Global Land-Cover Mapping With Weak Supervision: Outcome of the 2020 IEEE GRSS Data Fusion Contest

Caleb Robinson, Kolya Malkin, Nebojsa Jojic, Huijun Chen, Rongjun Qin, *Senior Member, IEEE*, Changlin Xiao, Michael Schmitt, *Senior Member, IEEE*, Pedram Ghamisi, *Senior Member, IEEE*, Ronny Hänsch, *Senior Member, IEEE*, and Naoto Yokoya, *Member, IEEE*

***Abstract*—This article presents the scientific outcomes of the 2020 Data Fusion Contest (DFC2020) organized by the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society. The 2020 Contest addressed the problem of automatic global land-cover mapping with weak supervision, i.e., estimating high-resolution semantic maps while only low-resolution reference data are available during training. Two separate competitions were organized to assess two different scenarios: 1) high-resolution labels are not available at all; and 2) a small amount of high-resolution labels are available additionally to low-resolution reference data. In this article, we describe the DFC2020 dataset that remains available for further evaluation of corresponding approaches and report the results of the best-performing methods during the contest.**

Manuscript received October 9, 2020; revised December 29, 2020; accepted February 20, 2021. Date of publication March 4, 2021; date of current version March 25, 2021. The work of Huijun Chen and Rongjun Qin was supported in part by the Office of Naval Research under Award N000141712928. *(Corresponding author: Naoto Yokoya.)*

Caleb Robinson was with the School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. He is now with the AI for Good Research Lab, Microsoft Research, Redmond, WA 98052 USA (e-mail: drobinson67@gatech.edu).

Kolya Malkin is with the Department of Mathematics, Yale University, New Haven, CT 06520 USA (e-mail: kolya_malkin@hotmail.com).

Nebojsa Jojic is with the Microsoft Research, Redmond, WA 98052 USA (e-mail: jojic@microsoft.com).

Huijun Chen is with the Department of Civil, Environmental and Geodetic Engineering and the Environmental Science Graduate Program, The Ohio State University, Columbus, OH 43210 USA (e-mail: chen.9317@osu.edu).

Rongjun Qin is with the Department of Civil, Environmental and Geodetic Engineering, the Department of Electrical, and Computer Engineering, and the Translational Data Analytics Institute, The Ohio State University, Columbus, OH 43210 USA (e-mail: qin.324@osu.edu).

Changlin Xiao is with the Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: changlinshaw@gmail.com).

Michael Schmitt is with the Department of Geoinformatics, Munich University of Applied Sciences, 80335 München, Germany (e-mail: michael.schmitt@hm.edu).

Pedram Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf, Machine Learning Group, Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany, and also with the Institute of Advanced Research in Artificial Intelligence, 1030 Vienna, Austria (e-mail: p.ghamisi@gmail.com).

Ronny Hänsch is with the German Aerospace Center, 82234 Weßling, Germany (e-mail: rww.haensch@gmail.com).

Naoto Yokoya is with the Department of Complexity of Science and Engineering, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan, and also with the RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan (e-mail: yokoya@k.u-tokyo.ac.jp).

Digital Object Identifier 10.1109/JSTARS.2021.3063849

***Index Terms*—Convolutional neural networks (CNNs), deep learning, image analysis and data fusion, land-cover mapping, multimodal, random forests (RFs), weak supervision.**

I. Introduction

HIGH-RESOLUTION global land-cover maps and their automatic updating allow us to understand the state and changes of the Earth's surface, yielding fundamental information for tackling global challenges such as climate change, natural disasters, and environmental conservation. Open satellite data, such as the ones provided by the Sentinel and Landsat missions, as well as small satellite constellations, have made it possible to obtain large-scale multimodal Earth observation data at high spatial and temporal resolutions covering the entire globe. Although machine and deep learning methods are effective for large-scale automated mapping, the high cost of labeled training data collection is a barrier to high-resolution high-accuracy global mapping.

Weakly supervised learning gained great attention both in theory and practice to reduce label data collection costs. In the field of remote sensing, low-resolution global maps are regularly updated and openly available though their accuracy have limitations. The task of achieving high-resolution and accurate land-cover classification from such low-resolution and noisy labels is a fundamental challenge, which can potentially lead to a paradigm shift in global mapping and facilitate the use of Earth observation data for the sustainable development goals [1].

A tremendous increase in the availability of remotely sensed data captured by different sensors, combined with their considerable heterogeneity (e.g., data types and resolutions), leads to a dramatic challenge for effective and efficient processing of such data [2]. On the other hand, the aforementioned increase in the volume of multimodal and multisensor data along with their ancillary products opens the possibility of utilizing multimodal datasets in a joint manner to further improve the performance of the processing approaches with respect to the applications at hand [3]. In this context, optical and synthetic aperture radar (SAR) data provide complementary information about the ground surface, and their synergistic use is an effective approach in terms of improving the frequency of observations as well as allowing for more accurate land-cover classification.

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 Earth observation, geospatial data fusion, and algorithms for image analysis. It aims at connecting people and resources, educating students and professionals, and promoting theoretical advances and best practices in image analysis and data fusion. Since 2006, the IADF TC has been organizing an annual challenge named the Data Fusion Contest (DFC) for fostering ideas and progress in remote sensing, distributing novel data, and benchmarking analysis methods [4]–[17]. The 2020 DFC (DFC2020) aimed at promoting research in automatic large-scale land-cover mapping from globally available multimodal satellite data with weak supervision. The contest serves as a benchmark to evaluate the best approaches for a fundamental task involving weakly supervised learning toward an increased generalization ability over the entire globe, which is a major open challenge in a wide range of fields, from Earth observation to computer vision and machine learning.

For 2020 Contest, the SEN12MS dataset [18] was employed for training land-cover classification models, which includes triplets of corresponding Sentinel-1 SAR data, Sentinel-2 multispectral imagery, and Moderate Resolution Imaging Spectroradiometer (MODIS)-derived low-resolution land-cover maps [19] sampled across the entire globe. While all data are provided at a ground sampling distance (GSD) of 10 m, the Sentinel images have a native resolution of about 10–20 m per pixel, whereas the MODIS-derived land cover has a native resolution of 500 m per pixel. For the contest, we use a simplified version of the International Geosphere-Biosphere Program (IGBP) classification scheme [20], which is a well-established land-cover scheme that has been used internationally for more than 20 years. Although it consists of generic globally applicable classes, we aggregated some of the classes characterized by just subtle distinctions (e.g., different types of forests) to create a simplified version of the IGBP scheme for slightly improved class balance and better accessibility for nongeography experts. For the validation and test phases of the contest, semimanually derived high-resolution land-cover maps of scenes that are not included in the SEN12MS dataset were produced and provided to the contest participants. In order to prevent contestants from hand-labeling these validation and test data, they were provided without geolocation information.

The DFC2020 consisted of two challenge tracks organized sequentially to promote innovation in two practical scenarios. In Track 1, semimanually derived high-resolution land-cover maps for the validation set were kept undisclosed. The objective was to predict land-cover labels at 10-m GSD using only MODIS-derived low-resolution and noisy labels for training. In Track 2, we disclosed high-resolution labels for the validation set, and the goal was to train models for land-cover mapping using both low-resolution noisy labels and a limited number of high-resolution clean labels. For both tracks, performance was assessed using the average accuracy of all classes. Average accuracy is the mean value of class accuracies (i.e., producer's accuracy) for all the classes. Participants submitted their prediction maps to the Codalab competition website,[1] where they could get instant evaluation and rank in the competition.

In this article, we describe the datasets used in DFC2020 in Section II and discuss the overall results of the competition in Section III. Then, we will focus in more detail on the approaches proposed by the first-ranked teams in both tracks: land-cover classification with low-resolution labels in Section IV and land-cover classification with low- and high-resolution labels in Section V. Finally, Section VI concludes this article.

II. Data and Baseline of the DFC2020

The data backbone of the DFC2020 is the SEN12MS dataset [18], which was provided for the training of weakly supervised machine learning models in both contest tracks. SEN12MS is one of the largest currently available remote sensing datasets and consists of 180 662 globally sampled patch triplets, where each patch is a multidimensional image tensor with a spatial extent of 256×256 pixels and a variable number of channel dimensions, depending on the three data modalities represented by each triplet.

1) The first patch of each triplet represents SAR data acquired by Sentinel-1 and contains two channels corresponding to the two available polarizations.
2) The second patch of each triplet represents a multispectral image tensor acquired by Sentinel-2. It contains 13 spectral bands.
3) The third patch of each triplet represents a tensor containing four different land-cover representations.

More details about the data and the preprocessing are described in Section II-A, while more information about the distribution of the classes in the dataset can be found in [21].

For the contest, we created an additional dataset consisting of 6114 patches collected from seven globally distributed cities (see Fig. 1). This DFC2020 dataset is basically sharing its attributes with the SEN12MS dataset, but additionally contains semiautomatically created land-cover annotations with a resolution of 10 m per pixel for use as reference during validation (Track 2) and testing (both Tracks 1 and 2). More information about the DFC2020 reference data is provided in Section II-D.

A. Sentinel-1 and Sentinel-2 Satellite Data

The Sentinel-1 mission [22] currently consists of two similar satellites, both equipped with C-band SAR sensors. Depending on which SAR imaging mode is used, resolutions down to 5 m with a wide coverage of up to 400 km can be achieved. Furthermore, Sentinel-1 provides dual polarization capabilities and very short revisit times of about six days at the equator.

For the *SEN12MS* dataset, Sentinel-1 images acquired in the most frequently available interferometric wide swath mode were used. They were downloaded in the form of ground-range-detected products and converted to σ^0 backscatter in decibel scale. While the resolution of such data originally is about 5 m in azimuth and 20 m in range, the images in the dataset were

[1] [Online]. Available: https://competitions.codalab.org/competitions/22289

Fig. 1. Seven regions of interest from which the DFC2020 data are sampled.

resampled to a square pixel spacing of 10 × 10 m. In order to exploit the full potential of Sentinel-1 data, *SEN12MS* contains both VV and VH polarized images.

B. Sentinel-2

The Sentinel-2 mission [23] currently also comprises two similar satellites in the same orbit, phased at 180° to each other. One of the mission's goals is to ensure continuity for multispectral imagery of the SPOT and LANDSAT kind, which have provided information about the land surfaces of our Earth for many decades. The *SEN12MS* dataset contains the full multispectral image tensors representing 13 spectral bands: ten surface-related bands (bands 2–4 and 8 at a resolution of 10 m; bands 5–7, 8A, 11, and 12 at a resolution of 20 m) and three atmosphere-related bands (bands 1, 9, and 10 at a resolution of 60 m). The images are extracted from the original precisely georeferenced Sentinel-2 granules after visually checking for the complete absence of cloud cover in the scene.

C. MODIS-Derived Land Cover Labels

The MODIS is the main instrument on board of the Terra and Aqua satellites. Based on calibrated MODIS reflectance data, annually updated global land-cover maps for the years 2001–2016 are provided as the MCD12Q1 V6 dataset at a GSD of 500 m [24].

SEN12MS contains four MODIS land-cover products for every patch. The data were created from 2016 data and upsampled to a pixel spacing of 10 m. The first of the provided products represents land cover following the IGBP classification scheme [20], while the remaining products contain the LCCS land-cover layer, the LCCS land-use layer, and the LCCS surface hydrology layer [25]. According to [24], the overall accuracies of the layers are about 67% (IGBP), 74% (LCCS land cover), 81% (LCCS land use), and 87% (LCCS surface hydrology), respectively. Together with the comparably low resolution of 500 m, this makes for a perfect example of weak supervision, given satellite data with a resolution in the 10-m domain.

D. High-Resolution Land-Cover Reference Labels

As described in [26], for the DFC2020, the IGBP classification scheme was aggregated to ten less fine-grained classes. This *simplified IGBP scheme* is similar to the classification scheme adopted by the authors of the FROM-GLC10 dataset [27]. Its classes are compared to the standard IGBP classes in Table I, while the distribution of classes is shown in Table II. The semiautomatic process for the generation of the high-resolution land-cover annotations as well as their validation is shortly described in the following.

1) Generation of the High-Resolution Land-Cover Annotations: For the generation of the high-resolution land-cover annotations, a semiautomatic shallow learning-based iterative approach was combined with a data fusion strategy. The procedure was carried out using the Google Earth Engine (GEE) [28] environment. For every scene, this procedure was implemented as follows.

1) Using the Google Earth aerial imagery basemap for visual comparison, several dozen samples for every class were selected manually.
2) Using those samples, a Random Forest (RF) classifier was trained. The input to the classifier was comprised of the following data sources:
   1) the VV and VH polarization channels of Sentinel-1;
   2) the ten surface-related bands of Sentinel-2. It was ensured by visual inspection that the data do not contain any clouds;
   3) the spectral indices NDVI, MNDWI, and BSI calculated from the relevant Sentinel-2 bands;

TABLE I
SIMPLIFIED IGBP LAND COVER CLASSIFICATION SCHEME

IGBP Class Number	IGBP Class Name	Aggregated Class Number	Simplified Class Name	Color
1 2 3 4 5	Evergreen Needleleaf Forest Evergreen Broadleaf Forest Deciduous Needleleaf Forest Deciduous Broadleaf Forest Mixed Forest	1	Forest	#009900
6 7	Closed Shrublands Open Shrublands	2	Shrubland	#c6b044
8 9	Woody Savannas Savanna	3	Savanna	#fbff13
10	Grasslands	4	Grassland	#b6ff05
11	Permanent Wetlands	5	Wetlands	#27ff87
12 14	Croplands Cropland / Natural Vegetation Mosaics	6	Croplands	#c24f44
13	Urban and Built-up Lands	7	Urban/Built-up	#a5a5a5
15	Permanent Snow and Ice	8	Snow/Ice	#69fff8
16	Barren	9	Barren	#f9ffa4
17	Water Bodies	10	Water	#1c0dff

TABLE II
REGIONS OF INTEREST AND CLASS DISTRIBUTION OF THE DFC2020 HIGH-RESOLUTION LAND-COVER REFERENCE DATA

ROI	Hemi-sphere	Season	Approx. Size (in km^2)	No. of Patches	Class distribution (in %)										ρ
Bandar Azali, Iran	North	Fall	70 × 88	676	8.9	0.0	0.0	0.3	5.4	34.1	3.4	0.0	0.1	44.6	0.71
Black Forest, Germany	North	Spring	102 × 150	1444	46.4	1.3	0.0	16.8	0.1	23.4	8.7	0.0	0.4	0.8	0.69
Cape Town, South Africa	South	Fall	102 × 120	1440	0.8	7.8	0.0	2.4	0.5	17.4	6.4	0.0	5.3	53.1	0.57
Khabarovsk, Russia	North	Summer	60 × 90	486	18.1	3.1	0.0	21.9	34.6	7.3	4.0	0.0	0.3	9.7	0.37
Kippa Ring, Australia	South	Winter	70 × 78	684	46.9	1.0	0.0	29.4	1.6	2.8	7.1	0.0	0.7	10.1	0.66
Mexico City, Mexico	North	Winter	60 × 63	484	24.4	13.7	0.0	8.5	1.5	11.3	35.9	0.0	3.8	0.7	0.40
Mumbai, India	North	Spring	80 × 84	900	13.2	14.4	0.0	0.2	4.5	20.2	9.7	0.0	5.5	32.3	0.75
TOTAL				6114	23.1	6.0	0.0	11.6	7.0	17.0	11.0	0.0	2.3	22.0	
				SEN12MS	11.3	6.9	23.6	16.8	1.1	17.9	10.6	0.0	5.2	6.5	

The color indicates the respective class. For color scheme, see Table I or Fig. 2. The last column (ρ) indicates the correlation between the respective scene and the full DFC2020 dataset. Note that the seasons are stated according to the respective hemisphere.

4) the MODIS-derived low-resolution simplified IGBP land-cover map;
5) the FROM-GLC10 high-resolution land-cover map; and
6) the spatial coordinates (X, Y) of each pixel.

The idea behind this feature selection was to provide as much information as possible to train a classifier, which adapts as good as possible to the current region of interest. MODIS-derived labels and FROM-GLC10 labels were supposed to provide guidance as a form of weak prior knowledge. On the one hand, the spatial coordinates regularize the RF so that it can distinguish between spatially disjunctive representations of the same class. On the other hand, they provide a spatial prior to enforce exploiting spatial correlations within the data.

3) The classifier was then applied to the current ROI to produce a high-resolution land-cover map.
4) The resulting land-cover map was then visually inspected and compared against all relevant data sources, in particular against the Google Earth aerial imagery basemap as a form of external information.
5) Steps 1–4 were repeated until convergence, i.e., until the RF-predicted land-cover map did not improve anymore despite additionally selected training samples.

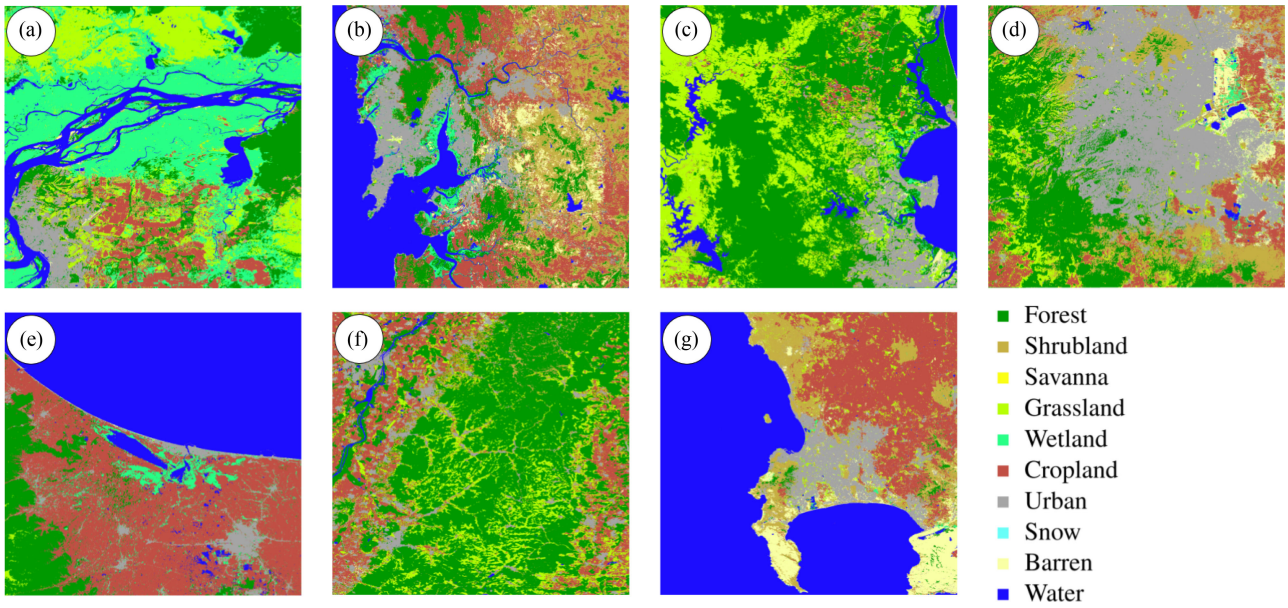

Fig. 2. Seven ROIs of the DFC2020 dataset: (A) Khabarovsk, Russia, (B) Mumbai, India, (C) Kippa Ring, Australia, (D) Mexico City, Mexico, (E) Bandar Anzali, Iran, (F) Black Forest, Germany, and (G) Cape Town, South Africa. Note that these images have been cropped to squares for visualization purposes; see Table II for the dimensions of each scene.

After this procedure was finished for an ROI, which usually took several dozen iterations and the selection of several hundreds of training samples, the Sentinel-1/-2 imagery, the MODIS-derived low-resolution land-cover maps, and the high-resolution land-cover annotations were exported from GEE and further processed similar to SEN12MS. This includes, in particular, the reprojection from the global WGS84 into regional UTM coordinate systems to obtain metric pixels as well as splitting the full scene images (cf. Fig. 2) into nonoverlapping patches of 256×256 pixels.

2) Statistics and Validation: The class distributions of the DFC2020 high-resolution land-cover reference data are compiled in Table II. It can be seen that there is a satisfying agreement between the SEN12MS dataset and the DFC2020 dataset, although there is a significantly larger share of *Forest*, *Wetlands*, and *Water* samples in DFC2020. On the other hand, the DFC2020 set does not contain any pixels of the *Savanna* class. This issue was already discussed in [21].

In order to provide an intuition about the reliability of the high-resolution reference data, an independent validation in Google Earth was carried out: More than 500 samples were randomly distributed over the seven ROIs and then visually inspected and compared to high-resolution aerial imagery. The agreement between the class in the reference data and the class choice of the visual inspector was recorded. The corresponding accuracies are summarized in Table III, and the confusion matrix is shown in Fig. 3. Over all ROIs, the average precision was 76.3%, the average recall 75.8%, and the overall accuracy 82.4%.

While it has to be noted that also the visual validation in Google Earth is error-prone (i.e., there is no validation against actual *ground truth*), the statistics reveal that the DFC2020 land-cover annotations can be considered as a satisfying reference. In particular, important classes such as *Water*, *Forest*, *Urban*, *Shrubland*, and *Grassland* are very accurate, with classwise accuracy (precision) $\gg 70\%$. More difficult (and less frequent) classes such as *Wetlands*, *Barren*, and *Cropland* are less accurate, which is mainly caused by confusions between *Shrubland* with *Cropland*, *Wetlands*, or *Grassland*; *Grassland* with *Cropland* or *Wetlands*; and *Barren* with *Urban* or *Cropland*.

TABLE III
OVERALL ACCURACY (BASED ON VISUAL INSPECTION) OF THE HIGH-RESOLUTION LAND-COVER LABELS OF THE DFC2020 DATASET

City	Overall Accuracy
Bandar Anzali, Iran	80.0%
Black Forest, Germany	89.4%
Cape Town, South Africa	81.7%
Khabarovsk, Russia	76.4%
Kippa-Ring, Australia	88.1%
Mexico City, Mexico	83.8%
Mumbai, India	73.4%

To provide another baseline, in [27], the FROM-GLC10 dataset was shown to have an overall accuracy of about 72%, with peak accuracies in the Beijing region in the 70–80% range [29].

E. Baseline Solutions

In [21], we have summarized a couple of baseline results to provide the participants of DFC2020 with an idea about the quality of their solutions. Those baselines were the following:

1) A comparison of MODIS-derived low-resolution labels against the high-resolution reference labels prepared for DFC2020. This is supposed to provide an estimate for the quality of globally available land-cover data, which can be used for weak supervision;

TABLE IV
CLASSWISE AND AVERAGE ACCURACIES ACHIEVED ON THE *DFC2020 VALIDATION* DATASET FOR DIFFERENT BENCHMARKS

Class	LR-HR	DLv3 S2 only	DLv3 S1+S2	Unet S2 only	Unet S1+S2	k-means S2 only	k-means S1+S2	RF S2 only	RF S1+S2
Forest	51.6%	71.4%	61.2%	67.3%	55.4%	2.4%	1.7%	77.1%	76.9%
Shrubland	7.7%	2.3%	3.8%	0.0%	3.7%	7.7%	5.9%	0.0%	0.0%
Savanna	–	–	–	–	–	–	–	–	–
Grassland	6.7%	64.4%	48.2%	76.7%	77.2%	11.2%	12.5%	90.3%	90.5%
Wetlands	0.6%	2.4%	3.8%	3.7%	3.2%	2.2%	0.3%	4.1%	4.0%
Croplands	64.4%	53.3%	61.9%	65.7%	50.7%	42.1%	13.4%	42.1%	39.6%
Urban	71.5%	71.0%	62.8%	80.9%	73.1%	0.0%	0.0%	0.0%	0.0%
Snow/Ice	–	–	–	–	–	–	–	–	–
Barren	0.3%	0.2%	1.0%	0.6%	0.8%	54.4%	6.2%	0.0%	0.0%
Water	95.1%	88.9%	95.8%	89.4%	92.7%	55.8%	68.9%	25.4%	34.5%
Average	**37.2%**	**44.2%**	**42.3%**	**48.1%**	**44.6%**	**22.0%**	**13.6%**	**29.9%**	**30.7%**

S2 only indicates that only Sentinel-2 data have been used for the prediction, whereas *S1+S2* indicates the case of Sentinel-1/Sentinel-2 data fusion. *LR-HR* indicates the baseline check of evaluating the MODIS-derived low-resolution labels against the high-resolution DFC2020 reference labels.

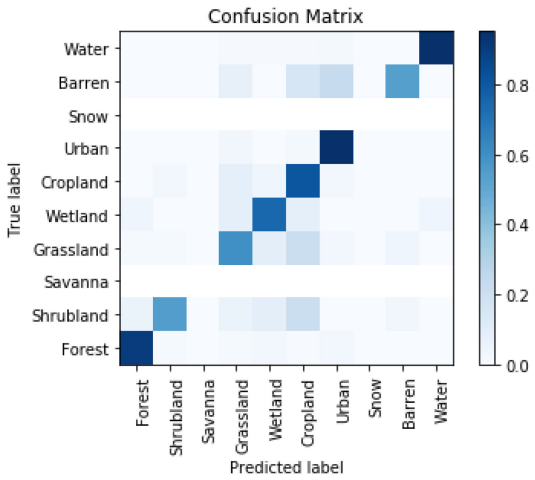

Fig. 3. Label validation against manually labeled check points (by visual inspection in Google Earth).

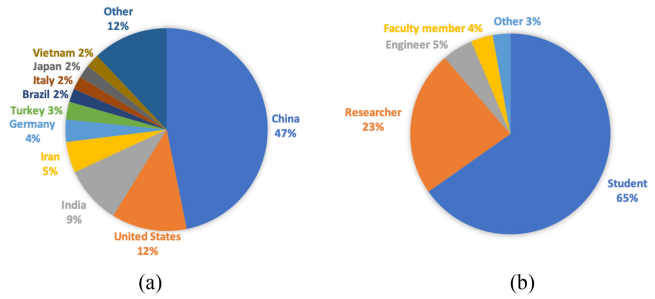

Fig. 4. (a) Geographic distribution and (b) position of the participants registered at IEEE DataPort for DFC2020 [31].

2) Two deep-learning-based semantic segmentation models, based on the DeepLabv3 and the Unet architectures to provide an idea about the capabilities of off-the-shelf convolutional neural network (CNN) approaches. Those models were trained and tested on either only Sentinel-2 input data or on Sentinel-1 and Sentinel-2 in a data fusion configuration;
3) An unsupervised shallow learning-based approach: k-means, also both with Sentinel-2 only or Sentinel-1 plus Sentinel-2. To make this unsupervised approach comparable to the supervised approaches, the number of clusters was set to $k = 8$ (i.e., according to the number of simplified IGBP classes encountered in the subsampled training data). The cluster segments were learned completely unsupervised, while the reordering of cluster labels was achieved with the Kuhn–Munkres algorithm [30]. For this purpose, the low-resolution MODIS-derived labels of the subsampled train split served as reference;
4) A supervised shallow learning-based approach, namely, RF, also both with Sentinel-2 only or Sentinel-1 plus Sentinel-2.

The accuracies achieved with those baselines are summarized in Table IV.

III. ORGANIZATION, SUBMISSIONS, AND RESULTS

There were 141 unique registrations at the IEEE DataPort website[2] for downloading the DFC2020 data from 22 countries. Fig. 4 shows the distribution of countries and affiliations. Forty-seven percent of the registrations were from China as similar to the previous editions, and students were the majority, indicating that DFC2020 was widely used for educational purposes. One hundred and fifty-nine teams registered at the Codalab competition websites during the development phase and 33 teams entered the test phase after screening the descriptions of their approaches submitted by the end of the development phase.

We received nearly 3k submissions during the development phase illustrating the active participation across all registered teams. After the initial development phase, the maximum number of submissions per team was limited to ten. Nevertheless, we received approximately 250 submissions for each track. The similar number of submissions in each track illustrates that both scenarios, i.e., having no or only a small amount of high-resolution labels, are of similar interest to the research community.

[2][Online]. Available: https://ieee-dataport.org/competitions/2020-ieee-grss-data-fusion-contest

TABLE V
TOP-RANKED TEAMS AND APPROACHES

Track	Rank	Team	Average Accuracy (%)	Affiliation	Approach						
					Preprocess.	Clustering	CNN	RF	Ensemble	Postprocess.	SEN12MS
1	1	calebrob6	57.49	Georgia Institute of Technology		✓	✓		✓	✓	
	2	WHU_YuXia	56.96	Wuhan University		✓	✓	✓		✓	
	3	Pineapples	56.88	German Aerospace Center	✓	✓		✓		✓	✓
	4	Antonia	56.76	The Ohio State University	✓			✓	✓	✓	
	5	BurningAllthing	56.14	Wuhan University	✓	✓	✓	✓	✓	✓	
	–	Baseline	47.74	Technical University of Munich			✓				✓
2	1	Antonia	61.42	The Ohio State University	✓			✓	✓	✓	
	2	Pineapples	61.36	German Aerospace Center	✓			✓		✓	
	3	dfchen	60.95	Xidian University		✓	✓	✓		✓	✓
	4	BurningAllthing	60.54	Wuhan University	✓	✓	✓	✓	✓	✓	
	5	mushroom1	60.50	Wuhan University			✓				✓

As a baseline, we provide the result of a CNN, which performed best on the validation set (cf. Table IV), i.e., a UNet trained the Sentinel-2 data of the SEN12MS training split without further pre- or postprocessing.

The first to fourth ranked teams in Track 1 and the first to third ranked teams in Track 2 were awarded as winners of the DFC2020 and presented their solutions during the 2020 IEEE International Geoscience and Remote Sensing Symposium. The seven winning teams are the following.

1) *First place of Track 1: calebrob6* team; Caleb Robinson, Kolya Malkin, Lucas Hu, Bistra Dilkina, and Nebojsa Jojic from the Georgia Institute of Technology, Yale University, the University of Southern California, and Microsoft Research, USA; a combination of iterative clustering and epitomic representations, followed by deep image prior postprocessing [32].
2) *Second place of Track 1: WHU_YuXia* team; Yu Xia, Yue Liao, Hongyan Zhang, and Guangyi Yang from Wuhan University, China; multibranch fusion of unsupervised multiresolution segmentation, RF classification of remote sensing indexes, and CNN predictions with postprocessing based on expert priors [33].
3) *Third place of Track 1: Pineapples* team; Daniele Cerra, Nina Merkle, Corentin Henry, Kevin Alonso, Pablo d'Angelo, Stefan Auer, Reza Bahmanyar, Xiangtian Yuan, Ksenia Bittner, Maximilian Langheinrich, Guichen Zhang, Miguel Pato, Jiaojiao Tian, and Peter Reinartz from the German Aerospace Center, Germany; automated label preprocessing, a Gaussian Naive Bayes classifier trained on cluster centroids, and classes obtained by k-means clustering and RFs with bag-of-words features, followed by classification refinement designed for specific classes [34].
4) *Fourth place of Track 1: Antonia* team; Huijun Chen, Changlin Xiao, Wei Liu, and Rongjun Qin from The Ohio State University, USA; automated label preprocessing, RFs, followed by classification refinement based on prior knowledge on class confusion [35].
5) *First place of Track 2: Antonia* team; Huijun Chen, Changlin Xiao, Wei Liu, and Rongjun Qin from The Ohio State University, USA; an ensemble of RFs trained on refined labels [36].
6) *Second place of Track 2: Pineapples* team; Daniele Cerra, Nina Merkle, Corentin Henry, Kevin Alonso, Pablo d'Angelo, Stefan Auer, Reza Bahmanyar, Xiangtian Yuan, Ksenia Bittner, Maximilian Langheinrich, Guichen Zhang, Miguel Pato, Jiaojiao Tian, and Peter Reinartz from the German Aerospace Center, Germany; as Track 1 third, but RFs trained on high-resolution labels and no use of topic vectors and bag-of-words features [37].
7) *Third place of Track 2: dfchen* team; Shuting Yin, Dafan Chen, Chengconghui Ma, and Yanchao Lian from Xidian University, China; a combination of RFs, k-means, and DeepLabv3++ with postprocessing and retraining [38].

Table V summarizes the teams ranked in the top five of both tracks and their approaches. The overall trend was that RF as a shallow supervised classification approach was used frequently by the winning teams. In more detail, eight approaches among the top five approaches of both tracks (ten approaches in total) investigated RF as a part of their classification framework. This shows that ensemble learning methods for classification are still found effective for large-scale land-cover classification. CNN (here as a deep supervised approach) was investigated in six approaches out of the top five approaches of both tracks.

Preprocessing and postprocessing were regularly utilized in the suggested frameworks mostly to refine weak labels and improve the quality of the classification maps, respectively.

IV. FIRST PLACE TEAM OF TRACK 1

The algorithm of the first place team in Track 1 [32] combined three approaches: 1) neighborhood-informed color clustering; 2) label super-resolution with epitomic representations; and 3) deep image prior postprocessing. We describe the three approaches in order.

A. Neighborhood-Informed Color Clustering

The first approach can be described as latent variable model of an image and label set, the inference in which involves clustering pixel intensities and assigning the clusters to the target classes. Precisely, at each pixel coordinate i—encoding both a sample's index in the image set and the coordinate within the image—we aim to infer the target class, ℓ_i. We introduce a latent cluster variable s_i placed at each coordinate, ranging from 1 to 32. This cluster variable generates the pixel intensities x_i from a learned diagonal-covariance Gaussian, i.e., $s_i \rightarrow x_i$ is

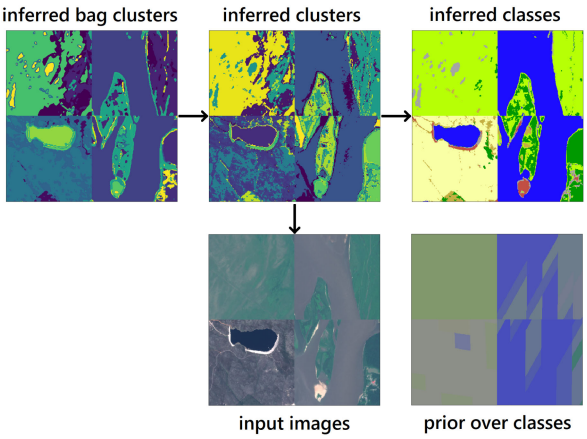

Fig. 5. Inferred color clusters, bag clusters, and labels in the neighborhood-informed clustering algorithm.

a Gaussian mixture model with 32 components. There is also a learned distribution $p(s|\ell)$, the probability of a pixel with a given label belonging to each color cluster. The inference through such a model consists in optimizing the Gaussians $p(x|s)$ and the categorical distributions $p(s|\ell)$ so as to maximize the likelihood of the image data; the final predictions are the marginal posterior distributions over the labels ℓ_i. To ground the inference of this model, we fix a prior $\mathbf{p}_i(\ell)$, which sets a weak belief about the target class of each pixel.[3] This prior is derived from the given low-resolution labels, as we explain below. Note that this algorithm does not separate training and testing sets: it reasons over the test images themselves.

To make the model sensitive to textures, we also introduce a bag-of-clusters variable b_i, also in the range $\{1, 2, \ldots, 32\}$, at each image coordinate. It is the mixture index in a categorical mixture model over the clusters s_i found in a 5×5 window around the coordinate i, i.e., it generates the 25 cluster variables in a neighborhood of i from a distribution $p(s|b)$. The cluster variables b_i indirectly inform the labels ℓ_i via the variable s_i. In summary, the model can be pictured as follows (see Fig. 5):

$$\begin{array}{ccccc} & & \ell & & \\ & & cat \downarrow & & \\ b & \longrightarrow & S & \longrightarrow & x \\ & & cat & & Gauss \end{array}$$

Given the input images $\{x_i\}$ and the prior $\mathbf{p}_i(\ell)$, the parameters $p(s|b)$, $p(s|\ell)$, the prior $p(b)$, and the Gaussian means and variances defining $p(x|s)$ are optimized to maximize the data likelihood

$$P = \sum_{\{b_j\},\{s_i\},\{\ell_i\}} \left(\prod_j p(b_j) \prod_{i \in W_j} p(s_j|b_i) p(x_i|s_i) \prod_i p(s_i|\ell_i) \mathbf{p}_i(\ell_i) \right)$$

[3]We use bold $\mathbf{p}$ to remind the reader that the prior is a fixed input to the inference algorithm, not a variable being optimized.

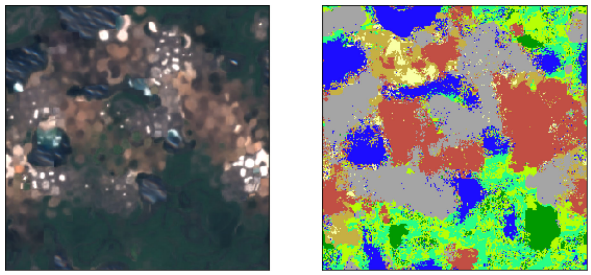

Fig. 6. Gaussian mean parameters in an epitome of validation set imagery and the inferred high-resolution label embedding.

where W_j denotes the set of image coordinates in a 5×5 window centered at j. The model can be optimized using a variational expectation–maximization (EM) algorithm, derived in a standard way by decoupling the posteriors $q(\{b_j\}, \{s_i\}, \{\ell_i\})$ to bound P from below and performing coordinate ascent.

The prior $\mathbf{p}_i(\ell)$ is derived from the given low-resolution labels as follows. [21, Fig. 3] provides us with the probabilities $p(\ell|c)$ of finding a high-resolution label ℓ at a point labeled with low-resolution class c. We first set $\mathbf{p}_i(\ell) = p(\ell|c_i)$, where c_i is the low-resolution label at position i, and then introduce uncertainty by smoothing (adding 0.05 to each $p(\ell|c_i)$ and renormalizing) and blurring over each 256×256 input image (the pointwise prior is mixed with the mean prior over the patch in a ratio of 10:1).

B. Epitomic Representations

The second approach, super-resolution with epitomic representations, is based on the work of [39]. We build a Gaussian mixture model of 7×7 image patches with a particular parameter-sharing parameterization—an epitome—and infer an assignment of labels in the latent variable space to produce a segmentation model.

The epitome is a 299×299 grid of means and variances for each spectral channel. It parameterizes a Gaussian mixture model of 7×7 image patches with 299^2 components: each 7×7 window in the epitome generates patches from a diagonal-covariance Gaussian with the corresponding mean and variance parameters. The epitome is trained to maximize the likelihood of all 7×7 patches in training data; we use the SGD-based training algorithm of [39] with self-diversification and posterior regularization. The Gaussian mean parameters of the resulting model are shown on the left of Fig. 6; each patch in the data is likely to be similar to some window in the epitome.

Denote the mixture index in this model—the position in the epitome—by s. By computing the posteriors over mixture components for a large sample of data patches, we derive a distribution over epitome positions for patches labeled with each low-resolution class c, $p(s|c)$. On the other hand, as described in the previous section, we are given $p(\ell|c)$, the probability of a pixel labeled as low-resolution class c belonging to high-resolution class ℓ. We infer a probabilistic assignment $p(\ell|s)$ of the high-resolution label to each epitome position s so as to minimize the relative entropy between the known $p(\ell|c)$ and

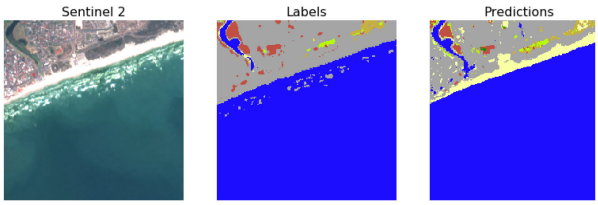

Fig. 7. Denoising of predictions by a small neural network.

TABLE VI
AVERAGE ACCURACY RESULTS OF EACH APPROACH ON THE VALIDATION AND TEST SETS

Method	Validation	Track 1 Test
Bag clustering	65.5%	50.4%
Epitome model	65.4%	49.4%
Ensemble	68.9%	50.0%
Neural smoothing	70.4%	53.6%
Final ensemble	–	57.5%

the inferred model $p_{\text{ep}}(\ell|c) = \sum_s p(\ell|s)p(s|c)$, an optimization problem that is straightforward to solve by an EM algorithm. The resulting $p(\ell|s)$ is shown on the right of Fig. 6.

The epitome can then be used as a segmentation model as follows: given a 7×7 patch of imagery x, we compute the posterior $p(s|x)$ over epitome positions s and then mix the labels $p(\ell|s)$ in a window around s, weighted by $p(s)$, to produce the predicted labels ℓ for the patch.

C. Deep Image Prior Postprocessing

Finally, inspired by the work of [40], we fit a small neural network—a fully convolutional network with five ReLU-activated layers of 64-channel 3×3 convolutions and a logistic regression classifier—to predict the output of the best-performing ensemble from the validation set imagery. We then use the trained model to make predictions over the same imagery, resulting in our final land-cover estimates. Because such a model has a small (11×11) receptive field and is sensitive only to local textures, it will not perfectly fit the outputs of the clustering and epitome algorithms. As shown in Fig. 7, it is not sensitive to certain types of errors made by those algorithms, such as speckled noise within uniform regions and boundaries between low-resolution class blocks—relics of the prior in the clustering algorithm.

D. Results and Discussion

We report the results of the three methods described in the above sections on the validation set and test set in Tables VI and VII. Specifically, we report results from five approaches: `Bag clustering`, the method described in Section IV-A; `Epitome model`, the method described in Section IV-B; `Ensemble`, an ensemble of the two previous methods; `Neural smoothing`, an application of the method described in Section IV-C to the *results* of `Ensemble`, and `Final ensemble`, an ensemble of the *results* of the previous methods based on the per class accuracy feedback we receive from the evaluation server.

TABLE VII
CLASSIFICATION ACCURACIES ON THE TEST SET FOR EACH METHOD USED BY THE FIRST PLACE TEAM IN TRACK 1

Class	Bag clustering	Epitome model	Ensemble	Neural smoothing	Final ensemble
Forest	**77.7%**	74.1%	76.8%	72.2%	74.3%
Shrubland	26.8%	24.9%	28.3%	**38.6%**	38.2%
Grassland	46.3%	28.5%	32.8%	52.0%	**53.4%**
Wetlands	21.7%	**38.9%**	28.9%	31.6%	35.8%
Croplands	57.6%	44.7%	**59.7%**	47.1%	53.6%
Urban	62.0%	70.6%	63.6%	71.2%	**81.3%**
Barren	12.3%	15.3%	10.9%	17.2%	**24.8%**
Water	**99.1%**	98.6%	99.0%	98.8%	98.6%
Average	50.4%	49.4%	50.0%	53.6%	**57.5%**

The highest accuracy per row is marked in bold.

Table VI compares the results on the validation set to those on the test set, before any high-resolution labels were known. We additionally evaluated the neural smoothing model that was trained on the validation set (i.e., the model that achieved a 70.4%) on the test set and found that it scored a 48.4%. Retraining the models on the imagery and labels generated by the Bag and Epitome methods from the test set improves the performance of this approach to 53.6%. Across both the validation and test sets, we find that the applying the neural smoothing results in a performance boost of roughly 3%.

Table VII compares the per class accuracy of each approach on the test set. Here, we see that the Bag clustering and Epitome models make complementary errors—for example, despite having similar average accuracy, the epitome model achieves a 17% higher performance on the Wetlands class than the bag model, and the bag model achieves a 18% higher performance on the Grassland class. This allows them to be ensembled effectively. In applied settings, per class leaderboard accuracy is obviously not available, but can be estimated by hand-labeling random samples of pixels from the study area.

V. FIRST PLACE TEAM OF TRACK 2

This section describes the algorithm developed by the first-place team of track 2 and reports the results. The algorithm is based on an ensemble of RF classifiers trained on refined samples. We first refine the low-resolution labels based on the prior knowledge of the confusion matrix of the low-resolution labels. Subsequently, initial classification results are generated from an ensemble of RFs, using spectral and textural features extracted from SAR and optical images. Finally, we implement a postprocessing step to fuse the classification results of classifiers trained with different features, which further improve the accuracy of the water class.

The algorithm follows the workflow summarized in Fig. 8, which comprises four steps: sample refinement, feature extraction, classification, and postprocessing. Each step is explained in detailed in the following sections (see Sections V-A– V-D).

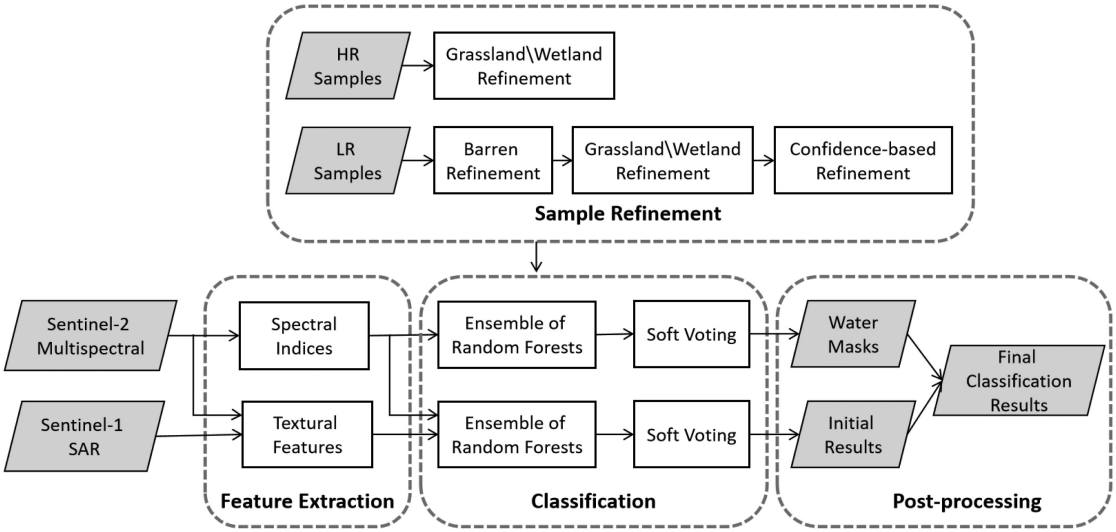

Fig. 8. Workflow of the proposed method.

A. Sample Refinement

The low-resolution labels from the global land-cover mapping products have been generated using a few (semi)automated processes, which are not very accurate. For example, the overall accuracy of the IGBP land-cover product, which the label is based on, is only 67% [24]. The average accuracy of the provided low-resolution labels in this contest is below 40% with respect to the high-resolution reference data (see Table IV). Moreover, some classes such as barren, shrublands, and wetlands are significantly unbalanced and associated with low-quality labels [21].

Since the results of the supervised machine learning methods highly rely on the quality of the training samples, we refine the samples based on the prior knowledge of the class confusion and the confidence of the label analyzed in [21]. We take following steps to correct common errors of the low-resolution labels based on empirical knowledge learnt from the training data.

1) Barren Refinement: We notice that barren labels are erroneous and only 8.8% of them are correct [21]. Since 39.9% of the shrubland samples are barren [21], we cluster the shrubland samples using k-means clustering algorithm into five clusters and choose one of them as the new barren samples.

2) Grassland and Wetland Refinement: We add the Savanna samples from the patches that contain water as wetland samples since we learn from the baseline paper [21] that 40.2% of the Savanna samples are wetlands. The rest of the Savanna samples are added as grassland samples. This step may produce errors in the wetland and grassland labels; thus, we try to alleviate this problem following step 3.

3) Confidence-Based Refinement: Refine the samples using a posterior confidence generated from a self-trained classifier. The modified low-resolution samples generated from the previous steps are used to train RF classifiers, and then, the classifiers are used to predict on the test set. The confidence of a sample is measured based on the maximal class probability among the eight classes. We only keep high confidence wetland and grassland samples newly added in the previous step.

As for the high-resolution samples, we performed an empirical sampling analysis, and we observed that the distributions of two classes (i.e., wetlands and grasslands) in the validation datasets and test datasets are quite different. We, therefore, exclude the wetlands and grasslands high-resolution samples.

B. Feature Extraction

The optical and SAR bands are first preprocessed before the feature extraction and classification. The Sentinel-1 SAR bands are clipped to the interval of $[-25, 0]$ and the Sentinel-2 optical bands are normalized to the range of 0–1 after truncating the digital numbers to the value of [0,10 000].

There are five vegetation-related classes including grasslands, croplands, wetlands, shrublands, and forest in the classification scheme. However, there are large confusions between these hard-to-distinguish vegetation classes with such a low-resolution label [21]. Therefore, in addition to the optical and SAR bands, we use the spectral indices to improve disparity between these classes. Furthermore, some classes, such as croplands and urban/built-up, have distinct textural patterns; therefore, we also extracted necessary textural features, which includes in total 36 features. These features are stacked and fed into our classifiers, consisting of ten multispectral bands, two SAR bands, 12 spectral features, and 12 textural features extracted from the SAR and RGB bands. The features are described in detail as follows.

1) Optical and SAR Bands: We preprocess ten Sentinel-2 bands whose original resolution is 10 m or 20 m and two Sentinel-1 SAR bands with the methods mentioned above and use them as features.

2) Spectral Features: Considering that empirical remote sensing indices are relatively reliable under different radiometric conditions, 12 spectral indices are computed from the Sentinel-2 multispectral bands. For more details, refer to [36].

3) Textural Features: Twelve gray-level co-occurrence matrix (GLCM) textural features are extracted, where six of them

are computed from the gray image of RGB bands and the other six features are generated from the VH polarized band from SAR images. The features are six attributes of GLCM: contrast, dissimilarity, homogeneity, energy, correlation, and angular second moment. A window size of 13×13 is selected through the validation test.

C. Classification

The RF classifier is widely used for addressing the remote sensing land-cover classification tasks. RF is essentially an ensemble learning method using decision tree classifiers. The voting strategy of multiple decision trees and the hierarchical examination of feature provide a good generalization capability and the ability to deal with high-dimensional feature spaces. Moreover, since the bagging strategy selects the training dataset by randomly drawing with replaceable examples, the RF is robust to noise. RF is particularly suitable for the classification task in this contest, because the low-resolution labels inherently contain many errors due to its mismatched resolution and its semiautomatic generation process. Thus, we select RFs for training on our refined labels.

The DFC2020 dataset contains more than five thousand patches, each with a size of 256×256, totaling more than 300 000 000 pixel-level training samples. To cope with a large number of training samples effectively within the acceptable memory and computation time, we use an ensemble of RF classifiers instead of a single RF classifier with a large number of trees.

To improve the generalization capability of our method and at the same time reduce training time, we set a large minimum sample size of 1000 and a small max depth of 60. The number of trees is set to 10. These hyperparameters are determined using the validation data. We adapt the class weight inversely proportional to the per-class sample numbers for 20 RF classifiers. We also train another 20 RF classifiers with equal weights, which summed up to 40 RF classifiers in total. Each RF was trained on 40 000 000 samples randomly drew from the refined training set which has more than 300 000 000 samples. In the testing phase, the classification results are generated via soft voting of the 40 RF classifiers. In the soft voting strategy, each base classifier n contributes to class probabilities with given weights $w_n . p_n(i, c)$ is the class probability of base classifier n for class c of pixel i. The soft voting class probability for class c of pixel i is

$$p_{\text{sv}}(i, c) = \frac{1}{N} \sum_{n=1}^{N} w_n \cdot p_n(i, c). \tag{1}$$

Then, the predicted class of pixel i is assigned as the class with the largest probability. We set equal weight for each base classifier.

D. Postprocessing

After the preliminary step, we generate the initial classification map from the trained ensemble of RFs. However, we find that models trained with texture features can sometimes mistakenly classify dynamic water surfaces (such as waves) as other classes. To address this problem and classify the pixels more accurately, we further implement a postprocessing step to refine the initial classification results. Since water pixels can be effectively detected by incorporating spectral information, we first train another 20 RF classifiers using only spectral bands and spectral indices to generate water masks. Subsequently, the final classification results are obtained by assigning the corresponding pixels of the initial classification maps in the water mask as water class.

TABLE VIII
EXPERIMENTAL SETTINGS FOR COMPARISON

Exp.	Labels	Features	Post-processing
#1	High	Spectral	No
#2	High + Low	Spectral	No
#3	High + Low	Spectral + Textual	No
#4	High + Low	Spectral + Textual	Yes

TABLE IX
CLASSIFICATION ACCURACIES OF THE RESULTS ON TEST SET

Class	Exp. #1	Exp. #2	Exp. #3	Exp. #4
Forest	78.45%	79.69%	**80.44%**	**80.44%**
Shrubland	**50.34%**	25.71%	24.81%	24.81%
Grassland	25.38%	47.49%	**50.28%**	50.27%
Wetlands	9.50%	**58.94%**	56.82%	56.28%
Croplands	28.95%	54.40%	**59.43%**	**59.43%**
Urban	72.97%	81.16%	**83.91%**	83.88%
Barren	**48.71%**	37.95%	37.32%	37.30%
Water	98.9 %	97.94%	97.85%	**98.96%**
Average	51.65%	60.41%	61.36%	**61.42%**

The result with the highest accuracy is marked bold for each class.

E. Results and Discussion

In this section, the experimental results of the proposed method on the testing dataset are reported. In order to further investigate the effectiveness of the sample refinement, the textural features and the postprocessing step, we evaluate our method using a few test settings shown in Table VIII. In the table, "Labels" means whether low-resolution labels are used to train the RFs; "Features" means which kind of features are used and "postprocessing" indicates whether postclassification processing is applied. The classification accuracies for each class on the test set are summarized in Table IX. The final average accuracy is 0.6142. Fig. 9 shows example results generated by our method.

By comparing the results of Exp. #1 and Exp. #2, we can see that by including the modified low-resolution labels from the test set, the average accuracy increases from 51.65% to 60.41%. This indicates that semantic information of the test set, in this case the low-resolution labels, is important to facilitate the classification on the test set, even though they contain many errors. As expected, for croplands and urban/built-up classes, which have distinguishable textural patterns, by incorporating textural features, the accuracy increases from 54.40% to 59.43% and from 81.16% to 83.91%, respectively. Compared with Exp. #3, Exp. #4 adds the postprocessing step, which aims to improve the

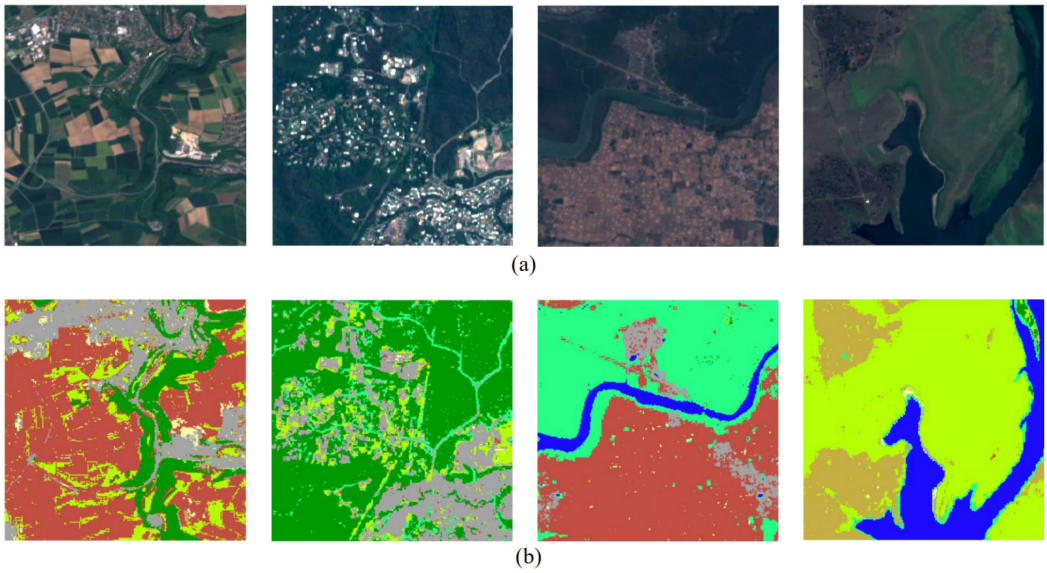

Fig. 9. Example results of the method of Team *Antonia*. Refer to Fig. 2 for legends. (a) RGB images. (b) Classification results.

accuracy of water class. We can observe that the water accuracy increased from 97.85% to 98.96% with the postprocessing. This also leads to the highest average accuracy of 61.42%.

VI. Conclusion

The automatic production of semantic maps of the Earth with high accuracy as well as high spatial and temporal resolution is one of the most important application areas of remote sensing and Earth observation. Since the tremendous amount of data make manual interpretation infeasible, corresponding systems have to rely on machine learning, i.e., supervised learning. This, however, requires not only the image data itself but also data of the desired system output, e.g., semantic class labels. Despite the abundance of available remote sensing imagery, the scarce availability of such reference data to train and evaluate machine-learning-based models remains a significant bottleneck. Highly accurate semantic labels produced by manual interpretation of the data itself or auxiliary images can only cover small geographic areas leading to models that usually do not generalize well to other parts of the world. On the other hand, semantic maps that cover larger areas (or even the whole globe) are notoriously of low quality with a significant amount of label noise caused either by being outdated, misaligned, of very low resolution—or a mixture of those. As a consequence, machine learning methods applied to remote sensing imagery cannot assume large training datasets with mostly correct labels. On the contrary, they do have to be capable to cope with low quality reference data and still be able to produce semantic maps of high quality (i.e., accurate and of high resolution).

In this article, we summarized the 2020 IEEE GRSS Data Fusion Contest, organized by the IEEE GRSS IADF TC, which addressed the task of weakly supervised learning. In particular, we described the challenge to create well-generalizing machine learning models for large-scale land-cover mapping if only noisy low-resolution labels and/or a small amount of high-resolution labels are available for training. To this aim, the contest built upon the SEN12MS dataset consisting of more than 180k image triplets of 256×256 pixels containing Sentinel-1 and -2 images as well as low-resolution semantic maps. Additionally, the contest provided more than 6k image triplets from seven globally distributed areas, which include not only the Sentinel-1 and -2 image data but also high-resolution labels that had been created semimanually. This allowed the participants of the contest to train on a combination of high-resolution images and low-resolution reference data (Track 1) or to use a small amount of additional high-resolution samples (Track 2), and to validate and evaluate on high-resolution labels. The winning approach in Track 1 used both a clustering algorithm and generative model to assign class labels to each high-resolution pixel based on the spectral values at that pixel and in neighboring pixels and then smoothed those class predictions with a neural network. This two-step approach (of assignment, then smoothing) provided better results than the straightforward approach of treating the low-resolution labels as if they were high-resolution labels and training a semantic segmentation network. Since this approach does not depend on high-resolution data at all, it can be applied globally by leveraging the comprehensive SEN12MS dataset. The winning approach in Track 2 refined first the low-resolution labels based on the prior knowledge of the confusion matrix of the low-resolution labels. Then, initial classification results were generated using an ensemble of RFs, using spectral and textural features extracted from SAR and optical images. Eventually, a postprocessing step was implemented to fuse the classification results of classifiers trained with different features to further improve the accuracy of the water class.

The results of the winning teams are interesting as it exposes an opportunity to more effectively train neural networks with low-resolution labels. Future research is needed to understand the limitations of these approaches in land-cover mapping, as well as other domains where it could be applied. We are excited to compare the results of these algorithms against new

benchmark land-cover datasets that take advantage of Sentinel-2 imagery such as LandCoverNet [41].

The four and three top-ranked solutions of both tracks presented their methods at IGARSS 2020, while the winning solution of each track is described in this article in detail and discusses further insights into the challenges of weakly supervised learning.

Similar to the previous editions, the DFC2020 attracted again global attention. Nearly, 150 registrations from more than 20 countries registered for downloading the data. From the nearly 160 teams that registered at the CodaLab page for the contest during the development stage (and uploaded nearly 3k solutions), more than 30 teams entered the test phase and provided approximately 250 solutions for each of the two tracks. This clearly illustrates the importance of the addressed research topic of weakly supervised learning. Furthermore, the majority of the participants are students showing that the DFC is introduced to early career scientists and used for educational purposes.

After the contest, the data have been made available again and will remain in open access for the benefit of the community. People interested can find all the related information on the IEEE GRSS website.[4] The SEN12MS dataset is available on the mediaTUM website,[5] and the validation and test datasets are available on the IEEE DataPort website.[6] The public leaderboard on the Codalab competition website[7] will remain open for future development so that one can submit prediction results to obtain the performance statistics, compare to other users, and hopefully improve the results presented in this article. We do believe that both the motivation of the contest and the corresponding datasets will continue to foster research toward large-scale land-cover mapping with modern machine learning models trained on existing land-cover data.

The DFC2020 provides one of the first benchmark datasets for large-scale weakly supervised learning in the context of global land-use/cover classification from multimodal data. Nevertheless, already now several extensions and variations can be foreseen that should—and hopefully will—be addressed by future contests and benchmarks. One example are types of label degradation other than low resolution and accuracy, e.g., semantic maps created by crowdsourcing such as OpenStreetMaps, which are often misaligned to the image data, outdated, or ambiguous. Moreover, it will be crucial to investigate solutions for situations where neither low- nor high-quality reference data are available in abundance. These approaches, often termed as self-supervised learning, aim at exploiting the abundance of available image data to solve at least parts of the mapping problem.

Future solutions have to address these issues to be able to create accurate maps of the surface of the Earth with a sufficiently high spatial and temporal resolution as required by many RS/EO workflows to model and understand geo-/biophysical processes as well as socioeconomic developments.

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

[5][Online]. Available: https://mediatum.ub.tum.de/1474000

[6][Online]. Available: https://ieee-dataport.org/competitions/2020-ieee-grss-data-fusion-contest

[7][Online]. Available: https://competitions.codalab.org/competitions/22289

REFERENCES

[1] K. Anderson, B. Ryan, W. Sonntag, A. Kavvada, and L. Friedl, "Earth observation in service of the 2030 agenda for sustainable development," *Geo-Spatial Inf. Sci.*, vol. 20, no. 2, pp. 77–96, 2017.

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

[3] P. Ghamisi *et al.*, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state-of-the-art," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 1, pp. 6–39, Mar. 2019.

[4] 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 Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3012–3021, Oct. 2007.

[5] 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.

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

[7] 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, Feb. 2012.

[8] 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, Feb. 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, Jun. 2013.

[10] 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.

[11] 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.

[12] M. Campos-Taberner *et al.*, "Processing of extremely high resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS Data Fusion Contest. Part A: 2D contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5547–5559, Dec. 2016.

[13] A.-V. Vo *et al.*, "Processing of extremely high resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS Data Fusion Contest. Part B: 3D contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5560–5575, Dec. 2016.

[14] 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.

[15] 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.

[16] Y. Xu *et al.*, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

[17] B. Le Saux, N. Yokoya, R. Hänsch, M. Brown, and G. Hager, "2019 Data Fusion Contest [Technical Committees]," *IEEE Geosci. and Remote Sens. Mag.*, vol. 7, no. 1, pp. 103–105, Mar. 2019.

[18] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS—A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," in *Proc. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, 2019, pp. 153–160.

[19] D. S.-M. M. Friedl "MCD12Q1 MODIS/Terra Aqua Land Cover Type Yearly L3 Global 500 m SIN Grid V006," 2015. [Online]. Available: https://lpdaac.usgs.gov/products/mcd12q1v006/

[20] T. R. Loveland and A. Belward, "The international geosphere biosphere programme data and information system global land cover data set (discover)," *Acta Astronautica*, vol. 41, no. 4-10, pp. 681–689, 1997.

[21] M. Schmitt, J. Prexl, P. Ebel, L. Liebel, and X. X. Zhu, "Weakly supervised semantic segmentation of satellite images for land cover mapping—Challenges and opportunities," in *Proc. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, 2020, vol. V-3-2020, pp. 795–802.

[22] R. Torres *et al.*, "GMES Sentinel-1 mission," *Remote Sens. Environ.*, vol. 120, pp. 9–24, 2012.
[23] M. Drusch *et al.*, "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sens. Environ.*, vol. 120, pp. 25–36, 2012.
[24] D. Sulla-Menashe, J. M. Gray, S. P. Abercrombie, and M. A. Friedl, "Hierarchical mapping of annual global land cover 2001 to present: The MODIS collection 6 land cover product," *Remote Sens. Environ.*, vol. 222, pp. 183–194, 2019.
[25] A. Di Gregorio, Land Cover Classification System: Classification Concepts and User Manual: LCCS. Rome, Italy: Food and Agriculture Org., 2005.
[26] N. Yokoya, P. Ghamisi, R. Hänsch, and M. Schmitt, "2020 IEEE GRSS Data Fusion Contest: Global land cover mapping with weak supervision," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 1, pp. 154–157, Mar. 2020.
[27] P. Gong *et al.*, "Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017," *Sci. Bull.*, vol. 64, no. 6, pp. 370–373, 2019.
[28] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, 2017.
[29] S. Dong, B. Gao, Y. Pan, R. Li, and Z. Chen, "Assessing the suitability of FROM-GLC10 data for understanding agricultural ecosystems in China: Beijing as a case study," *Remote Sens. Lett.*, vol. 11, no. 1, pp. 11–18, 2020.
[30] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–39, 1957.
[31] N. Yokoya, P. Ghamisi, R. Hänsch, and M. Schmitt, "Report on the 2020 IEEE GRSS Data Fusion Contest—Global land cover mapping with weak supervision," *IEEE Geosci. Remote Sens. Mag*., vol. 8, no. 4, pp. 134–137, Dec. 2020.
[32] C. Robinson, N. Malkin, L. Hu, B. Dilkina, and N. Jojic, "Weakly supervised semantic segmentation in the 2020 IEEE GRSS Data Fusion Contest," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020.
[33] Y. Xia, Y. Liao, H. Zhang, and G. Yang, "Land cover mapping based on multi-branch fusion of object-based and pixel-based segmentation with filtered labels," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020.
[34] D. Cerra *et al.*, "Stepwise refinement of low resolution labels for earth observation data: Part 1," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020.
[35] H. Chen, W. Liu, C. Xiao, and R. Qin, "Large-scale land cover mapping of satellite images using ensemble of random forests—IEEE Data Fusion Contest 2020 Track 1," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020.
[36] H. Chen, C. Xiao, W. Liu, and R. Qin, "Large-scale land cover mapping of satellite images using ensemble of random forests—IEEE Data Fusion Contest 2020 Track 2," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020.
[37] D. Cerra *et al.*, "Stepwise refinement of low resolution labels for earth observation data: Part 2," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020.
[38] S. Yin *et al.*, "Weakly supervised land cover classification method for large-scale multi-resolution labeled satellite images data sets," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020.
[39] K. Malkin, A. Ortiz, and N. Jojic, "Mining self-similarity: Label super-resolution with epitomic representations," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 531–547.
[40] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9446–9454.
[41] S. H. Alemohammad *et al.*, "LandCoverNet: A global land cover classification training dataset," 2020. [Online]. Available: https://doi.org/10.34911/rdnt.d2ce8i

Caleb Robinson received the B.Sc. degree in computer science from the University of Mississippi, Oxford, MS, USA, in 2015, and the Ph.D. degree in computational science and engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2020.

His dissertation work was on large-scale machine learning for geospatial problems in computational sustainability. Since 2020, he has been a Data Scientist with the AI for Good Research Lab, Microsoft Research, Redmond, WA, USA. His current research interests include self-supervised methods for training deep learning models with large amounts of unlabeled remotely sensed imagery, and change detection methods for use with aerial imagery.

Kolya Malkin received the B.S. degree in mathematics from the University of Washington, Seattle, WA, USA, in 2015. He is currently working toward the Ph.D. degree with Yale University, New Haven, CT, USA.

His research interests in machine learning include deep learning for weakly supervised segmentation, hybrid neural and Bayesian graphical models and their application to computer vision and natural language processing, and algorithms for land-cover mapping and change detection.

Nebojsa Jojic received the B.S. degree in electrical engineering from the University of Belgrade, Belgrade, Serbia, in 1995, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2001.

Since 2000, he has been with Microsoft Research, Redmond, WA, USA, where he is currently a Senior Principal Researcher. His research interests include machine learning with applications in computer vision, computational biology, computational immunology, signal processing, and natural language processing.

Huijun Chen received the bachelor's degree in remote sensing science and technology and the master's degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016 and 2019, respectively. She is currently working toward the Ph.D. degree with the Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, USA, and also with the Environmental Sciences Graduate Program.

Her research interests include deep learning, transfer learning, and high-resolution land-cover mapping.

Rongjun Qin (Senior Member, IEEE) received the B.S. degree in computational mathematics and the M.S. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009 and 2011, respectively, and the Ph.D. degree in photogrammetry and remote sensing from ETH Zürich, Zürich, Switzerland, in 2015.

He is currently a Faculty Member with the Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, USA, where he is also with the Department of Electrical and Computer Engineering. He serves as a Reviewer for more than 15 international journals in the field photogrammetry and remote sensing. He has authored rational polynomial coefficient stereo processor and multistereo processor used for reconstructing 3-D information from 2-D images with high quality. His research interests include photogrammetric 3-D reconstruction, remote sensing image classification, unmanned aerial vehicle (UAV) image rocessing, image dense matching, and change detection. His research seeks for computational solutions to various geometric and interpretation problems in an urban context using imaging sensors such as aerial/UAV imagery, LiDAR, and satellite multispectral/hyperspectral images.

Prof. Qin was a recipient of the First Prize of Mathematical Modeling Contest and several other prominent scholarship awards. He is an Associate Editor for the *Photogrammetric Engineering and Remote Sensing Journal*. He is also chairing the Working Group "Satellite Constellation for Remote Sensing" of the International Society for Photogrammetry and Remote Sensing Commission.

Changlin Xiao received the B.S. degree in measuring, testing technologies and instruments and the M.S. degree in measurement technologies and instruments from the University of Electronic Science and Technology of China, Chengdu, China, in 2009 and 2012, respectively, and the Ph.D. degree in civil and environmental engineering and geodetic science from The Ohio State University, Columbus, OH, USA, in 2017.

He is currently a Senior Researcher with a smart city related company and tying to implement the 3-D reconstruction of the reality at a large scale by satellite and other remote sensing data. His research interests include computer vision and 3-D photogrammetry, including image classification, object detection, and visual tracking; land-cover classification, tree detection, and change detection; 3-D building model reconstruction, building facade parsing and modeling, 3-D localization; and deep learning neural networks and transfer learning.

Michael Schmitt (Senior Member, IEEE) received the Dipl.-Ing. (Univ.) degree in geodesy and geoinformation, the Dr.-Ing. degree in remote sensing, and the Habilitation in data fusion from the Technical University of Munich (TUM), Munich, Germany, in 2009, 2014, and 2018, respectively.

Since 2020, he has been a Full Professor of Applied Geodesy and Remote Sensing with the Department of Geoinformatics, Munich University of Applied Sciences, Munich. From 2015 to 2020, he was a Senior Researcher and the Deputy Head with the Professorship for Signal Processing in Earth Observation, TUM. In 2019, he was additionally appointed as an Adjunct Teaching Professor with the Department of Aerospace and Geodesy, TUM. In 2016, he was a Guest Scientist with the University of Massachusetts at Amherst, Amherst, MA, USA. His research interests include image analysis and machine learning applied to the extraction of information from multimodal remote sensing observations. In particular, he is interested in remote sensing data fusion with a focus on synthetic aperture radar (SAR) and optical data.

Dr. Schmitt is the Co-Chair of the Working Group "SAR and Microwave Sensing" of the International Society for Photogrammetry and Remote Sensing and also of the Working Group "Benchmarking" of the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society. He frequently serves as a Reviewer for a number of renowned international journals and conferences and has received several best reviewer awards. He is an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

Pedram Ghamisi (Senior Member, IEEE) received the M.Sc. degree (Hons.) in remote sensing with the K. N. Toosi University of Technology, Tehran, Iran, in 2012, and the Ph.D. degree in electrical and computer engineering with the University of Iceland, Reykjavik, Iceland, in 2015.

He is currently the Head of the Machine Learning Group, Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Freiberg, Germany, the Chief Technology Officer and Co-Founder of VasoGnosis Inc., Milwaukee, WI, USA., and a Visiting Professor with the Institute of Advanced Research in Artificial Intelligence, Vienna, Austria.

Dr. Ghamisi is the Vice-Chair of the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society.

Ronny Hänsch (Senior Member, IEEE) received the undergraduate degree in computer science and the Ph.D. degree from the Technische Universität Berlin, Berlin, Germany, in 2007 and 2014, respectively.

He is currently with the German Aerospace Center, Weßling, Germany. His current research interests include ensemble methods for image analysis.

Dr. Hänsch is the Co-Chair (2017–2021) of the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society and the Co-Chair of the International Society for Photogrammetry and Remote Sensing Working Group II/1 (Image Orientation).

Naoto Yokoya (Member, IEEE) received the M.Eng. and Ph.D. degrees in aerospace engineering from the University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively.

He is currently a Lecturer with the University of Tokyo and a Unit Leader with the RIKEN Center for Advanced Intelligence Project, Tokyo, where he leads the Geoinformatics Unit.

Dr. Yokoya is the Chair (2019–2021) of the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society.