Deep Learning in Remote Sensing: A Review

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, Friedrich Fraundorfer

Abstract

This is the pre-acceptance version, to read the final version please go to IEEE Geoscience and Remote Sensing Magazine on IEEE XPlore.

Standing at the paradigm shift towards data-intensive science, machine learning techniques are becoming increasingly important. In particular, as a major breakthrough in the field, deep learning has proven as an extremely powerful tool in many fields. Shall we embrace deep learning as the key to all? Or, should we resist a "black-box" solution? There are controversial opinions in the remote sensing community. In this article, we analyze the challenges of using deep learning for remote sensing data analysis, review the recent advances, and provide resources to make deep learning in remote sensing ridiculously simple to start with. More importantly, we advocate 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.

X. Zhu and L. Mou are with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Germany and with Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Germany, E-mails: xiao.zhu@dlr.de; lichao.mou@dlr.de.

D. Tuia was with the Department of Geography, University of Zurich, Switzerland. He is now with the Laboratory of GeoInformation Science and Remote Sensing, Wageningen University of Research, the Netherlands. E-mail: devis.tuia@wur.nl.

G.-S Xia and L. Zhang are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. E-mail:guisong.xia@whu.edu.cn; zlp62@whu.edu.cn.

F. Xu is with the Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan Univeristy. E-mail: fengxu@fudan.edu.cn.

F. Fraundorfer is with the Institute of Computer Graphics and Vision, TU Graz, Austria and with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Germany. E-mail: fraundorfer@icg.tugraz.at.

The work of X. Zhu and L. Mou are supported by the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No [ERC-2016-StG-714087], Acronym: *So2Sat*), Helmholtz Association under the framework of the Young Investigators Group "SiPEO" (VH-NG-1018, www.sipeo.bgu.tum.de) and China Scholarship Council. The work of D. Tuia is supported by the Swiss National Science Foundation (SNSF) under the project NO. PP0P2 150593. The work of G.-S. Xia and L. Zhang are supported by the National Natural Science Foundation of China (NSFC) projects with grant No. 41501462 and No. 41431175. The work of F. Xu are supported by the National Natural Science Foundation of China (NSFC) projects with grant No. 61571134.

Index Terms

Deep learning, remote sensing, machine learning, big data, Earth observation

I. MOTIVATION

Deep learning is the fastest-growing trend in big data analysis and has been deemed one of the 10 breakthrough technologies of 2013 [1]. It is characterized by neural networks (NNs) involving usually more than two layers (for this reason, they are called *deep*). As their shallow counterpart, deep neural networks exploit feature representations learned exclusively from data, instead of hand-crafting features that are mostly designed based 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. Recent advances in the field have proven deep learning a very successful set of tools, sometimes even able to surpass human ability to solve highly computational tasks (see, for instance, the highly mediatized Go match between Google's AlphaGo AI and the World Go Champion Lee Sedol. Motivated by those exciting advances, deep learning is becoming the model of choice in many fields of application. For instance, convolutional neural networks (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., spatially shrinking the feature maps layer by layer). Recent studies indicate that the feature representations learned by CNNs are greatly effective in large-scale image recognition [2–4], object detection [5, 6], and semantic segmentation [7, 8]. Furthermore, as an important branch of the deep learning family, recurrent neural networks (RNNs) have been shown to be very successful on a variety of tasks involved in sequential data analysis, such as action recognition [9, 10] and image captioning [11].

Following this wave of success and thanks to the increased availability of data and computational resources, the use of deep learning in remote sensing is finally taking off in remote sensing as well. Remote sensing data bring some new challenges for deep learning, since satellite image analysis raises some unique questions that translate into challenging new scientific questions:

- Remote sensing data are often *multi-modal*, e.g. from optical (multi- and hyperspectral) and synthetic aperture radar (SAR) sensors, where both the imaging geometries and the 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 is to develop novel architectures for the matching of images taken from different perspectives and even different imaging modality, preferably without requiring an existing 3D model. Also, besides conventional decision fusion, an alternative is to investigate the transferability of trained networks to other imaging modalities.

- Remote sensing data are *geo-located*, 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 GIS layers, geo-tagged images from social media, or simply other sensors (as above). On one hand, this fact allows tackling of data fusion with non-traditional data modalities while, on the other hand, it opens the field to new applications, such as pictures localization, location-based services or reality augmentation.
- Remote Sensing data are *geodetic measurements* with controlled quality. This enables us to retrieve geo-parameters with confidence estimates. However, differently from purely data-driven approaches, the role of prior knowledge about the sensors adequacy and data quality becomes even more crucial. For example, to retrieve topographic information, even at the same spatial resolution, interferograms acquired using single-pass SAR system are considered to be more important than the ones acquired in repeat-pass manner.
- *The time variable* is becoming increasingly in the field. The Copernicus program guarantees continuous data acquisition for decades. For instances, 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 for optimally exploiting 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, and often on a global scale. For example, even if they were launched in 2014, Sentinel satellites have already acquired about 25 Peta Bytes 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. On the other hand, 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 *geo-physical or bio-chemical 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 that is

traditionally used as priors for the estimates. This particularity suggests that the so-far dogma of expert-free fully automated deep learning should be questioned for remote sensing and physical models should be re-introduced into the concept, as, for example, in the concept of emulators [12].

Remote sensing scientists have exploited the power of deep learning to tackle these different challenges and started a new wave of promising research. In this paper, we review these advances. After the introductory Section II detailing deep learning models (with emphasis put on convolutional neural networks), we enter sections dedicated to advances in hyperspectral image analysis (Section III-A), synthetic aperture radar (Section III-B), very high resolution (Section III-C, data fusion (Section III-D), and 3D reconstruction (Section III-E). Section IV then provides the tools of the trade for scientists willing to explore deep learning in their research, including open codes and data repositories. Section V concludes the paper by giving an overview of the challenges ahead.

II. From Perceptron to Deep Learning

Perceptron is the basic of the earliest NNs [13]. It is a bio-inspired 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.

A. Autoencoder models

1) Autoencoder and Stacked Autoencoder (SAE): An autoencoder [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:

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

where $\boldsymbol{\Theta}$ 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 $\boldsymbol{y}$:

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

where $\boldsymbol{\Theta}'$ is usually constrained to be the form of $\boldsymbol{\Theta}' = \boldsymbol{\Theta}^T$, i.e., the same weight is used for encoding the input and decoding the latent representation. The reconstruction error is defined

as the Euclidian distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ that is constrained to approximate the input data $\boldsymbol{x}$ (i.e., making $\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \rightarrow 0$). The parameters of the autoencoder are generally optimized by stochastic gradient descent (SGD).

An SAE is a neural network consisting of multiple layers of autoencoders in which the outputs of each layer are wired to the inputs of the following one.

2) Sparse Autoencoder: The conventional autoencoder 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 $\boldsymbol{X} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^N\}$, training a sparse autoencoder [15] boils down to finding the optimal parameters by minimizing the following loss function:

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

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 $\hat{\rho}_j$. KL-divergence is a standard function for measuring how similar two distributions are:

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

In the sparse autoencoder 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[1] value, and $\hat{\rho}$ indicates the average activation value of hidden neuron $\boldsymbol{h}_j$ over the training samples. Similar to the autoencoder, the optimization of a sparse autoencoder can be achieved via back-propagation and SGD.

3) Restricted Boltzmann Machine (RBM) & Deep Belief Network (DBN): Unlike the deterministic network architectures, such as autoencoders or sparse autoencoders, an RBM (cf. Fig. 1) is a stochastic undirected graphical model consisting of a visible layer and a hidden layer, and

[1]An activation corresponds to how much a region of the image reacts when convolved with a filter. In the first layer, for example, each location in the image receives a value that corresponds to a linear combination of the original bands 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.

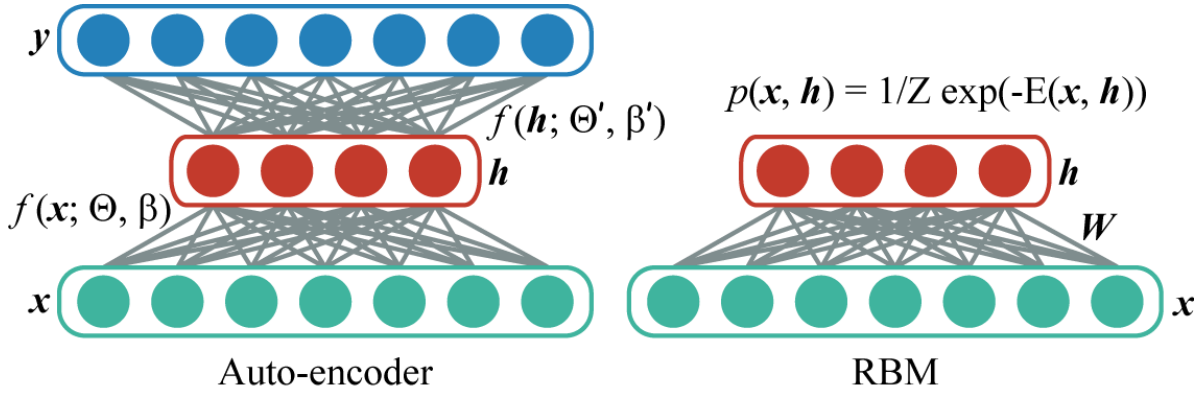


Fig. 1. Schematic comparison of an autoencoder (left) versus a restricted Boltzmann Machine (right).

it has symmetric connections between these two layers. No connecting exists 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 $\boldsymbol{x}$ is also named as the visible random variable, which is denoted as $\boldsymbol{v}$ in [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 $\boldsymbol{h}$ are fixed.

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

B. Convolutional neural networks (CNNs).

Unsupervised deep neural networks have been under the spot in the recent year. The leading model is the convolutional neural network (CNN), which learns the filters performing convolutions in the image domain. Here, we briefly review some successful CNN architectures proposed in computer vision in the recent years. For a comprehensive introduction on CNNs, we invite the reader to consider the excellent book by Goodfellow and colleagues [17].

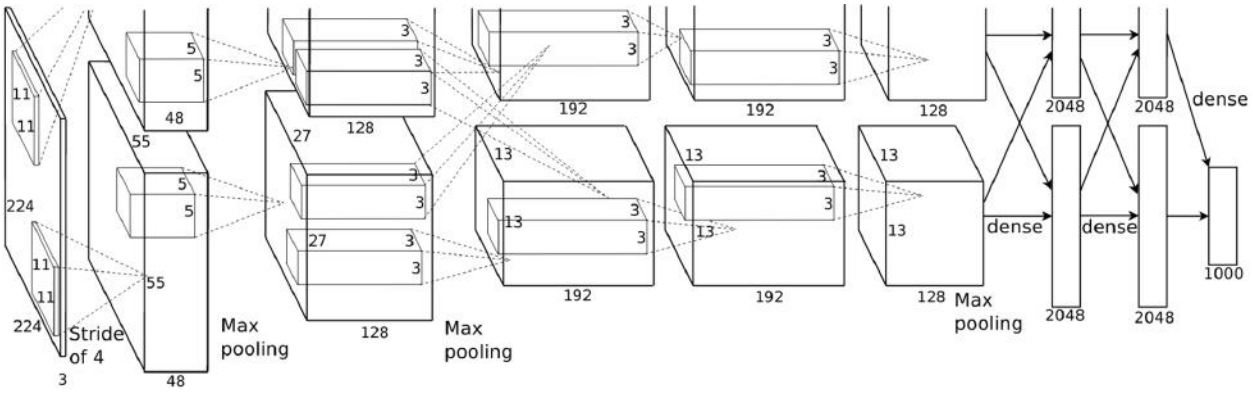


Fig. 2. Architecture of AlexNet, as shown in [2].

1) AlexNet: In 2012, Krizhevsky *et al.* [2] created AlexNet, which is a "large, deep convolutional neural network" that won the 2012 ILSVRC (ImageNet Large-Scale Visual Recognition Challenge). The year 2012 is marked as the first year where a CNN was used to achieve a top 5 test error rate of 15.4%.

AlexNet (cf. Fig. 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 more complicated objects. The contributions of AlexNet are as follows:

- Using rectified linear units (ReLU) as nonlinearity functions that are capable of decreasing training time, as ReLU is several times faster than the conventional hyperbolic tangent function.
- Implementing dropout layers in order to avoid the problem of overfitting.
- Using data augmentation techniques to artificially increase the size of the training set (and see a more diverse set of situations). From this, the training patches are translated and reflected on the horizontal and vertical axes.

One of the keys of the success of AlexNet is that the model was trained on GPUs. Since GPUs can offer a much larger number of cores than CPUs, it allows much faster training, which in turn allows one to use larger datasets and bigger images.

2) VGG Net: The design philosophy of the VGG Nets [3] is simplicity and depth. In 2014, Simonyan and Zisserman created VGG Nets that strictly makes use of 3×3 filters with stride and padding of 1, along with 2×2 max-pooling layers with stride 2. The main points of VGG Nets are that they:

- Use filters with small receptive field of 3×3, rather than using 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 is one of the most influential CNN models, as it reinforces the notion that CNNs with deeper architectures can promote hierarchical feature representations of visual data, which in turn improves the classification accuracy. A drawback is that, to train such a model from scratch, one would need large computational power and a very large labeled training set.

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

4) FCN: The fully convolutional network (FCN) [7] is the most important work in deep learning for semantic segmentation, which is 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 pixels size as the input (contrarily to the 'single class per image' of the aforementioned models). FCN introduces many significant ideas:

- End-to-end learning of the upsampling algorithm via an encoder/decoder structure that first downsamples the activations size and then upsamples it again.
- Using fully convolutional architecture allows the network to take images of arbitrary size as input since 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 multi-scale inference.

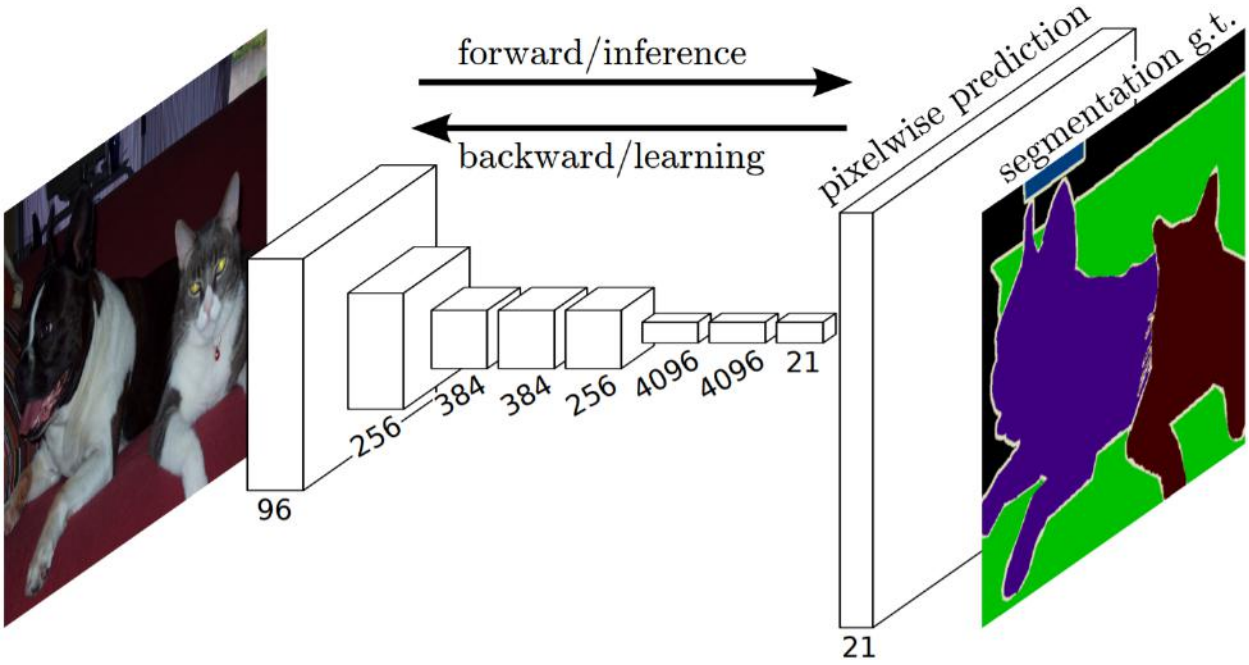


Fig. 3. Architecture of FCN [7].

Fig. 3 shows the architecture of FCN.

III. Remote Sensing Meets Deep Learning

Deep learning is taking off in remote sensing, as shown in Fig. 4, which summarizes the number of papers on the topic since 2014. Their 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 3D reconstruction.

A. 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 high importance in many practical applications, such as land cover/use classification or change and object detection. Also, because high quality hyperspectral satellite data is becoming available, e.g., via the launch of EnMAP, planned in

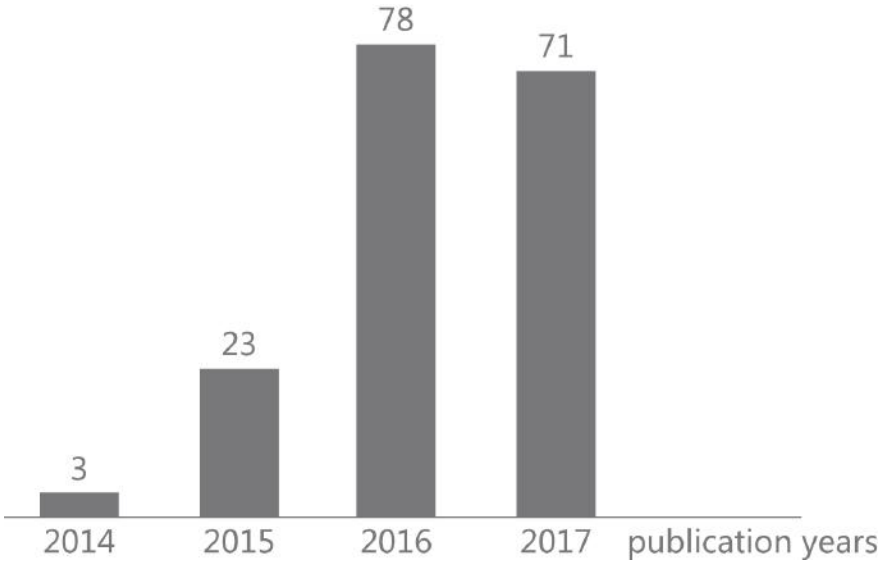


Fig. 4. Statistics on papers related to deep learning in remote sensing. [source: ISI web of Science; status: September 2017]

2020, and DESIS, planned in 2017, hyperspectral image analysis has been one of the most active research directions in the remote sensing community over the last decade.

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 into this field. In this section, we would like to review two application cases, namely, land cover/use classification (III-A1) and anomaly detection (III-A2).

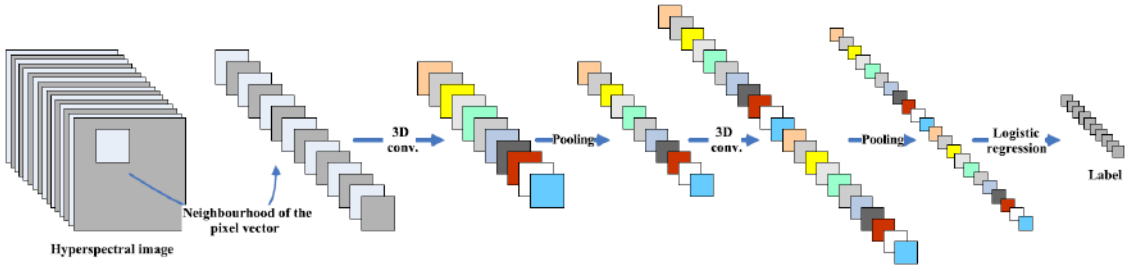


Fig. 5. Flowchart of the 3D CNN architecture proposed in [19] for spectral-spatial hyperspectral image classification.

1) 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 the 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, inter-class similarity, and intra-class variability result in the hyperspectral imaging procedure being inherently nonlinear. It is believed that, in comparison to the aforementioned "shallow" models, deep learning architectures are able to extract high-level, hierarchical, and abstract features, which are generally more robust to the nonlinear input data.

i) Autoencoders for hyperspectral data classification: A first attempt in this direction can be found in [22], where authors make use of a stacked autoencoder to extract hierarchical features in the spectral domain. Subsequently, in [23], authors employ DBM. Similarly, Tao *et al.* [24] use sparse stacked autoencoder to learn an effective feature representation from unlabeled data, and then the learned features are fed into a linear SVM for hyperspectral data classification.

ii) Supervised CNNs: In [25], authors train a simple 1D CNN that contains five layers, namely, an input layer, a convolutional layer, a max pooling layer, a fully connected layer, and an output layer – and directly classify the hyperspectral images in spectral domain.

Makantasis *et al.* [26] exploited a 2D CNN to encode spectral and spatial information, followed by a multi-layer perceptron performing the actual classification. In [27], authors attempt to carry out classification of crop types using 1D CNN and 2D CNN. They concluded that the 2D CNNs can outperform the 1D CNNs, but some small objects in the final classification map provided by 2D CNN are smoothed and misclassified. To avoid overfitting, Zhao and Du [28] propose a spectral-spatial feature-based classification framework, which jointly makes use of a local discriminant embedding-based dimension reduction algorithm and a 2D CNN. In [21], authors propose a self-improving CNN model, which combines a 2D CNN with a fractional order Darwinian particle swarm optimization algorithm to iteratively select the most informative bands that are suitable for training the designed CNN. Santara *et al.* [29] propose an end-to-end band-adaptive spectral-spatial feature learning network to address the problems of the curse of dimensionality. In [30], to allow CNN appropriately trained using limited labeled data, authors present a novel pixel-pair CNN to significantly augment the number of training samples.

Following recent vision developments in 3D 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 3D CNN, convolution operations are performed spatial-spectrally while in 2D CNNs they are done only spatially. Compared to 1D and 2D CNNs, 3D CNNs can model spectral information better owing to 3D convolution operations. Authors in [19] introduced a supervised, ℓ_2 regularized 3D CNN-based model (see Fig. 5). While authors of [32] followed a

similar idea for spatial-spectral classification.

(a) (b)

Fig 6. Object detection maps using learned filters of the first residual block in the unsupervised Residual Conv-Deconv network [33, 34], in which some "neurons" own good description power for semantic visual patterns in the object level. For example, the feature maps activated by the convolutional filters # 52 and # 03 in the first residual block can be used to precisely capture (a) metal sheets and (b) vegetative covers, respectively.

iii) Unsupervised Deep Learning: To be less dependent on the existence of large annotated collections of labeled data, unsupervised feature extraction remains of great interest. Authors of [35] propose an unsupervised convolutional network for learning spectral-spatial features using sparse learning to estimate the network weights in a greedy layer-wise fashion instead of end-to-end learning. Mou *et al.* [33, 34] propose a network architecture called fully Residual Conv-Deconv network for unsupervised spectral-spatial feature learning of hyperspectral images. They report an extensive study of the filters learned (cf. Fig. 6).

iv) RNN for Hyperspectral Image Classification: In [36], authors propose a RNN model with a new activation function and modified gated recurrent unit for hyperspectral image classification, which can effectively analyze hyperspectral pixels as sequential data and then determine information categories via network reasoning (cf. Fig. 7).

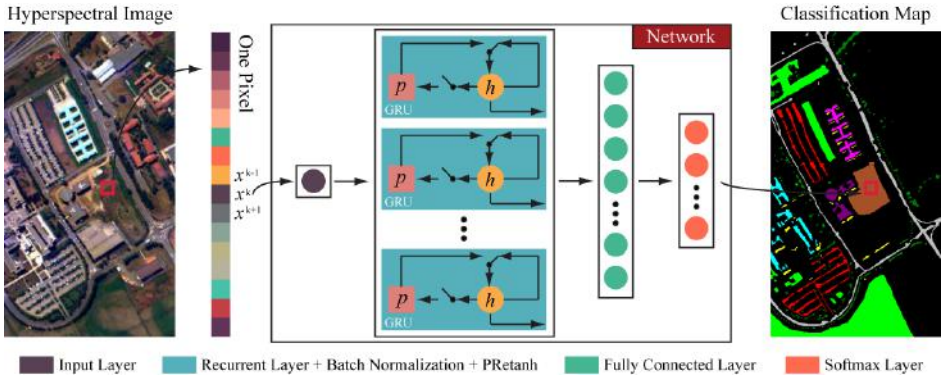


Fig. 7. RNN proposed for hyperspectral image classification task in [36].

2) Anomaly Detection: In a hyperspectral image, the pixels whose spectral signatures are significantly different from the global background pixels are considered anomalies. Since 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 mark addressing this problem via deep learning can be found in [37]. Li *et al.* [37] propose an anomaly detection framework, in which a multi-layer CNN is trained by using the differences in gray 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 the multi/hyper-spectral 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 has just made its mark [37, 38]. Some potential problems to be further explored include nonlinear spectral unmixing, hyperspectral image enhancement, hyperspectral time-series analysis, etc.

B. Interpretation of SAR Images

Over the past few years, there have been many publications in deep learning-related studies for SAR image analysis. Among these studies, deep learning techniques have been mostly applied 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.

1) 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 much 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 found its marks.

Chen *et al.* [40] introduce CNN into SAR ATR and tested on the standard ATR dataset MSTAR [41]. The major issue is found to be the lack of sufficient training samples as compared to optical images. It might cause severe overfitting and, therefore, greatly limits the capability of generalizing the model. 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 superior on the MSTAR dataset (i.e., a state-of-the-art accuracy of 99.1% on standard operating condition (SOC)). Extensive experiments are conducted to test the generalization capability of the so-called AConvNets and they are found to be quite robust in several extended operating conditions (EOC). The removal of the fully-connected layers, which is originally designed to be a trainable classifier, might be justifiable in this case, because the limited number of target types can be seen as the feature templates that the AConvNets is extracting.

Many authors applied CNN to SAR ATR and tested on MSTAR dataset, e.g., [43–46], etc. 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 are proposed, including translation, rotation, interpolation, etc.

Cui *et al.* [47] introduce DBN to SAR ATR, where stacked RBM are used to extract features and then fed to trainable classifier.

Wagner [48] proposes to use CNN to first extract feature vectors and then fed them to a 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.

More studies applying CNNs to the ART problem are found. Bentes *et al.* [49] apply CNN to ship-iceberg discrimination and tested on TerraSAR-X StripMap images. Schwegmann *et al.* [50] apply a specific type of deep neural networks, the highway networks, to ship discrimination in SAR imagery and achieved promising results. Ødegaard *et al.* [51] apply CNN to detect ships in harbor background in SAR images. To address the issue of lack of training samples, they employed a simulation software to generate simulated data for training. Song *et al.* [52] follow this idea and introduce a deep generative neural network for SAR ATR. A generative deconvolutional NN is first trained to generate simulated SAR image from a given target label, during which a feature space is 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 which is capable of interpreting a previously unseen target in the context of all known targets.

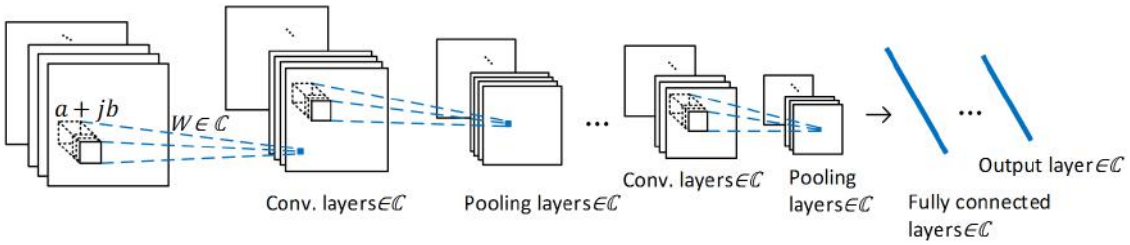


Fig. 8. Structure of complex-valued CNN (adapted from [53]).

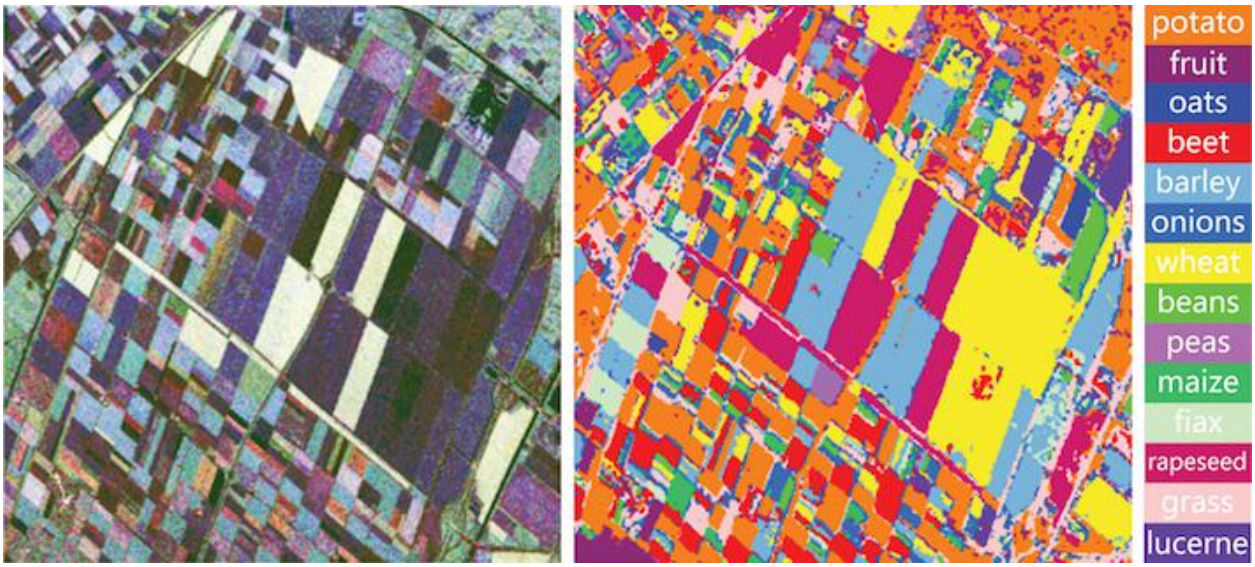


Fig. 9. Flevoland dataset. Left: Pauli RGB of the PolSAR dataset; Right: classification result from [53].

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

One large stream of studies will employ at least one type of unsupervised generative graphical models, such as DBN, SAE or RBM.

Xie *et al.* [56] first introduce multi-layer feature learning for PolSAR classification, where SAE is employed to extract useful features from a channel PolSAR image.

Geng *et al.* [57] propose a deep convolutional autoencoder (DCAE) to extract features and conduct classification automatically. The DCAE consists of a hand-crafted first layer of convolution, which contains kernels, such as gray-level cooccurrence matrix and Gabor filters, and a hand-crafted second layer of scale transformation, which integrates correlated neighbor pixels. The rest layers are trained SAE. This approach is tested on high-resolution single-polarization TerraSAR-X images. Geng *et al.* [58] later propose a similar framework, called deep supervised and contractive neural network (DSCNN), for SAR image classification, which further includes the histogram of oriented gradient (HOG) descriptors as hand-crafted kernels. The trainable AE layers employ a supervised penalty, which captures the relevant information between features and labels, and a contractive restriction, which enhances local invariance. An interesting finding of Geng *et al.* [58] is that speckle reduction yields the worse performance and the authors suspect that speckle reduction might smooth out some useful information.

Lv *et al.* [59] test DBN on urban land use and land cover classification using PolSAR data. Hou *et al.* [60] propose SAE combined with superpixel for PolSAR image classification. Multiple layers of AE are trained on a pixel-by-pixel basis. Superpixels are formed based on Pauli-decomposed pseudo-color image. The output of SAE is used as a feature in the final step of k-nearest neighbor clustering of superpixels. Zhang *et al.* [61] apply stacked sparse AE to PolSAR image classification. Qin *et al.* [62] apply adaptive boosting of RBMs to PolSAR image classification. Zhao *et al.* [63] propose discriminant DBN (DisDBN) for SAR image classification, in which the discriminant features are learned by combining ensemble learning with a deep belief network in an unsupervised manner.

Jiao and Liu [64] propose 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 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] apply CNN to PolSAR image classification, where covariance matrix is extracted as 6-real-channel data input. Duan *et al.* [66] propose to replace the conventional-pooling layer in CNN by a wavelet-constrained pooling layer. The so-called convolutional-wavelet neural network is then used in conjunction with superpixels and Markov Random Field (MRF) to produce the final segmentation map.

Zhang *et al.* [53] propose a complex-valued (CV) CNN (cf. Fig. 8) specifically designed to process complex values in PolSAR data, i.e., the off-diagonal elements of coherency or covariance matrix. CV-CNN not only takes complex numbers as input but also employs complex weights and complex operations throughout different layers. A complex-valued backpropagation algorithm is also developed to train it. Fig. 9 shows an example of PolSAR classification using CV-CNN.

3) Parameter inversion: Authors in [67] apply CNN to estimate ice concentration using SAR image during melt season. The labels are produced by visual interpretation by ice experts. It is tested on dual-pol RadarSat-2 data. Since the problem considered is regression of a continuous value, the loss function is selected as mean squared error. The final results suggest that CNN can produce a more detailed result then operational products.

C. Interpretation of High-resolution Satellite Images

1) 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 the 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, land resource management, etc. However, due to the high spatial resolutions, different scene images may contain the same kinds of objects or share similar spatial arrangement. For example, both residential area and commercial area 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, to train a complicated nonlinear classifier is

time-consuming. Hence, to extract a holistic and discriminative feature representation is the most significant part for scene classification. Traditional approaches are mostly based on the Bag-of-Visual-Words (BoVW) model [78, 79], but their potential for improvement was limited by the ability of experts to design the feature extractor and the expressive power encoded.

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

- Using pre-trained networks. The pre-trained deep CNN on the natural image dataset, e.g., OverFeat [88], GoogLeNet [89], etc., have led to impressive results on scene classification of high-resolution satellite images by directly extracting the features from the intermediate layers to form global feature representations [81–83, 87]. For example, [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 local feature extractor and combine it with feature coding techniques, such as BoVW [78] and Vector of Locally Aggregated Descriptors (VLAD) to generate the final image representation.
- Making a pre-trained model adapt to the specific conditions observed in a dataset under study, one can decide to fine-tune it on a smaller labeled dataset of satellite images. For example, [82] and [86] fine-tune some high-level layers of the GoogLeNet [89], etc., using the UC-Merced dataset [90] (see Section IV-C), obtaining better results than directly using only the pre-trained CNNs. This can be explained because the features learned are more oriented to the satellite images after fine-tuning, which can help to exploit the intrinsic characteristic of satellite images. Nonetheless, compared with the natural image dataset that consists of more than ten millions of samples, the scales of public satellite image datasets (i.e., UC-Merced dataset [90], RSSCN7 dataset [80], WHU-RS19 [91] are fairly small – for example, up to several thousands – for which we cannot fine-tune the whole CNNs to make them more adaptive to satellite images.
- Training new networks. In addition to the above two ways to use deep learning methods for classifying satellite images, some researchers train the network using satellite images from scratch. For example, [82] and [86] train the networks by only using the existing

satellite image dataset, which suffers a drop in classification accuracy compared with using the pre-trained networks as global feature extractors or fine-tuning the pre-trained networks. The reason may lies in the fact that the large-scale networks usually contain millions of parameters to be learned. Thus, training them using the small-scale satellite image datasets 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 images [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 dataset, the network need to be retrained.

2) 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 building, vehicle, aircraft, etc.) within a satellite image and predict their corresponding categories as shown in Fig. 10. Due to the powerful ability of learning high-level (more abstract and semantically meaningful) feature representations, the deep CNNs are being explored in object detection systems in opposition to the more traditional proposals 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.

Jin *et al.* [96] propose a vector-guided vehicle detection approach for IKONOS satellite imagery using a morphological shared-weight neural network, which learns the implicit vehicle model and incorporates both spatial and spectral characteristics, and classifies pixels into vehicles and non-vehicles. To address the problem of large-scale variance of objects, Chen *et al.* [97] propose a hybrid deep CNN model for vehicle detection in satellite images, which divides all feature maps of the last convolutional and max-pooling layer of CNN into multiple blocks of variable receptive field size or pooling size, to extract multi-scale features. Jiang *et al.* [98] propose a CNN-based vehicle detection approach, where 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 pre-trained CNNs for object detection. Zhou *et al.* [99] propose a weakly supervised learning framework to train an object detector, where a pre-trained 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] propose a hierarchical oil tank detector, which combines deep

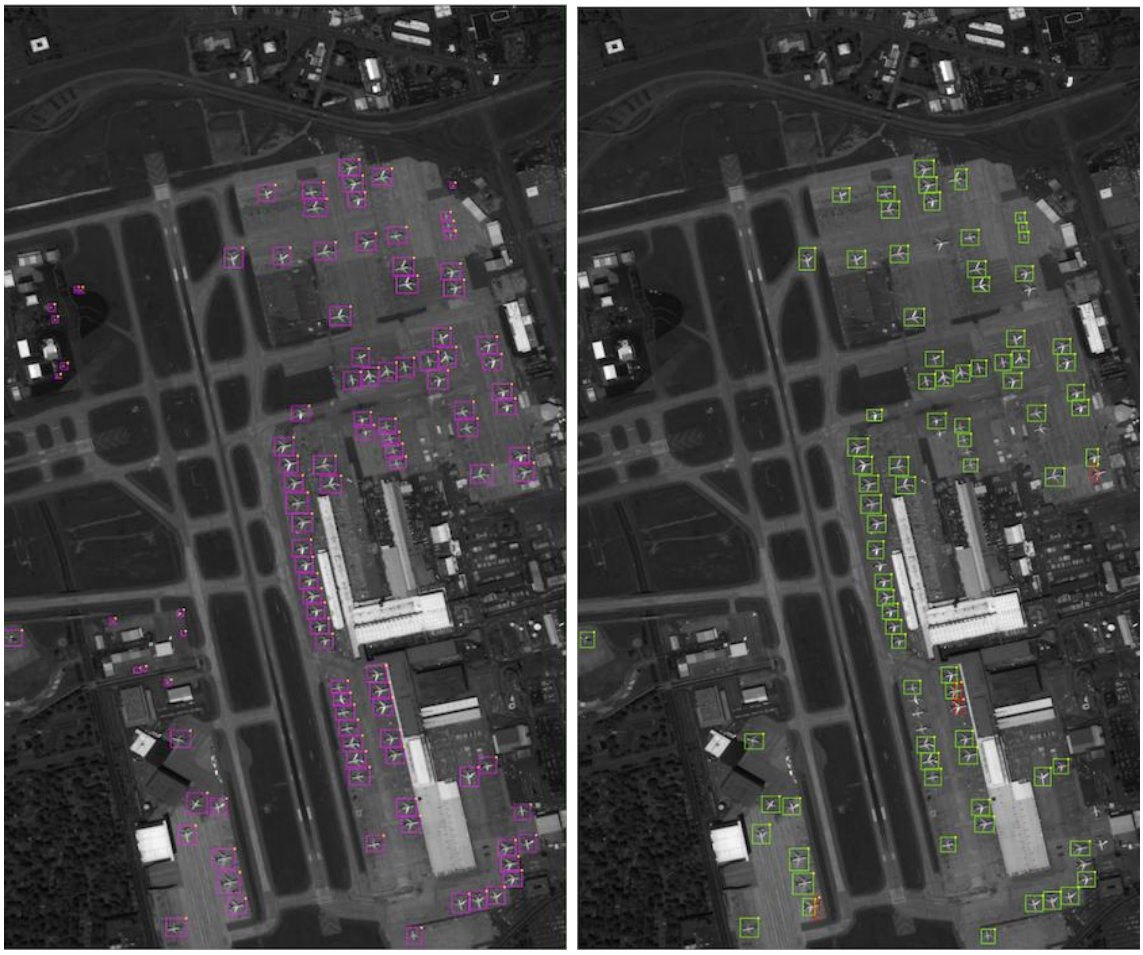


Fig. 10. Illustration of a typical object detection result within a high-resolution satellite image. The left: the annotated ground-truth of targets of interests (airplanes); The right: the airplanes detected by a CNN-based detector.

surrounding features, which are extracted from the pre-trained CNN model with local features (histogram of oriented gradients [101]). The candidate regions are selected by an ellipse and line segment detector. Salberg [102] proposes to extract features from the pre-trained AlexNet model and applies the deep CNN features for automatic detection of seals in aerial images. Ševo *et al.* [103] propose a two-stage approach for CNN training and develop an automatic object detection method based on a pre-trained CNN, where the GoogLeNet is first fine-tuned twice on UC-Merced dataset, using different fine-tuning options, and then the fine-tuned model is utilized for sliding-window object detection. To address the problem of orientation variations of objects, Zhu *et al.* [104] employ the pre-trained CNN features that are extracted from combined layers and implement orientation-robust object detection in a coarse localization framework.

Zhang *et al.* [105] propose a weakly supervised learning approach using coupled CNNs for aircraft detection. The authors employ 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, which is composed of a candidate region proposal network, and a localization network are developed to generate region proposals and locate the aircrafts 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] propose an effective approach to learn 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 which enforces the training samples before and after being rotating to share the similar features to guarantee the rotation-invariant ability of RICNN model.

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

3) Image Retrieval: Remote sensing image retrieval aims at retrieving images with 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 on the descriptive capability of image features to a large degree. Building image representation via feature coding methods (e.g., BoVW and VLAD) using low-level hand-crafted 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 pre-trained CNN models, and the deep features prove to

perform better than low-level features regardless of the retrieval system. Zhou *et al.* [112] propose a CNN architecture followed by a three-layer perceptron, which is trained on a large remote sensing dataset and able to achieve remarkable performance even with low dimensional deep features. Jiang *et al.* [113] present a sketch-based satellite image retrieval method by learning deep cross-domain features, which enables us to retrieve satellite images with hand-free sketches only.

Although there is still a lack of sufficient study of exploiting deep learning approaches for remote sensing image retrieval at present, in consideration of the great potential in learning high-level features of 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 that how to integrate the feedback from users into the deep learning retrieval scheme.

D. Multimodal Data Fusion

Data fusion is one of the fast-moving areas of remote sensing [114–116]: due to the recent increases in availability of sensor data, the perspectives of using big and heterogeneous data to study environmental processes have become more tangible.

Of course, when data are big and relations to be unveiled are complex, one would favour high capacity models: in this respect, deep neural networks are natural candidates to tackle the challenges of modern data fusion in remote sensing. Below, we review three areas of remote sensing image analysis where data fusion tasks have been approached with deep learning: pansharpening (Sec. III-D1), feature and decision-level fusion (Sec. III-D2), and fusion of heterogeneous sources (Sec. III-D3).

1) Pansharpening and Super-Resolution: Pansharpening is the task of improving the spatial resolution of multispectral data by fusing it with data characterized by sharper spatial information. It is a special instance of the more general problem of super-resolution. 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 neural networks for pansharpening multispectral images is certainly an interesting concept, since most image acquired by satellite 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 authors use a shallow network to upsample the intensity component obtained after the IHS of color images (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 dataset of QuickBird images for their analysis. Even though this is interesting, in this paper, authors simply replace one operation (the nearest neighbours or bicubic convolution) with a CNN.

In [121], 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] authors use a super-resolution CNN trained on natural images [123] as a pre-trained model and fine-tune it on a dataset of hyperspectral images. By doing so, they make an attempt at transfer leaning [124] between the domains of color (three bands, large bandwidths) and hyperspectral images (many bands, narrow bandwidths). Fine-tuning existing architectures, which have been trained on massive datasets with very big models, is often a relevant solution, since one makes use of discriminative strong features and only injects task-specific knowledge.

In [125], authors learn an upsampling of the panchromatic band via a stack of autoencoders: 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, therefore being upsampled using the data relationships learned from the panchromatic images.

2) Feature and Decision-level Fusion for image classification: Most current remote sensing literature, dealing with deep neural networks, 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, etc.). 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 will also review works using different data sources to tackle separate, but related predictive tasks, or *multi-task* problems.

i) Feature-level fusion: using multiple sources simultaneously in a network. As most image processing techniques, deep neural networks 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 neural networks are numerous and, in [126], authors compared networks trained on color RGB data (fine tuned from existing architectures) with networks, including a DSM channel on the 2015 Data Fusion Contest dataset over the city of Zeebruges [127][2]. They use the CNN as a feature extractor and then use the features to train a SVM, predicting a single semantic class for the entire patch. They then apply the classifier in a sliding window manner.

Parallel research 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 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] on the ISPRS Vaihingen and Potsdam benchmark datasets[3] stacking color infrared (CIR: infrared, red and green channels) and a normalized digital elevation model. The architectures compared and some zoomed results are reported in Figs. 11 and 12, respectively. Other strategies to spatial upsampling have been proposed in the recent literature, including the direct use of upsampled activation maps as features to train the final classifier [128]. In [129], 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.

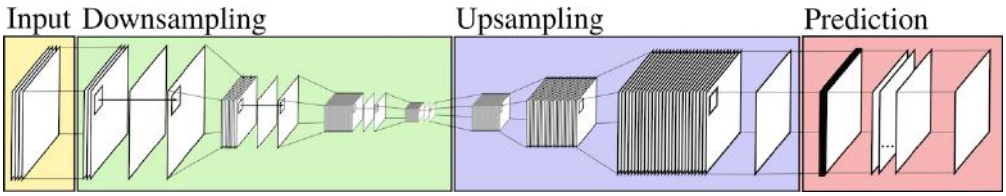


Fig. 11. Deconvolution network proposed in [92]. The yellow and green part correspond to a fully convolutional network with a 9×9 pixels bottleneck; then a deconvolutional block (purple) leads to predictions of the same size as the input image (in [92], 65×65 pixels).

[2]Data are available at http://www.grss-ieee.org/community/technical-committees/data-fusion/, also see Section IV-C

[3]Available at http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html, also see Section IV-C

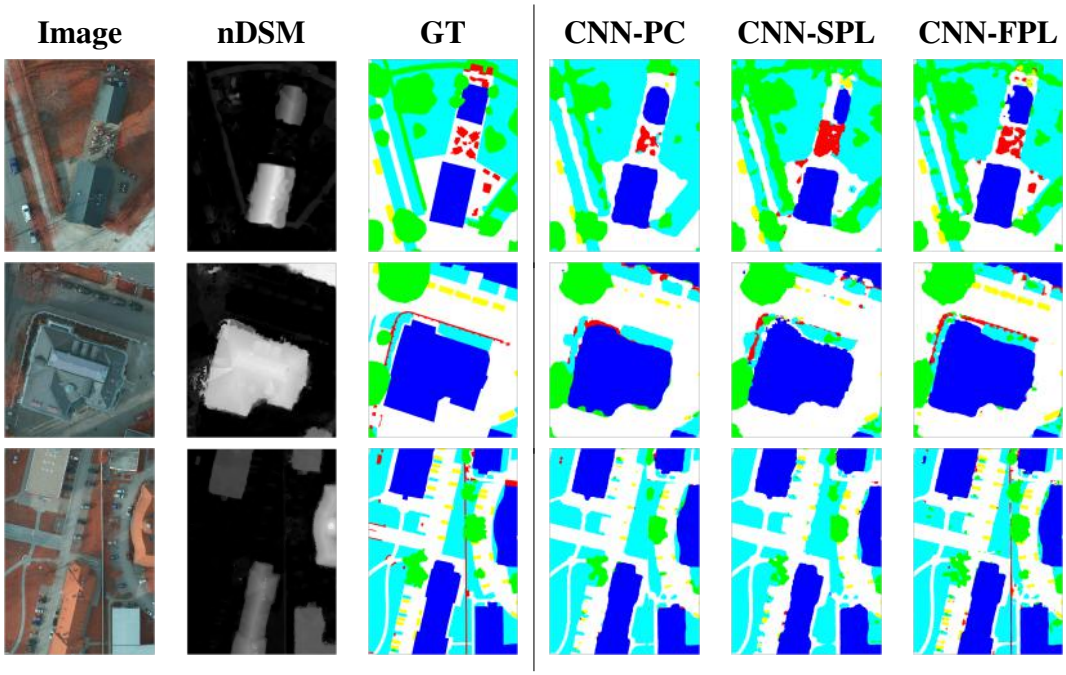


Fig. 12. Image classification results on the Potsdam datasets, considering 65×65 pixels patches (from [92]). **CNN-PC**: patch-based CNN, predicting single labels per patch and using a sliding window approach; **CNN-SPL**: fully convolutional CNN, predicting a 9×9 output, then upsampled to the original size via interpolation; **CNN-FPL**: deconvolutional network predicting the 65×65 output at full resolution.

Besides dense predictions, other strategies have been proposed to include spatial information in deep neural networks: for example, authors of [58] extract different types of spatial filters and stack them in a single tensor; the tensor is then used to learn a supervised stack of autoencoders. They apply their models on the classification of SAR images, so the fusion here is to be considered between different types of spatial information. The neural network is then followed by a conditional random field, to decrease the effect of speckle noise inherent in SAR images. In [131], authors learn 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, extracting therefore higher level information in an automatic way as deep neural networks do. I.e., it learns the right filters parameters (along with their combinations) instead of learning the filters coefficients themselves.

Data fusion is also a key component in change detection, where one would like to extract joint features from a bi-temporal sequence. The aim is to learn a joint representation, where both (co-registered) images can be compared: this area is especially interesting when methods can align data from multiple sensors (see [132, 133]). Three studies employ deep learning to

this end:

- [134], where authors learn a joint representation of two images with Deep Belief Networks. 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. Their approach is applied on optical images from the Chinese GaoFen-1 satellite and WorldView-2.
- [135], where the joint representation is learned via a stack of autoencoders 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 datasets, including pairs of optical and SAR images and an example performing change detection between optical and SAR images.
- More recent work, instead, addresses the transferability of deep learning for change detection, while analyzing data of long time series for large-scale problems. For example, in [38], authors make use of an end-to-end RNN to solve the multi/hyper-spectral change detection task, since RNN is well known to be good at processing sequential data. In their framework, an RNN based on long short-term memory (LSTM) is employed to learn joint spectral feature representations from a bi-temporal image sequence. In addition, authors also show that their network can detect multi-class changes and has a good transferability for change detection in a new scene without fine-tuning. Authors of [136] introduce a RNN-based transfer learning approach to detect annual urban dynamics of four cities (Beijing, New York, Melbourne, and Munich) from 1984 to 2016, using Landsat data. The main challenge here is that training data in such large-scale and long-term image sequence are very scarce. By combining RNN and transfer learning, they are able to transfer the feature representations learned from few training samples to new target scenes directly. Some zoomed results are reported in Fig. 13.

Another view on feature fusion can be found when considering neural networks fusing features obtained from different inputs: two (or more) networks are trained in parallel and their activations are then fused at a later stage, for example by feature concatenation. The author of [137] studies a solution in this direction which fuses two CNNs: the first considers CIR images of the Vaihingen dataset and passes them through the pre-trained VGG network to learn color features, while the

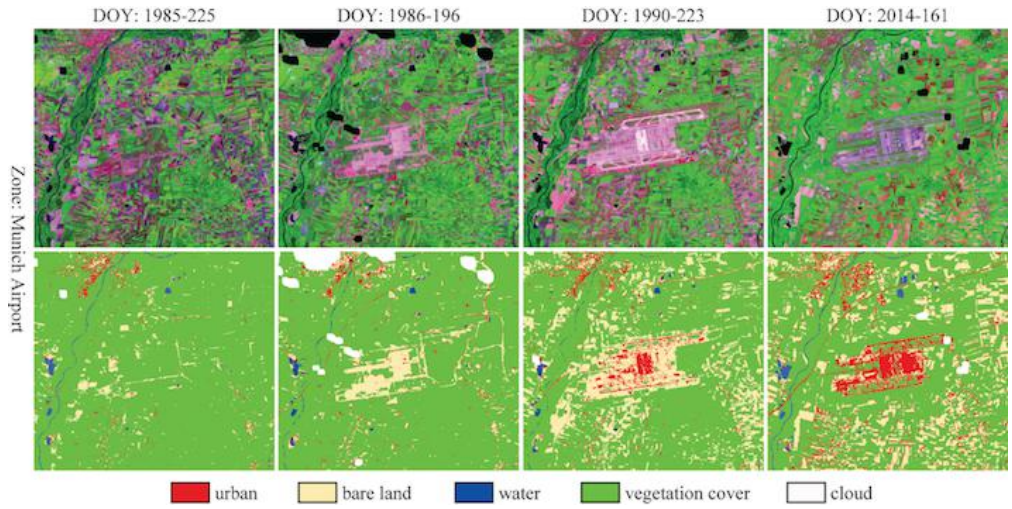


Fig. 13. Using large-scale and long-term multi-temporal image sequence, a deep learning-based system helps us to analyze how land cover changes. This example shows how Munich airport was built over the past 30 years.

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 authors learn a fully connected layer performing the fusion between networks learned at different spatial scales. They apply their model on tasks of buildings and road detection. In [139], authors train a two-stream CNN with two separate, yet identical convolutional streams, which 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 similar network architecture and contrastive loss function, authors of [140] learn a network for the identification of corresponding patches in SAR and Optical imagery of urban scenes.

ii) Decision-level fusion: fusing CNN (and other) outputs. If the works reviewed above 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 line of works has studied ways of performing decision fusion with deep learning. Even though the distinction with the models reviewed above might seem artificial, we review here 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 post-processing 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], authors fuse a classification map obtained by a CNN with another obtained by a random forest classifier which is trained using hand-crafted features. Both models use CIR, DSM and normalized DSM inputs from the Vaihingen dataset. 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, authors in [133] consider learning an ensemble of CNNs and then averaging their predictions: their proposed pipeline has two main streams, one processing the CIR data and another processing the DSM. They train several CNNs, using them as inputs the activation maps of each layer of the main model, as well as one fusing the CIR and DEM 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 form different data sources and levels of abstraction. Compared to the previous one in this section, this model has the advantage of being entirely learned in an end-to-end fashion, but also incurs in extreme computational load and a complex architecture involving many skip connections and fusion layers.

- *Decision fusion learned in the network*: an alternative to an ad-hoc fusion (multiplication or averaging of the posterior maps), one could learn the optimal fusion. In [142], authors perform the fusion between two maps obtained by pre-trained 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 favouring one or the other map. Their results show that such a learned fusion outperforms the, yet more intuitive, simple averaging of the posterior probabilities.

iii) Using CNN for solving different tasks. So far, only literature dealing with a single task (image classification) has been reviewed. But, besides it, one might want to predict other quantities or use the image classification results to improve the quality of related tasks as image registration. In this case, predicting different outputs jointly allows one to tighten feature representations with different meanings, therefore leading to another type of data fusion with respect to the one seen until now (that was mainly concerned by fusing different inputs). Summarizing, we will talk about *fusing outputs*. Below, we discuss three examples from recent literature, where alternative tasks are learned together with image classification.

- *Edges.* In the last section, we discussed the work of Marmanis *et al.* [133], where authors

were producing and fusing 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 desirable) classification maps. In [143], authors learn a CNN to separately output edges 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 model are very accurate but the computational load involved becomes very high: authors report models involving up to 800 millions of parameters to be learned.

- *Depth.* Most approaches discussed above 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 form image data would indeed be very valuable, since 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, 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 back-propagation by alternating over the two losses. Note that, in this case, the DSM is used as an output (contrarily with most approaches above) and is, therefore, not needed at prediction time.
- *Registration.* When performing change detection, one expects perfect co-registration of the sources. But, especially when working at very high resolution, this is difficulty to achieve. For instance, think of urban areas, where buildings are tilted by the viewing angle. In their entry to the IEEE GRSS data fusion contest 2016[4], authors of [146] learn jointly the registration between the images, the land use classification of each input and a change detection map with a conditional random field model. The land use classifier used is a two-layers CNN trained from scratch and the model is applied successfully either to pairs of VHR images or to datasets composed of VHR images and video frames from the International Space Station.

[4]Data are available on http://www.grss-ieee.org/community/technical-committees/data-fusion/

3) Fusing Heterogeneous Sources: Data fusion is not only about fusing image data with the same viewpoint. Multimodal remote sensing data exceed these restrictive boundaries and approaches to tackle new, exciting problems with remote sensing are appearing in the literature. A strong example is the joint use of ground-based and aerial images [147]: services, such as 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 recreation of virtual environments. In the following, we review a series of applications in this direction.

In [148], 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 a energy function to detect trees jointly in all sources, but also avoiding multiple and illogical detections (e.g. trees in the middle of a street lane). 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 fine-grained classification of the trees species on the detections, with impressive results. Authors of [150] take advantage of an approach that combines CNN and MRF and can estimate fine grained categories (e.g., road, sidewalk, background, building and parking) by doing joint inference over both monocular aerial imagery and ground images taken from a stereo camera on top of a car.

Many papers in geospatial computer vision work towards cross-view image localization: when presented to a ground picture, it would be relevant to be able to locate it in space. This is very important for photo-sharing platforms, for which only a fraction of the uploaded photos comes with geo-location. Authors of [151, 152] worked towards this aim, by training a cross-view Siamese network [153] to match ground images and aerial views. Siamese networks have also been recently applied to detect changes between matched ground panoramas and aerial images in [147]. Back to more traditional CNNs, authors of [154, 155] study the specificity of images to refer to a given city: they study how much images of Charleston resemble those from 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 the above, where the likelihood of localization is given by a similarity score between the features of the fully connected layer of Places CNN.

E. 3D Reconstruction

3D data generation from image data plays an important role for remote sensing. 3D data (e.g., in form of a DSM or DTM) 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 3D data generation workflow, two main steps must be performed. First, camera orientation, which means computing the position and orientation of the cameras that produced the image. This can be computed from the image data itself, by identifying and matching tie-points and then performing camera-resectioning. The second step is triangulation with calculates the 3D measurements for point correspondences that get established through stereo matching. The fundamental algorithms in this pipeline are of geometrical nature, the implementations are based on analytical calculations. So far, machine learning did not play a big role in this pipeline. However, there are steps in this pipeline which recently could be improved significantly by using machine learning techniques.

1) Tie points identification and matching: For instance, during camera orientation, the identification and matching of tie-points has long been done 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 re-detect 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 of the last mile (i.e., they are still less accurate than humans). This is a domain where machine learning and, in particular, convolutional neural networks are employed to learn, based on a huge amount of correct tie-point matches and point locations what characterizes tie-points and what is the best way of computing the similarity between them is. In the area of tie-point detection and matching, Fischer *et al.* [159] used CNN to learn a descriptor for image patch matching from training examples, similar to the well-known SIFT descriptor. In this work, authors trained a CNN with 5 convolutional layers and 2 fully connected layers. The trained network computes a descriptor for a given image patch. In the experiments on standard data sets, authors could show that the trained descriptors outperform engineered descriptors (i.e. SIFT) significantly in an tie-point matching task. Similar successes are described in other works 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 proposed a deep CNN to detect tie-point locations in an image and output a descriptor vector for each tie-point.

2) Stereo processing using convolutional neural networks: 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 semi-global matching (SGM) approach by Hirschmueller [164] acted as the gold-standard method for some considerable time.

Since 2002, the progress on stereo processing is tracked by the Middlebury stereo evaluation benchmark[5]. The benchmark allows to compare 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 done 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, typically some heuristic filtering is 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 proposed to train a CNN to compute the similarity measure between image patches (instead of using NCC or the Census transform). This change proved to be significant. Compared to SGM, which is often considered as a baseline method the proposed method achieved a significantly lower error rate. For SGM the error rate was still 18.4% while for the MC-CNN method the error rate was only 8%. After that, other variants of CNN-based stereo methods have been proposed and the best ranking method, today, has an error rate of only 5.9%. In table the error rates of the top-ranking CNN-based methods are listed.

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 for finding the depths (e.g. based on

[5] http://vision.middlebury.edu/stereo/

TABLE I

TOP RANKING STEREO METHODS FROM THE MIDDLEBURY STEREO EVALUATION BENCHMARK AS OF MAY 2017. CNN BASED METHODS ARE LEADING THE BOARD.

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

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] proposed such a paradigm shifting design for stereo processing. In their proposal, the stereo processing problem is solely modeled as a CNN. The proposed CNN takes two images of a stereo pair as and input and directly outputs the final disparity map. A single CNN architecture replaces all the individual algorithms 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 fact of the proposed method is that the stereo algorithm itself can be learned from data only. The network architecture itself does not define the algorithm but the data and the end-to-end training defines what type of processing the network should perform.

3) Large scale semantic 3D 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 3D information is generated from image data. The traditional 3D generation process did neglect object information. 3D data was generated from geometric constraints only. 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 3D generation process. CNN-based classification allows one to assign class labels to aerial imagery with unprecedented accuracy[170]. Pixels in images are then assigned labels like vegetation, road, building etc. This semantic information can now be used to steer the 3D data generation process. Class label specific parameters can be chosen for

the 3D data generation process.

The latest proposal in this area, however, is a joint reconstruction of 3D and semantic information. This has been proposed in the work of Haene *et al.* [171], where 3D reconstruction is performed with a volumetric method. The area to be reconstructed is partitioned into small cells, the size of it defining the resolution of the 3D reconstruction. The reconstruction algorithm now finds the optimal partitioning of this voxel grid into occupied and non-occupied voxels which fits to the image data. The result is a 3D reconstruction of the scene. The work of Haene *et al.* also jointly assigns the 3D reconstruction to a class label for each voxel, e.g. vegetation, building, road, sky. Each generated 3D data point now also has a class label. The 3D reconstruction is semantically interpretable. This process is a joint process, the computation of the occupied and non-occupied voxels takes 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. On the other hand, if a set of voxels are stacked on top of each other, it is likely that these belong to some building, i.e. the probability for assigning the label class of building is increased for this structure. This semantic 3D reconstruction method has been successfully applied to 3D reconstruction from aerial imagery by Blaha *et al.* [172, 173]. In their work they achieved a semantic 3D reconstruction of cities on large scales. The 3D model not only contains 3D data but also class labels. E.g., 3D structure that represents buildings gets the class label of building. Even more, every building has even the roof structures labeled as roof. Fig. 14 shows an image of a semantic 3D reconstruction produced by the method of [172].

In summary, it can be said that CNNs quickly took on a significant role in 3D data generation. Utilizing CNNs for stereo processing significantly boosted the accuracy and precision of depth estimation. On the other hand, the availability of reliable class labels extracted from CNNs classifiers opened the possibility of creating semantic 3D reconstructions, a research areas which is about to grow significantly.

IV. Deep Learning in Remote Sensing Made Ridiculously Simple to Start with

To make an easy start for researchers who attempt to work on deep learning in remote sensing, we list some available resources, including tutorials (Sec. IV-A) and open-source deep learning frameworks (Sec. IV-B). In addition, we provide a selected list of open remote sensing data for

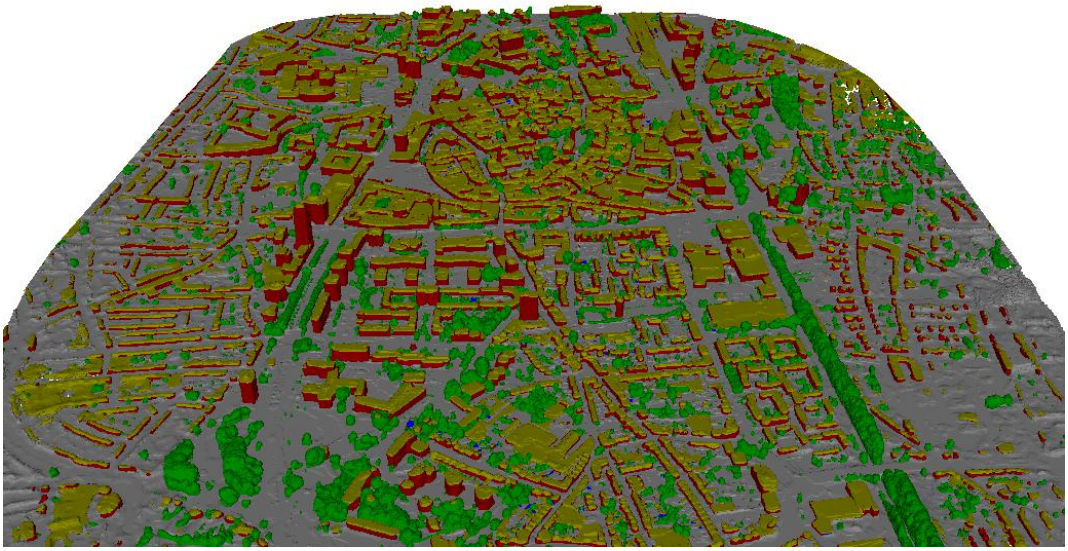


Fig. 14. Semantic 3D reconstruction from the Enschede aerial image data set computed with the method of [172]. The different colors represent different class labels: ground (gray), building (red), roof (yellow), vegetation (green) and clutter (blue). (Figure provided by the authors of [172])

training deep learning models (Sec. IV-C), as well as some showcasing examples with source codes developed using different deep learning frameworks (Sec. IV-D).

A. *Tutorials*

Some valuable tutorials for early deep learners, 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] which provide some general recommendations for the choice of the parameters.

B. *Open-source Deep Learning Frameworks*

When diving deep into deep learning, choosing an open-source framework is of great importance. Fig. 15 shows the most popular open-source deep learning frameworks, such as Caffe, Torch, Theano, TensorFlow, and Microsoft-CNTK. Since the field and surrounding technologies are relatively new and have been developing rapidly, the most common concerns amongst 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

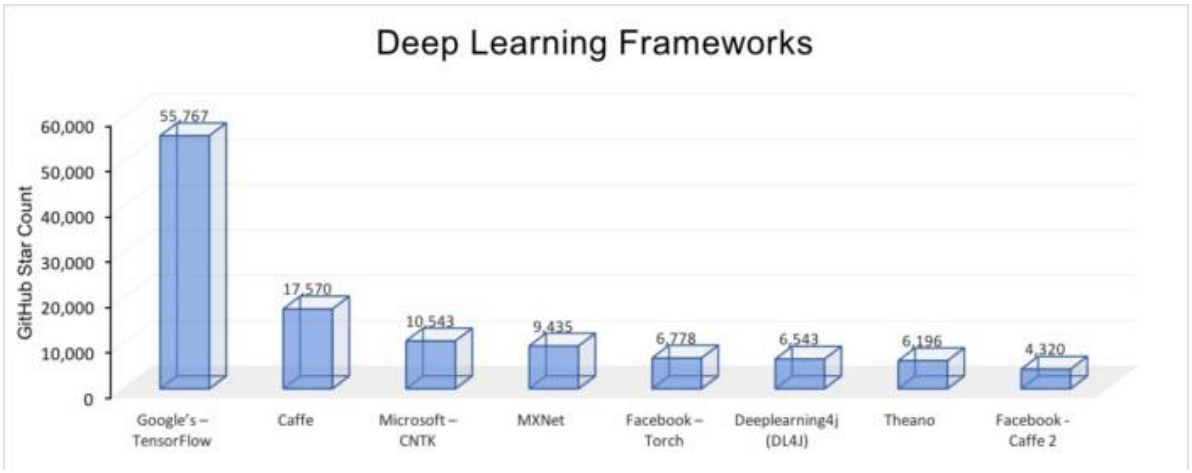


Fig. 15. Popular open-source deep learning frameworks. The ranking is based on the number of stars awarded by developers in GitHub. (Image source: http://www.cio.com/article/3193689/artificial-intelligence/which-deep-learning-network-is-best-for-you.html)

learning frameworks can be found at http://www.cio.com/article/3193689/artificial-intelligence/which-deep-learning-network-is-best-for-you.html.

C. Remote Sensing Data for Training Deep Learning Models

To train deep learning methods with good generalization abilities, one needs large datasets. This is true for both fine-tuning models and training small networks from scratch, while if we consider training large architectures, one should preferably resort to pre-trained methods [176]. In recent years, there is several datasets that have been made public and that can be used to train deep neural networks. Below is a non-exhaustive list.

1) Scene classification (one image is classified into a single label):

- *The UC Merced dataset* [177]. This dataset is a collection of aerial images (256×256 pixels in RGB space) depicting 21 land use classes. Each class comprises 100 images. Since each image comes with a single label, the dataset can be only used for image classification purposes, i.e., to classify the whole image into a single land use class. The dataset can be downloaded at http://vision.ucmerced.edu/datasets/landuse.html.
- *The AID dataset* [74]. This dataset is a collection of 10,000 annotated aerial images distributed in 30 land use scene classes and can be used for image classification purpose. In comparison with the UC Merced dataset, AID contains much more images and covers a wider range of scene categories. Thus it is in line with the data requirements of modern deep

learning. The dataset can be downloaded at http://www.lmars.whu.edu.cn/xia/AID-project.html.

- *The NWPU-RESISC45 dataset* [178]. This dataset contains 31,500 aerial images spread over 45 scene classes. So far, it is the largest dataset for land use scene classification in terms of both total number of images and number of scene classes. The dataset can be obtained at http://www.escience.cn/people/JunweiHan/NWPU-RESISC45.html.

2) Image classification (each pixel of an image is classified into a label):

- *The Zurich Summer Dataset* [179]. This dataset 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 0.6m resolution and 8 land use classes are presented. All images are released, along with their ground truths. The dataset can be obtained at https://sites.google.com/site/michelevolpiresearch/data/zurich-dataset.
- *Zeebruges, or the Data Fusion Contest 2015 dataset* [127] In 2015, the Image Analysis and Data Fusion Technical Committee of the IEEE GRSS organized a data processing competition aiming at 5-centimeter resolution land mapping. To do so, the organizers provided both a RGB aerial image and a dense (65 pts/m^2) lidar point cloud over the harbour of Zeebruges (Belgium). The data are organized on seven $10'000 \times 10'000$ pixels tiles. All the tiles have been labeled densely in 8 land classes, including land use (building, roads) and objects (vehicle, boats) classes [126]. The data can be obtained from the Data and Algorithm Standard Evaluation Website (DASE) 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.
- The ISPRS 2D semantic labeling challenge. The working group II/4 of the ISPRS '3D Scene Reconstruction and Analysis' provided a sub-decimeter resolution dataset over the two cities of Vaihingen and Potsdam. The data are similar to those of the Zeebruges data above, with the difference that the height information is provided as a digital surface model at the same resolution of the image data. Moreover, images are provided with an infrared channel. The dataset is also fully labeeld into six classes, including land classes (roads, meadows) and objects (cars). It also comes with a clutter class gathering all unknown objects. The Vaihingen dataset comes with 33 tiles of average size of 2000×3000 pixels. Half of the tiles come with labels. The other 17 tiles come with no labels and participants

must upload classification maps for evaluation. The Potsdam dataset comes with 24 labeled tiles (6000×6000 pixels) and 14 unlabeled ones. Both datasets can be obtained from http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html.

3) Registration / matching:

- *The SARptical Dataset* [180]. With the growing attention on very high resolution SAR data, the fusion of optical and SAR images in dense urban area has become an emerging and timely topic. Lying at the base of such fusion topic is the challenging task of the co-registration of SAR and optical images. Such two images are acquired with intrinsically different imaging geometries, and thus are nearly impossible to be co-registered without a precise 3D model of the imaged scene. SARptical is a unique dataset 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 co-registered. They are generated based on co-registered 3D InSAR point clouds (which are reconstructed by SAR tomography using tens of TerraSAR-X high resolution spotlight images), and 3D optical point clouds (which are reconstructed by structure from motion followed by dense stereo matching using several UltraCam images with a ground spacing of 20cm). This dataset can be downloaded from https://www.sipeo.bgu.tum.de/downloads.

D. Showcasing

Starting to work with CNNs from zero 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[6]. 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 Fig. 11. It exploits the MatConvNet library for MATLAB (http://www.vlfeat.org/matconvnet/). It provides a pre-trained network for both the Vaihingen and Potsdam datasets described above. The initial models are specific to remote sensing data and have been trained on each dataset separately. This example is mostly meant to show how to fine-tune an existing model in MatConvNet by training few

[6]All these examples are provided 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.

extra iteration 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 datasets is also provided. Overall, it allows one to reproduce the results in [92] which are similar to the last column of Fig. 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 Fig. 12. In both cases, the models must be re-trained (or at least heavily fine tuned).

The code can be downloaded from https://sites.google.com/site/michelevolpiresearch/codes/dense-labeling.

- *Fully convolutional (SegNet) architecture in Caffè.* The second example is released by the authors of [142] and exploits the Caffè library (http://caffe.berkeleyvision.org/). The model exploits the SegNet architecture from Kendall *et al.* [181]. Authors release the pre-trained model to reproduce the results of [142] on the Vaihingen dataset. 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 and demonstrates via the MSTAR dataset. 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.* The last example is released by the authors of [33, 34] and shows how to build up a residual Conv-Deconv network for unsupervised spectral-spatial feature learning of hyperspectral data. It exploits TensorFlow (https://www.tensorflow.org/) and Keras (https://keras.io/) libraries. One can transfer the trained network for their own classification purpose by fine-tuning on the target data sets or obtain "free" object detection 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.

V. Conclusion and Future Trends

In this paper, we 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 also surfing the wave of deep neural networks and observing a similar trend as in other fields: deep nets are solid models that tend to improve over classical approaches using hand crafted 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 ahead:

- What are the further applications in remote sensing which can potentially benefit from deep learning? In general, deep nets are particularly beneficial for remote sensing problems whose physical models are complicated, e.g., nonlinear, or even not yet well understood, or/and cannot be generalized. Yet, so far, in varies remote sensing fields, most deep learning-related research has been focused on classification and detection-related tasks using a number of benchmark data sets.
- Is the transferability of deep nets sufficient to extract geo-information on a global scale? Complex light scattering mechanisms in natural objects, various atmospheric scattering conditions, intra-class 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], self-taught learning [184], etc.
- How to tackle problems raised by very limited annotated data in remote sensing?
  - Is possible to learn deep hierarchical models for remote sensing image understanding in a weakly-supervised, semi-supervised or even unsupervised way? Here, we list a few inspiring work in machine learning and computer vision: [185], [186], and [34].
  - How do we benchmark the fast-growing deep-learning algorithms in remote sensing applications? Some recent initiatives include 2017 IEEE GRSS Data Fusion Contest dataset[7] and Functional Map of the World Challenge dataset[8].
- Fusion of physics-based modeling and deep neural network is a promising direction. Remote sensing imagery is a direct product of physics processes, such as light reflection, microwave scattering, etc. It has to resort to a synergy of the physics-based models which describe the

[7] http://www.grss-ieee.org/2017-ieee-grss-data-fusion-contest/

[8] https://www.iarpa.gov/challenges/fmow.html

a priori knowledge of the process behind imagery and newly develop 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 which remote sensing can make a difference. In this context, deep learning remains an incredible toolbox that allows researcher in remote sensing to exceed the boundaries of the field, to move beyond traditional small-scale benchmarking task and tackling large-scale, real-life problems with implicit models that generalize well. The data are now here, the hardware is ready, deep learning frameworks are openly available and it is now time to design models that are tailored to big remote sensing data and their multi-modal, geo-located, multi-aspect and multi-temporal aspects that were raised in the introduction.

On the other hand, commercial players are on the march to remote sensing and Earth observation. For example, Planet has launched about 140 small satellites which map the whole Earth daily. Standing on the paradigm shift from computational science to data-driven science, we, remote sensing experts, shall appropriately position ourselves among other data scientists, who are also trying to use deep learning for innovative remote sensing applications. This requires us, in turn and as mentioned before, to bring our domain expertises into deep learning to provide prior knowledge that is tailored to specific remote sensing problems.

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

REFERENCES

[1] *MIT Technology Review*, 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 *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *IEEE International Conference on Learning Representation (ICLR)*, 2015.

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

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 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 *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[8] H. Noh, S. Hong, and B. Han, "Learning deconvolutional network for semantic segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[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 *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[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 *IEEE International Conference on Machine Learning (ICML)*, 2015.

[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 Sensing*, vol. 7, pp. 9347–9370, 2015.

[13] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *International Joint Conference on Neural Networks (IJCNN)*, 1989.

[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," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[15] A. Ng, "Sparse autoencoder," https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf, online.

[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*. MIT Press, 2016, http://www.deeplearningbook.org.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the 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 Transactions on Geoscience and Remote Sensing*, 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 Proc. Mag.*, vol. 31, 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 Geoscience and Remote Sensing Letters*, 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 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 Geoscience and Remote Sensing Letters*, 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," *Journal of Sensors*, vol. 2015, no. 258619, 2015.

[26] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2015.

[27] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, 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 Transactions on*

Geoscience and Remote Sensing, 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 Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5293–5301, 2017.

[30] W. Li, G. Wu, F. Zhang, and Q. D. and, "Hyperspectral image classification using deep pixel-pair features," *IEEE Transactions on Geoscience and Remote Sensing*, 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 *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[32] Y. Li, H. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery with 3d convolutional neural network," *Remote Sensing*, 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 *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017.

[34] L. Mou, P. Ghamisi, and X. Zhu, "Unsupervised spectral-spatial feature learning via deep residual conv-deconv network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, in press.

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

[36] L. Mou, P. Ghamisi, and X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 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 Geoscience and Remote Sensing Letters*, 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 Sensing*, vol. 8, no. 6, p. 506, 2016.

[39] D. E. Dudgeon, R. T. Lacoss, and A. Moreira, "An overview of automatic target

recognition," *The Lincoln Laboratory Journal*, vol. 6, pp. 3–10, 1993.

[40] S. Chen and H. Wang, "SAR target recognition based on deep learning," in *International Conference on Data Science and Advanced Analytics*, 2014.

[41] E. R. Keydel, S. W. Lee, and J. T. Moore, "MSTAR extended operating conditions: a tutorial," in *Proc. SPIE 2757, Algorithms for Synthetic Aperture Radar Imagery III*, 1996.

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

[43] D. Morgan, "Deep convolutional neural networks for ATR from SAR imagery," in *Proc. SPIE 9475, Algorithms for Synthetic Aperture Radar Imagery XXII*, 2015.

[44] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geoscience and Remote Sensing Letters*, 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 Sensing Letters*, vol. 7, no. 9, pp. 895–904, 2016.

[46] M. Wilmanski, C. Kreucher, and J. Lauer, "Modern approaches in deep learning for SAR ATR," in *Proc. SPIE 9843, Algorithms for Synthetic Aperture Radar Imagery XXIII*, 2016.

[47] Z. Cui, Z. Cao, J. Yang, and H. Ren, "Hierarchical recognition system for target recognition from sparse representations," *Mathematical Problems in Engineering*, vol. 2015, no. 527095, 2016.

[48] S. A. Wagner, "SAR ATR by a combination of convolutional neural network and support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, 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 *European Conference on Synthetic Aperture Radar (EUSAR)*, 2016.

[50] C. Schwegmann, W. Kleynhans, B. Salmon, L. Mdakane, and R. Meyer, "Very deep learning for ship discrimination in Synthetic Aperture Radar imagery," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016.

[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 *IEEE Radar Conference (RadarConf)*, 2016.

[52] Q. Song, F. Xu, and Y. Q. Jin, "Deep SAR image generative neural network and auto-construction of target feature space," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017.

[53] Z. Zhang, H. Wang, F. Xu, and Y. Q. Jin, "Complex-valued convolutional neural networks and its applications to PolSAR image classification," *IEEE Transactions on Geoscience and Remote Sensing*, in press.

[54] Y. Q. Jin and F. Xu, "Polarimetric scattering and SAR information retrieval," *Wiley-IEEE*, 2013.

[55] F. Xu, Y. Q. Jin, and A. Moreira, "A preliminary study on SAR advanced information retrieval and scene reconstruction," *IEEE Geoscience and Remote Sensing Letters*, 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 *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2014.

[57] J. Geng, J. Fan, H. Wang, X. Ma, B. Li, and F. Chen, "High-resolution SAR image classification via deep convolutional autoencoders," *IEEE Geoscience and Remote Sensing Letters*, 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 Transactions on Geoscience and Remote Sensing*, 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," *Journal of Sensors*, vol. 2015, no. 538063, 2015.

[60] B. Hou, H. Kou, and L. Jiao, "Classification of polarimetric SAR images using multilayer autoencoders and superpixels," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 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 Geoscience and Remote Sensing Letters*, 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 Sensing Letters*, 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 Recognition*, vol. 61, pp. 686–701, 2017.

[64] L. Jiao and F. Liu, "Wishart deep stacking network for fast PolSAR image classification," *IEEE Transactions on Image Processing*, 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 Geoscience and Remote Sensing Letters*, 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 Recognition*, vol. 64, pp. 255–267, 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 Transactions on Geoscience and Remote Sensing*, 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 Processing*, 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," *International Journal of Remote Sensing*, 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. Geoscience and Remote Sensing*, 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 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. Geoscience and Remote Sensing*, 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 Sensing*, vol. 8, no. 7, p. 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 Transactions on Geoscience and Remote Sensing*, 2017.

[75] S. Bhagavathy and B. S. Manjunath, "Modeling and detection of geospatial objects using texture motifs," *IEEE Transactions on Geoscience and Remote Sensing*, 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 Transactions on Geoscience and Remote Sensing*, 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 Sensing of 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 *IEEE International Conference on Computer Vision (ICCV))*, 2003.

[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 Sensing 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 Geoscience and Remote Sensing Letters*, 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 *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[82] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neuralnetworks," *arXiv:1508.00092*, 2015.

[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 Sensing*, 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 Transactions on Geoscience and Remote Sensing*, 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 Geoscience and Remote Sensing Letters*, 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 Recognition*, vol. 61, pp. 539–556, 2016.

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

[88] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *IEEE International Conference on Learning Representations (ICLR)*, 2014.

[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 *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[90] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010.

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

[92] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, 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 Journal of Photogrammetry and Remote Sensing*, vol. 117, pp. 11–28, 2016.

[94] N. Yokoya and A. Iwasaki, "Object detection based on sparse representation and hough voting for optical remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 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 Journal of Photogrammetry and Remote Sensing*, vol. 89, pp. 37–48, 2014.

[96] X. Jin and C. H. Davis, "Vehicle detection from high-resolution satellite imagery using morphological shared-weight neural networks," *Image and Vision Computing*, 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 Geoscience and remote sensing letters*, 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 *International Symposium on Bioelectronics and Bioinformatics*, 2015.

[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 Systems and Signal Processing*, 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 10, pp. 4895–4909, 2015.

[101] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[102] A.-B. Salberg, "Detection of seals in remote sensing images using features extracted from deep convolutional neural networks," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2015.

[103] I. Ševo and A. Avramović, "Convolutional neural network based automatic object detection on aerial images," *IEEE Geoscience and Remote Sensing Letters*, 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 *IEEE International Conference on Image Processing (ICIP)*, 2015.

[105] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Transactions on Geoscience and Remote Sensing*, 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 Transactions on Geoscience and Remote Sensing*, 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," *Neurocomputing*, 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 Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3325–3337, 2015.

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

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

[111] P. Napoletano, "Visual descriptors for content-based retrieval of remote sensing images," *arXiv:1602.00970*, 2016.

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

[113] T. Jiang, G.-S. Xia, and Q. Lu, "Sketch-based aerial image retrieval," in *IEEE International Conference on Image Processing (ICIP)*, 2017.

[114] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proceedings of the 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 Geoscience and Remote Sensing Magazine*, 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 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 grs-s data-fusion contest," *IEEE Transactions on Geoscience and Remote Sensing*, 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*

Geoscience and Remote Sensing Letters, 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 Geoscience and Remote Sensing Magazine*, 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," *Sensing and Imaging*, vol. 17, 2016.

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

[122] Y. Yuan, S. Zheng, and X. Lu, "Hyperspectral image superresolution by transfer learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 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 *European Conference on Computer Vision (ECCV)*, 2014.

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

[125] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geoscience and Remote Sensing Letters*, 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 International Geoscience and Remote Sensing Symposium (IGARSS)*, 2015.

[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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 12, pp. 5547–5559, 2016.

[128] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution semantic labeling with convolutional neural networks," *arXiv:1611.01962*, 2017.

[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 *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2016.

[130] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," *arXiv:1506.02142*, 2015.

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

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

[133] D. Marcos, R. Hamid, and D. Tuia, "Geospatial correspondence for multimodal registration," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[134] M. Gong, T. Zhan, P. Zhang, and Q. Miao, "Superpixel-based difference representation learning for change detection in multispectral remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 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 Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 24–41, 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 Sensing of Environment*, in revision.

[137] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," *arXiv:1606.02585*, 2016.

[138] A. Marcu and M. Leordeanu, "Dual local-global contextual pathways for recognition in aerial imagery," *arXiv:1605:05462*, 2016.

[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 *Joint*

Urban Remote Sensing Event (JURSE), 2017.

[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 *Joint Urban Remote Sensing Event (JURSE)*, 2017.

[141] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. van den Hengel, "Semantic labeling of aerial and satellite imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 7, pp. 2868–2881, 2016.

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

[143] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *arXiv:1612.01337*, 2017.

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

[145] S. Srivastava, M. Volpi, and D. Tuia, "Joint height estimation and semantic labeling of monocular aerial images with CNNs," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017.

[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 *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016.

[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," *Proceedings of the 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 c urban trees," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[150] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun, "Hd maps: Fine-grained road segmentation by parsing ground and aerial images," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[151] T. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[152] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *European Conference on Computer Vision (ECCV)*, 2016.

[153] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[154] S. Workman and N. Jacobs, "On the location dependence of convolutional neural network features," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2015.

[155] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

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

[157] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

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

[159] P. Fischer, A. Dosovitskiy, and T. Brox, "Descriptor matching with convolutional neural networks: a comparison to SIFT," *arXiv:1406.6909*, 2014.

[160] A. Handa, M. Blösch, V. Patraucean, S. Stent, J. McCormac, and A. J. Davison, "gvnn: Neural network library for geometric computer vision," *arXiv:1607.07405*, 2016.

[161] K. Lenc and A. Vedaldi, "Learning covariant feature detectors," *arXiv:1605.01224*, 2016.

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

[163] K. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned invariant feature transform," in

European Conference on Computer Vision (ECCV), 2016.

[164] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 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 *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[166] L. Li, X. Yu, S. Zhang, X. Zhao, and L. Zhang, "3D cost aggregation with multiple minimum spanning trees for stereo matching," *Applied Optics*, 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 *International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing*, 2017.

[168] H. Park and K. M. Lee, "Look wider to match image patches with convolutional neural networks," *IEEE Signal Processing Letters*, 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 *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[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 *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2016.

[171] C. Häne, C. Zach, A. Cohen, R. Angst, and M. Pollefeys, "Joint 3D scene reconstruction and class segmentation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[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 *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[173] M. Blahá, C. Vogel, A. Richard, J. D. Wegner, T. Pock, and K. Schindler, "Towards integrated 3D reconstruction and semantic interpretation of urban scenes," in *Dreiländertagung*

der SGPF, DGPF und OVG : Lösungen für eine Welt im Wandel : Vorträge, 2016.

[174] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," *arXiv:1206.5533*, 2012.

[175] G. Montavon, G. B. Orr, and K.-R. Müller, *Neural networks: Tricks of the trade*. Springer, 2012.

[176] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," *arXiv:1508.00092*, 2015.

[177] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS)*, 2010.

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

[179] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on EarthVision*, 2015.

[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 Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 14–26, 2017.

[181] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv:1511.02680*, 2015.

[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," *Annals of 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 *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2015.

[184] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *IEEE International Conference on Machine Learning (ICML)*, 2007.

[185] T. Durand, N. Thome, and M. Cord, "Weldon: Weakly supervised learning of deep convolutional neural networks," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[186] R. Johnson and T. Zhang, "Supervised and semi-supervised text categorization using lstm for region embeddings," in *IEEE International Conference on Machine Learning (ICML)*, 2016.

Xiao Xiang Zhu (S'10-M'12-SM'14) received the bachelor degree in space engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2006. She received the Master (M.Sc.) degree, her doctor of engineering (Dr.-Ing.) degree and her "Habilitation" in the field of signal processing from Technical University of Munich (TUM), Munich, Germany, in 2008, 2011 and 2013, respectively.

She is currently the Professor for Signal Processing in Earth Observation (since 2015) at Technical University of Munich (TUM) and German Aerospace Center (DLR), the head of the Team Signal Analysis (since 2011) at the Remote Sensing Technology Institute, DLR, and the head of the Helmholtz Young Investigator Group "SiPEO" (since 2013), DLR and TUM. Prof. Zhu was a guest scientist or visiting professor at the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan and University of California, Los Angeles, United States in 2009, 2014, 2015 and 2016, respectively. Her main research interests are: advanced InSAR techniques such as high dimensional tomographic SAR imaging and SqueeSAR; computer vision in remote sensing including object reconstruction and multi-dimensional data visualization; big data analysis in remote sensing, and modern signal processing, including innovative algorithms such as sparse reconstruction, nonlocal means filter, robust estimation and deep learning, with applications in the field of remote sensing such as multi/hyperspectral image analysis.

Dr. Zhu is an associate Editor of IEEE Transactions on Geoscience and Remote Sensing.

Devis Tuia (S'07, M'09, SM'15) received the Ph.D. from University of Lausanne in 2009. He was a Post-doc at the University of Valéncia, the University of Colorado, Boulder, CO and EPFL Lausanne. Between 2014 and 2017, he was Assistant Professor at the University of Zurich. He is now Associate Professor at the GeoInformation Science and Remote Sensing Laboratory at Wageningen University, the Netherlands. He is interested in algorithms for information extraction and data fusion of geospatial data (including remote sensing) using machine learning and computer vision. More info on http://devis.tuia.googlepages.com/

Lichao Mou (S'16) received the Bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012 and the Master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), China, in 2015. In 2015 he spent six months at the Computer Vision Group at the University of Freiburg in Germany. He is currently working toward the Ph.D. degree at the German Aerospace Center (DLR), Wessling, Germany, and the Technical University of Munich (TUM), Munich, Germany. His research interests include remote sensing, computer vision, and machine learning, especially remote sensing video analysis and deep networks with their applications in remote sensing.

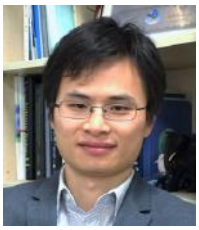
He was the recipient of the first place in the 2016 IEEE GRSS Data Fusion Contest.

Gui-Song Xia (M'10-SM'15) received the B.S. degree in electronic engineering and the M.S. degree in signal processing from Wuhan University, Wuhan, China, in 2005 and 2007, respectively, and the Ph.D. degree in image processing and computer vision from the CNRS LTCI, TELECOM ParisTech, Paris, France, in 2011. Since 2011, he has been a Post-Doctoral Researcher with the Centre de Recherche en Mathmatiques de la Decision, CNRS, Paris-Dauphine University, Paris, for one and a half years.

He is currently a Professor with the State Key Laboratory of Information Engineering, Surveying, Mapping and Remote Sensing, Wuhan University, China. His current research interests include mathematical image modeling, texture synthesis, image indexing and content-based retrieval, structure from motion, perceptual grouping, and remote sensing maging.

Liangpei Zhang (M'06-SM'08) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xian Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xian, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, 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. He holds 15 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a fellow of the Institution of Engineering and Technology, an Executive Member (Board of Governors) of the China National Committee of 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 regularly serves as a Co-Chair of the series SPIE Conferences on Multispectral Image Processing and Pattern Recognition, Conference on Asia Remote Sensing, and many other conferences. He edits several conference proceedings, issues, and geoinformatics symposiums. He also serves as an Associate Editor of the International Journal of Ambient Computing and Intelligence, the international Journal of Image and Graphics, the International Journal of Digital Multimedia Broadcasting, the Journal of Geo-spatial Information Science, and the Journal of Remote Sensing, and the Guest Editor of the Journal of Applied Remote Sensing and the Journal of Sensors. He is serving as an Associate Editor of the IEEE Transactions on Geoscience and Remote Sensing.



Feng Xu (S06-M08-SM14) received the B.E. with honor in Information Engineering from Southeast University, Nanjing, China and the Ph.D. with honor in Electronic Engineering from Fudan University, Shanghai, China, in 2003 and 2008, respectively. From 2008 to 2010, he was a postdoctoral fellow with the NOAA Center for Satellite Application and Research (STAR), Camp Springs, MD. From 2010 to 2013, he was with at Intelligent Automation Inc. Rockville MD, while partly working for NASA Goddard Space Flight Center, Greenbelt, MD as a research scientist. In 2012, he was selected into Chinas Global Experts Recruitment Program, and subsequently returned to Fudan University in June 2013, where he currently is a professor in the school of information science and technology and the vice director of the MoE Key Lab for Information Science of Electromagnetic Waves.

He has published more than 30 papers in peer-reviewed journals, co-authored 2 books, and 2 patents, among many conference papers. Among other honors, he was awarded the second-class National Nature Science Award of China in 2011. He was the 2014 recipient of the Early Career Award of IEEE Geoscience and Remote Sensing Society and the 2007 recipient of the SUMMA graduate fellowship in the advanced electromagnetics area. He currently serves as the associate editor for IEEE Geoscience and Remote Sensing Letters. He is the founding chair of IEEE GRSS Shanghai Chapter. His research interests include electromagnetic scattering theory, SAR information retrieval and radar system development.

Friedrich Fraundorfer (S'10-M'12-SM'14) received the Ph.D. degree in computer science from Graz University of Technology, Austria in 2006 working at the Institute of Computer Graphics and Vision headed by Franz Leberl and Horst Bischof.

He is currently Assistant Professor at Graz University of Technology, Austria. Prior to this he had post-doc stays at the University of Kentucky (US), at the University of North Carolina at Chapel Hill (US) and at ETH Zrich (Switzerland). From 2012 to 2014 he acted as Deputy Director of the Chair of Remote Sensing Technology at the Faculty of Civil, Geo and Environmental Engineering at the Technische Universit?t Mnchen. His main research areas are 3D Computer Vision, Robot Vision, Multi View Geometry, Visual-Inertial Fusion, Micro Aerial Vehicle, Autonomous Systems, Aerial Imaging. He is the author of a well perceived two-part tutorial about visual odometry in the IEEE Robotics and Automation Magazine. His work on autonomous UAVs was awarded Best Paper Finalist at IEEE IROS 2012.