

Open Data for Global Multimodal Land Use Classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest

Naoto Yokoya¹, *Member, IEEE*, Pedram Ghamisi², *Member, IEEE*, Junshi Xia³, Sergey Sukhanov, Roel Heremans, Ivan Tankoyeu, Benjamin Bechtel⁴, Bertrand Le Saux⁵, *Member, IEEE*, Gabriele Moser⁶, *Senior Member, IEEE*, and Devis Tuia⁷, *Senior Member, IEEE*

Abstract—In this paper, we present the scientific outcomes of the 2017 Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society. The 2017 Contest was aimed at addressing the problem of local climate zones classification based on a multi-temporal and multimodal dataset, including image (Landsat 8 and Sentinel-2) and vector data (from OpenStreetMap). The competition, based on separate geographical locations for the training and testing of the proposed solution, aimed at models that were accurate (assessed by accuracy metrics on an undisclosed reference for the test cities), general (assessed by spreading the test cities across the globe), and computationally feasible (assessed by having a test phase of limited time). The techniques proposed by the participants to the Contest spanned across a rather broad range of topics, and of mixed ideas and methodologies deriving from computer vision and machine learning but also deeply rooted in the specificities of remote sensing. In particular, rigorous atmospheric correction, the use of multitemp images, and the use of ensemble methods fusing results obtained from different data sources/time instants made the difference.

Index Terms—Convolutional neural networks (CNNs), crowdsourcing, deep learning (DL), ensemble learning, image analysis and data fusion (IADF), multimodal, multiresolution, multisource, OpenStreetMap (OSM), random fields.

I. INTRODUCTION

LAND use/land cover classification at the global scale is one of the challenges of geospatial analysis. Providing a unified categorization of types of human habitats, as well as of land cover in rural areas, faces great challenges, since the structures of cities vary greatly depending on architectural, cultural, and environmental local conditions. Nonetheless, the payback for a successful characterization would be enormous, since this would allow a better calibration of climatic models [1], successful intercities comparisons [2] and, in general, an objective way of describing cities and their impacts on the environment.

Despite the existence of global built-up layers, such as the Global Urban Footprint [3] and the Global Built-up Density of the ESA Urban Thematic Exploitation Platform [4] and the Global Urban Human Settlements Layer of the Joint Research Center of the European Commission [5], or of regional/continental initiatives such as the Copernicus Urban Atlas in the European Union [6], we are still lacking a unified view of land use in multiple categories describing how the urban space is structured. When considering land use, it becomes important to consider densities, layouts, and volumes, since most categories will be characterized by buildings and trees, and what will differentiate them is *how they are organized*.

Recently, the concept of *local climate zones* (LCZs) [1] has been proposed to provide a land use/land cover description in this direction. LCZs are a generic, climate-based typology of urban and natural landscapes, which delivers information on basic physical properties of an area that can be used by planners or climate modelers. They are generally applied at a coarse spatial scale (typically grids of resolution 100 or 200 m), in order to be able to catch this sense of urban structure that cannot be perceived when working at single-pixel scale at very high resolution.

LCZs have taken momentum in the recent GIScience and remote sensing literature [7]–[9], but most efforts have been

Manuscript received November 20, 2017; revised January 22, 2018; accepted January 23, 2018. Date of publication April 15, 2018; date of current version May 1, 2018. The work of B. Bechtel was supported by the Cluster of Excellence “CliSAP” (EXC177), University of Hamburg, funded through the German Science Foundation (DFG). The work of D. Tuia was supported by the Swiss National Science Foundation under Grant PP00P2-150593. (*Corresponding author: Devis Tuia.*)

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

P. Ghamisi is with the Department of Signal Processing in Earth Observation, Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling 82234, Germany (e-mail: pedram.ghamisi@dlr.de).

J. Xia is with the Department of Advanced Interdisciplinary Studies, The University of Tokyo, Tokyo 153-8904, Japan (e-mail: xiajunshi@gmail.com).

S. Sukhanov, R. Heremans, and I. Tankoyeu are with the AGT International, Darmstadt 64295, Germany (e-mail: ssukhanov@agtinternational.com; rheremans@agtinternational.com; itankoyeu@agtinternational.com).

B. Bechtel is with the Center for Earth System Research and Sustainability, Universität Hamburg, Hamburg 20146, Germany (e-mail: benjamin.becht@uni-hamburg.de).

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

G. Moser is with the Department of Electrical, Electronic, Telecommunications Engineering and Naval Architecture, University of Genoa, Genoa I-16145, Italy (e-mail: gabriele.moser@unige.it).

D. Tuia is with the Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, Wageningen 6700 HB, The Netherlands (e-mail: devis.tuia@wur.nl).

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.2018.2799698

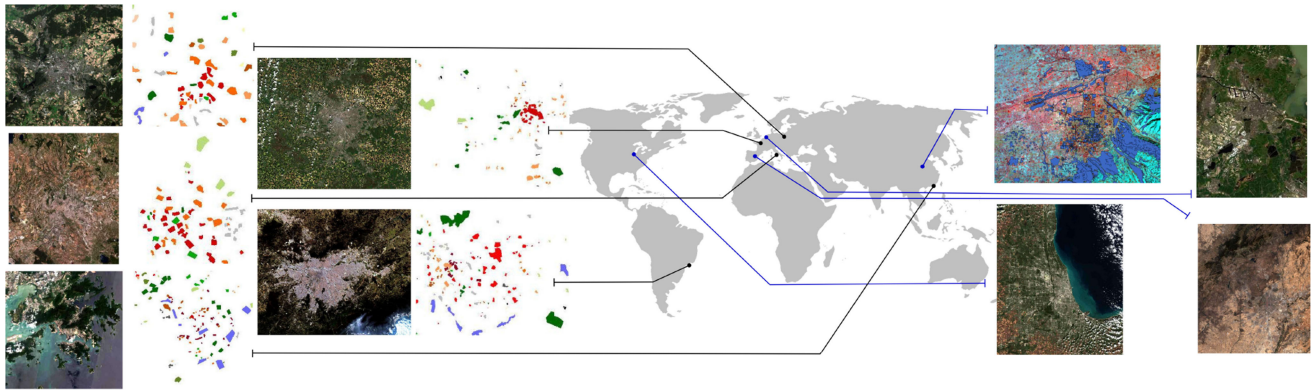


Fig. 1. Global worldwide distribution of DFC 2017 data. Training cities (on the left: Berlin, Hong Kong, Paris, Rome, and Sao Paulo) come with Landsat 8/Sentinel-2 images and OSM layers as well as LCZ maps (classes and colors are defined in Table I). For test cities (on the right: Amsterdam, Chicago, Madrid, and Xi'an), only satellite images and OSM layers are provided while the LCZ labels remain undisclosed.

put into classifying single cities into the LCZ categories [10]–[14]. In other words, so far accurate ground references for each individual considered city are needed to provide accurate LCZ maps. In this respect, several international efforts, such as GeoWiki¹ and the World Urban Database and Portal (WUDAPT²), have been organized by researchers to gather high-quality land cover/land use information worldwide, typically via crowdsourcing [15], [16], games [17], or other challenges.

But if the data collection efforts are being very successful, no models designed explicitly to generalize to additional cities are available to date. This means that in order to classify a new city into LCZs, one should lead a separate LCZ campaign in that same city with volunteers. Moreover, there is evidence that training samples from unexperienced volunteers can result in inaccurate classification results [18]. The success of models trained on some cities when applied to others remains at best unclear at the moment. Such problem is known in machine learning as domain adaptation [19] and has been tackled in remote sensing image processing with various technical solutions including feature selection, feature extraction, classifier adaptation, and active learning [20].

The Data Fusion Contest 2017 (DFC17) tackles exactly these open questions in LCZ classification. More precisely, it aims at designing new LCZ classification solutions based on open data, both from remote sensing and GIS, with particular attention to the issue of generalizing results to new urban areas unseen during model training. It follows a tradition of yearly data processing competitions [21]–[30] organized by the Image Analysis and Data Fusion Technical Committee (IADF TC³) of the IEEE Geoscience and Remote Sensing Society (IEEE GRSS): Every year since 2006, a dataset has been released to the scientific community and participants have been invited to perform a task of interest, which, in the case of the DFC17 was LCZ classification over several cities using open, global, and multimodal data.

The LCZ classification problem was cast as a 17-classes classification problem following the definitions in [1] (see Section II): LCZ reference data, as well as satellite and GIS layers from OpenStreetMap (OSM), were provided for five cities (the *training cities* hereafter). The participants could then train models and prepare for the blind classification round, for which a new set of four cities (the *test cities* hereafter) were provided without any ground reference. Participants were asked to submit their LCZ maps for the test cities on the Data and Algorithm Standard Evaluation website (DASE⁴) [31], a platform developed by IEEE GRSS with the company Ticinum Aerospace S.r.l. (Pavia, Italy).

In this paper, we report the outcomes of the competition: After describing the dataset (see Section II), we will discuss first 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-place and second-place teams (Sections IV and V, respectively). Finally, conclusions are drawn in Section VI.

II. DATA OF THE DFC17

Following the idea of an open data contest, free and open data from different sources were preprocessed and provided for the five training cities (Berlin, Hong Kong, Paris, Rome, and Sao Paulo) and the four test cities (Amsterdam, Chicago, Madrid, and Xi'an), also geographically represented in Fig. 1.

Fig. 2 showcases the data types for the case of Rome, Italy. In particular, multispectral data from Landsat 8 and Sentinel-2 as well as data from OSM were included. The satellite data were provided on the target grid at 100-m resolution, whereas the OSM data were both provided as vector layers and partly also rasterized on a 5-m grid. The preprocessing was conducted in SAGA GIS [32]. The detailed information about satellite and OSM data used in the contest can be found on the IEEE GRSS website.⁵ Additionally, participants were encouraged to use and share auxiliary data from free and open sources.

¹<http://geo-wiki.org>

²<http://www.wudapt.org>

³<http://www.grss-ieee.org/community/technical-committees/data-fusion/>

⁴<http://dase.ticinumaerospace.com>

⁵<http://www.grss-ieee.org/community/technical-committees/data-fusion/2017-ieee-grss-data-fusion-contest-2/>

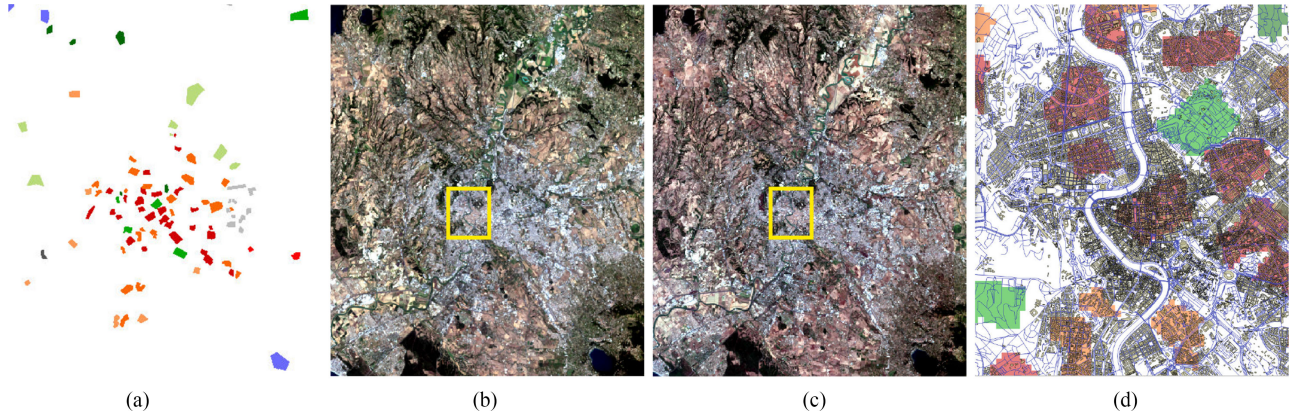


Fig. 2. DFC 2017 data for the city of Rome, Italy. (a) Ground truth of the LCZs: see color legend in Table I; (b) Landsat 8 natural composite image (bands 4-3-2), 1 out of 4 dates available for this city; (c) Sentinel-2 natural composite image; (d) zoom in the city center [corresponding to the rectangle area in (b) and (c)] with LCZ, OSM street layer (in blue) and OSM building layers (in gray).

A. Landsat Images

Landsat 8, launched in 2013, is the latest mission of the USGS and NASA focused on moderate resolution land satellites. As compared to the previous Landsat missions, it has additional spectral bands in the blue and cirrus cloud-detection wavelengths, improved signal-to-noise ratio and radiometric resolution, and can collect more images per day [33]. Several scenes per city [from 2 to 5 different dates, including the one in Fig. 2(b)] were downloaded from the USGS EarthExplorer portal. Then, the visible, short-wave, and long-wave infrared bands [therefore excluding the atmospheric band (9) and the panchromatic band (8)] were resampled to the 100-m target grid using an area weighted average.

B. Sentinel-2 Images

Sentinel-2A, launched in 2015, is a European Commission and ESA mission that provides global multispectral high-resolution observations of land surfaces. In particular, it aims at the systematic and frequent provision of high-resolution multispectral imagery for continuity and enhancement of the SPOT satellite series of the French Space Agency (CNES). These data aim to provide the basis for the next generation of land-cover and cover change maps, and operational products on geophysical variables describing the land surface [34]. For this study, we downloaded Sentinel-2 data via the Amazon Web Services Archive and resampled, to 100-m resolution, nine multispectral bands including the visible, vegetation red edge, and short-wave infrared wavelengths and excluding the atmospheric bands [1, 9, and 10, see Fig. 2(c)]. One date, corresponding to 1–5 Sentinel-2 tiles, was selected for each site. Additionally, direct links to the original data (10–20 m) were provided to encourage use of the additional spatial details included in the full resolution imagery.

C. OSM Layers

OSM is a volunteered geographic information project aiming at providing open, user-generated maps [35]. The information in OSM is in vector format (point, line, and polygon geometries)

with linked attribute data. In particular, OSM includes information on roads, railways, points of interest, natural features, water areas, land use, and buildings, among others. Previous studies proved the effectiveness of OSM to train models for land cover classification [36], [37], and recently OSM was also explored for adding value to LCZ classification [38]. The OSM data were downloaded in shapefile format from the Geofabrik portal (<http://www.geofabrik.de>). For some cities, several administrative areas had to be merged. Building footprints (polygon), land use (polygon), water areas (polygon), and road network (line) were provided as vector data. Additionally, building footprints, land use, and water layers were rasterized to a 5-m grid, which was superimposable with the satellite images.

D. LCZ Ground Truth

For the training cities, we provided a ground truth of the various LCZ classes on several areas of the considered cities. The LCZ classes are defined in Table I. These samples were initially extracted from the WUDAPT database and thoroughly revised to ensure the highest possible correctness. The training data were provided as raster layers at 100-m resolution, superimposable to the satellite images. The ground truth for the test set remained (and still remains) undisclosed and was used for evaluation of the results in DASE.

It is worth noting that the classes in the ground truth were severely imbalanced (see Table II). In the training set, the class sizes ranged from 323 to 17716 samples (i.e., the largest class had nearly 55 times more samples than the smallest one), and the average and median sizes were 4814 and 2819 samples, respectively. On one hand, this distribution in the training set approximated the proportions of the LCZs in the considered scenes. On the other hand, it implied an additional challenge for the classification algorithms, some of which (e.g., support vector machines) are known to be affected by imbalance issues. The test set was similarly imbalanced as well. Furthermore, the ratio between the numbers of training and test samples of each class ranged from 0.3 to 6.8 (average = 1.9, median = 1.2). This variability in the training/test balance was a direct consequence of

TABLE I
LCZ CLASSES


















Built types					Land cover types					
#	class			#	class			#	class	
1	Compact high-rise			6	Open low-rise			101	Dense trees	
2	Compact midrise			7	Light low-rise			102	Scattered trees	
3	Compact low-rise			8	Large low-rise			103	Bush, scrub	
4	Open high-rise			9	Sparsely built			104	Low plants	
5	Open midrise			10	Heavy industry			105	Bare rock / paved	
								106	Bare soil / sand	
								107	Water	

TABLE II
DISTRIBUTION OF THE TRAINING AND TEST SET SIZES OF THE LCZ CLASSES

Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
# samples train	1642	6103	5738	2098	4759	8891	0	4889	1156	449	17 716	2819	1741	14 457	323	503	8561
# samples test	242	4892	1535	2270	2255	8265	0	11 230	1072	920	3170	4528	1284	12 994	1104	391	4454

the training and test cities being located in different countries and continents. It also was consistent with the general goal of the DFC17 of investigating classification approaches that aimed at generalizing across diverse geographical areas.

III. SUBMISSIONS AND RESULTS

The ranking of the submitted classification maps was based on an overall accuracy (OA) over the ensembles of cities, in order to reward the capacity to handle multiple classes as much as the adaptability to various geographic contexts. Specifically, we used the average over the ensemble of labeled points from the four test cities. Let C be the set of classes (LCZ) and X^j be the test images to classify ($1 \leq j \leq 4$). Let also $X_c^j \subset X^j$ be the set of points classified with label $c \in C$ in the map uploaded for image j and let X_c^{j*} be the ground truth for the same label c in the same image j . OA is the proportion of the correctly classified points over all the classes and all images

$$OA = \frac{1}{\sum_{j=1}^4 \sum_{c \in C} |X_c^j|} \sum_{j=1}^4 \sum_{c \in C} |X_c^j \cap X_c^{j*}|. \quad (1)$$

The evaluation took place only on those ground-truth regions, but since the location of ground-truth samples was undisclosed, participants had to submit fully classified maps. Although the final ranking was based on the OA, we also measured the Cohen's *Kappa* [39] ($\kappa = \frac{p(a) - p(e)}{1 - p(e)}$, where $p(a)$ is the observed accuracy—quantified through the OA—and $p(e)$ is the accuracy expected by chance and computed using the confusion matrix), and the number of actually predicted classes for obtaining additional insight on the results. In particular, the number of actual classes in the maps was a relevant indicator because the LCZ classes in the dataset were quite imbalanced, as discussed in Section II, the classes were imbalanced and it was important to identify the possible presence of missed classes.

The four teams that submitted the best-ranked classification maps were awarded. Their solutions were presented during the 2017 IEEE International Geoscience and Remote Sensing Symposium in Fort Worth, TX, USA. Starting from the top ranked and then in descending order, the four teams are as follows.

- 1) *WXYZ team*: N. Yokoya, P. Ghamisi, and J. Xia from the University of Tokyo, Japan, DLR, and TU München, Germany: *Multimodal, multitemporal, and multisource global data fusion for LCZs classification based on ensemble learning* [40].
- 2) *AGT team*: S. Sukhanov, R. Heremans, I. Tankoyeu, J. Louradour, D. Trofimova, and C. Debes from AGT International, Germany: *Multilevel ensembling for LCZs classification* [41].
- 3) *Camilasa team*: C. S. dos Anjos Lacerda, M. Gonçalves Lacerda, L. do Livramento Andrade, and R. Neves Salles from the Institute of Advanced Studies of the Brazilian Air Force, Brazil: *Classification of urban environments using feature extraction and random forest* [42].
- 4) *Nanjingxxy team*: Yong Xu, Fan Ma, Deyu Meng, Chao Ren, and Yee Leung from the Chinese University of Hong Kong and the Xi'an Jiaotong University, China: *A co-training approach to the classification of LCZs with multisource data* [43].

In Table III, we provide details about the ten best performing teams of the leaderboard, as recorded after the three weeks of the evaluation phase. We group these methods according to their main characteristics. Namely, we identify random forest (RF) type methods (also including rotation forests (RoFs) and decision tree approaches), boosting (Bo.), deep learning (DL) [mostly convolutional neural networks (CNNs)], and expert handcrafted features (Exp.) as main components. These models were sometimes combined in multiple classifier systems by some teams. Finally, we also note the use of additional, open-access data to augment training data (denoted by Add. in the table).

In the end, the best approaches (which reached values of OA higher than 70%) were based on ensemble methods: RFs (or recent related developments such as canonical correlation forests (CCFs) [44]), boosting (with also recent evolutions such as XGBoost [45]), and multiple classifier systems that aggregate the outputs of several classifiers. This is in line with what has been observed in several recent competitions, as well as in a few past IEEE GRSS data fusion contests [23], [24], [27]: Using ensemble models averaging over single classifiers helps filling

TABLE III
TOP 10 RESULTS WITH PERFORMANCE MEASURES AND TYPE OF APPROACH USED: RF, Bo., DL, Exp., ADDITIONAL TRAINING DATA (ADD.)

#	Team	Approach					OA (%)	Kappa	No. classes in the submitted maps
		RF	Bo.	DL	Exp.	Add.			
1	WYZZ	✓			✓		74.94	0.71	15
2	AGT	✓	✓	✓	✓	✓	72.63	0.68	14
3	Camilasa	✓			✓		72.34	0.68	16
4	nanjingxyy	✓			✓		69.89	0.65	13
5	FIMO, National University, Vietnam	✓	✓		✓	✓	67.37	0.62	16
6	Wuhan						66.59	0.62	16
7	aboulch, ONERA, France			✓		✓	64.30	0.59	14
8	on_by, Xidian University, China			✓			62.22	0.55	7
9	rainbow1						60.76	0.54	9
10	Sonic, Wuhan University, China			✓			60.22	0.55	14

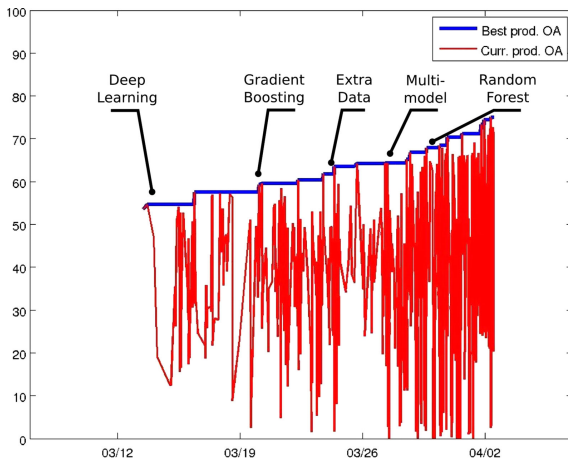


Fig. 3. Evolution of the performances of the submissions during the evaluation phase: OA over time (red) and best OA (blue).

the last mile of the performance gap, since they allow classifiers to complement one another and partly cope with relatively small numbers of training instances. Indeed, the number of training samples in the dataset of DFC17 was not small but not extremely large either, which was probably the reason why no pure DL approaches ranked among the winners—a difference in scientific outcome with respect to classification results observed among the best-ranking submissions of the previous two editions of the DFC [28], [30].

Fig. 3 displays the evolution of the performances of the maps submitted over the time of the evaluation phase. It shows that DL models were ready to use soon after the release of the test data (first submissions were received less than 12 h after opening the server) thus establishing an acceptable baseline with 51.4% of OA. After the first week however, the lead was taken on by teams exploiting ensemble methods like extreme gradient boost or RFs, showing that these approaches can take full advantage of imbalanced and sparse data once the right hyperparameters have been found by tuning.

Figs. 4 and 5 show the prediction results for two of the test cities: Amsterdam and Chicago, respectively. In both the maps, the city contour could be well distinguished, and few errors were made between urban and natural categories. However, the

different algorithms led to visually very diverse maps: for example, some confused dense and scattered vegetation, and others simply did not retrieve some classes. In Fig. 5(b) or (d), one can appreciate the benefits of fusing multisource data to obtain maps that comprehensively detect all considered land covers and land uses: there is no mis- or unclassified area as in Fig. 5(a) and (c).

In the next sections, the solutions proposed by the first and second ranked teams are presented. In these sections, the authors will detail their design, as well as provide additional visual results and insight of their proposed solutions.

IV. FIRST-PLACE TEAM: WXYZ

This section describes the algorithm developed by the first-place team and reports the results with a special focus on analyzing feature importance and the impact of sampling methods for constructing the training datasets. The algorithm is based on decision tree ensemble classifiers, namely CCFs [44] and RoFs [46], using spatial and spectral features extracted from Landsat 8 and OSM data.

A. Proposed Framework

The algorithm follows an information flow summarized in Fig. 6, which comprises four steps: preprocessing, feature extraction, classification, and postprocessing. Each step is detailed in the following sections (Sections IV-A to IV-D). The number of datasets and the size of training data provided in the contest were 16 Landsat 8 images, and 81 845 pixels, respectively. Therefore, a particular emphasis in the proposed framework was dedicated to fast, automatic, yet effective approaches to achieve accurate classification maps in an acceptable CPU processing time.

1) *Preprocessing*: Among the multimodal data mentioned in Section II, only Landsat 8 and OSM data were used as inputs. There are three reasons why Sentinel-2 data were not used in the proposed framework: 1) there are no long-wave infrared bands; 2) there is only one temporal image for each city; and 3) scattered clouds are included in the scenes of Hong Kong, Paris, and Xi'an. Owing to the multitemporality of the Landsat data, temporal-spectral variability can be taken into account to train the classifiers effectively. The original Landsat 8

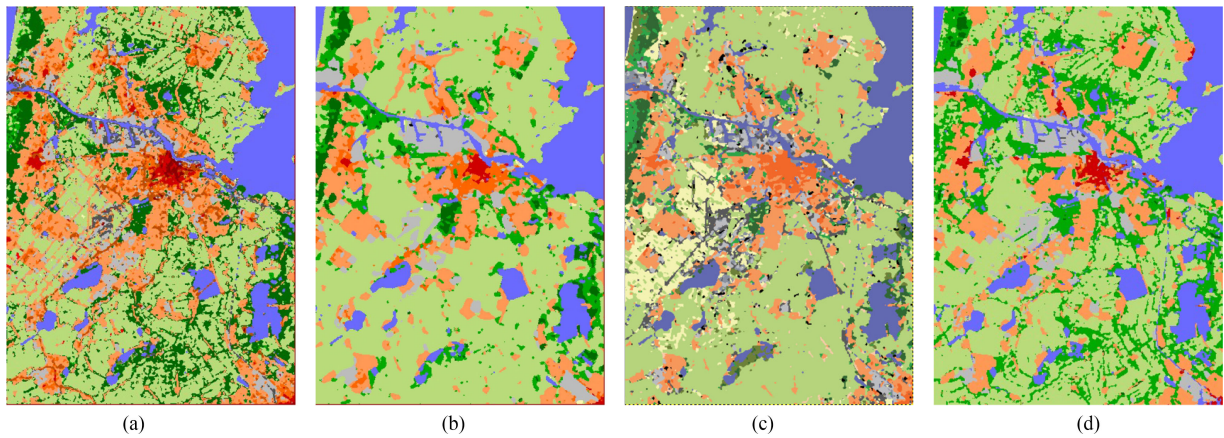


Fig. 4. Predicted classification maps for Amsterdam. From left to right: WXYZ, AGT, Camilasa, and nanjingxyy teams.

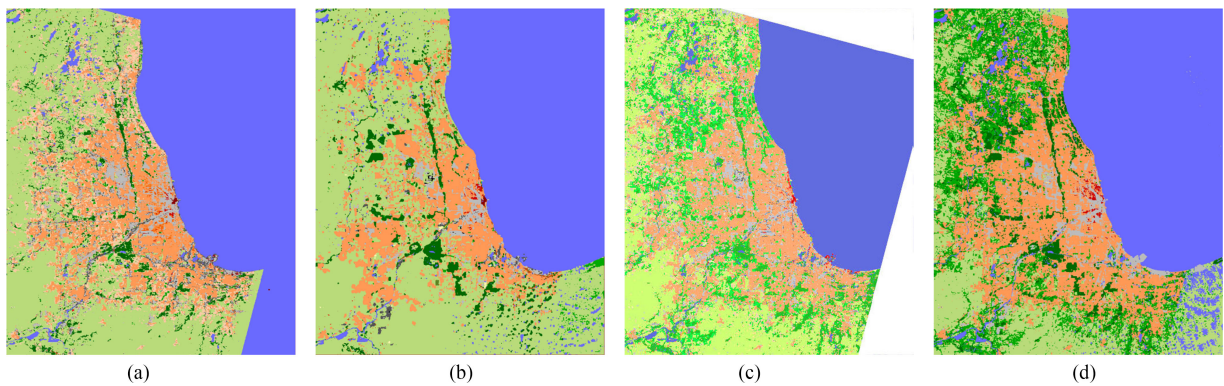


Fig. 5. Predicted classification maps for Chicago. From left to right: WXYZ, AGT, Camilasa, and nanjingxyy teams.

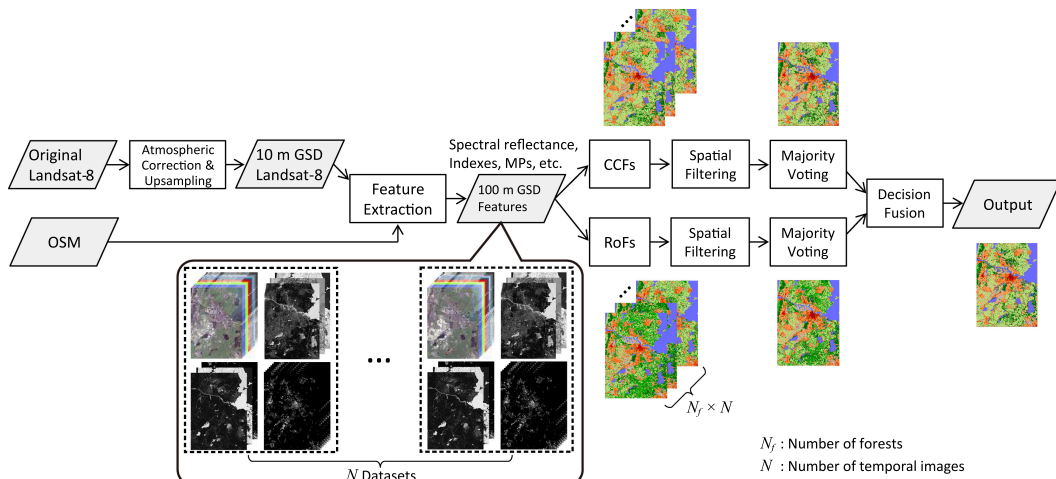


Fig. 6. Flowchart of the algorithm developed by the first ranked team (WXYZ).

images were downloaded via Amazon Simple Storage Service (Amazon S3). Ground sampling distances (GSDs) of the Landsat 8 data are 15, 30, and 100 m for panchromatic, visible and short-wave infrared, and long-wave infrared bands, respectively. All 11 bands were taken into account as inputs, as detailed below.

Atmospheric correction and haze removal were performed to eliminate atmospheric effects in the original Landsat 8 multi-

spectral bands (bands 1–7 and 9) using ATCOR-2/3, version 9.0.0 [47]. The panchromatic and long-wave infrared bands (bands 8, 10, and 11) were normalized between 0 and 1 in order to have input spaces of comparable numeric ranges. All bands were upsampled at a GSD of 10 m based on bicubic interpolation so that feature extraction could be easily performed at a GSD of 100 m. OSM layers were binary, where 0 and 1 mean absence and presence of elements, respectively, and spatially

downsampled to a GSD of 10 m to reduce the computational complexity in the subsequent processing steps.

2) *Feature Extraction*: Handcrafted features suitable for the classification of LCZs were extracted from the Landsat 8 and OSM images. Since decision tree ensemble classifiers can identify important features from high-dimensional input features, in the proposed framework various features considered or expected to be effective for the LCZ classification problem were used as inputs. A total of 43 features, including spectral reflectance (22), spectral indices (6), OSM features (3), and spatial features (12), were extracted at the 100-m GSD: as described in detail in the following.

- 1) Mean and standard deviation were computed for each patch of 10×10 pixels for all bands of the 10-m GSD Landsat 8 data, leading to 22 features. The vector of mean values represents the mean spectral signature at the 100-m GSD, and that of standard deviation values indicates the degree of spectral variability. Although box filtering was exactly the same as the one used for the data provided in the contest, it was reprocessed on the reflectance data in order to use physical values as inputs.
- 2) Three spectral indices were computed from the 10-m GSD Landsat 8 data, namely the normalized difference vegetation index (NDVI), the normalized difference water index (NDWI), and the bare soil index (BSI). The advantage of using these indices was already shown for LCZ classification [8]. In particular, NDVI is known to be effective to distinguish the compact and open LCZs. In the same way as spectral reflectance, mean and standard deviation were also computed for each patch of 10×10 pixels of NDVI, NDWI, and BSI, resulting in six features.
- 3) The OSM images were also downsampled at the 100-m GSD by box filtering. Since the OSM layers are binary, the three features give the proportions of “buildings,” “land use,” and “water” at each pixel.
- 4) Spatial information was extracted from the 10-m GSD NDVI and OSM “building” images by calculating morphological profiles (MPs) [48] composed of opening and closing by reconstruction. A circular structuring element with the sizes of 3×3 , 5×5 , and 7×7 was taken into account, and thus 12 features were obtained. All MPs were spatially downsampled at the 100-m GSD by box filtering.

In the proposed framework, multitemporal Landsat 8 images are treated as different data samples, i.e., if there are N Landsat 8 images for one training city that have P pixels as ground truth, we have $N \times P$ training samples. In the same manner, for each test city, we can obtain N classification maps using each base classifier (RoF or CCF). As there is no available multitemporal OSM data for the studied areas, OSM was utilized individually for each city together with all multitemporal Landsat 8 data samples. Note that some Landsat 8 images that included scattered clouds were not used.

3) *Classification*: Two decision tree ensemble methods, namely CCFs⁶ [44] and RoFs [46], were used for classification. Generally, decision tree ensemble methods have the following advantages:

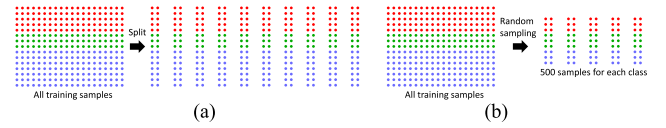


Fig. 7. (a) Uniform sampling and (b) imbalance-corrected sampling for constructing training datasets.

- 1) they are robust to the application to high-dimensional input features;
- 2) they are robust to missing data;
- 3) they perform out-of-sample prediction rapidly;
- 4) they require only slight parameter tuning;
- 5) they are capable of analyzing feature importance.

These advantages are suitable for the LCZ classification contest that includes the following challenges:

- 1) various types of features, such as the aforementioned handcrafted features, could be extracted from the input multimodal dataset, thus resulting in a high-dimensional input image;
- 2) there are missing data in the OSM layers;
- 3)4) the time for testing was limited;
- 5) it is preferable to recognize which features are important.

RoFs are a decision tree ensemble classifier based on data transformation (or feature extraction) and random subspaces [46]. Unlike RF, RoFs use random splits of features and unsupervised principal component analysis for rotation of feature axes before constructing decision trees. The rotation of the feature axes aims at improving the accuracy and diversity of individual base classifiers simultaneously. CCFs were proposed recently as a decision tree ensemble method based on *supervised* feature extraction [44]. CCFs first use bagging like RF and then perform canonical correlation analysis between features and labels on each training subset for the rotation of the feature axes. Finally, hyperplane splits of each decision tree are calculated in a rotated-feature space directed by the label information. The superior performance of these two ensemble classifiers over RF was already proven in the remote sensing community in terms of classification accuracy and generalization capability with acceptable computational complexity [49]–[51].

For both methods, the number of trees was set to 20 with reference to the work reported in [51]. A total of 15 training datasets were constructed. The first ten sets were prepared by splitting the whole training data into ten subsets without replacement [i.e., uniform sampling; see Fig. 7(a)]. The other five sets were created by extracting the same number of training samples (i.e., 500) randomly for all classes [i.e., imbalance-corrected sampling; see Fig. 7(b)]. In this way, it was possible to increase the diversity of the forests, which played an important role to improve the classification performance of ensemble classifiers. Although classes 10, 15, and 16 had fewer than 500 pixels for training, the total number of training samples for each class was more than 500 since each city had multitemporal Landsat 8 images and different temporal observations were used as independent samples. Finally, 15 different forests were built for each of the CCFs and RoFs based on all training sets.

⁶The source code is available at <https://bitbucket.org/twgr/ccf>

TABLE IV
FINAL OAS (%) OBTAINED BY ROFS, CCFs, AND THE WXYZ RESULT

	Amsterdam	Chicago	Madrid	Xi'an	All cities
RoFs	64.56	71.34	81.28	51.95	71.63
CCFs	71.65	73.83	80.15	59.37	73.96
First ranked	71.65	73.83	81.28	59.37	74.94

4) *Postprocessing*: A 3×3 majority filter was applied to all classification maps to reduce the labeling uncertainty and salt and pepper appearance of the labeled pixels. The final classification map was obtained for each ensemble method using majority voting on $15 \times N$ classification maps, where N is the number of multitemporal Landsat 8 images for each city. The visual comparison of the classification maps revealed that CCF substantially outperformed RoF on Amsterdam, Chicago, and Xi'an. However, CCF caused considerable misclassification on large areas of the Madrid dataset, and therefore, we chose the classification map obtained by RoF for this particular city.

B. Results and Discussion

Figs. 4(a) and 5(a) show the LCZ classification maps obtained by the presented algorithm for Amsterdam and Chicago, respectively. The WXYZ team achieved a value of OA of 74.94% and a kappa coefficient of 0.71, as shown in Table III. The OAs obtained by RoFs, CCFs, and the submitted (and first place) result for each city are shown in Table IV.⁷ Overall, CCFs tended to outperform RoFs; however, RoFs showed a slightly better result for Madrid, consistently with the aforementioned remark on the misclassifications on the data of this city.

Table V shows the confusion matrix with producer's (PA) and user's accuracies (UA). As in [40], the PAs range largely. The considered classifiers could achieve high accuracies for the classes with sufficient training data (e.g., classes 6, 8, 11, 14, and 17). Several class pairs (e.g., classes 2 and 3, classes 4 and 5, classes 9 and 14) were confused due to similar spectral-spatial features while they would require height information for further accuracy improvement.

One of the important findings in [40] was that the classification accuracy was improved by 4.65% in terms of OA and 0.05 in terms of kappa coefficient by integrating the classification results obtained by using the five sets of imbalance-corrected training data: Table VI shows the change of the confusion matrix when adding such data to the training set. Each cell corresponds to the difference between the number of samples before and after adding the imbalance-corrected training data. The cell's background color indicates the proportion of the increased (or decreased) number of samples, divided by the total number of test samples per class, which ranges from -1 to 1 where -1 , 0 , and 1 correspond to red, white, and blue, respectively. Blue cells on the diagonal of the confusion matrix show the improved classes (true positives) with the imbalance-corrected training

TABLE V
CONFUSION MATRIX OF THE WXYZ TEAM, MODIFIED FROM [40]

		Ground truth																	UA (%)
		1	2	3	4	5	6	8	9	10	11	12	13	14	15	16	17		
Prediction	1	167	151	15	1	1	2	20	0	24	0	0	0	0	0	1	0	43.72	
	2	2	3033	321	97	166	41	16	0	3	0	0	0	0	0	0	0	82.44	
	3	0	181	624	125	6	230	375	0	46	0	0	0	84	67	0	0	35.90	
	4	12	350	18	1243	121	52	81	0	8	1	15	0	1	0	0	9	65.04	
	5	0	469	202	242	1540	1223	74	0	1	5	114	0	1	11	0	0	39.67	
	6	1	343	168	250	252	6432	1004	86	85	65	227	12	330	25	0	16	69.19	
	8	59	294	149	290	154	89	9055	0	625	25	66	0	51	429	15	80	79.56	
	9	0	0	0	2	0	13	124	18	1	10	10	0	7	13	0	11	8.61	
	10	0	0	0	15	1	5	305	1	92	5	21	0	3	107	16	23	15.49	
	11	0	0	0	0	0	6	14	28	20	2690	201	11	207	2	0	3	84.54	
	12	0	2	2	0	6	50	7	0	0	190	3230	131	454	300	0	6	73.78	
	13	0	11	2	0	7	66	3	10	0	2	391	940	57	11	0	0	62.67	
	14	1	57	18	2	1	55	96	929	6	177	233	190	11788	29	26	28	86.45	
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	
	16	0	0	16	0	0	0	45	0	9	0	0	0	1	78	294	6	65.48	
	17	0	1	0	3	0	1	11	0	0	0	20	0	10	32	39	4272	97.33	
	PA (%)		69.01	62.00	40.65	54.76	68.29	77.82	80.63	1.68	10.00	84.86	71.33	73.21	90.72	0.00	75.19	95.91	74.94
		OA (%)																	74.94

Background color indicates the proportion of the number of samples divided by the total number of test samples for each class, ranging from 0 (white) to the maximum (black).

TABLE VI
CHANGES IN THE CONFUSION MATRIX WHILE ADDING
IMBALANCE-CORRECTED TRAINING DATA

		Ground truth																	UA (%)
		1	2	3	4	5	6	8	9	10	11	12	13	14	15	16	17		
Prediction	1	8	20	14	-10	0	2	4	0	7	0	0	0	0	0	1	0	-3.60	
	2	0	342	4	83	55	0	7	0	-5	0	0	0	0	0	0	0	-1.84	
	3	-1	-187	-13	-100	-43	-11	-56	0	-33	0	0	-28	-11	0	0	0	7.22	
	4	11	336	13	994	95	25	3	0	8	0	-4	0	1	-1	0	9	5.90	
	5	-1	58	50	-28	263	91	11	0	0	4	83	0	1	11	0	0	1.43	
	6	-3	-400	-61	-853	-295	-146	-202	-7	-22	-20	-359	-3	-38	-28	0	-6	13.16	
	8	-13	-238	5	-83	-83	-33	-153	0	-48	-6	-7	-2	-10	-87	-9	-11	3.83	
	9	0	0	0	2	0	5	124	18	1	10	10	0	7	13	0	11	8.61	
	10	0	0	0	15	1	5	296	1	91	5	20	0	3	101	16	23	9.61	
	11	0	0	0	1	0	0	-5	0	0	-8	-213	-1	-9	0	0	-1	5.65	
	12	0	1	1	0	6	30	-6	0	0	23	526	71	22	11	0	3	0.50	
	13	0	11	2	0	7	66	3	10	0	1	385	939	55	11	0	0	52.67	
	14	1	57	-2	0	-6	-32	-14	-22	2	-9	-439	-1004	-3	-6	0	-1	8.43	
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	
	16	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	-4	-2	0.13	
	17	-2	0	0	-19	0	-2	-12	0	-1	0	-2	0	0	-14	-4	-25	1.20	
	PA (%)		3.31	7.0	-1.20	43.79	11.66	-1.77	1.36	1.68	9.89	0.25	11.62	73.13	-0.02	0.00	-1.02	-0.56	—
OA (%)																	4.65		

The cell color indicates the change of accuracy or the true/false positives (divided by the total number of test samples for each class), ranging from -1 to 1 (red: -1 , white: 0 , blue: $+1$).

data. On the other hand, red cells other than on the diagonal indicate mitigated confusions (false positives). We observed the following.

- 1) False positives for classes that have large numbers of training samples, such as classes 6, 8, and 14, were mitigated. This implies that the classification boundaries of these major classes are expanded in the feature space due to the imbalance of the number of training samples, and this expansion was reduced with the use of the imbalance-corrected training data.
- 2) Classes 4 and 13 show significant improvements, followed by classes 2, 5, 10, and 12. Among them, classes 4, 10, 12, and 13 have limited numbers of training samples. The imbalance-corrected training data contributed to improving classification accuracies of these small classes.

Fig. 8 shows the ranking of feature importance for the ensemble of CCFs using all training datasets. Green, blue, and yellow colors correspond to spectral reflectance, spectral indices, and OSM, respectively. Note that MPs of NDVI and the OSM building layer are categorized into spectral indices and OSM, respectively. We observed the following.

⁷The OAs of RoF and CCF on the test samples were separately computed for this paper after the end of the contest by submitting the corresponding maps to the DASE web platform.

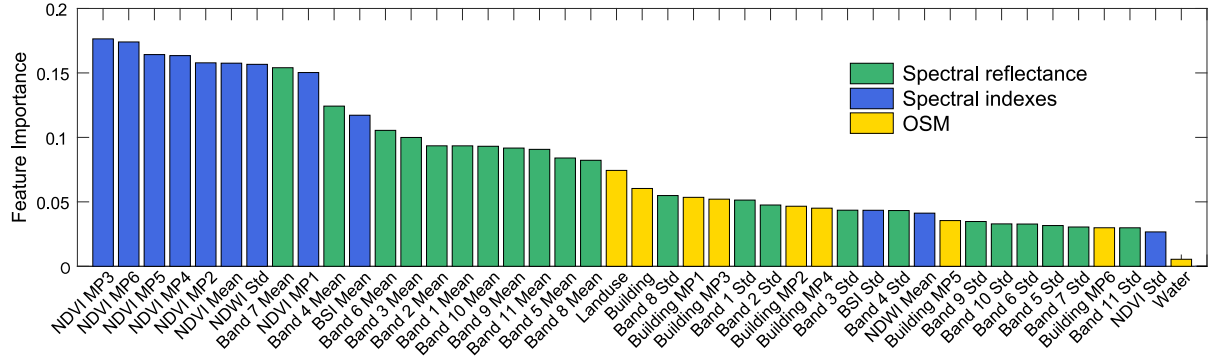


Fig. 8. Feature importance obtained by the CCF ensemble.

TABLE VII
LCZ MAP GENERATION TIMES FOR TESTING CITIES (WXYZ TEAM)

City	No. of images	Feature extr. time (s)	Prediction time (s)	Postprocessing time (s)
Amsterdam	4	33.4	379.8	1.1
Chicago	3	250.3	2646.5	8.5
Madrid	4	128.5	1718.7	4.0
Xi'an	4	59.1	767.4	2.5

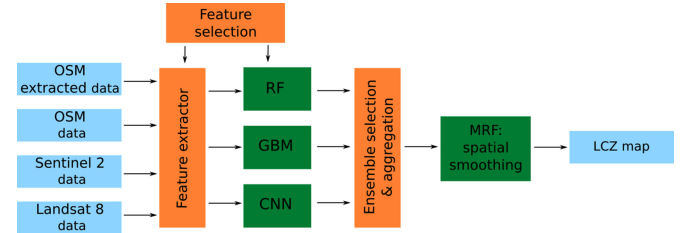


Fig. 9. LCZ classification pipeline of the AGT team.

- 1) Spectral indices show the highest importance among the three categories. In particular, NDVI is found as the most important feature, followed by NDWI and BSI.
- 2) Spectral reflectance information shows the second highest importance in the categories. The mean values that represent spectral signatures at a GSD of 100 m are more important than the standard deviation values that indicate the degree of spectral variability—a comment consistent with the limited spatial texture that can be appreciated at 100-m resolution in images of urban areas and of the surrounding suburban and vegetated regions.
- 3) The OSM-derived features were of lesser importance than the satellite-derived features for the trained classifiers. Among the three layers, “land use” shows the highest importance, followed by “building.” “Water” was identified as the least useful feature, most probably because it was not available for all cities (for instance, the “water” layer was unavailable for Berlin).

Table VII summarizes the processing times of feature extraction, prediction, and postprocessing on the test cities. Calculations were made on a PC with 4-core Intel(R) Core(TM) i7 CPU @3.1 GHz. It is shown that the most time-consuming part was the prediction step, accounting for approximately 90% of the calculation time of these three steps. This is because the prediction was repeated $15 \times N$ times for each city using 15 different forests and N Landsat 8 images. By adding five forests learned with the imbalance-corrected training datasets, the aforementioned improvement in classification accuracy (i.e., 4.65% of OA and 0.05 of kappa coefficient) was achieved in exchange for 1.5 times the prediction time.

V. SECOND-PLACE TEAM: AGT

In this section, we describe the ensemble system for LCZ classification proposed by the second-place team. This system was developed to rigorously address all challenges of the given dataset and the LCZ mapping problem in general: severe class imbalance (see class distribution in Table II Section II), heterogeneous nature of data, noise and varying quality of the multispectral images (MSI), nonconsistent number of samples across timesteps, limited number of annotated data as compared to the number of classes and the sample variability. The proposed method is based on the fusion of the provided image data (MSI from the Landsat 8 and Sentinel-2 satellites) with the additional information obtained from the OSM layers. Three types of classifiers were used within the classification ensemble: CNN, RF, and gradient boosting machines (GBM). RF and GBM were trained using handcrafted features following an automatic feature selection process, whereas CNNs were applied to the raw data directly. At the end, spatial smoothing using a Markov random field (MRF) was applied to enhance the resulting map.

A. Proposed Framework

The overall classification framework is depicted in Fig. 9 and contains the following modules: data collection/enrichment, feature extraction and selection, machine learning model learning and validation, model ensembling, and postprocessing of the classification results. In the following, we detail each of these steps.

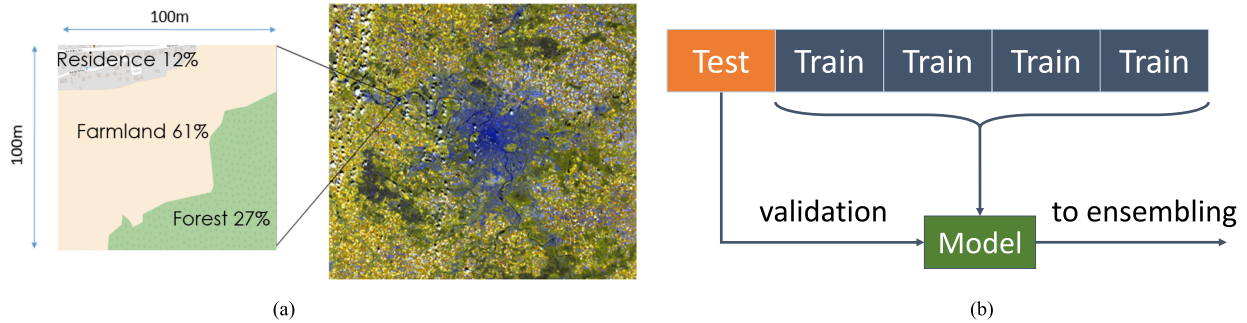


Fig. 10. Example of calculated relative area of OSM layers for a pixel and the cross-validation strategy used.

1) *Data Enrichment*: Additional Landsat 8 images were gathered from the EarthExplorer⁸ system and were combined with the given MSI (these images were only used for CNN training). A standard set of spectral features was extracted from these MSI (see below). The OSM service was used to extract additional crowdsourced site information data to provide additional informative features along with the OSM layers already provided by the contest organizers. To select the most important features, the feature importance property of the RF classifier was used.

2) *Features*: Several features were extracted either from the MSI stacks or from the OSM datasets.

- 1) *OSM*: Features extracted from OSM have been proven in the past to be a valuable type of information for LCZ classification problems [52]. On one hand, for some well-annotated metropolitan areas (e.g., Berlin) the incorporation of this information may provide a significant boost to the classification accuracy for many traditional classifiers. On the other hand, there are many areas where detailed coverage with OSM information is not available or is very noisy due either to human factors or to rapid site development. In such cases, the performance of a classifier can significantly degrade by the discrepancies between data sources and ground references. After an initial analysis of the OSM information that was provided as a part of training data, we realized that the quality of the OSM maps was not optimal and did not incorporate all the valuable information that is possible to extract from OSM (e.g. natural, land use, and waterway; see also Section IV-B). Motivated by potential gains in classification performance, we exploited other classification-relevant layers of OSM such as “leisure,” “military,” “nature,” “office,” “shop,” and “waterway.” These layers were represented in the feature set as relative coverage areas. To this end, for each pixel in the dataset, the relative coverage area of each of the available OSM layers was calculated, as depicted in Fig. 10(a). Moreover, in order to enrich the dataset with information about building height and elevation, we engineered the following features: amount of buildings located in the polygon, average and maximum floor number of the buildings, and average and maximum height of the buildings.

- 2) *Spectral Features*: Additionally to the OSM features, a set of standard spectral features was extracted from the MSI. They included the NDVI, NDWI, and BSI already mentioned in Section IV-A2, as well as the normalized difference moisture index [53], the advanced vegetation index [53], the shadow index [53], the spectral angle mapper (SAM) [54], and the minimum noise fraction (MNF) [55]. The extraction of each of these features generated a two-dimensional (2-D) matrix except for the last two features, which corresponded to 3-D matrices where the third dimension was the number of bands for MNF and the number of classes for SAM. As a reminder, the SAM feature extraction calculates the spectral angle between the image spectra and a known spectra (or *endmember*). It is robust to differences in illumination, since it uses the vector direction rather than the vector length. The MNF transform computes the normalized linear combination of the original bands that maximizes the signal-to-noise ratio. For each class, a SAM feature was calculated. The reference signature was calculated as the mean spectrum over all the pixels belonging to the respective class.

Since only one Sentinel-2 image was provided per site, while two or more Landsat 8 images were made available, we decided to stack the Sentinel-2 image with each available Landsat 8 image. This way, an extended 3-D data cube was constructed with a fixed spatial dimension per site and a spectral band dimension equal to $N = N_L + N_S$, where N_L and N_S are the numbers of spectral bands from Landsat 8 and Sentinel-2, respectively.

- 3) *Classification and Ensembling*: Ensemble methods in classification are generally used to leverage the power of multiple diverse models and achieve higher prediction performance. Based on the bias-variance tradeoff concept, ensemble methods have proven their high operational efficiency in many classification scenarios, including remote sensing [56] and image analysis. We applied the idea of ensembles in two ways: by using an ensemble-based classifier to select relevant features and by combining the output of several classifiers in order to fuse the single LCZ maps.

To select the most discriminative features, we trained an RF classifier using all features extracted from the training dataset and additional data, and we collected and monitored the feature importance value provided by the RF classifier. We additionally included an extra feature that was generated randomly (fol-

⁸<https://earthexplorer.usgs.gov>

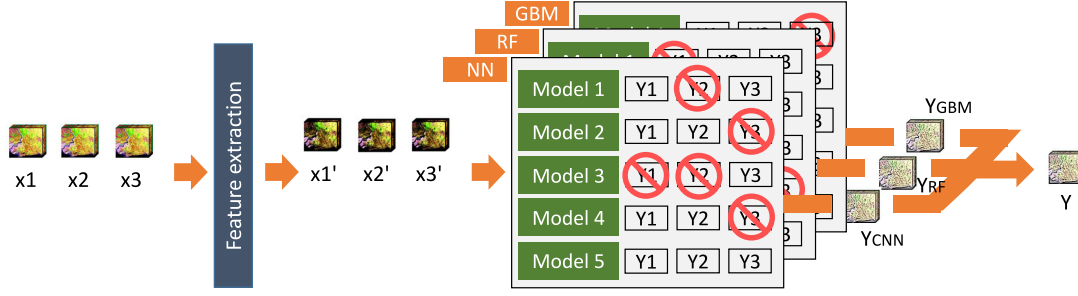
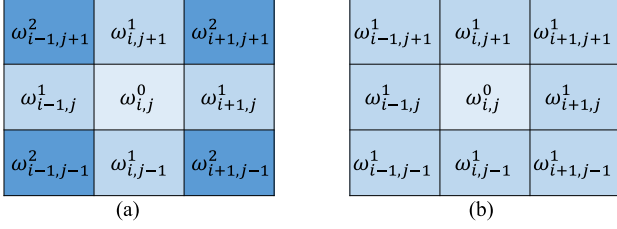


Fig. 11. Schematic flowchart of the LCZ classification pipeline of AGT.

Fig. 12. 3×3 kernels used in the convolutional layer of our CNN. (a) Kernel with two weights. (b) Kernel with three weights.TABLE VIII
EXPERIMENTAL RESULTS BEFORE AND AFTER MRF-BASED POSTPROCESSING

	Overall accuracy	Kappa	Number of classes predicted
Without postprocessing	65.53%	0.60	16
Final results with postprocessing	72.63%	0.68	14

lowing a uniform distribution on the interval $[0, 1]$) to get an intuition of the usefulness of the derived features. We selected a feature k for further model training only if its importance was greater than the importance of the random extra feature, i.e., $I_k > I_{\text{random}}$. After performing such filtering, we obtained a total of 90 features.

To establish a cross-validation framework and learn the models, we split the training dataset into five folds, each one using four cities for training and the fifth for validation. On every iteration, a model for every classifier was trained using the training data folds and validated on the validation one. The used cross-validation framework is depicted in Fig. 10(b).

The citywise split was dictated by the desire to leverage the advantage of CNN that allows the processing of spatial data as it takes neighboring pixels into account for classification. Obviously, splitting in this way imposes some limitations comparing to pixelwise splitting.

The design of the classification ensemble consisted of three layers: the site temporal samples combiner, the first-level model combiner, and the second-level model combiner. Fig. 11 presents the schematic flowchart of the LCZ classification pipeline in the case of three temporal inputs. Here, X_1, X_2, X_3 are the raw multitemporal MSIs, X_1', X_2', X_3' are the input multitemporal features, Y_1, Y_2, Y_3 are the predicted LCZ maps from a single model, $Y_{\text{CNN}}, Y_{\text{RF}}, Y_{\text{XGB}}$ are the combined LCZ

maps from the corresponding models, and Y is the resulting and final classification map.

- 1) The site temporal samples combiner is responsible for the aggregation of quality classification maps that are coming from the same site (city). We estimated the quality of a predicted map based on the average entropy of the multinomial class distribution predicted by a classifier: A classifier that is confident in its prediction shows low entropy. The entropy

$$-\sum_{\text{class}} p_{\text{class}} \cdot \log(p_{\text{class}}) \quad (2)$$

was computed for each spatial sample of every temporal prediction map. For the combination, we picked only the top $\lambda\%$ of the predicted maps showing the highest average entropy, where λ is a predefined value adapted experimentally for every model type (i.e. $\lambda = 5\%$ for the CNN, $\lambda = 30\%$ for the RF, and $\lambda = 50\%$ for the GBM).

- 2) On the first-level model combiner, we aggregated five LCZ maps coming from the site temporal samples combiners.
- 3) The final combination was done on the second-level model combiner, where three LCZ maps, issued by every type of model, are combined. The combination on all levels was done by averaging classifier posterior probabilities, while in the general case a weighted average approach could be applied (however, a reliable optimization of weights would be required in that case).

We chose RF, GBM, and CNN as the base models for the ensemble since RF is a well-known robust and stable classifier, GBM is the state of the art in many classification tasks based on boosting approaches and is not prone to overfitting, and CNN because they can efficiently capture spatial relationships. The three classifiers are considered to be the state of the art in the computer vision domain. With respect to classifier settings, the following peculiarities are worth mentioning.

- 1) We used a weighted version of RF, where we assigned individual weights to every class, calculated according to the inverse class frequencies in the training data. With that approach, we were aiming to tackle the class imbalance problem by making rare classes more significant during the training phase, so that they have more chances to be discovered during the classification step.
- 2) In order to select the optimal set of parameters for both RF and GBM, we employed a grid search technique by sys-

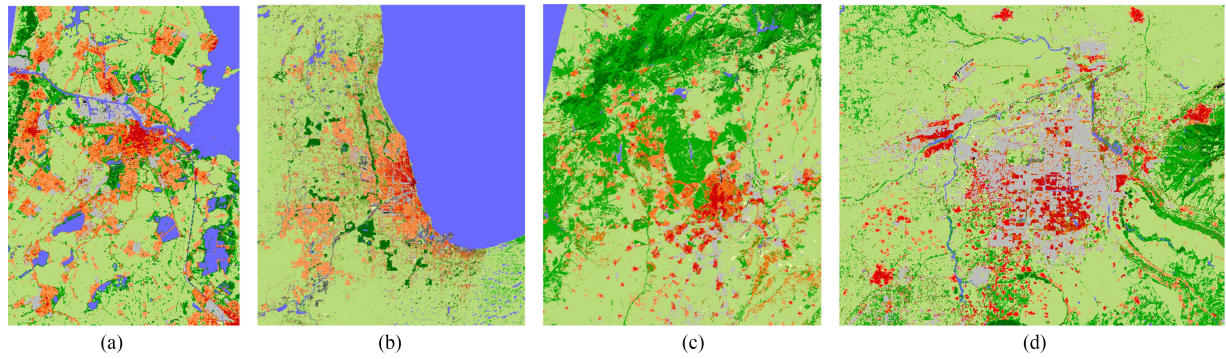


Fig. 13. Resulting LCZ maps for the test cities before postprocessing. (a) Amsterdam. (b) Chicago. (c) Madrid. (d) Xi'an.

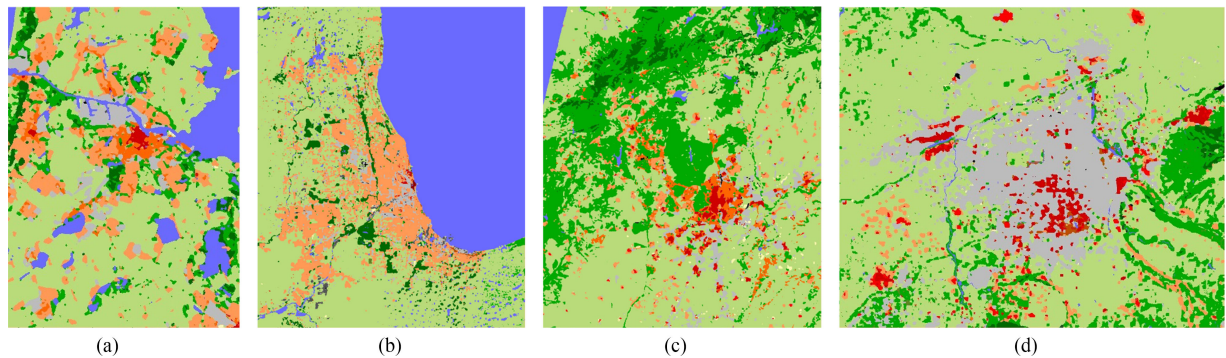


Fig. 14. Resulting LCZ maps for the test cities after postprocessing. (a) Amsterdam. (b) Chicago. (c) Madrid. (d) Xi'an.

tematically going over all possible parameter candidates with a particular step and selected the ones that provided the highest performance on the validation set. Note should be taken on the fact that the hyperparameter tuning showed minor effects on the resulting performance.

- 3) For CNN, we trained several architectures in order to find the one that suited the best. The architectures we considered were shallow, containing only three layers to prevent overfitting. The first layer was a 1×1 convolution with batch normalization and a tanh or ReLU nonlinearity. For the second layer, we designed a special type of convolutional kernel with high degree of parameter sharing. The kernel used was symmetric with respect to the origin in order to share more parameters and achieve rotation invariance (as shown in [57]; see Fig. 12). With these kernel topologies, we could achieve the highest performance minimizing the risk of overfitting. The third layer was the softmax layer that was connected directly to the convolutional layer and produced posterior probability estimates for all the LCZ classes.
- 4) During the cross-validation phase for the CNN, we performed early stopping and model selection based on the accuracy values on the validation set.

4) *MRF*: As discussed above, RF and GBM are not able to capture spatial relationships in the MSIs and are prone to produce noisy and unstable LCZ outputs. On the contrary, CNN inherently utilizes neighborhood information and outputs more homogeneous class labels within neighboring regions. The combination of RF, GBM, and CNN for some areas of MSI neutral-

izes this positive effect of CNN, resulting in LCZ maps covered with salt and pepper patterns.

To overcome this issue even further, we applied a spatial smoothing by modeling the resulting label field as an MRF. Using an MRF model allows considering the class-conditional probabilities of a classifier (or ensemble of classifiers in case of probabilistic outputs) and reassigning class labels based on their spatial context and classification uncertainty. In previous works [56], [58] it was shown that, under an MRF model assumption, the iterated conditional modes (ICM) algorithm could be successfully applied, starting from an initial map, in order to suppress mislabeled pixels and obtain a smooth and stable output. One of the practical challenges when applying MRF with ICM is to find the attraction parameter that acts as the regularization, directly affecting the bias-variance tradeoff. It should be chosen with much care since too small values do not bring significant improvement while too large values can lead to oversmoothing, resulting in a degraded map quality.

We optimized the attraction parameter for every class using the Tikhonov regularization approach [56] based on our cross-validation framework described above.

B. Results and Discussion

Table VIII reports the aggregated OA, the corresponding kappa measure, and the number of predicted classes for the testing cities.

It is clear that the MRF-based postprocessing was able to significantly increase the performance of the overall approach providing a boost in OA of 7.1% and for kappa of 0.08. A

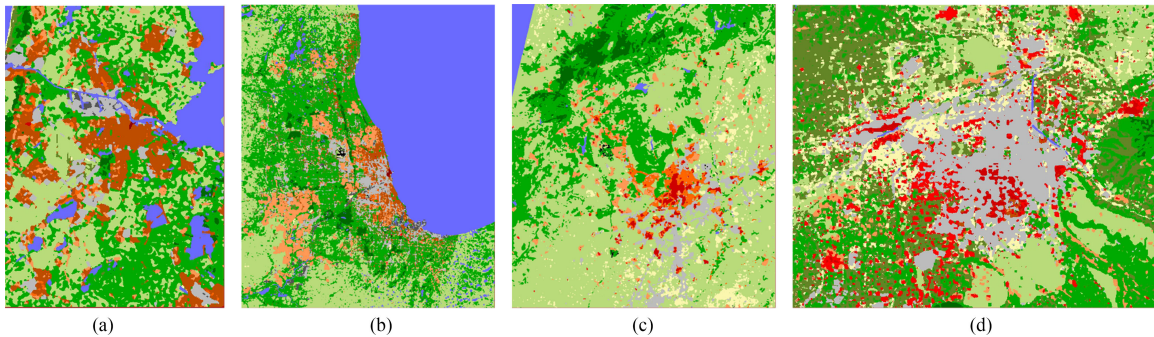


TABLE IX
CONFUSION MATRIX OF THE AGT TEAM

		Ground truth																	UA (%)
		1	2	3	4	5	6	8	9	10	11	12	13	14	15	16	17		
Prediction	1	227	25	13	4	0	0	0	0	0	0	2	0	0	1	0	11	80.21	
	2	4	3666	344	475	743	108	3	0	0	0	10	0	0	0	0	14	68.31	
	3	0	66	145	0	12	24	15	0	4	0	0	0	0	0	0	0	65.59	
	4	0	34	8	284	0	30	34	0	4	0	3	0	0	0	4	30	64.51	
	5	0	332	145	234	1104	712	16	0	5	95	0	1	1	0	0	8	11.61	
	6	0	200	243	87	205	6312	876	467	0	57	122	0	0	20	8	4	73.90	
	8	10	520	547	1181	160	131	9652	0	616	26	8	0	58	598	200	121	69.80	
	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	
	10	0	0	0	0	0	0	48	0	153	0	0	0	0	13	0	0	71.50	
	11	0	0	0	0	0	0	5	11	0	2186	91	11	298	0	0	21	83.34	
	12	0	12	9	1	5	251	23	0	1	586	3857	272	164	277	0	174	68.48	
	13	0	0	2	0	0	0	0	0	0	3	0	0	0	190	0	0	0	
	14	0	15	61	4	25	696	509	654	55	303	322	1001	12269	70	41	37	76.39	
	15	2	0	8	0	0	0	5	0	0	0	0	0	0	30	0	0	81.08	
	16	0	8	18	0	1	1	6	0	0	0	0	0	0	79	101	107	61.97	
	17	0	12	0	0	0	0	38	0	91	0	18	0	13	11	38	4033	94.80	
	PA (%)		93.80	74.94	9.45	12.51	48.96	76.37	85.95	0	16.63	68.96	85.18	0	94.42	2.72	25.83	90.55	
																		OA (%)	72.63

general overview of the postprocessing effect can be seen by comparing the classification maps before (see Fig. 13) and after postprocessing (see Fig. 14).

On one hand, one can see that the effect of reassigning the labels based on the neighborhood made the classification maps smoother and allowed suppressing misclassifications due to noise and other data-related issues. On the other hand, post-processing suppressed small classes significantly, resulting in the prediction of only 14 of the 17 classes. Generally, while designing our approach, we could observe the tradeoff between the OA value and the number of detected classes. Such dependence imposed additional limitations on the way we tackled the class imbalance problem: resampling-based methods such as the synthetic minority oversampling technique (SMOTE) [59] or adapted versions of resampling as in [60] did not bring improvements to the OA though allowed to discover more underrepresented classes. In Fig. 15, classification maps obtained using SMOTE are presented where originally underrepresented classes became prominent and in some cases led to wrong classifications. As we were to maximize OA, we did not include these balancing methods into the final approach.

While training, we observed that due to the limited amount of annotated data, the CNNs were prone to overfit, which made it challenging to find the correct model parameters. The CNNs were considerably outperformed by the tree-based models (RF and GBM), but still provided smooth maps that, in the end, were beneficial for the ensemble. In order not to lose spatial structure of the data, we employed a geographical approach as a cross-validation strategy, i.e., “onefold—one city.” Stratified cross validation of independent pixels could be another strategy that would allow folds to contain data from all training cities, thus inducing more accurate models. We provide the confusion matrix for the prediction on the four testing cities in Table IX.

As can be seen from this confusion matrix, the number of samples for the testing set varied a lot and we were able to quite accurately detect samples coming from most large classes. More than 50% of all testing samples corresponded to classes 6, 8, and 14 indicating that discovering only these three classes would provide already 50% of OA. Conversely, small classes (e.g., class 1) did not significantly contribute to the OA. In particular, there are two distinct “superclasses” that can be seen from the confusion matrix: $\{1, 2, 3, 4, 5, 6, 8, 9\}$ and $\{11, 12, 13, 14, 15\}$. Most errors occurred within these two groups, thus indicating the need for more informative features and better preprocess-

ing and noise reduction techniques. The first group corresponds to the *built types* within the LCZ classes, whereas the second group corresponds to the *land cover types*. In order to reduce the confusion within the *built types*, additional features representing the height of the buildings would significantly improve the recognition rate. A first choice to extract height information is to consult the OSM data sources. Other experiments, omitted for brevity, confirmed that this choice indeed helped the classifier to distinguish the *built types* classes. A second choice to extract the height would be based on social media sources, although a big challenge here would be to filter out the abundant noise within this type of data. To reduce the confusion within the *land cover* group, one could benefit from a feature that would contain agricultural information for the corresponding pixels. In this case, the LCZ classes 13 and 14 corresponding to agricultural zones could be disentangled from classes 11 and 12 that are dense and scattered forest, respectively. In particular, regional agricultural GIS systems could be employed to extract this kind of information, though it requires a large coordination effort and additional costs. As a conclusive remark, we provide Table X that reports the processing times for every step of LCZ map generation: feature extraction, prediction, ensembling, and postprocessing. All the calculations were done on a PC with 4-core Intel Xeon CPU E5620 @2.40 GHz. From the table, it is clear that the feature extraction and postprocessing steps are consuming the most of the computation time and are comparable in absolute values. Although the postprocessing applied practically doubles the overall computation time for each city, it is capable of providing more than 7% improvement of OA and 0.08 increase of kappa measure according to Table VIII.

TABLE X
LCZ MAP GENERATION TIMES FOR TESTING CITIES (AGT TEAM)

City	No. of images	Feature extr. time (s)	Prediction time (s)	Ensemb. time (s)	Postprocessing time (s)
Amsterdam	5	141.5	77.0	1.9	95.8
Chicago	4	972.3	553.3	12.6	893.9
Madrid	5	582.6	262.7	12.4	429.9
Xi'an	4	208.1	103.7	1.8	198.2

VI. CONCLUSION

In this paper, we summarized the organization and presented the scientific results of the 2017 IEEE GRSS Data Fusion Contest, organized by the IEEE GRSS IADF TC. We described the dataset and the overall outcomes of the competition, by first presenting the overall results of the ten top-ranked teams and then focusing on the strategies proposed by the first-place and second-place teams. These teams made use of both the image and OSM data available and developed methodologies rooted in the latest advances in computer vision and machine learning. Special focus was given to ensemble methods to fuse classification maps obtained by different methodologies or with different data sources.

By observing the evolution of the outcomes during the three weeks of the test phase, we noticed that participants first tried to use recent computer vision methods (DL in particular) to solve the classification problem proposed to the community. However, given the limited amount of training data and the specificities of remote sensing problems, these results were quickly overrun by methods encoding priors about remote sensing data (such as, for instance, the thorough atmospheric correction of the Landsat scenes applied by the winners) or by the use of extra open data (such as the full resolution scenes used by both the first-place and second-place teams), or sets of extra images and OSM layers (used by the second-place team). This shows that, to be successful for complex tasks such as LCZ classification, one needs to use all the available types of information that certainly includes multimodal remote sensing data [61], but also vector data or ground information [62]. A true multimodal system for the classification is needed and, through this competition, we showed examples of how beneficial it could be.

The data will remain downloadable for free from the IEEE GRSS website.⁹ Ground references were made available for the training cities, and the DASE evaluation server will remain open and welcomes new submissions to improve the results reported in this paper. We hope that these data will serve to push remote sensing data fusion not only to further improve methodologically, but also to explore out of the purely image domain and to contribute to the LCZ mapping community.

ACKNOWLEDGMENT

The authors would like to thank the WUDAPT (<http://www.wudapt.org>) and GeoWIKI (<http://geo-wiki.org>) initiatives for

providing the data packages used in this study, the DASE benchmarking platform (<http://dase.ticnumaerospace.com>), and the IEEE GRSS IADF TC. For their contributions to the LCZ samples, the authors thank in particular C. Ren, D. Fenner, D. Milosevic, G. Dumas, M. De Fatima Andrade, M. Foley, O. Brousse, and R. Wang. Landsat 8 data are available from the U.S. Geological Survey (<https://www.usgs.gov>). OSM Data OpenStreetMap contributors are available under the Open Database Licence (<http://www.openstreetmap.org/copyright>). Original Copernicus Sentinel Data 2016 are available from the ESA (<https://sentinel.esa.int>). The authors also acknowledge I. Friedrich for her support in the preparation of the DFC 2017 dataset. The AGT team would like to thank their colleagues J. Louradour, D. Trofimova, and C. Debes. D. Tuia acknowledges the support of the Swiss National Science Foundation.

REFERENCES

- [1] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bull. Amer. Meteorol. Soc.*, vol. 93, pp. 1879–1900, 2012.
- [2] H. Taubenböck, T. Esch, A. Felbier, M. Wiesner, A. Roth, and S. Dech, "Monitoring urbanization in mega cities from space," *Remote Sens. Environ.*, vol. 117, pp. 162–176, 2012.
- [3] T. Esch *et al.*, "Urban footprint processor—Fully automated processing chain generating settlement masks from global data of the TanDEM-X mission," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 6, pp. 1617–1621, Nov. 2013.
- [4] T. Esch *et al.*, "Earth observation-supported service platform for the development and provision of thematic information on the built environment—The TEP-Urban project," in *Proc. 2017 Joint Urban Remote Sens. Event*, Dubai, UAE, Mar. 6–8, 2017.
- [5] M. Pesaresi *et al.*, "A global human settlement layer from optical HR/VHR RS data: Concept and first results," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 5, pp. 2102–2131, Oct. 2013.
- [6] European Commission, "Mapping guide for a European Urban Atlas," Eur. Space Res. Inst., Eur. Space Agency, Frascati, Italy, Tech. Rep. ITD-0421-GSELand-TN-01, 2008.
- [7] B. Bechtel and C. Daneke, "Classification of local climate zones based on multiple earth observation data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 4, pp. 1191–1202, Aug. 2012.
- [8] B. Bechtel *et al.*, "Mapping local climate zones for a worldwide database of the form and function of cities," *ISPRS Int. J. Geo-Inf.*, vol. 4, no. 1, pp. 199–219, 2015.
- [9] B. Bechtel *et al.*, "Beyond the urban mask: Local climate zones as a generic descriptor of urban areas: Potential and recent developments," in *Proc. Joint Urban Remote Sens. Event*, Dubai, UAE, 2017.
- [10] O. Danylo, L. See, B. Bechtel, D. Schepaschenko, and S. Fritz, "Contributing to WUDAPT: A local climate zone classification of two cities in Ukraine," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 5, pp. 1841–1853, May 2016.
- [11] M.-L. Verdonck, A. Okujeni, S. van der Linden, M. Demuzere, R. De Wulf, and F. Van Coillie, "Influence of neighbourhood information on 'local climate zone' mapping in heterogeneous cities," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 62, pp. 102–113, 2017.
- [12] N. Perera and R. Emmanuel, "A 'local climate zone' based approach to urban planning in Colombo, Sri Lanka," *Urban Clim.*, vol. 23, pp. 188–203, 2018.
- [13] Y. Xu, C. Ren, M. Cai, N. Y. Y. Edward, and T. Wu, "Classification of local climate zones using ASTER and landsat data for high-density cities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3397–3405, Jul. 2017.
- [14] B. Bechtel, L. See, G. Mills, and M. Foley, "Classification of local climate zones using SAR and multispectral data in an arid environment," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 7, pp. 3097–3105, Jul. 2016.
- [15] L. See *et al.*, "Comparing the quality of crowdsourced data contributed by expert and non-experts," *PLoS One*, vol. 8, 2013, Art. no. e69958.
- [16] G. M. Foody, L. See, M. van der Velde, C. Perger, C. Schill, and D. S. Boyd, "Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project," *Trans. GIS*, vol. 17, no. 6, pp. 847–860, 2013.

⁹<http://www.grss-ieee.org/community/technical-committees/data-fusion>, under the "Past Contests" tab.

- [17] J. C. L. Bayas *et al.*, "Crowdsourcing in-situ data on land cover and land use using gamification and mobile technology," *Remote Sens.*, vol. 8, no. 11, p. 905, 2016.
- [18] B. Bechtel *et al.*, "Quality of crowdsourced data on urban morphology—The human influence experiment (HUMINEX)," *Urban Sci.*, vol. 1, no. 2, p. 15, 2017.
- [19] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015.
- [20] D. Tuia, C. Persello, and L. Bruzzone, "Recent advances in domain adaptation for the classification of remote sensing data," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 41–57, Jun. 2016.
- [21] 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.
- [22] 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 GRS-S data fusion contest," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 331–335, Jul. 2008.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] W. Liao *et al.*, "Processing of thermal hyperspectral and digital color cameras: Outcome of the 2014 data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2984–2996, Jun. 2015.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] O. Conrad *et al.*, "System for automated geoscientific analyses (SAGA) v. 2.1.4," *Geosci. Model Develop.*, vol. 8, no. 7, pp. 1991–2007, 2015.
- [33] D. P. Roy *et al.*, "Landsat-8: Science and product vision for terrestrial global change research," *Remote Sens. Environ.*, vol. 145, pp. 154–172, 2014.
- [34] 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.
- [35] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 12–18, Oct.–Dec. 2008.
- [36] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6054–6068, Nov. 2017.
- [37] 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.
- [38] P. Lopes, C. Fonte, L. See, and B. Bechtel, "Using OpenStreetMap data to assist in the creation of LCZ maps," in *Proc. Joint Urban Remote Sens. Event*, 2017, pp. 1–4.
- [39] R. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2009.
- [40] N. Yokoya, P. Ghamisi, and J. Xia, "Multimodal, multitemporal, and multi-source global data fusion for local climate zones classification based on ensemble learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Fort Worth, TX, USA, 2017, pp. 1197–1200.
- [41] S. Sukhanov, I. Tankoyeu, J. Louradour, R. Heremans, D. Trofimova, and C. Debes, "Multilevel ensembling for local climate zones classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Fort Worth, TX, USA, 2017, pp. 1201–1204.
- [42] C. Souza dos Anjos, M. Goncalves Lacerda, L. do Livramento Andrade, and R. Neves Salles, "Classification of urban environments using feature extraction and random forest," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Fort Worth, TX, USA, 2017, pp. 1205–1208.
- [43] Y. Xu, F. Ma, D. Meng, C. Ren, and Y. Leung, "A co-training approach to the classification of local climate zones with multi-source data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Fort Worth, TX, USA, 2017, pp. 1209–1212.
- [44] T. Rainforth and F. Wood, "Canonical correlation forests," arXiv:1507.05444, 2015.
- [45] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. SIGKDD*, 2016, pp. 785–794.
- [46] J. J. Rodriguez and L. I. Kuncheva, "Rotation forest: A new classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1619–1630, Oct. 2009.
- [47] R. Richter and D. Schlöpfer, *Atmospheric/Topographic Correction for Satellite Imagery: ATCOR 2/3 Users Guide*, German Aerosp. Center (DLR), Wessling, Germany, 2016.
- [48] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.
- [49] J. Xia, P. Du, X. He, and J. Chanussot, "Hyperspectral remote sensing image classification based on rotation forest," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 239–243, Jan. 2014.
- [50] J. Xia, J. Chanussot, P. Du, and X. He, "Spectral-spatial classification for hyperspectral data using rotation forests with local feature extraction and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2532–2546, May 2015.
- [51] J. Xia, N. Yokoya, and A. Iwasaki, "Hyperspectral image classification with canonical correlation forests," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 421–431, Jan. 2017.
- [52] C. Ren, M. Cai, R. Wang, Y. Xu, and E. Ng, "Local climate zone (LCZ) classification using the world urban database and access portal tools (WU-DAPT) method: A case study in Wuhan and Hangzhou," in *Proc. Int. Conf. Countermeasure Urban Heat Islands*, Singapore, May 2016.
- [53] "Spaceanalyzer—indices." [Online]. Available: www.spaceanalyzer.com
- [54] S. Oshigami *et al.*, "Mineralogical mapping of southern Namibia by application of continuum-removal MSAM method to the HyMap data," *Int. J. Remote Sens.*, vol. 34, no. 15, pp. 5282–5295, 2013.
- [55] J. B. Lee, A. S. Woodyatt, and M. Berman, "Enhancement of high spectral resolution remote-sensing data by a noise-adjusted principal components transform," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 3, pp. 295–304, May 1990.
- [56] A. Merentitis, C. Debes, and R. Heremans, "Ensemble learning in hyperspectral image classification: Toward selecting a favorable bias-variance tradeoff," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1089–1102, Apr. 2014.
- [57] D. Marcos, M. Volpi, N. Komodakis, and D. Tuia, "RotEqNet: Rotation equivariant vector field networks," in *Proc. Int. Conf. Comput. Vis.*, Venice, Italy, 2017.
- [58] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by Markov modeling of spatial-contextual information," *Proc. IEEE*, vol. 101, no. 3, pp. 631–651, Mar. 2013.
- [59] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "SMOTE: Synthetic minority oversampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
- [60] S. Sukhanov, A. Merentitis, C. Debes, J. Hahn, and A. M. Zoubir, "Bootstrap-based SVM aggregation for class imbalance problems," in *Proc. Eur. Signal Process. Conf.*, 2015, pp. 165–169.
- [61] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.
- [62] S. Lefèvre, D. Tuia, J. D. Wegner, T. Prodiut, and A. S. Nassar, "Towards seamless multi-view scene analysis from satellite to street-level," *Proc. IEEE*, vol. 105, no. 10, pp. 1884–1899, Oct. 2017.