Advanced Multi-Sensor Optical Remote Sensing for Urban Land Use and Land Cover Classification: Outcome of the 2018 IEEE GRSS Data Fusion Contest

Yonghao Xu, *Student Member, IEEE*, Bo Du, *Senior Member, IEEE*, Liangpei Zhang, *Fellow, IEEE*, Daniele Cerra, *Member, IEEE*, Miguel Pato, Emiliano Carmona, Saurabh Prasad, *Senior Member, IEEE*, Naoto Yokoya, *Member, IEEE*, Ronny Hänsch, *Member, IEEE*, and Bertrand Le Saux, *Member, IEEE*

***Abstract*—This paper presents the scientific outcomes of the 2018 Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society. The 2018 Contest addressed the problem of urban observation and monitoring with advanced multi-source optical remote sensing (multispectral LiDAR, hyperspectral imaging, and very high-resolution imagery). The competition was based on urban land use and land cover classification, aiming to distinguish between very diverse and detailed classes of urban objects, materials, and vegetation. Besides data fusion, it also quantified the respective assets of the novel sensors used to collect the data. Participants proposed elaborate approaches rooted in remote-sensing, and also in machine learning and computer vision, to make the most of the available data. Winning approaches combine convolutional neural networks with subtle earth-observation data scientist expertise.**

***Index Terms*—Convolutional neural networks (CNN), deep learning, hyperspectral (HS) imaging (HSI), image analysis and data fusion, multimodal, multiresolution, multisource, multispectral light detection and ranging (LiDAR).**

Manuscript received December 14, 2018; revised March 14, 2019; accepted April 7, 2019. The work of Y. Xu, B. Du, and L. Zhang was supported in part by the National Natural Science Foundation of China under Grants 41431175, 61822113, 41871243, and 61471274, in part by the National Key R & D Program of China under Grant 2018YFA0605501, and in part by the Natural Science Foundation of Hubei Province under Grant 2018CFA05. *(Corresponding author: Bertrand Le Saux.)*

Y. Xu and L. Zhang are with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: yonghaoxu@ieee.org; zlp62@whu.edu.cn).

B. Du is with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: remoteking@whu.edu.cn).

D. Cerra, M. Pato, and E. Carmona are with the German Aerospace Center (DLR), Remote Sensing Technology Institute (MF-PBA), 82234 Weßling, Germany (e-mail: daniele.cerra@dlr.de; miguel.figueiredovazpato@dlr.de; emiliano.carmona@dlr.de).

S. Prasad is with the Electrical and Computer Engineering Department, University of Houston, Houston, TX 77004 USA (e-mail: saurabh.prasad@ieee.org).

N. Yokoya is with the RIKEN Center for Advanced Intelligence Project, RIKEN, 103-0027 Tokyo, Japan (e-mail: naoto.yokoya@riken.jp).

R. Hänsch is with the Computer Vision and Remote Sensing Department, Technical University of Berlin, 10587 Berlin, Germany (e-mail: r.haensch@tu-berlin.de).

B. Le Saux is with DTIS, ONERA, University Paris Saclay, F-91123 Palaiseau, France (e-mail: bertrand.le_saux@onera.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2019.2911113

I. Introduction

OBSERVATION and monitoring of urban centers is a major challenge for remote sensing and geospatial analysis with tremendous needs for working solutions and many potential applications. Urban planning benefits from keeping track of city center evolution or knowing how the land is used (for public facilities, residential or commercial areas, etc.). Quantifying impervious surfaces and how much space is dedicated to vegetation is as crucial for environmental problems as identifying allergenic tree species or quantifying car traffic is for health issues.

Nowadays, multiple sensor technologies can be used to measure scenes and objects from the air, including sensors for multispectral and hyperspectral (HS) imaging (HSI), synthetic aperture radar (SAR), and light detection and ranging (LiDAR). They bring different and complementary information—spectral characteristics which may help to distinguish between various materials, height of objects and buildings to differentiate, e.g., between different types of settlement, and intensity or phase information. With very high-resolution (VHR) data, object shape and relationships between objects become more meaningful in order to understand the content of the observed scene.

The Image Analysis and Data Fusion Technical Committee (IADF TC) of the IEEE Geoscience and Remote Sensing Society (GRSS) is an international network of scientists working on remote sensing image analysis, geo-spatial data fusion, and algorithms. It aims at connecting people and resources, educating students and professionals, and fostering innovation in multimodal earth-observation data processing. Since 2006, it has been organizing the Data Fusion Contest (DFC) every year, which brings new challenges to the community in order to evaluate existing techniques and foster the progress of new approaches.

Two clear contest objectives were pursued previously. The first one consists in delivering previously unseen types of data captured by novel sensors and multiple sensor fusion including pansharpening [1], multi-temporal SAR and optical data [2], HS data which have become reference datasets [3]–[5], multiangular data [6], or videos from space with optical data at multiple resolutions [7]. The second goal is the release of multimodal data

(possibly coupled with ground truth) at a larger scale than the current state-of-the-art. This aims at enabling new families of algorithms to emerge. It includes change detection [8], large-scale fusion of optical, SAR, and LiDAR data [9], classification [4], [5], and large-scale classification and domain adaptation [10].

The 2018 DFC actually belongs to both categories. It proposed data captured by an innovative LiDAR system, which operates at several wavelengths and is capable of recording a diversity of spectral reflectances from objects [11]. It also tackled the problem of automatic classification of multi-modal optical remote sensing data to monitor urban land use and land cover (LULC). A dataset over a large extent of Central Houston (up to 5 km^2) was released, which comprised very high-resolution data for every sensor and an associated semantic reference data with a very diverse taxonomy.

Specifically, the following data were gathered, co-registered, and annotated: multispectral LiDAR point-cloud; HS data; and VHR color imagery. The land use classification task was cast as a 20-class problem, which comprises more detailed urban classes than usual. For example, buildings are either commercial or residential, while vegetation comprises stressed and healthy grass, evergreen and deciduous trees. To test the limits of current sensors, rare objects which correspond to specific man-made materials were also included—cars, trains, railways, and stadium seats.

The competition was framed as three challenges: Two single-sensor tracks for HS and LiDAR and a data fusion track for a combination of at least two sources of data. It took place in two phases: First, participants got access to an area in Central Houston as well as to the corresponding reference data for training. Second, only optical multi-source data were released for a blind classification round. The considered area was also in Central Houston, but larger and with more diverse content. Participants were asked to submit their classification maps on the IEEE GRSS Data and Algorithm Standard Evaluation website (DASE [1]) [12], [13], where they could get instant evaluation and rank in the competition.

In this paper, we report the outcomes of the competition. After describing the dataset (see Section II), first we will discuss the overall results of the contest as a whole (see Section III). Then, we will focus in more detail on the approaches proposed by the first and second place teams (see Sections IV and V, respectively). Finally, conclusions will be drawn in Section VI.

II. Data of the DFC 2018

The following multimodal optical remote sensing datasets were preprocessed and provided to the participants:

1) Multispectral LiDAR (MS-LiDAR) point cloud data, the rasterized intensity and digital surface model (DSM) at a 0.5-m ground sampling distance (GSD);
2) HS data at a 1-m GSD;
3) VHR color imagery at a 5-cm GSD.

The datasets were acquired by the National Center for Airborne Laser Mapping (NCALM) at the University of Houston (UH) on February 16, 2017, between 16:31 and 18:18 GMT, covering the University of Houston campus and its surrounding urban areas. The MS-LiDAR data provided in the contest are the first benchmark multispectral LiDAR data made freely available to the remote sensing community.

The three remote sensing datasets and the corresponding reference data for the training area [the red area in Fig. 1(a)] were provided on January 15, 2018. The remote sensing datasets covering the test area [the entire imagery except red in Fig. 1(a)] were disclosed on March 13, 2018, followed by the 12-day test phase. Fig. 1(b)–(g) show visual examples of reference data, the color composite of MS-LiDAR, the DSM, the color composite of HS data, and the VHR imagery, respectively. Image registration was performed on the three multimodal remote sensing data using ground control points. A particular care was brought so that all the sensors are lined up exactly, such that the centers of pixels from HSI match the color and LiDAR layers.

A. Multispectral LiDAR

The multispectral LiDAR data were acquired by an Optech Titan MW (14SEN/CON340) with an integrated camera. This MS-LiDAR sensor was operated at three different laser wavelengths, i.e., 1550 (#1, near infrared), 1064 (#2, near infrared), and 532 nm (#3, green). The point cloud data from first return for all channels were made available. Seven LiDAR-derived rasters were produced—three intensity rasters for each wavelength and four elevation models representing the elevation in meters above sea level. In particular, elevation rasters include: 1) first surface model (i.e., DSM) generated from first returns detected on Titan channels #1 and #2; 2) bare-earth digital elevation model (DEM) generated from returns classified as ground from all Titan sensors; 3) bare-earth DEM with void filling for manmade structures; and 4) a hybrid ground and building DEM, generated from returns that where classified as coming from buildings and the ground by all Titan sensors. All rasters were resampled to a 0.5-m grid—intensity rasters were interpolated using inverse distance weighting to a power two with a search radius of 3 m while elevation rasters were generated using Kriging, with a search radius of 3–5 m. The size of the rasterized datasets is 8344×2404 pixels.

B. HS Data

The HS imagery was collected by an ITRES CASI 1500 sensor, covering a 380–1050 nm spectral range with 48 bands at a 1-m GSD. This HS data cube has been orthorectified and radiometrically calibrated to units of spectral radiance (milli-SRU). The sampling of HS imagery is mostly aligned with the VHR imagery, even if a few, residual errors can remain due to various factors—camera parameters, image parallax or distortion, or sensor trajectory. The dataset was distributed in radiance and the image size is 4172×1202 pixels.

C. VHR Color Imagery

The VHR color imagery was obtained by a DiMAC ULTRALIGHT+ camera with a 70-mm focal length. Processing steps

[1] http://dase.grss-ieee.org/

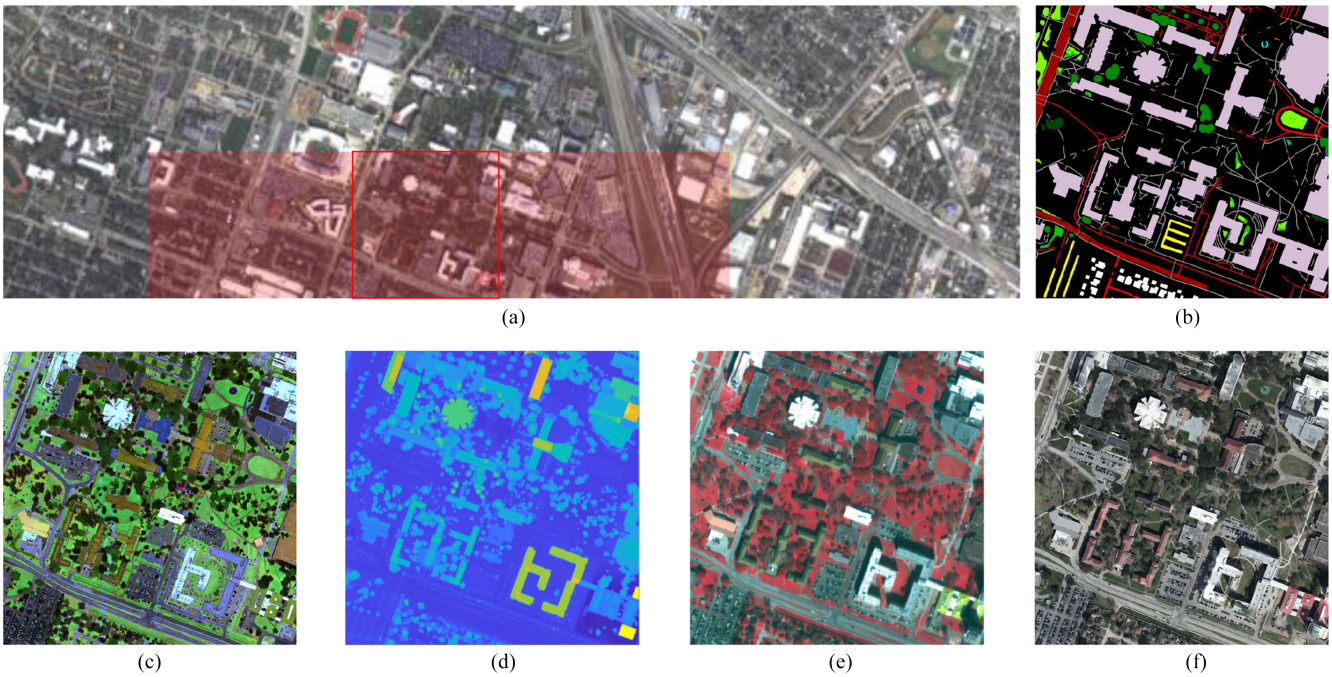


Fig. 1. Dataset overview. (a) Training (red) and test (entire imagery except red) areas, examples of (b) ground truth, (c) color composite of multispectral LiDAR intensity, (d) DSM, (e) color composite of HS imagery, and (f) VHR color imagery.

were applied—optimization of white balance, calibration with respect to plane instruments, and orthorectification geolocalization. Given large parallax differences, the creation of seamless images is extremely difficult around large buildings, resulting in a few artifacts (data voids) around larger structures such as the UH main stadium. The final image product was resampled at a 5-cm GSD with the size of 83440 $\times$ 24040 pixels. The image was distributed after being divided into 14 (i.e., 7 $\times$ 2) tiles with each tile having the size of 11920 $\times$ 12020 pixels.

D. Reference Data

For the training area [the red area in Fig. 1(a)], we provided reference data of the 20 LULC classes. Table I defines the LULC classes with the number of training and test samples. The reference data were prepared by the organizer based on a field survey, open map information (e.g., OpenStreetMap), and visual inspection of the datasets distributed in the contest. The reference data were provided only for the training area as a raster image at a 0.5-m GSD. The reference data for testing remain undisclosed and were used for the evaluation of the submitted results at a 0.5-m GSD for all the tracks in DASE.

As shown in Table I, the distribution of the classes is imbalanced for training, while that of the test area is better balanced by resampling. The training and test areas were fully separated into different regions with a ratio of 4 to 10 to assess the generalization ability of classification systems. Different from the 2013 DFC, where the ground truth was sparse, the dense reference data provided for training during 2018 DFC were made available to promote the advancement of deep learning-based approaches, leading to the imbalance issue. For testing, the reference data were created in the same way as for the training area but the samples were randomly resampled from the entire test area to balance the numbers of test samples for different classes.

TABLE I
LULC CLASSES

#	class	# of training samples	# of test samples
1	Healthy grass	39196	20000
2	Stressed grass	130008	20000
3	Artificial turf	2736	20000
4	Evergreen trees	54322	20000
5	Deciduous trees	20172	20000
6	Bare earth	18064	20000
7	Water	1064	1628
8	Residential buildings	158995	20000
9	Non-residential buildings	894769	20000
10	Roads	183283	20000
11	Sidewalks	136035	20000
12	Crosswalks	6059	5345
13	Major thoroughfares	185438	20000
14	Highways	39438	20000
15	Railways	27748	11232
16	Paved parking lots	45932	20000
17	Unpaved parking lots	587	3524
18	Cars	26289	20000
19	Trains	21479	20000
20	Stadium seats	27296	20000

III. Submissions and Results

There are 374 unique registrations for downloading the data and 95 teams participated in the contest. We have received a total of 1334 submissions, divided into 538, 347, and 449 submissions for the data fusion, multispectral LiDAR, and HS tracks, respectively. The ranking of the submitted classification results

TABLE II
TOP RANKED TEAMS WITH CLASSIFICATION PERFORMANCE IN OVERALL ACCURACY (OA), COHEN'S KAPPA, AND AVERAGE PRODUCER'S ACCURACY (AA)

Rank	Track	Team	OA (%)	Kappa	AA (%)	Affiliation
1	MS LiDAR	Gaussian	81.07	0.80	72.01	Wuhan Univ.
2	Data Fusion	Gaussian	80.78	0.80	71.66	Wuhan Univ.
3	Data Fusion	dlrpba	80.74	0.80	76.32	DLR
4	Data Fusion	AGTDA	79.79	0.79	76.15	AGT
5	Data Fusion	IPIU	79.23	0.78	74.4	Xidian Univ.
6	MS LiDAR	AGTDA	78.05	0.77	71.54	AGT
7	MS LiDAR	IPIU	78.01	0.77	70.83	Xidian Univ.
8	Data Fusion	challenger	77.9	0.77	75.99	Xidian Univ.
9	HS	challenger	77.39	0.73	75.78	Xidian Univ.
10	Data Fusion	XudongKang	76.45	0.75	71.26	Hunan Univ.
11	HS	XudongKang	76.37	0.75	71.59	Hunan Univ.
12	MS LiDAR	GaoLei	75.82	0.74	70.27	Xidian Univ.

was automatically computed on DASE based on the overall accuracy (OA) for each track. The evaluation was carried out at a 0.5-m GSD for all tracks. The average accuracy (i.e., average of producer's class accuracies) and Cohen's kappa were also measured to provide additional insights into the results. Table II provides an overview of the twelve best performing teams of the leaderboard among all the tracks. As expected, the data fusion-based results were competitive, occupying six out of the top 12. It is worth noting that the best result was obtained by using only the multispectral LiDAR data, implying the great potential of multispectral LiDAR for the complex LULC classification.

The best ranked team for each track (*Gaussian* for both the data fusion and multispectral LiDAR tracks and *challenger* for the HS track) and additional top-ranking teams (*dlrpba* and *AGTDA*) among all the tracks were awarded. The top two teams (*Gaussian* and *dlrpba*) were determined based on OA. The solutions of the four top-ranked teams were presented during the 2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in Valencia, Spain. The four teams are given as follows.

1) First place: *Gaussian* team; Yonghao Xu, Bo Du, and Liangpei Zhang from Wuhan University, China; multi-source remote sensing data classification via fully convolutional networks and post-classification processing [14].
2) Second place: *dlrpba* team; Daniele Cerra, Miguel Pato, Emiliano Carmona, Jiaojiao Tian, Seyed Majid Azimi, Rupert Müller, Ksenia Bittner, Corentin Henry, Eleonora Vig, Franz Kurz, Reza Bahmanyar, Pablo d'Angelo, Kevin Alonso, Peter Fischer, and Peter Reinartz from German Aerospace Center, Germany; combining deep and shallow neural networks (NNs) with *ad hoc* detectors for the classification of complex multi-modal urban scenes [15].
3) Third place: *challenger* team; Shuai Fang, Dou Quan, Shuang Wang, Lei Zhang, and Ligang Zhou from Xidian University, China; a two-branch network with semi-supervised learning for HS classification [16].
4) Third place, ex aequo: *AGTDA* team; Sergey Sukhanov, Dmitrii Budylskii, Ivan Tankoyeu, Roel Heremans, and Christian Debes from AGT International, Germany; fusion of LiDAR, HS, and RGB data for urban LULC classification [17].

The best performing approaches are based on deep NNs together with post-processing and/or object detection techniques. In the history of the DFC classification benchmarks, this is the first time that deep learning-based approaches occupied the leaderboard so much and demonstrated the capability of dealing with complex urban LULC classification. Indeed, there is a shift in the way data fusion is processed; not anymore using ensemble methods to fuse features, including deep learning ones, as in [10], but directly with deep networks. This can be attributed to the unprecedented size of the dataset and the availability of numerous training samples for all classes. It is worth noting that the top two teams achieved the best results with the use of *ad hoc* post-processing and/or object detection techniques to boost the classification performance, which yields an improvement of around 15% accuracy. This trend is consistent with the DFC editions in 2013 and 2014 [4], [5], where classification refinement by post-processing played a key role to address the specific classification tasks.

Fig. 2 shows the classification maps of the four winning teams over the entire scene. Although there are some minor differences, the maps in the data fusion and multispectral LiDAR tracks [see Fig. 2(a), (b), and (d)] are consistent while the one in the HS track [see Fig. 2(c)] shows a major difference (e.g., many pixels were misclassified as water). This implies that multispectral LiDAR data play a significant role in the classification task. Though, it is worth noting that these results were obtained using only the derived data, i.e., derived DSM and intensity rasters. Deeper analysis might be reached by processing the original point cloud.

As derived from the overall results, vegetation classes were relatively easy to be distinguished. In particular, evergreen and deciduous trees were well discriminated using MS LiDAR rather than HS data. Various types of roads (i.e., classes #10–14) were often confused with each other since they have similar spatial-spectral characteristics. Highways (class #14) required specific post-processing to be discriminated from the other road classes as reported in the winning solutions (see Sections IV and V). Even with *ad hoc* detectors, it was challenging to detect crosswalks because their materials are the same as roads, sidewalks, and major thoroughfares. It was not possible to identify unpaved parking lots due to intra-class variance and inter-class similarity.

In Sections IV and V, we present the solutions proposed by the first and second ranked teams, respectively. We will detail the winning classification methodologies and provide in-depth analysis of the pros and cons of the solutions.

IV. FIRST PLACE: WUHAN UNIVERSITY TEAM

In this section, we describe the algorithm proposed by the first-place team in detail. The algorithm is based on a fully convolutional network (FCN) [18], named as Fusion-FCN. With well designed network architecture, hierarchical features can be learnt from three different types of data including LiDAR data, HS images, and VHR images simultaneously. Besides, we further implement post-classification processing with the topological relationship among different objects based on the result yielded by the proposed Fusion-FCN, which helps to correct the confusions between some similar categories such as different types of roads.

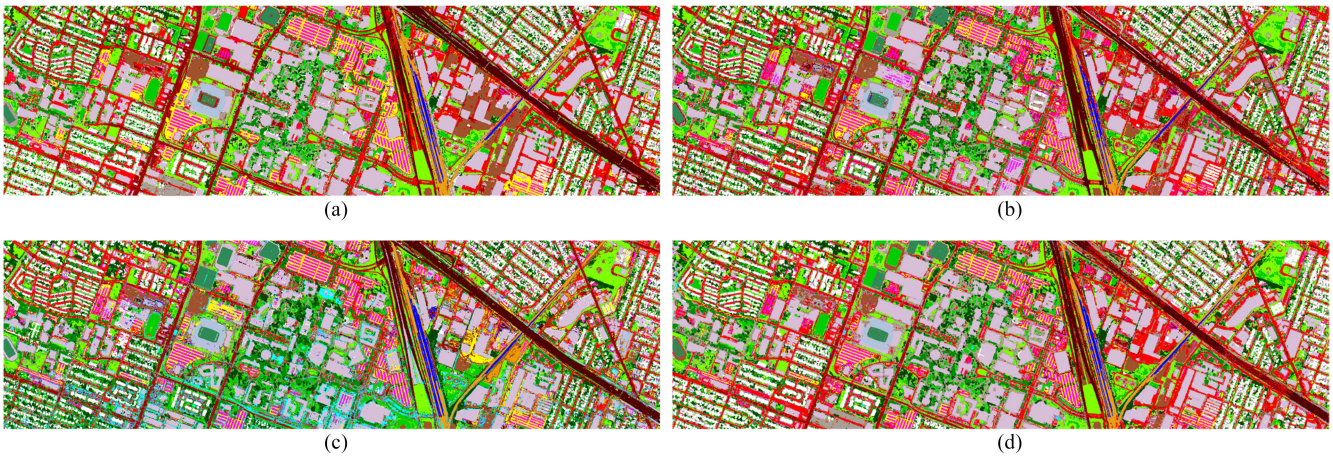


Fig. 2. Classification maps of the winners: (a) Gaussian in data fusion track, (b) dlrpba in data fusion track, (c) challenger in hs track, and (d) AGTDA in data fusion track.

A. Preprocessing

The data preprocessing techniques utilized in our experiments are described as follows.

1) *Resampling*: Since the classification results are expected to be at a 0.5-m GSD, both the HS image and the VHR image are resampled at a 0.5-m resolution with the nearest neighbor method.
2) *Outlier correction*: We find that there are some outliers in the original LiDAR intensity raster data and the DSM data, which may be detrimental to the classification. Here, we simply apply a filtering process to these data. Those pixel values that are greater than a threshold τ are replaced with the minimum value in the data. We set τ as $1e4$ and $1e10$ for LiDAR intensity raster data and DSM data, respectively.
3) *Normalized DSM*: In order to obtain the real height of the object from the LiDAR data, we calculate the normalized DSM (NDSM) value with the following equation:

$$\text{NDSM} = \text{DSM} - \text{DEM}. \tag{1}$$

4) *Data normalization*: For all the data utilized in our experiments, we normalize each feature dimension in the data into a range of $[0, 1]$.
5) *Image partitioning*: Considering the limited memory of the GPU utilized in our experiments, we conduct image partitioning to decrease the memory cost. In the training phase, the full training image is partitioned into 40 subimages with a size of 1202×300. During the test phase, since there is no need to restore the gradient of the network anymore, the full test image is partitioned into 15 subimages with a size of 2404×600.

B. Fusion-FCN

Following the great success of deep learning in computer vision field [19]–[21], many deep models have been proposed to address the remote sensing image classification task [22]–[28]. In this subsection, we describe the proposed Fusion-FCN for the interpretation of multi-sensor remote sensing data in detail. Compared with previous FCN-based approaches [29]–[32], our method can well maintain the boundaries of different objects and decrease the risk of spatial information loss.

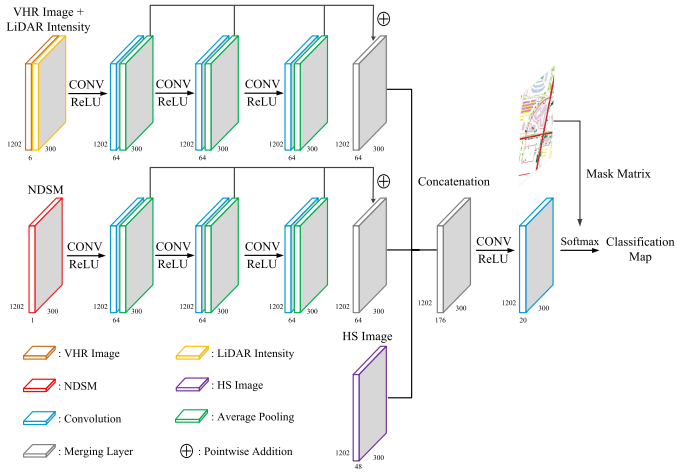


Fig. 3. Architecture of the proposed Fusion-FCN. There are three branches in the network. Each branch acts as a feature extractor for a corresponding type of data. A concatenation layer is adopted to implement the feature fusion.

1) *Overview of the proposed network:* As shown in Fig. 3, the proposed Fusion-FCN consists of three branches. The VHR image and LiDAR intensity raster data are fed into the first branch to learn the hierarchical spatial features. The NDSM data are fed into the second branch to learn the hierarchical elevation features. Both these two branches share the same architecture including three 3×3 convolutional layers and three 2×2 average pooling layers. Those three pooling layers in each branch are further merged into a merging layer with a point-wise addition. This process will make the network possesses the property of multi-scale, which may be beneficial to the remote sensing data classification, where different targets usually tend to have different sizes [33]. Notice that the zero padding is utilized in both convolutional and pooling layers to process

the pixels in the boundary. In this way, the convolutional and pooling features will share a consistent spatial size with the input images. Then, the merging layers in the previous two branches are further concatenated with the third branch (i.e., the original HS image) for the purpose of feature fusion. A 1×1 convolutional layer and the softmax function are adopted to accomplish the pixel-wise image classification.

2) *Optimization:* Let $\hat{y}(u,v)$ and $y(u,v)$ denote the predicted label and real label of the pixel with location (u,v) in the image. Then, the loss function of the network can be defined as the cross entropy between the predicted labels and the real ones

$$\mathcal{L} = -\frac{1}{rc}\sum_{u=1}^{r}\sum_{v=1}^{c}\left[y(u,v)\cdot\log\left(\hat{y}(u,v)\right) + \left(1-y(u,v)\right)\cdot\log\left(1-\hat{y}(u,v)\right)\right] \tag{2}$$

where r and c are numbers of rows and columns of the data, respectively.

The stochastic gradient descent algorithm with the Adam optimizer [34] is utilized to train the network.

C. Post-Classification Processing

Up to now, we can get a preliminary classification map from the trained Fusion-FCN. We find that there are still some misclassifications between similar subclasses, such as different types of roads, since these subclasses share very similar spectral characteristics. To this end, we further implement some post-classification processing with the topological relationship among different objects based on the result yielded by the proposed Fusion-FCN. In order to avoid the phenomenon that some pixels may end up without any class label in this process, we adopt the reclassification/relabeling strategy. We first design some target-specific criteria according to the properties of different objects. If the pixels satisfy these criteria, they will then be relabeled into the corresponding category. Otherwise, their class label will be kept unchanged. The correction for highway objects and the paved parking lot objects is presented as an example.

1) *Correction for highway objects:* We first extract the mixture results of different types of road objects including class No. 10 (roads), class No. 13 (major thoroughfares), and class No. 14 (highways). It can be seen from Fig. 4(a) that most of the highway regions in the mixture results are misclassified as roads or major thoroughfares. In order to remove those tiny connected components, the opening and closing operations are applied to this road network map with a 5×5 square structure element, as shown in Fig. 4(b). Then, the Hough transformation [35] is utilized to implement the line detection. The detected straight lines are colored blue in Fig. 4(c). The final detection results for highway objects are obtained with an empirical criterion that the width of the highway object should be greater than 150 pixels, as shown in Fig. 4(d).

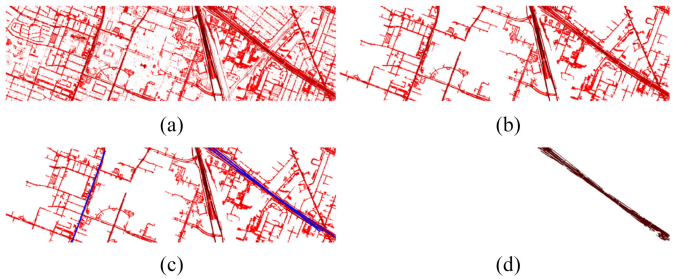


Fig. 4. Illustrations of the correction for highway objects. (a) Mixture map of different types of roads. (b) Road network map after opening and closing operations. Connected components that contain fewer than $1e6$ pixels are removed. (c) Line detection results (colored blue) with Hough transformation. (d) Final map for highway objects with a criterion that the width of the highway object should be greater than 150 pixels.

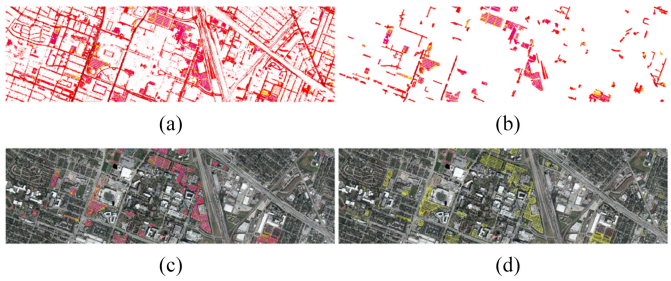


Fig. 5. Illustrations of the correction for paved parking lots objects. (a) Mixture map of roads, major thoroughfares, paved parking lots, and cars. (b) Map after erosion and dilation operations. Connected components that contain fewer than $1e3$ pixels are removed. (c) Detection map for the paved parking lots with a criterion to enforce the car pixels in the connected components should account for more than 15%. (d) Final map for the paved parking lot objects which are colored in yellow.

2) *Correction for paved parking lot objects:* Considering the misclassification between parking lots and different types of roads, we first generate a mixture map with pixels classified as class No. 10 (roads), class No. 13 (major thoroughfares), class No. 16 (paved parking lots), and class No. 18 (cars), as shown in Fig. 5(a). Then, morphological operations including erosion and dilation are utilized to remove those tiny connected components, as shown in Fig. 5(b). The detection for the paved parking lots is achieved with a criterion to enforce that the car pixels in the connected components should account for more than a threshold τ_{car}. In order to select a suitable τ_{car}, we first choose the upper left parking lot in the training image as the observed region. Both the area of the parking lot and the number of car pixels inside this region are counted. Based on these statistics, we calculate the car occupancy of this parking lot and the result is approximately 27%. Considering that the observed parking lot we chose from the training image is almost fully occupied by cars, the threshold used in the post-processing step is supposed to be smaller than this value, so that the less-occupied parking lots can also be considered. On the other hand, a too small threshold may also lead to confusion for those real road objects since the car occupancy for road regions is

TABLE III
EXPERIMENTAL RESULTS WITH DIFFERENT STRATEGIES (REPORTED IN PRODUCER'S ACCURACY, WITH BEST RESULTS SHOWN IN BOLD)

#	class	HS-FCN	LiDAR-FCN	Fusion-FCN	LiDAR-FCN-post	Fusion-FCN-post
1	Healthy grass	88.79	93.72	88.70	93.26	**94.52**
2	Stressed grass	81.69	58.45	**87.30**	61.70	82.40
3	Artificial turf	4.77	13.52	64.14	82.97	**84.26**
4	Evergreen trees	91.42	96.59	97.05	97.37	**97.45**
5	Deciduous trees	25.19	69.48	**73.02**	71.97	71.96
6	Bare earth	0	50.77	27.64	**96.22**	92.87
7	Water	**43.98**	6.63	9.15	7.43	11.24
8	Residential buildings	26.22	88.52	75.03	**89.10**	78.27
9	Non-residential buildings	88.34	89.11	**93.55**	88.88	91.94
10	Roads	**71.29**	70.84	62.44	68.47	68.97
11	Sidewalks	48.05	**75.79**	68.52	75.58	61.55
12	Crosswalks	0.04	**16.58**	7.46	12.18	4.06
13	Major thoroughfares	29.42	32.68	**59.94**	48.82	45.24
14	Highways	11.54	18.80	17.95	**96.60**	93.98
15	Railways	6.55	76.16	80.46	89.68	**90.78**
16	Paved parking lots	24.41	63.74	60.80	94.21	**96.01**
17	Unpaved parking lots	0	0	0	0	0
18	Cars	10.43	64.32	64.33	68.05	**71.29**
19	Trains	31.97	72.56	50.94	**98.08**	96.73
20	Stadium seats	61.41	58.99	41.97	99.64	**99.74**
	OA (%)	41.09	62.37	63.28	**81.07**	80.78
	Kappa	0.37	0.60	0.61	**0.80**	**0.80**

usually much smaller. Based on the above-discussed analysis, we empirically set 15% as the final threshold. In this way, those pure road regions can thereby be filtered, as shown in Fig. 5(c). The final map for paved parking lot objects is shown in Fig. 5(d).

Other techniques utilized in the post-classification processing are briefly summarized as follows.

1) *Artificial turf:* The classification of this class is improved by relabeling those road regions whose NDVI value is greater than 0.75 into the artificial turf category. Morphological operations including opening and closing are also used in this step.
2) *Bare soil:* The erosion and dilation operations with a 7×7 square structure element are adopted to preprocess the union set of both road and bare soil categories. Those connected components whose area is greater than 5000 pixels are relabeled into the bare soil category.
3) *Train:* Pixels having an NDSM value between 3 to 6 m are first extracted from the NDSM layer. Those connected components with a roundness value less than 0.1 are relabeled into the train category.
4) *Stadium seats:* An elevation constraint is applied on the road categories and those pixels having an NDSM value between 3 to 9 meters are relabeled into the stadium seats category.

Finally, the majority voting with a window size of 5×5 is also utilized to further smooth the classification map.

D. Results and Discussion

In this section, we report the experimental results of the proposed method. In order to further investigate the influence of various components in the approach, such as different types of remote sensing data and post-classification processing, we also conduct an ablation study. A brief introduction about the comparing methods are given as follows.

1) *HS-FCN*: A modified version of the proposed Fusion-FCN which only utilizes HS image. It contains two branches. The first branch acts as a spatial feature extractor, where the first three principal components of the HS image are input. The original HS image is fed into the second branch.
2) *LiDAR-FCN*: A modified version of the proposed Fusion-FCN which only utilizes the LiDAR data. It also contains two branches. The first branch acts as a spatial feature extractor, where the LiDAR intensity rasters are input. The second branch acts as an elevation feature extractor which receives the NDSM data.
3) *Fusion-FCN*: The proposed approach, which utilizes the information from VHR images, LiDAR data, and HS image, is shown in Fig. 3.
4) *LiDAR-FCN-post*: The proposed LiDAR-FCN with post-classification processing.
5) *Fusion-FCN-post*: The proposed Fusion-FCN with post-classification processing.

As we can see from Table III, using HS image alone a high accuracy can be hardly obtained with the proposed FCN approach. By contrast, owing to the detailed elevation information contained in the LiDAR data, LiDAR-FCN yields an OA of 62.37%, which outperforms the result of HS-FCN over 20%. Therefore, elevation information plays a significant role in the urban LULC classification task. Combining both HS image and LiDAR data along with the VHR image, the performance can be further improved to 63.28%.

One of the advantages of the proposed approach is the small receptive field adopted in the FCN architectures, which helps to yield a very detailed base map where the

boundaries of different objects are well maintained. This property enables us to implement post-classification processing for those misclassified categories. As shown in Table III, with the help of post-classification processing, the OA of Fusion-FCN can be improved greatly to 80.78%. We also conduct similar post-classification processing to the result of LiDAR-FCN as a comparison. The quantitative result shows that the OA of LiDAR-FCN can also be improved significantly to 81.07% (even slightly better than Fusion-FCN-post), which demonstrates that the proposed post-processing steps are not sensitive to different baseline methods. Compared to the result of Fusion-FCN-post, the slight advantage of LiDAR-FCN-post mainly comes from the classification of residential buildings (89.10% versus 78.27%). This phenomenon also indicates that the LiDAR data plays a significant role in the identification of the building category, and simply stacking more features from other sensors may mislead the classification for this category.

The results in Table III also show some limitations of the proposed methods. First, although Fusion-FCN can yield a higher accuracy on most of the categories compared with the single-sensor-based FCN, it performs much worse on the water class than HS-FCN. Thus, the architecture of Fusion-FCN can be further improved to achieve a better fusion for different types of data. Besides, most of the post-classification techniques utilized in our experiments still rely on the expert knowledge from the designer, and the hyper-parameters need to be tuned manually. How to incorporate these techniques into the network training would be an interesting topic in our future work.

V. Second Place: DLR Team

Recently, classifiers based on deep learning are being extensively used in remote sensing [24]. On the one hand, they are simple to operate if pre-trained or given enough available training data, are able to capture the relevant features from a wide variety of classes, and are robust to overfitting [36]. On the other hand, a deep network often resembles a black box in which it is difficult to understand which features (or their combinations) are driving the decision process. Furthermore, these classifiers may give too much importance to higher order interactions between pixels of the same object. Shallow NNs may sometimes have higher generalization power [37], [38] and, in the specific case of image classification, usually give more diverse predictions when compared to deeper networks [39].

A comparison in [36] concludes that deep networks outperform shallow ones for objects which can be described at different scales and have peculiar features for each such scale. By contrast, classes which are driven by their spectral characteristics, and often exhibit a stationary texture relevant for a single scale, may be equally or better represented by a shallow network. This group of objects may include natural classes, such as grass and bare earth, as opposed to man-made objects often driven by context and for which a multi-scale analysis may yield a better characterization.

For the 2018 DFC, we tested both architectures and verified that a shallow network yielded indeed more homogeneous results on natural classes, including grass, trees, water, and bare earth. These classes were slightly underrepresented in the classification results of a deep network, which on the other hand yielded a significantly superior performance in recognizing more complex structures such as different types of roads and trains.

Based on these considerations, our approach combined the output of both deep and shallow networks. The final classification was derived by overlaying the output of dedicated detectors for specific classes which, for their characteristics, needed to be analyzed with different strategies. The complete workflow is reported in Fig. 6, with its single steps being discussed in the next sections.

A. Preprocessing and Feature Extraction

The multimodal dataset underwent the following preprocessing steps before the feature extraction and classification stages.

1) The LiDAR-derived digital surface models (first and last pulse) were normalized by subtracting the available digital terrestrial model, previously blurred using a Gaussian filter. Additional noise and abnormal values were then removed from the normalized digital surface models (NDSMs).
2) The MS-LiDAR intensity images exhibited both periodic and non-periodic noise. To reduce this noise a 5×5 median filter was applied since it produced better results than notch filters in the Fourier domain.
3) The HS dataset was resampled to 50 cm GSD using an order-3 spline and 42 (out of 48) spectral bands were selected as input for the next stages.

Subsequently, the following features have been extracted from the available datasets.

1) *Topics*: High-level features are captured by the so-called topic vectors, derived from multi-modal latent Dirichlet allocation (mmLDA) [40] and the bag-of-words (BoW) model. These features are computed on image patches extracted from the HS (1-m GSD) and RGB (50-cm GSD) images, with each image element finally represented as a mixture of 50 topics discovered by mmLDA. Fig. 7 codes the dominant topic for each pixel with a different color, showing the strong correlation between some topics and the different classes of interest. For further details, see [15].
2) *Vegetation indices*: In order to separate healthy from stressed grass, both narrow- and broad-band vegetation indices, such as the red edge inflection point (REIP) and the normalized differential vegetation index (NDVI), have been extracted from the upsampled HS image.

The input stack for both shallow and deep networks (see Section V-B) are generated at 50-cm GSD, with each pixel represented as a 100-D vector composed by 48 spectral bands (42 HS, three RGB, and three MS LiDAR bands), the two NDSMs, and the 50-D topic vector.

B. Classification

The scene provided for the contest covers a complex urban environment with a large set of heterogeneous classes. The

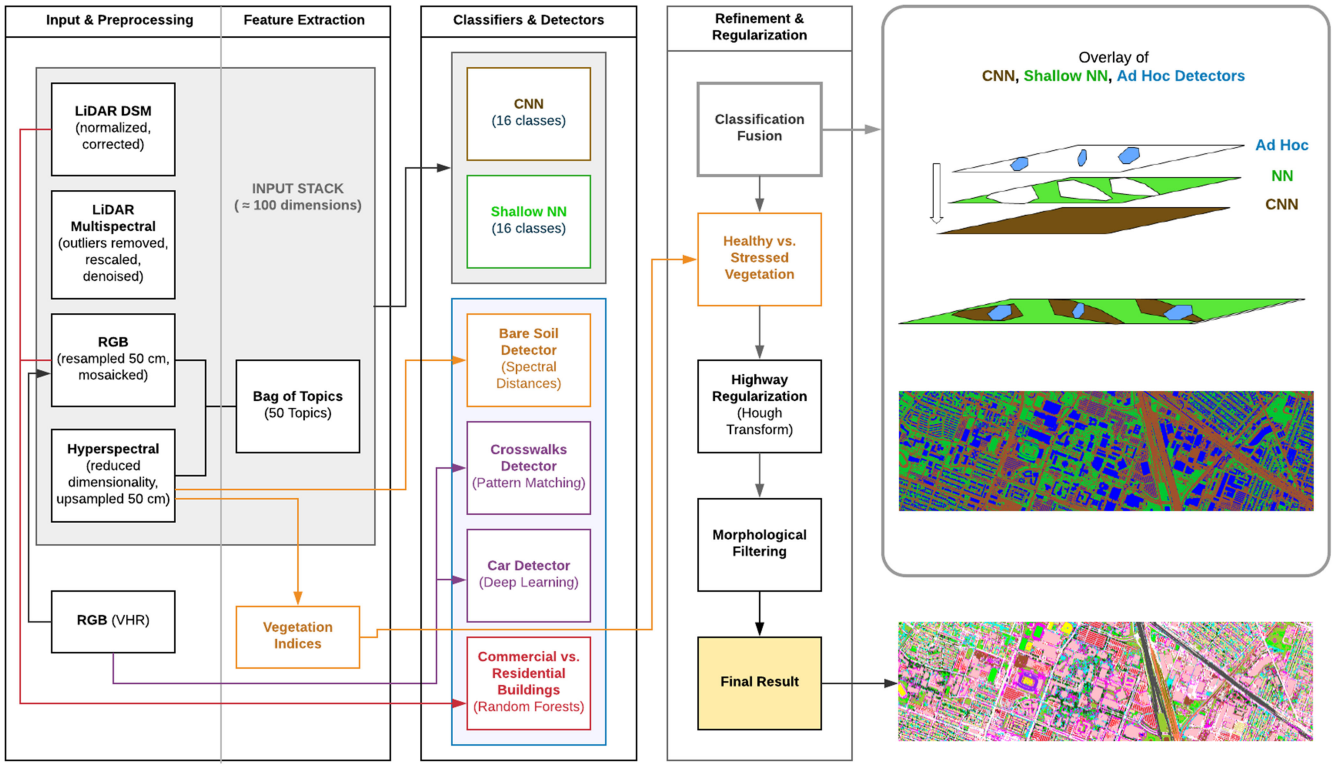


Fig. 6. Workflow of the classification procedure. The classification fusion step is performed according to the top right map, showing the contribution to the final classification results from the deep convolutional NN (CNN) (sienna), shallow fully connected NN (green) and *ad hoc* detectors (blue). Further details in Fig. 10.

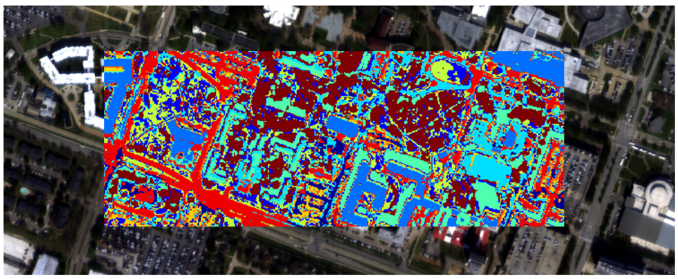


Fig. 7. Illustration of the extracted topic features (insert). The colors represent the dominant topic for each pixel.

classes are not only diverse and inhomogeneous in terms of scale, shape, context, and spectral properties, but their distribution is also highly imbalanced in the training set (cf. Table I). If high accuracies are to be attained for all or most classes, such a challenging scene calls for an integrated approach combining generic classifiers and class-wise tailored detectors in a complementary fashion, as opposed to a unified approach with a single generic classifier. With this in mind, our classification strategy strove to combine: 1) base classifiers, trained on a simplified set of classes to achieve a first generic but accurate classification map, and 2) a number of *ad hoc* detectors, specifically tailored to identify one or two classes thereby refining the results of the base classifiers. The next paragraphs detail our implementation of the base classifiers, *ad hoc* detectors as well as their combination to form our final classifier.

1) Base Classifiers: Classifying the 20 classes of interest listed in Table I at the same time is very challenging, because of the different features driving the recognition of specific classes. For example, shape features are dominant for the class "cars," while spectral features are less important as the color of a car can vary a lot. The opposite is true for the class "water." Therefore, it is considerably easier to work with a restricted set of classes where semantically similar classes are merged, while others are altogether excluded. There is however a tradeoff between restricting the set of classes and obtaining a good overall result in the classification task. After several trials during the training phase of the contest, we defined a simplified set of 16 classes where grass (classes 1 and 2) and buildings (classes 8 and 9) are merged, while crosswalks and cars (classes 12 and 18) are excluded. The merging of road-like classes proved disadvantageous, as we did not manage to obtain an *ad hoc* road-like detector outperforming the base classifiers.

It was in the simplified set of 16 classes described above that our base classifiers were trained. In an effort to exploit the potential of deep learning and at the same time the simplicity of traditional classifiers, we adopted two complementary base classifiers—a deep CNN and a shallow fully connected NN. A multi-class support vector machine with linear kernels was also used but discarded early on due to its inferior performance. The

Keras API [41] with TensorFlow backend was used to implement and train both CNN and NN.

The structure of the CNN can be summarized as eight convolutional, two fully connected, and a final softmax layer. For the classification of a pixel, the network uses as input a matrix of 25 pixels $\times$ 50-D features. The 25 pixels are obtained from the patch of 5 $\times$ 5 pixels around the pixel of interest, while the 50-D features correspond to the first half of the 100-D feature vector previously introduced. Only the 50-D topic vector of the pixel under classification is used in the final steps of the CNN (incorporated to the first fully connected layer). The convolutions are selectively applied along the spatial (one dimension), spectral (one dimension), or combined (two dimensions) dimensions of the input data. The design of the network was chosen after investigating different configurations and contains 1.324×10^6 trainable parameters. At the final steps of the network, two fully connected layers are used before the softmax layer that uses a categorical cross-entropy loss function for the classification into the simplified set of 16 classes. The CNN makes use of the Adam optimizer [34] with the *amsgrad* option. Amsgrad uses non-increasing step sizes, and this may avoid convergence problems which are present in the Adam algorithm [42]. In our preliminary tests, amsgrad showed on average lower training errors. During training, special care was paid to reduce overfitting given the limited amount of training data. For this reason, L2 regularization is introduced in all convolutional layers, a 25% dropout is added between the two fully connected layers and the network training is stopped after a small number of epochs.

The structure of the NN consisted of a two-hidden-layer fully connected NN with a final softmax layer and a categorical cross-entropy loss function. Considering the results obtained on the training set, we opted for 128 $\times$ 64 hidden nodes with rectified linear unit activations, stochastic gradient descent optimizer with batch size of 128, and early stopping after five epochs. In order to handle the imbalanced distribution of classes in the training set (cf. Table I), weights inversely proportional to the number of class samples were applied during training. This ensured that the network learned the features of even the most underrepresented classes. The NN base classifier was fed with different combinations of features, with the final results obtained with the 100-D feature vector containing HS, RGB, MS LiDAR, NDSMs, and topics described in Section V-A. The network contains a total of $\sim$ 22 k trainable parameters. Ensembles of five and ten NN classifiers, merged with majority voting, have also been tested. These led to mild and negligible improvements in training and testing accuracies, respectively, so they were not used to produce the final results.

2) Ad Hoc Detectors: The base classifications described in the previous section were complemented with dedicated detectors for bare earth, residential and non-residential buildings, crosswalks and cars. These *ad hoc* detectors are briefly illustrated in the following paragraphs; see also [15] for a complementary description of the methods used in each detector.

1) *Bare earth*: This class, driven by spectral features, was improved by applying a spectral angle mapper classifier to the HS data, complemented by two cycles of morphological openings and closings (with a disk of radius two as the structuring element). The resulting map was overlaid on the base classification.
2) *Residential and non-residential buildings*: Both types of buildings were segmented by thresholding the NDSM without making a distinction between residential and non-residential. This separation was achieved with a random forest classifier using features extracted from RGB and NDSM [43], and later refined by overlaying the output of a fully CNN (same input features) for the residential buildings class only.
3) *Crosswalks*: A limited number of samples for crosswalk patterns was selected in the 5-cm RGB ground truth and used to train a detector based on normalized cross correlation. Fig. 8 illustrates the details of the results for this dedicated crosswalk detector.
4) *Cars*: After extending the labeling of cars in the training set in a semi-automatic way, a pre-trained fully CNN [44] was trained on the 5-cm RGB dataset. The resulting network was then used to perform pixel-wise car segmentation as shown in Fig. 9. The car mask was improved by applying morphological opening and dilation (with a disk of radius one as the structuring element) and by masking out cars on the highways which were yielding some false alarms.

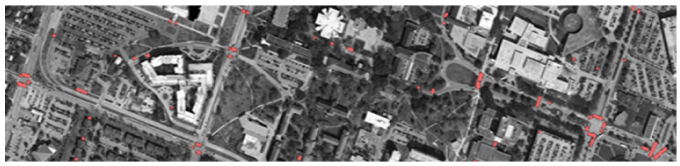


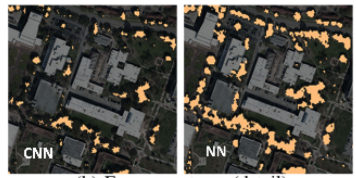
Fig. 8. Detail of crosswalk detection in VHR RGB data. The detected crosswalks are highlighted in red.

Fig. 9. Detail of pixel-wise car segmentation in VHR RGB data. The detected cars are highlighted in blue.

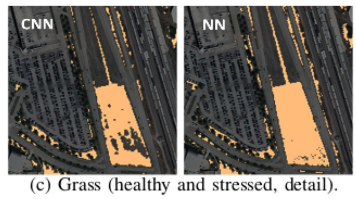
3) Final Classifier: The results of the base classifiers and *ad hoc* detectors need to be carefully combined to retain the merits of each individual method. Fig. 6 details the adopted end-to-end workflow of our final classifier, including the classification fusion step. Fig. 10 shows instead, from left to right: 1) the accuracy obtained in the training phase using different input datasets; 2) the results of the base classifiers (CNN and NN) and their combination on the 16-class problem detailed above; 3) how these are improved by *ad hoc* detectors and post-processing; and 4) subsets of classification results that help justifying our

#	Class	NN train acc. [%]			Test accuracies [%]				
		HS	LiDAR	All	CNN NN Adhoc	CNN NN Adhoc	CNN NN Adhoc	CNN NN Adhoc	CNN NN Adhoc
1	Healthy grass	94.9	93.6	97.6	99.2	99.4	99.7	**95.6**	**94.5**
2	Stressed grass				0.0	0.0	0.0	**84.6**	**88.7**
3	Artificial turf	100	99.5	100	21.3	94.6	94.6	20.8	95.7
4	Evergreen trees	96.8	96.7	99.0	95.6	97.9	98.1	94.8	96.5
5	Deciduous trees	93.8	95.0	99.1	59.7	82.9	83.0	61.7	81.6
6	Bare earth	99.5	77.1	99.6	5.4	44.0	45.4	**93.4**	**94.0**
7	Water	99.9	97.7	100	96.7	77.5	84.5	93.6	90.8
8	Residential buildings	76.7	87.1	94.4	0.0	0.0	0.0	**84.3**	**83.1**
9	Non-residential buildings				95.1	92.0	93.6	**91.0**	**90.6**
10	Roads	36.2	9.1	45.6	76.0	53.0	59.9	72.2	70.4
11	Sidewalks	86.7	66.5	88.4	69.5	75.4	66.0	63.2	60.3
12	Crosswalks	–	–	–	0.0	0.0	0.0	**30.6**	**30.6**
13	Major thoroughfares	40.3	15.2	53.2	33.3	41.9	30.7	35.9	35.7
14	Highways	91.6	70.5	98.2	30.3	25.9	30.3	**72.4**	**72.4**
15	Railways	99.1	71.6	99.7	85.4	7.2	77.6	93.2	93.2
16	Paved parking lots	95.5	68.0	97.0	58.4	69.4	72.3	53.8	65.6
17	Unpaved parking lots	100	95.9	100	0.0	0.0	0.0	0.0	0.0
18	Cars	–	–	–	0.0	0.0	0.0	**96.9**	**97.0**
19	Trains	93.3	96.0	99.3	89.6	94.1	89.6	89.1	93.4
20	Stadium Seats	99.4	81.9	99.8	85.1	50.9	86.9	**92.2**	**92.4**
OA [%]		74.3	71.6	86.5	51.2	54.5	58.6	74.3	80.7
Kappa		0.66	0.61	0.81	0.48	0.52	0.56	0.73	0.80
AA [%]		87.7	76.3	91.9	50.0	50.3	55.6	71.0	76.3

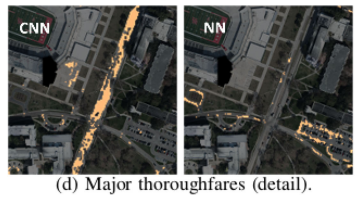
(a) Classification accuracies and final classifier components.



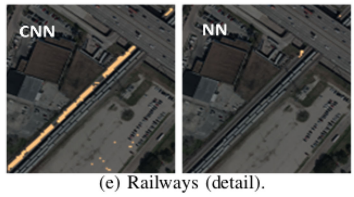
(b) Evergreen trees (detail).



(c) Grass (healthy and stressed, detail).



(d) Major thoroughfares (detail).



(e) Railways (detail).

Fig. 10. Summary of classification results. (a) Classification accuracies (producer's accuracies) and final classifier components. The training accuracies for sample NN base classifications are shown when using HS only (column 4), LiDAR only (column 5), and HS, LiDAR, and RGB altogether (column 6). The per-class test accuracy of the base classifier is reported for different combinations of NN and CNN (with and without the output of *ad hoc* detectors) in columns 7–10, with the dominant base classifier color-coded in the background and values reported in blue wherever *ad hoc* detectors or post-processing played a relevant role in recognizing or improving a specific class. The final results are reported in the last column. The overall accuracy, Cohen's Kappa, and average accuracy for all classifiers are reported in the last three lines. (b)–(e) Details of sample classification maps using the CNN (left) and NN (right) base classifiers for evergreen trees, grass (healthy and stressed), major thoroughfares, and railways. Such differences are mostly confirmed by the performances of NN and CNN on the undisclosed test samples.

choices for the classification fusion. Please refer also to [15] for additional details regarding our classification procedure.

Overall, the NN base classifier performs better for natural classes such as grass, trees, or artificial turf. These are classes for which pixel-wise information is usually enough—without taking into account more complex contexts—to achieve a satisfactory classification. Note nevertheless that the NN base classifier does consider spatial interactions to some degree through the extracted topic features, which can be useful to characterize stationary textures such as tree crowns for evergreen trees. Fig. 10(b) explicitly shows that NN outperforms CNN for this class. The same happens for grass (healthy and stressed), cf. Fig. 10(c). In contrast, CNN outperforms NN for man-made structures including buildings, roads, and trains. These are classes where context and shape information—at which deep convolutional networks excel—are crucial for classification. The superior performance of CNN is evident for major thoroughfares [see Fig. 10(d)] and railways [see Fig. 10(e)].

The relative advantages of NN and CNN were analyzed during the training phase, and have been at the basis of the classification fusion strategy shown in the top right of Fig. 6. In particular, our final classifier consisted of a sequential overlay of three components:

1) the full CNN classification map;
2) the NN classification map for selected classes (see Fig. 10 for selection);
3) the *ad hoc* detector maps for the corresponding classes.

The dominant classifiers for each class are identified in the table of Fig. 10(a) (columns 7 through 10). As the CNN output is used as a bottom layer for the final classification map, final results contain no unlabeled pixels.

4) Classification Refinements and Post-processing: In order to get our final classification results the following refinements were applied.

1) *Stadium seats*: A dedicated stadium seat detector based on the architecture of the NN base classifier but using a restricted set of input data was designed and trained to improve the prediction for this class.
2) *Healthy and stressed grass*: At first, the REIP was used as a discriminative feature since it has been shown to be more effective at detecting vegetation stress than broad-band indices such as NDVI [45]. Nevertheless, these first attempts failed, as the central wavelengths of the available bands differed significantly from the optimal spectral features needed to correctly compute the REIP, which employs narrow bands and is very sensitive to such variations.

Fig. 11. Overview of test scene and corresponding classification. Top: RGB mosaic of the whole University of Houston scene. Middle: Classes belonging to the *ad hoc* detectors and classifiers—bare earth (sienna), residential buildings (yellow), non-residential buildings (pink), crosswalks (cyan), and cars (red), overlaid on the image directly above. Bottom: Final classification results.

Therefore, in the end a simpler approach using NDVI has been preferred. The grass detected by the NN base classifier was separated into healthy and stressed components with an NDVI threshold of 0.535.

3) *Highways*: The confusion between highways and similar classes was reduced by extracting the three main highway directions with the help of the Hough transform. Samples formerly classified as roads or major thoroughfares close to the extracted highway directions were reclassified as highways.

4) *Morphological filtering*: Morphological openings and closings (with a disk of radius two as the structuring element) were applied three times to all classes except cars and crosswalks.

C. Discussion

Our final classification map is presented in Fig. 11 along with the original RGB scene and the outcome of our *ad hoc* detectors. As detailed in Table II and Fig. 10(a), our last submission

achieved an overall accuracy of 80.74%, a Cohen's Kappa of 0.80, and an average accuracy of 76.32%. Given the complexity of the scene and the detailed list of LULC classes, we consider these to be rather satisfactory results. The high average accuracy obtained (cf. Table II) is particularly noteworthy. As mentioned in Section V-B, our classification strategy was designed to learn the features of all the classes evenly, in an effort to maximize the average classification accuracy. This necessarily implied the overweighting of underrepresented classes (e.g., water) in the training set. Therefore, a better overall accuracy could have been obtained with the same classification scheme at the expense of an inferior average accuracy.

Before examining the strengths and pitfalls of our approach on a class-by-class basis, it is worth pointing out that we have only participated in the data fusion track of the contest. The importance of data fusion for our classification strategy is evident when considering the sample training accuracies for the NN base classifier in Fig. 10(a) (columns 4–6). The three columns show the training accuracies per class when using only HS data, only LiDAR data, and HS, LiDAR, and VHR RGB data. For all classes, the addition of data acquired from different sensors yields improvements ranging from mild (for natural classes such as grass or trees) to substantial (for man-made objects such as buildings or roads). The overall and average training accuracies increase significantly, as does Cohen's Kappa from 0.66 (HS only) or 0.61 (LiDAR only) to 0.81 (all). Our classification procedure thus clearly benefits from the availability of multimodal data for training (and eventually testing) and it would yield poorer results for single-source datasets. Although the relevance of data fusion is by no means surprising, it is important to explicitly show it for the classification of complex scenes as the one considered here.

The performance of CNN, NN, and their combination on the 16-class problem is reported in columns 7–9 of Fig. 10(a). Merging the CNN and NN base classifiers yields an improvement of 7.4% with respect to the use of CNN alone. If the four missing classes were ignored, the joint classifier (column 9) would yield an overall accuracy around 73%. Even though NN clearly outperforms CNN for trains, major thoroughfares, and sidewalks, the user's accuracy (not reported) is much lower in NN results with respect to CNN, as the false alarms increase at least by a factor of two. Therefore, we believe that adopting CNN as the classifier of choice was correct also in these cases.

The results of applying post-processing steps and overlaying the *ad hoc* detectors are reported in columns 10–11, for the cases of CNN alone and the combined use of CNN and NN, respectively. Also here, the overall accuracy improves considerably (6.4%) if the output of both classifiers is used. This confirms that the classification procedures of CNN and NN are complementary, and both contribute significantly to the final performance.

The test accuracies obtained for our final submission, reported in the last column in Fig. 10(a), show several interesting trends. First, the *ad hoc* detectors performed very well, with test accuracies above or very close to 90%, including cars (97.0%), bare earth (94.0%), non-residential buildings (90.6%), and residential buildings (83.1%). The exception is crosswalks with an accuracy of 30.6%. The main difficulty in recognizing this class correctly was the difference in shape, size, and color of the crosswalks across the scene—the set used for training the template matching algorithm could capture all these variations only partially. Second, the CNN and NN base classifiers excelled with accuracies over 90% for artificial turf (95.7%, NN), trains (93.4%, CNN), railways (93.2%, CNN), and water (90.8%, NN). The performance for artificial turf and water is remarkable given their reduced number of samples in the training set (cf. Table I). Moreover, the NN classification of evergreen and deciduous trees (96.5% and 81.6%, respectively) was effective without the need for an *ad hoc* detector. The refinements applied to the final classifier also proved effective as attested by the test accuracies for healthy grass (94.5%), stressed grass (88.7%), and stadium seats (92.4%).

However, the performance of our classification scheme shows some limitations. Apart from crosswalks (discussed above), the other cases with test accuracies below 80% are road-like classes (roads, sidewalks, major thoroughfares, and highways) and parking lots (paved and unpaved). The task of identifying and separating between roads (70.4%), sidewalks (60.3%), major thoroughfares (35.7%), and highways (72.4%) proved very difficult even for the CNN base classifier. Our results could certainly be improved with dedicated graph-based road segmentation algorithms. On the other hand, despite several attempts during the submission phase of the contest, our classifiers performed poorly for paved parking lots (65.6%) and completely missed unpaved parking lots (0.0%). We could not pinpoint the reason for this shortcoming in the test phase.

The presented approach shows the advantages of combining different strategies for the classification of complex scenes acquired by multimodal sensors. On the one hand, context-driven classes are better characterized by deeper NNs. On the other hand, for natural classes a shallower network yields more homogeneous results as the focus is shifted from an object to a single image element. Finally, classes demanding specific detectors have been analyzed separately, and for the case of cars a pre-trained deep network went a long way in improving detection results. The use of such different techniques introduces nevertheless additional problems—the parameters to be adjusted and the computational resources increase considerably, hindering an automatic or semi-automatic production of final classification results comparable to the ones presented here.

VI. Conclusion

In this paper, we summarized the organization and we presented the scientific results of the 2018 IEEE GRSS DFC, organized by the IEEE GRSS IADF Technical Committee. We described the multi-source data and the outcomes of the land-use/land-cover classification competition. We analyzed the algorithms used by the participants, with a focus on the two winning strategies.

Regarding the algorithms, given the variety of classes (20) and the amount of available data for training, convolutional and shallow NNs performed extremely well. They also prove to be handy for data fusion, even if particular care is required for the design of the architecture. This is a change with respect to previous DFC [5], [10] where limited labeled data led to the use

of other algorithms such as random forests or boosting. Indeed, it shows how our community can benefit from extended, labeled datasets and should pursue the development of such public resources.

It is also worth noting that for both winning entries, *ad hoc* classifiers and post-processing also made the difference, allowing a 15% increase of the overall accuracy. While decision fusion methods were already proposed in this paper, much work remains to be done for integration and fusion of expert knowledge into the NNs, especially to do it automatically. Moreover, such expertise usually makes sense for everyone and validates the decision. Further research to make CNN explainable will be profitable to help the public approval and diffusion of these methods. With respect to the data, fusion of multiple sources and even multi-spectral LiDAR alone prove to be especially informative since the best LULC classifications were obtained with such sensors (accuracies over 80% overall and 71% on average). Also, LiDAR information was processed using rasterized 2.5D only. This suggests promising paths for developing approaches able to process and classify real 3-D outputs of the sensors.

After the contest, the data has been made available again and will remain in open access for the benefit of the community. People interested can find all the relative information on the IEEE GRSS website.[2] After registering on the IEEE GRSS DASE server,[3] one can download the training data with the corresponding labels or the test data and then submit classification results to obtain the performance statistics, compare with other users and hopefully, improve the results presented in this paper. We do believe this dataset might have a great impact for fostering research in data fusion, but also for development of single-sensor processing, since it is the largest freely available HS dataset, with ten times more labeled data than the widely used Salinas or Pavia datasets [46], or the first available multispectral-LiDAR dataset.

Acknowledgment

B. Le Saux would like to thank R. Daudt for the help with building the ground-truth. S. Prasad would like to thank Dr. J. F. Diaz for preprocessing and preparing the data, as well as F. F. Shahraki and S. Mukherjee for their contributions in the preparation of the ground truth. D. Cerra, M. Pato, and E. Carmona would like to thank the remaining authors of the DLR IGARSS conference paper [15], without whom the results of the team would have not been possible.

References

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

[2] F. Pacifici, F. Del Frate, W. J. Emery, P. Gamba, and J. Chanussot, "Urban mapping using coarse SAR and optical data: Outcome of the 2007 GRSS Data Fusion Contest," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 331–335, Jul. 2008.

[2] http://www.grss-ieee.org/community/technical-committees/data-fusion, under the 'Past Contests' tab.

[3] http://dase.grss-ieee.org/

[3] G. Licciardi *et al.*, "Decision fusion for the classification of hyperspectral data: Outcome of the 2008 GRSS Data Fusion Contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3857–3865, Nov. 2009.

[4] C. Debes *et al.*, "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[5] W. Liao *et al.*, "Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 IEEE GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2984–2996, Jun. 2015.

[6] F. Pacifici and Q. Du, "Foreword to the special issue on optical multiangular data exploitation and outcome of the 2011 GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 3–7, Jan. 2012.

[7] L. Mou *et al.*, "Multi-temporal very high resolution from space: Outcome of the 2016 IEEE GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3435–3447, Aug. 2017.

[8] N. Longbotham *et al.*, "Multi-modal change detection, application to the detection of flooded areas: Outcome of the 2009–2010 Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 331–342, Jan. 2012.

[9] C. Berger *et al.*, "Multi-modal and multi-temporal data fusion: Outcome of the 2012 GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 3, pp. 1324–1340, Mar. 2013.

[10] N. Yokoya *et al.*, "Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1363–1377, May 2018.

[11] J. C. Fernandez-Diaz *et al.*, "Capability assessment and performance metrics for the Titan multispectral mapping LiDAR," *Remote Sens.*, vol. 8, no. 11, 2016, Art. no. 936. [Online]. Available: http://www.mdpi.com/2072-4292/8/11/936

[12] F. Dell'Acqua *et al.*, "The IEEE GRSS standardized remote sensing data website: A step towards science 2.0 in remote sensing," in *Proc. Living Planet Symp.*, Prague, Czech Republic, 2016, vol. SP-740.

[13] F. Dell'Acqua, G. C. Iannelli, J. Kerekes, G. Moser, L. Pierce, and E. Goldoni, "The IEEE GRSS data and algorithm standard evaluation (DASE) website: Incrementally building a standardized assessment for algorithm performance," in *Proc. Int. Geosci. Remote Sens. Symp.*, Jul. 2017, pp. 2601–2608.

[14] Y. Xu, B. Du, and L. Zhang, "Multi-source remote sensing data classification via fully convolutional networks and post-classification processing," in *Proc. Int. Geosci. Remote Sens. Symp.*, Valencia, Spain, 2018, pp. 3852–3855.

[15] D. Cerra *et al.*, "Combining deep and shallow neural networks with ad hoc detectors for the classification of complex multi-modal urban scenes," in *Proc. Int. Geosci. Remote Sens. Symp.*, Valencia, Spain, 2018, pp. 3856–3859.

[16] S. Fang, D. Quan, S. Wang, L. Zhang, and L. Zhou, "A two-branch network with semi-supervised learning for hyperspectral classification," in *Proc. Int. Geosci. Remote Sens. Symp.*, Valencia, Spain, 2018, pp. 3860–3863.

[17] S. Sukhanov, D. Budylskii, I. Tankoyeu, R. Heremans, and C. Debes, "Fusion of LiDAR, hyperspectral and RGB data for urban land use and land cover classification," in *Proc. Int. Geosci. Remote Sens. Symp.*, Valencia, Spain, 2018, pp. 3864–3867.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[19] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2018.

[20] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2019.

[21] X. Lu, Y. Chen, and X. Li, "Hierarchical recurrent neural hashing for image retrieval with hierarchical convolutional features," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 106–120, Jan. 2018.

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

[23] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[24] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[25] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral-spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5893–5909, Oct. 2018.

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

[27] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

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

[29] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

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

[31] N. Audebert, B. L. Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271617301818

[32] N. Audebert, B. L. Saux, and S. Lefèvre, "Joint learning from earth observation and OpenStreetMap data to get faster better semantic maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Honolulu, United States, 2017, pp. 1552–1560. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01523573

[33] Y. Xu, B. Du, F. Zhang, and L. Zhang, "Hyperspectral image classification via a random patches network," *ISPRS J. Photogramm. Remote Sens.*, vol. 142, pp. 344–357, 2018.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15, [Online]. Available: https://arxiv.org/pdf/1412.6980.pdf

[35] R. O. Duda and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Commun. ACM*, vol. 15, no. 1, pp. 11–15, 1972.

[36] H. Mhaskar, Q. Liao, and T. Poggio, "When and why are deep networks better than shallow ones?" in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2343–2349.

[37] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2654–2662. [Online]. Available: http://papers.nips.cc/paper/5484-do-deep-nets-really-need-to-be-deep.pdf

[38] H. Mhaskar and T. A. Poggio, "Deep vs. shallow networks : An approximation theory perspective," 2016, arXiv:1608.03287. [Online]. Available: http://arxiv.org/abs/1608.03287

[39] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surface of multilayer networks," 2014, arXiv:1412.0233. [Online]. Available: http://arxiv.org/abs/1412.0233

[40] R. Bahmanyar, D. Espinoza-Molina, and M. Datcu, "Multisensor earth observation image classification based on a multimodal latent Dirichlet allocation model," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 459–463, Mar. 2018.

[41] F. Chollet *et al.*, "Keras," 2015. [Online]. Available: https://github.com/fchollet/keras

[42] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–23. [Online]. Available: https://openreview.net/forum?id=ryQu7f-RZ

[43] J. Tian, S. Cui, and P. Reinartz, "Building change detection based on satellite stereo imagery and digital surface models," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 406–417, Jan. 2014.

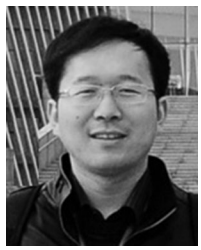
[44] S. M. Azimi, P. Fischer, M. Körner, and P. Reinartz, "Aerial LaneNet: Lane marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks," Mar. 2018, arXiv:1803.06904.

[45] J. U. Eitel *et al.*, "Broadband, red-edge information from satellites improves early stress detection in a New Mexico conifer woodland," *Remote Sens. Environ.*, vol. 115, no. 12, pp. 3640–3646, 2011.

[46] P. Ghamisi *et al.*, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.

Yonghao Xu (S'16) received the B.S. degree in photogrammetry and remote sensing in 2016 from Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing.

His research interests include hyperspectral image processing, computer vision, and machine learning.



Bo Du (M'10–SM'15) received the B.S. and Ph.D. degrees in photogrammetry and remote sensing from the State Key Lab of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Computer Science, Wuhan University. He has published more than 60 research papers in the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, etc. His research interests include pattern recognition, hyperspectral image processing, and signal processing.

Prof. Du is currently an Associate Editor for *Pattern Recognition* and *Neurocomputing*.

Liangpei Zhang (M'06–SM'08–F'19) received the B.S. degree in physics from Hunan Normal University, Changsha, China, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1982, 1988, and 1998, respectively.

He was the Head of the Remote Sensing Division, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. He has authored or coauthored more than 700 research papers and six books. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is currently an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Daniele Cerra (S'07—M'10) received the B.Sc. and M.Sc. degrees in computer engineering from Roma Tre University, Rome, Italy, in 2003 and 2005, respectively, and the Ph.D. degree in signal and image processing from Télécom ParisTech University, Paris, France, in 2010.

Since 2007, he has been with the Department of Photogrammetry and Image Analysis, German Aerospace Center (DLR), Wessling, Germany. From April to September 2018, he was a Visiting Professor of remote sensing with the University of Osnabrück, Osnabrück, Germany. His research interests include hyperspectral remote sensing, algorithmic information theory, and data compression.

Dr. Cerra was ranked first, second, and third in the IEEE GRSS Data Fusion Contests in 2019, 2018, and 2013, respectively.

Miguel Pato received the M.Sc. degree in physics engineering from the Technical University of Lisbon, Lisbon, Portugal, in 2007, and the Ph.D. degree in physics from the University of Padua, Italy, and Paris Diderot University, France, in 2011.

From 2011 to 2016, he was a Postdoctoral Researcher in the field of astroparticle physics with the University of Zurich, Zurich, Switzerland, Technical University of Munich, Munich, Germany, and Stockholm University, Stockholm, Sweden. Since 2017, he has been with the Department of Photogrammetry and Image Analysis, Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany. His research interests include hyperspectral image analysis and he is currently working on the development of the processing chain for the Environmental Mapping and Analysis Program (EnMAP).

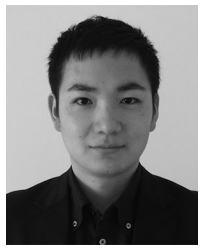
Emiliano Carmona received the M.Sc. degree in physics and the Ph.D. degree in astroparticle physics from the University of Valencia, Valencia, Spain, in 1997 and 2004, respectively.

From 2005 to 2013, he was a Postdoctoral Researcher with the Max-Planck-Institute for Physics, Munich, Germany, and a Researcher with CIEMAT—Centre for Energy, Environment and Technology, Madrid, Spain. Since 2014, he has been with the Department of Photogrammetry and Image Analysis, German Aerospace Center (DLR), Wessling, Germany. His research interests focus on hyperspectral remote sensing.

Saurabh Prasad (S'05–M'09–SM'14) received the B.S. degree from Jamia Millia Islamia, New Delhi, India, in 2003, the M.S. degree from Old Dominion University, Norfolk, VA, USA, in 2005, and the Ph.D. degree from Mississippi State University, Starkville, MS, USA, in 2008, all in electrical engineering.

He is currently an Assistant Professor with the Electrical and Computer Engineering Department, University of Houston, Houston, TX, USA, where he leads a research group on image processing and machine learning.

Dr. Prasad was the recipient of two Research Excellence Awards, in 2007 and 2008, during his Ph.D. studies at Mississippi State University, including the university-wide Outstanding Graduate Student Research Award. He was the recipient of the Best Student Paper Award at the IEEE International Geoscience and Remote Sensing Symposium 2008, held in Boston, MA, USA, the State Pride Faculty Award at Mississippi State University for his academic and research contributions in 2010, the NASA New Investigator (Early Career) Award 2014, and the Junior Faculty Research Award at the University of Houston in 2017. He is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.



Naoto Yokoya (S'10–M'13) received the M.Sc. and Ph.D. degrees in aerospace engineering from the University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively. He is a Unit Leader at the RIKEN Center for Advanced Intelligence Project, Tokyo, Japan, where he leads the Geoinformatics Unit. He is a Visiting Associate Professor with the Tokyo University of Agriculture and Technology, Tokyo, Japan. His research interests include image processing and data fusion for remote sensing and geoinformatics.

Ronny Hänsch (M'14) received the Ph.D. degree from Technische Universität Berlin, Berlin, Germany, in 2014.

He worked in the field of image-based classification with a focus on remote sensing data. He has organized multiple tutorials regarding machine learning and computer vision at international conferences. His recent research interests focus on the development of methods for 3-D reconstruction as well as ensemble methods. His research interests include computer vision, remote sensing, neural networks, and ensemble theory.

Dr. Hänsch is the Co-Chair of the IEEE Geoscience and Remote Sensing Symposium Image Analysis and Data Fusion Technical Committee and of the International Society for Photogrammetry and Remote Sensing WG II/1 Image Orientation.

Bertrand Le Saux (M'17) received the M.Eng. and M.Sc. degrees from INP Grenoble in 1999, and the Ph.D. degree from Versailles University/Inria Rocquencourt in 2003.

He is a Research Scientist with the Information Processing and Systems Department at ONERA, the French Aerospace Laboratory. He also has been a Research Fellow with CNR Pisa (2004), University of Bern (2005) and ENS Cachan (2006–2007). He teaches machine learning and pattern recognition at Institut d'Optique Graduate School and ENSTA ParisTech. His research objective is visual understanding by means of data-driven techniques. His current research interests include the development of machine learning and deep learning methods for remote sensing, (flying) robotics, and 3-D vision.

Dr. Le Saux has been the Chair of the Image Analysis and Data Fusion Technical Committee of the IEEE GRSS, since 2017, and was previously the Co-Chair from 2015 to 2017.