was the 2014 recipient of the Early Career Award of the IEEE Geoscience and Remote Sensing Society (GRSS) and the 2007 recipient of the SUMMA Foundation graduate fellowship in the advanced electromagnetics area. He serves as the associate editor for *IEEE Geoscience and Remote Sensing Letters* and is the founding chair of the IEEE GRSS Shanghai Chapter. His research interests include electromagnetic scattering theory, synthetic aperture radar information retrieval, and radar system development.

Friedrich Fraundorfer (fraundorfer@icg.tugraz.at) received his M.S. and Ph.D degrees in computer science from the Graz University of Technology, Austria, where he is currently an assistant professor. He has held postdoctoral positions at the University of Kentucky, Lexington, the University of North Carolina at Chapel Hill, and the Swiss Federal Institute of Technology in Zurich. From 2012 to 2014, he acted as the deputy director of remote sensing technology with the Faculty of Civil, Geo, and Environmental Engineering at the Technical University of Munich, Germany. His main research areas are three-dimensional computer vision, robot vision, multiview geometry, visual-inertial fusion, microaerial vehicles, autonomous systems, and aerial imaging. His work on autonomous unmanned aerial vehicles was a finalist for the Best Paper Award at the 2012 IEEE International Conference on Intelligent Robots.

REFERENCES

[1] MIT Technology Review. (2013). 10 breakthrough technologies 2013. [Online]. Available: https://www.technologyreview.com/lists/technologies/2013/

[2] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1115.

[3] K. Simonyan and A. Zisserman. (2014). Very deep convolutional networks for large-scale image recognition. arXiv. [Online]. Available: https://arxiv.org/pdf/1409.1556.pdf

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

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, 2016.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[8] H. Noh, S. Hong, and B. Han, "Learning deconvolutional network for semantic segmentation," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 1520–1528.

[9] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2625–2634.

[10] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1110–1118.

[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. IEEE Int. Conf. Machine Learning (ICML)*, 2015, pp. 2048–2057.

[12] J. P. Rivera, J. Verrelst, J. Gomez-Dans, J. Muñoz-Marí, J. Moreno, and G. Camps-Valls, "An emulator toolbox to approximate radiative transfer models with statitistical learning," *Remote Sens.*, vol. 7, no. 7, pp. 9347–9370, 2015.

[13] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, 1989, pp. 445–448.

[14] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learning Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[15] A. Ng. (2010). Sparse autoencoder. [Online]. Available: https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf

[16] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based

on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, 2016.

[20] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification," *IEEE Signal Process. Mag*, vol. 31, no. 1, pp. 45–54, 2014.

[21] P. Ghamisi, Y. Chen, and X. Zhu, "A self-improving convolution neural network for the classification of hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 10, pp. 1537–1541, 2016.

[22] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, 2014.

[23] Y. Chen, X. Zhao, and X. Jia, "Spectra-spatial classification of hyperspectral data based on deep belief network," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, 2015.

[24] C. Tao, H. Pan, Y. Li, and Z. Zou, "Unsupervised spectral-spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 6, pp. 2381–2392, 2015.

[25] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, 2015. doi: 10.1155/2015/258619

[26] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, 2015, pp. 4959–4962.

[27] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.* doi: 10.1109/LGRS.2017.2681128.

[28] W. Zhao and S. Du, "Spectral-spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, 2016.

[29] A. Santara, K. Mani, P. Hatwar, A. Singh, A. Garg, K. Padia, and P. Mitra, "Bass net: Band-adaptive spectral-spatial feature learning neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5293–5301, 2017.

[30] W. Li, G. Wu, and F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, 2017.

[31] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[32] Y. Li, H. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[33] L. Mou, P. Ghamisi, and X. X. Zhu, "Fully conv-deconv network for unsupervised spectral-spatial feature extraction of hyperspectral imagery via residual learning," in *Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, to be published.

[34] L. Mou, P. Ghamisi, and X. Zhu, "Unsupervised spectral-spatial feature learning via deep residual conv-deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.* doi:10.1109/TGRS.2017.2748160.

[35] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, 2016.

[36] L. Mou, P. Ghamisi, and X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, 2017.

[37] W. Li, G. Wu, and Q. Du, "Transferred deep learning for anomaly detection in hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.* doi: 10.1109/LGRS.2017.2657818.

[38] H. Lyu, H. Lu, and L. Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sens.*, vol. 8, no. 6, pp. 506, 2016.

[39] D. E. Dudgeon, R. T. Lacoss, and A. Moreira, "An overview of automatic target recognition," *Lincoln Laboratory J.*, vol. 6, no. 1, pp. 3–10, 1993.

[40] S. Chen and H. Wang, "SAR target recognition based on deep learning," in *Proc. Int. Conf. Data Science and Advanced Analytics*, 2014, pp. 541–547.

[41] E. R. Keydel, S. W. Lee, and J. T. Moore, "MSTAR extended operating conditions: A tutorial," *Proc. SPIE-Int. Soc. Opt. Eng.*, vol. 2757, 1996. doi: 10.1117/12.242059.

[42] S. Chen, H. Wang, F. Xu, and Y. Q. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, 2016.

[43] D. Morgan, "Deep convolutional neural networks for ATR from SAR imagery," *Proc. SPIE-Int. Soc. Opt. Eng.*, vol. 9475, 2015. doi: 10.1117/12.2176558.

[44] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, 2016.

[45] K. Du, Y. Deng, R. Wang, T. Zhao, and N. Li, "SAR ATR based on displacement- and rotation-insensitive CNN," *Remote Sens. Lett.*, vol. 7, no. 9, pp. 895–904, 2016.

[46] M. Wilmanski, C. Kreucher, and J. Lauer, "Modern approaches in deep learning for SAR ATR," *Proc. SPIE-Int. Soc. Opt. Eng.*, vol. 9843, 2016. doi: 10.1117/12.2220290.

[47] Z. Cui, Z. Cao, J. Yang, and H. Ren, "Hierarchical recognition system for target recognition from sparse representations," *Math. Problems Eng.*, vol. 2015, 2015. doi: 10.1155/2015/527095

[48] S. A. Wagner, "SAR ATR by a combination of convolutional neural network and support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 2861–2872, 2016.

[49] C. Bentes, A. Frost, D. Velotto, and B. Tings, "Ship-iceberg discrimination with convolutional neural networks in high resolution SAR images," in *Proc. European Conf. Synthetic Aperture Radar (EUSAR)*, 2016, pp. 1–4.

[50] C. Schwegmann, W. Kleynhans, B. Salmon, L. Mdakane, and R. Meyer, "Very deep learning for ship discrimination in synthetic aperture radar imagery," in *Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, 2016, pp. 104–107.

[51] N. Ødegaard, A. O. Knapskog, C. Cochin, and J. C. Louvigne, "Classification of ships using real and simulated data in a

convolutional neural network," in *Proc. IEEE Radar Conf.*, 2016, pp. 1–6.

[52] Q. Song and F. Xu, "Zero-shot learning of SAR target feature space with deep generative neural networks," *IEEE Geosci. Remote Sens. Lett.*, to be published.

[53] Z. Zhang, H. Wang, F. Xu, and Y. Q. Jin, "Complex-valued convolutional neural networks and its applications to PolSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, to be published.

[54] Y. Q. Jin and F. Xu, *Polarimetric Scattering and SAR Information Retrieval*. Hoboken, NJ: Wiley, 2013.

[55] F. Xu, Y. Q. Jin, and A. Moreira, "A preliminary study on SAR advanced information retrieval and scene reconstruction," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 10, pp. 1443–1447, 2016.

[56] H. Xie, S. Wang, K. Liu, S. Lin, and B. Hou, "Multilayer feature learning for polarimetric synthetic radar data classification," in *Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, 2014, pp. 2818–2821.

[57] J. Geng, J. Fan, H. Wang, X. Ma, B. Li, and F. Chen, "High-resolution SAR image classification via deep convolutional autoencoders," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2351–2355, 2015.

[58] J. Geng, H. Wang, J. Fan, and X. Ma, "Deep supervised and contractive neural network for SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2442–2459, 2017.

[59] Q. Lv, Y. Dou, X. Niu, J. Xu, J. Xu, and F. Xia, "Urban land use and land cover classification using remotely sensed SAR data through deep belief networks," *J. Sensors*, vol. 2015, 2015. doi: 10.1155/2015/538063

[60] B. Hou, H. Kou, and L. Jiao, "Classification of polarimetric SAR images using multilayer autoencoders and superpixels," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 7, pp. 3072–3081, 2016.

[61] L. Zhang, W. Ma, and D. Zhang, "Stacked sparse autoencoder in PolSAR data classification using local spatial information," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 9, pp. 1359–1363, 2016.

[62] F. Qin, J. Guo, and W. Sun, "Object-oriented ensemble classification for polarimetric SAR imagery using restricted Boltzmann machines," *Remote Sens. Lett.*, vol. 8, no. 3, pp. 204–213, 2017.

[63] Z. Zhao, L. Jiao, J. Zhao, J. Gu, and J. Zhao, "Discriminant deep belief network for high-resolution SAR image classification," *Pattern Recognit.*, vol. 61, pp. 686–701, May 2017.

[64] L. Jiao and F. Liu, "Wishart deep stacking network for fast PolSAR image classification," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3273–3286, 2016.

[65] Y. Zhou, H. Wang, F. Xu, and Y. Q. Jin, "Polarimetric SAR image classification using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1935–1939, 2016.

[66] Y. Duan, F. Liu, L. Jiao, P. Zhao, and L. Zhang, "SAR image segmentation based on convolutional-wavelet neural network and Markov random field," *Pattern Recognit.*, vol. 64, pp. 255–267, Apr. 2017.

[67] L. Wang, K. A. Scott, L. Xu, and D. A. Clausi, "Sea ice concentration estimation during melt from dual-Pol SAR scenes using deep convolutional neural networks: A case study," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4524–4533, 2016.

[68] W. Yang, D. Dai, B. Triggs, and G. Xia, "SAR-based terrain classification using weakly supervised hierarchical Markov aspect models," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4232–4243, 2012.

[69] W. Shao, W. Yang, and G.-S. Xia, "Extreme value theory-based calibration for multiple feature fusion in high-resolution satellite scene classification," *Int. J. Remote Sens.*, vol. 34, no. 3, pp. 8588–8602, 2013.

[70] W. Yang, X. Yin, and G. Xia, "Learning high-level features for satellite image classification with limited labeled samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4472–4482, 2015.

[71] F. Hu, G. S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2015–2030, 2015.

[72] B. Zhao, Y. Zhong, G. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, 2016.

[73] F. Hu, G. Xia, J. Hu, Y. Zhong, and K. Xu, "Fast binary coding for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 8, no. 7, pp. 555, 2016.

[74] G.-S. Xia, J. Hu, B. Shi, X. Bai, Y. Zhong, X. Lu, and L. Zhang, "AID: A benchmark dataset for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, 2017.

[75] S. Bhagavathy and B. S. Manjunath, "Modeling and detection of geospatial objects using texture motifs," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 12, pp. 3706–3715, 2006.

[76] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, 2016.

[77] X. Chen, H. Zhao, P. Li, and Z. Yin, "Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes," *Remote Sens. Environment*, vol. 104, no. 2, pp. 133–146, 2006.

[78] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2003, p. 1470.

[79] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, 2016.

[80] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, 2015.

[81] O. Penatti, K. Nogueira, and J. Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR) Workshop*, 2015, pp. 44–51.

[82] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva. (2015). Land use classification in remote sensing images by convolutional neural networks. arXiv. [Online]. Available: https://arxiv.org/abs/1508.00092

[83] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14 680–14 707, 2015.

[84] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, 2016.

[85] F. Luus, B. Salmon, F. Bergh, and B. Maharaj, "Multiview deep learning for land-use classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2448–2452, 2015.

[86] K. Nogueira, O. Penatti, and J. Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, May 2017.

[87] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, 2016.

[88] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv. [Online]. Available: https://arxiv.org/pdf/1312.6229.pdf

[89] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[90] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM SIGSPATIAL Int. Conf. Advances in Geographic Information Systems*, 2010, pp. 270–279.

[91] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau., H. Sun, and H. Maitre, "Structural high-resolution satellite image indexing," in *Proc. Symp.: 100 Years ISPRS—Advancing Remote Sensing Science*, Vienna, Austria, 2010, pp. 298–303.

[92] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, 2017.

[93] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 117, pp. 11–28, July 2016.

[94] N. Yokoya and A. Iwasaki, "Object detection based on sparse representation and hough voting for optical remote sensing imagery," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2053–2062, 2015.

[95] J. Han, P. Zhou, D. Zhang, G. Cheng, L. Guo, Z. Liu, S. Bu, and J. Wu, "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," *ISPRS J. Photogrammetry Remote Sens.*, vol. 89, pp. 37–48, Mar. 2014.

[96] X. Jin and C. H. Davis, "Vehicle detection from high-resolution satellite imagery using morphological shared-weight neural networks," *Image Vis. Comput.*, vol. 25, no. 9, pp. 1422–1431, 2007.

[97] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, 2014.

[98] Q. Jiang, L. Cao, M. Cheng, C. Wang, and J. Li, "Deep neural networks-based vehicle detection in satellite images," in *Proc. Int. Symp. Bioelectronics and Bioinformatics*, 2015, pp. 184–187.

[99] P. Zhou, G. Cheng, Z. Liu, S. Bu, and X. Hu, "Weakly supervised target detection in remote sensing images based on transferred deep features and negative bootstrapping," *Multidimensional Syst. Signal Process.*, vol. 27, no. 4, pp. 925–944, 2016.

[100] L. Zhang, Z. Shi, and J. Wu, "A hierarchical oil tank detector with deep surrounding features for high-resolution optical satellite imagery," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 10, pp. 4895–4909, 2015.

[101] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[102] A.-B. Salberg, "Detection of seals in remote sensing images using features extracted from deep convolutional neural networks," in *Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, 2015, pp. 1893–1986.

[103] I. Ševo and A. Avramović, "Convolutional neural network based automatic object detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 5, pp. 740–744, 2016.

[104] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2015, pp. 3735–3739.

[105] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, 2016.

[106] J. Tang, C. Deng, G.-B. Huang, and B. Zhao, "Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1174–1185, 2015.

[107] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomput.*, vol. 70, no. 1, pp. 489–501, 2006.

[108] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, 2015.

[109] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, 2013.

[110] S. Özkan, T. Ates, E. Tola, M. Soysal, and E. Esen, "Performance analysis of state-of-the-art representation methods for geographical image retrieval and categorization," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 11, pp. 1996–2000, 2014.

[111] P. Napoletano. (2016). Visual descriptors for content-based retrieval of remote sensing images. arXiv. [Online]. Available: https://arxiv.org/abs/1602.00970

[112] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sens.*, vol. 9, no. 5, pp. 489, 2017.

[113] T. Jiang, G.-S. Xia, and Q. Lu, "Sketch-based aerial image retrieval," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, to be published.

[114] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, 2015.

[115] M. Schmitt and X. X. Zhu, "Data fusion and remote sensing: An ever-growing relationship," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 4, pp. 6–23, 2016.

[116] L. Mou, X. X. Zhu, M. Vakalopoulou, K. Karantzalos, N. Paragios, B. L. Saux, G. Moser, and D. Tuia, "Multi-temporal very high resolution from space: Outcome of the 2016 IEEE GRSS Data Fusion Contest," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3435–3447, 2017.

[117] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce, "Comparison of pansharpening algorithms: Outcome of the 2006 GRSS Data-Fusion Contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3012–3021, 2007.

[118] D. Fasbender, D. Tuia, M. Kanevski, and P. Bogaert, "Support-based implementation of Bayesian data fusion for spatial enhancement: Applications to aster thermal images," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 589–602, 2008.

[119] L. Loncan, L. B. Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. A. Licciardi, M. Simoes, J. Y. Tourneret, M. A. Veganzones, G. Vivone, Q. Wei, and N. Yokoya, "Hyperspectral pansharpening: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 27–46, 2015.

[120] J. Zhong, B. Yang, G. Huang, F. Zhong, and Z. Chen, "Remote sensing image fusion with convolutional neural network," *Sens. Imaging*, vol. 17, no. 1, 2016.

[121] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, pp. 594, 2016.

[122] Y. Yuan, S. Zheng, and X. Lu, "Hyperspectral image superresolution by transfer learning," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1963–1974, 2017.

[123] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. European Conf. Computer Vision (ECCV)*, 2014, pp. 184–199.

[124] D. Tuia, C. Persello, and L. Bruzzone, "Recent advances in domain adaptation for the classification of remote sensing data," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 41–57, 2016.

[125] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pansharpening method with deep neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1037–1041, 2015.

[126] A. Lagrange, B. L. Saux, A. Beaupere, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, and M. Ferecatu, "Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks," in *IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, 2015, pp. 4173–4176.

[127] M. Campos-Taberner, A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. L. Saux, A. Beaupère, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, M. Ferecatu, M. Shimoni, G. Moser, and D. Tuia, "Processing of extremely high resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS Data Fusion Contest. Part A: 2D contest," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5547–5559, 2016.

[128] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. (2017). High-resolution semantic labeling with convolutional neural networks. arXiv. [Online]. Available: https://arxiv.org/abs/1611.01962

[129] 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 *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR) Workshops*, 2016, pp. 1–9.

[130] Y. Gal and Z. Ghahramani. (2015). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. arXiv. [Online]. Available: https://arxiv.org/abs/1506.02142

[131] D. Tuia, N. Courty, and R. Flamary, "Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions," *ISPRS J. Photogrammetry Remote Sens.*, vol. 105, pp. 272–285, July 2015.

[132] M. Volpi, G. Camps-Valls, and D. Tuia, "Spectral alignment of cross-sensor images with automated kernel canonical correlation analysis," *ISPRS J. Photogrammetry Remote Sens.*, vol. 107, pp. 50–63, Sept. 2015.

[133] D. Marcos, R. Hamid, and D. Tuia, "Geospatial correspondence for multimodal registration," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5091–5100.

[134] M. Gong, T. Zhan, P. Zhang, and Q. Miao, "Superpixel-based difference representation learning for change detection in multispectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2658–2673, 2017.

[135] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 116, pp. 24–41, June 2016.

[136] H. Lyu, H. Lu, L. Mou, W. Li, X. Li, X. Li, J. Wang, X. X. Zhu, L. Yu, and P. Gong, "A deep information based transfer learning method to detect annual urban dynamics of four developed cities from 1984–2016 by Landsat data," *Remote Sens. Environment*, to be published.

[137] J. Sherrah. (2016). Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. arXiv. [Online]. Available: https://arxiv.org/abs/1606.02585

[138] A. Marcu and M. Leordeanu. (2016). Dual local-global contextual pathways for recognition in aerial imagery. arXiv. [Online]. Available: https://arxiv.org/abs/1605.05462

[139] J. Hu, L. Mou, A. Schmitt, and X. X. Zhu, "FusioNet: A two-stream convolutional neural network for urban scene classification using PolSAR and hyperspectral data," in *Proc. Joint Urban Remote Sensing Event (JURSE)*, to be published.

[140] L. Mou, M. Schmitt, Y. Wang, and X. X. Zhu, "A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes," in *Proc. Joint Urban Remote Sensing Event (JURSE)*, to be published.

[141] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. van den Hengel, "Semantic labeling of aerial and satellite imagery," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 7, pp. 2868–2881, 2016.

[142] N. Audebert, B. L. Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Proc. Asian Conf. Computer Vision (ACCV)*, 2016, pp. 180–196.

[143] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. (2017). Classification with an edge: Improving semantic image segmentation with boundary detection. arXiv. [Online]. Available: https://arxiv.org/abs/1612.01337

[144] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2366–2374.

[145] S. Srivastava, M. Volpi, and D. Tuia, "Joint height estimation and semantic labeling of monocular aerial images with CNNs," in *Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, to be published.

[146] M. Vakalopoulou, C. Platias, M. Papadomanolaki, N. Paragios, and K. Karantzalos, "Simultaneous registration, segmentation and change detection from multisensor, multitemporal satellite image pairs," in *Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, 2016, pp. 1827–1830.

[147] S. Lefèvre, D. Tuia, J. D. Wegner, T. Produit, and A. S. Nassar, "Towards seamless multi-view scene analysis from satellite to street-level," *Proc. IEEE*. doi: 10.1109/JPROC.2017.2684300.

[148] J. D. Wegner, S. Branson, D. Hall, and P. Perona, "Cataloging public objects using aerial and street-level images—Urban trees," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 6014–6023.

[149] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 91–99.

[150] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun, "Hd maps: Fine-grained road segmentation by parsing ground and aerial images," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3611–3619.

[151] T. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5007–5015.

[152] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *Proc. European Conf. Computer Vision (ECCV)*, 2016, pp. 494–509.

[153] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 539–546.

[154] S. Workman and N. Jacobs, "On the location dependence of convolutional neural network features," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR) Workshops*, 2015, pp. 70–78.

[155] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 3961–3969.

[156] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Advances in Neural Information Systems (NIPS)*, 2014, pp. 487–495.

[157] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[158] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[159] P. Fischer, A. Dosovitskiy, and T. Brox. (2014). Descriptor matching with convolutional neural networks: A comparison to SIFT. arXiv. [Online]. Available: https://arxiv.org/abs/1406.6909

[160] A. Handa, M. Blösch, V. Patraucean, S. Stent, J. McCormac, and A. J. Davison. (2016). gvnn: Neural network library for geometric computer vision. arXiv. [Online]. Available: https://arxiv.org/abs/1607.07405

[161] K. Lenc and A. Vedaldi. (2016). Learning covariant feature detectors. arXiv. [Online]. Available: https://arxiv.org/abs/1605.01224

[162] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3279–3286.

[163] K. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned invariant feature transform," in *Proc. European Conf. Computer Vision (ECCV)*, 2016, pp. 467–483.

[164] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, 2008.

[165] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1592–1599.

[166] L. Li, X. Yu, S. Zhang, X. Zhao, and L. Zhang, "3D cost aggregation with multiple minimum spanning trees for stereo matching," *Appl. Opt.*, vol. 56, no. 12, pp. 3411–3420, 2017.

[167] S. Drouyer, S. Beucher, M. Bilodeau, M. Moreaud, and L. Sorbier, "Sparse stereo disparity map densification using hierarchical image segmentation," in *Proc. Int. Symp. Mathematical Morphology and Its Applications to Signal and Image Processing*, 2017, pp. 172–184.

[168] H. Park and K. M. Lee, "Look wider to match image patches with convolutional neural networks," *IEEE Signal Process. Lett.* doi: 10.1109/LSP.2016.2637355.

[169] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks

for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4040–4048.

[170] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of fully convolutional neural networks," in *Proc. ISPRS Ann. Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2016, pp. 473–480.

[171] C. Häne, C. Zach, A. Cohen, R. Angst, and M. Pollefeys, "Joint 3D scene reconstruction and class segmentation," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 97–104.

[172] M. Bláha, C. Vogel, A. Richard, J. D. Wegner, T. Pock, and K. Schindler, "Large-scale semantic 3D reconstruction: An adaptive multi-resolution model for multi-class volumetric labeling," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3176–3184.

[173] M. Bláha, C. Vogel, A. Richard, J. D. Wegner, T. Pock, and K. Schindler, "Towards integrated 3D reconstruction and semantic interpretation of urban scenes," in *Proc. Dreiländertagung der SGPF, DGPF und OVG: Lösungen für eine Welt im Wandel: Vorträge*, 2016, pp. 44–53.

[174] Y. Bengio. (2012). Practical recommendations for gradient-based training of deep architectures. arXiv. [Online]. Available: https://arxiv.org/abs/1206.5533

[175] G. Montavon, G. B. Orr, and K.-R. Müller, *Neural Networks: Tricks of the Trade*. New York: Springer-Verlag, 2012.

[176] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva. (2015). Land use classification in remote sensing images by convolutional neural networks. arXiv. [Online]. Available: https://arxiv.org/abs/1508.00092

[177] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM SIGSPATIAL Int. Conf. Advances in Geographic Information Systems (ACM GIS)*, 2010, pp. 270–279.

[178] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*. doi: 10.1109/JPROC.2017.2675998.

[179] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR) Workshop on EarthVision*, 2015, pp. 1–9.

[180] Y. Wang, X. X. Zhu, B. Zeisl, and M. Pollefeys, "Fusing meter-resolution 4-D InSAR point clouds and optical images for semantic urban infrastructure monitoring," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 14–26, 2017.

[181] A. Kendall, V. Badrinarayanan, and R. Cipolla. (2015). Bayesian SegNet: Model uncertainty in deep convolutional encoder–decoder architectures for scene understanding. arXiv. [Online]. Available: https://arxiv.org/abs/1511.02680

[182] P. Gong, L. Yu, C. Li, J. Wang, L. Liang, X. Li, L. Ji, Y. Bai, Y. Cheng, and Z. Zhu, "A new research paradigm for global land cover mapping," *Ann. GIS*, vol. 22, no. 2, pp. 1–16, 2016.

[183] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling, "Never-ending learning," in *Proc. Conf. Artificial Intelligence (AAAI)*, 2015, pp. 2302–2310.

[184] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. IEEE Int. Conf. Machine Learning (ICML)*, 2007, pp. 759–766.

[185] T. Durand, N. Thome, and M. Cord, "Weldon: Weakly supervised learning of deep convolutional neural networks," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4743–4752.

[186] R. Johnson and T. Zhang, "Supervised and semi-supervised text categorization using LSTM for region embeddings," in *Proc. IEEE Int. Conf. Machine Learning (ICML)*, 2016, pp. 526–534.

[187] ISI. (2017, Sept.). Web of science. [Online]. Available: https://webofknowledge.com

[188] IEEE GRSS. (2017). Image analysis and data fusion. [Online]. Available: http://www.grss-ieee.org/community/technical-committees/data-fusion/

[189] ISPRS. (2017). 2D semantic labeling contest. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

[190] M. De Felice. (2017, May 4). Which deep learning network is best for you? *IDG Communications*. [Online]. Available: http://www.cio.com/article/3193689/artificial-intelligence/which-deep-learning-network-is-best-for-you.html

[191] NSF. (2010, Oct. 28). UC merced land use dataset. *UC Merced*. [Online]. Available: http://vision.ucmerced.edu/datasets/landuse.html

[192] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, and L. Zhang. (2017, Feb. 27). AID: A benchmark dataset for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* [Online]. Available: http://www.lmars.whu.edu.cn/xia/AID-project.html

[193] EScience. (2016). NWPU-RESISC45 datasets. [Online]. Available: http://www.escience.cn/people/JunweiHan/NWPU-RESISC45.html

[194] M. Volpi. (2017). Zurich summer dataset. Google Sites. [Online]. Available: https://sites.google.com/site/michelevolpiresearch/data/zurich-dataset.

[195] M. Volpi. (2017). Dense semantic labeling. Google Sites. [Online]. https://sites.google.com/site/michelevol piresearch/codes/dense-labeling

[196] IEEE GRSS. (2017). 2017 IEEE GRSS data fusion contest. [Online]. Available: http://www.grss-ieee.org/2017-ieee-grss-data-fusion-contest/

[197] Office of the Director of National Intelligence. (2017). Functional map of the world challenge. *IARPA*. [Online]. Available: https://www.iarpa.gov/challenges/fmow.html

GRS