The Outcome of the 2021 IEEE GRSS Data Fusion Contest—Track MSD: Multitemporal Semantic Change Detection

Zhuohong Li, Fangxiao Lu, Hongyan Zhang, *Senior Member, IEEE*, Lilin Tu, Jiayi Li, *Senior Member, IEEE*, Xin Huang, *Senior Member, IEEE*, Caleb Robinson, Nikolay Malkin, Nebojsa Jojic, Pedram Ghamisi, *Senior Member, IEEE*, Ronny Hänsch, *Senior Member, IEEE*, and Naoto Yokoya, *Member, IEEE*

***Abstract*—We present here the scientific outcomes of the 2021 Data Fusion Contest (DFC2021) organized by the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society. DFC2021 was dedicated to research on geospatial artificial intelligence (AI) for social good with a global objective of modeling the state and changes of artificial and natural environments from multimodal and multitemporal remotely sensed data toward sustainable developments. DFC2021 included two challenge tracks: "Detection of settlements without electricity" and "Multitemporal semantic change detection." This article mainly focuses on the outcome of the multitemporal semantic change detection track. We describe in this article the DFC2021 dataset that remains available for further evaluation of corresponding approaches and report the results of the best-performing methods during the contest.**

Manuscript received November 11, 2021; revised January 4, 2022; accepted January 12, 2022. Date of publication January 25, 2022; date of current version February 14, 2022. The work of Lilin Tu, Jiayi Li, and Xin Huang was supported by the National Natural Science Foundation of China under Grant 41971295, and the CAS Interdisciplinary Innovation Team under Grant JCTD-2019-04. *(Corresponding author: Naoto Yokoya.)*

Zhuohong Li, Fangxiao Lu, and Hongyan Zhang are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430079, China (e-mail: ashelee@whu.edu.cn; fangxiaolu@whu.edu.cn; zhanghongyan@whu.edu.cn).

Lilin Tu and Jiayi Li are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: tulilin0312@163.com; zjjercia@whu.edu.cn).

Xin Huang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, and also with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: xhuang@whu.edu.cn).

Caleb Robinson is with the Microsoft AI for Good Research Lab, Redmond, WA 98052, USA (e-mail: calebrob6@gmail.com).

Nikolay Malkin is with the Mila - Université de Montréal, Montreal, QC H2S 3H1, Canada (e-mail: kolya_malkin@hotmail.com).

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

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

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

Naoto Yokoya is with the Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, University of Tokyo, 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.2022.3144318

***Index Terms*—Convolutional neural networks, deep learning, image analysis and data fusion, land cover change detection, multimodal, random forests, weak supervision.**

I. Introduction

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 organizes an annual challenge named the Data Fusion Contest (DFC) for fostering ideas and progress in remote sensing, distributing novel data, and benchmarking analysis methods [1]–[15]. DFC2021 promotes interdisciplinary research on geospatial artificial intelligence (AI) for social good. The global objective was to build models for understanding the state and changes of artificial and natural environments from multimodal and multitemporal remote sensing data toward sustainable developments. The contest is designed as a benchmark competition following previous editions [15]–[18]. DFC2021 includes the two following tracks, which were run in parallel:

1) Track DSE: Detection of Settlements without Electricity;
2) Track MSD: Multitemporal Semantic change Detection.

This article focuses mainly on the outcome of Track MSD. Track MSD was co-organized together with Microsoft AI for Good with a particular focus on automatic land cover change detection and classification from multitemporal, multiresolution, and multispectral imagery. The main task of this track is to produce bitemporal high-resolution land cover maps inputting only low-resolution and noisy land cover labels for training.

The multisource datasets were captured over the U.S. state of Maryland and consisted of 1) 1-m multispectral aerial imagery for 2013 and 2017 from USDA National Agriculture Imagery Program (NAIP) data, 2) 30-m multispectral satellite imagery (Landsat-8) for five points in time between 2013 and 2017, and 3) 30-m noisy low-resolution land cover labels for 2013 and 2016 from USGS National Land Cover Database (NLCD) data (see Fig. 1). Participants needed to infer high-resolution (1-m GSD)

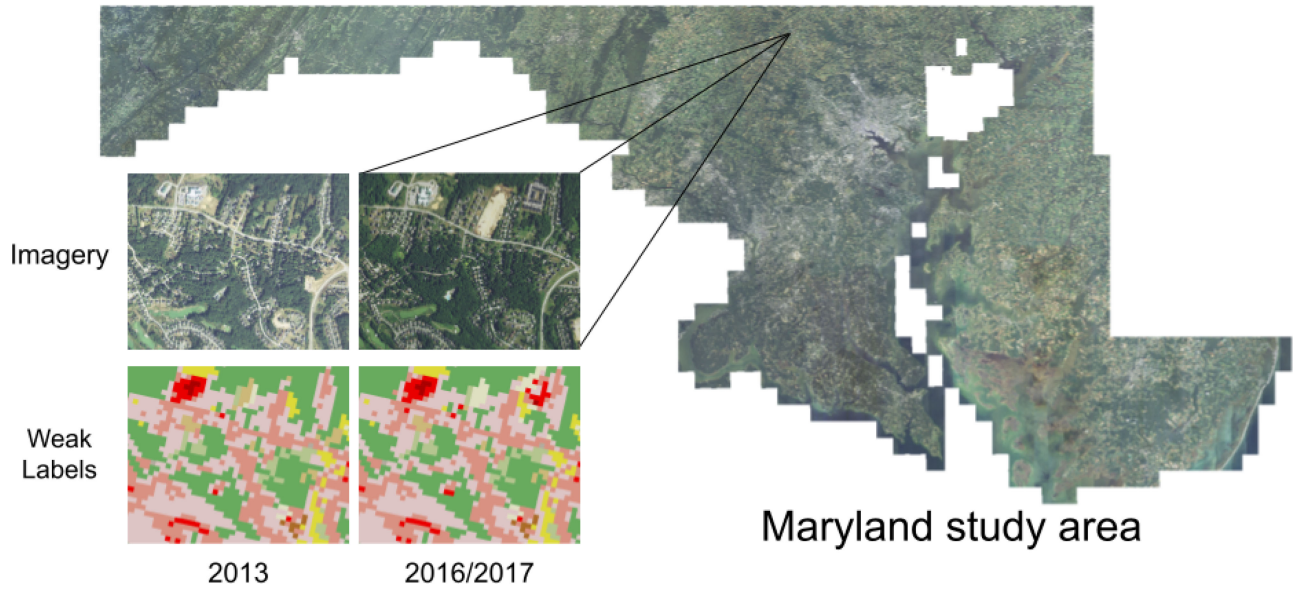

Fig. 1. Study area and example data for the MSD track of the 2021 IEEE GRSS Data Fusion Contest.

land cover maps that identify changes between 2013 and 2017. The performance was evaluated based on the intersection-over-union metric averaged over eight change types. The challenge was twofold: detecting which parts had changed between two high-resolution aerial images and identifying the class of change based on weak labels.

Both tracks of the DFC2021 addressed the following real-world social problems: 1) analysis of multisensor, multiresolution, and multitemporal data, and 2) learning from low-quality labeled samples (weak supervision). The aforementioned problems are major open challenges in a wide range of fields, from Earth observation to computer vision and machine learning [19]. The main feature of the contest is that it tackles directly some of the most pressing social issues, e.g., energy equality and environmental conservation. In other words, the results of the contest will impact not only in terms of technological development but also as a tool for solving actual social problems [19].

The rest of this article is organized as follows. We describe the datasets used in DFC21 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 of track MSD in Sections IV and V. Finally, Section VI concludes this article.

II. Data and Baseline of the Data Fusion Contest 2021

The dataset for the MSD track of the DFC2021 includes nine layers of data covering the U.S. state of Maryland spanning from 2013 to 2017. Specifically, there are two layers of high-resolution (1-m GSD) aerial imagery from the NAIP for 2013 and 2017, five layers of low-resolution (30-m GSD) annual composites of Landsat 8 multispectral imagery for each year from 2013 to 2017, and two layers of noisy low-resolution land cover labels from the NLCD for 2013 and 2016. Details about each type of layer is given in the following.

1) **NAIP**. The NAIP layers are four-band—red, green, blue, and near infrared (NIR)—aerial imagery at a 1-m spatial resolution. This imagery is cloud-free and is captured in good weather conditions independently on a state-by-state basis every two to three years.
2) **Landsat 8**. The Landsat 8 layers are nine-band multispectral satellite imagery at a 30-m spatial resolution. Using Google Earth Engine, we generate a median composite of all Landsat 8 surface reflectance scenes intersecting Maryland for each year between 2013 and 2017. Before taking the median, we mask each scene with per-pixel cloud and cloud shadow estimates generated by the CF-MASK algorithm.
3) **NLCD**. The NLCD layers contain 16-class land cover data at a 30-m spatial resolution for 2013 and 2016 (from the April 2019 data release). These data are created by the Multi-Resolution Land Characteristics Consortium with a consistent methodology that allows change to be inferred between the different years of data.

The dataset is broken up into 2250 nonoverlapping *tiles*, each of which covers an area of approximately 4 km $\times$ 4 km. Each layer is resampled to 1 m/px (using bilinear interpolation for the Landsat data and nearest neighbor interpolation for NLCD data), projected into a Web Mercator projection[1] and stored as a GeoTIFF following the tile definitions.

The task of the competition is to predict loss/gain of four types of land cover classes between the two NAIP layers at a 1-m resolution: "water," "tree canopy," "low vegetation," or "impervious surfaces." The 16-class low-resolution labels are the only label data that participants have access to. The participants are also given a table of the joint class frequencies of the low-resolution and high-resolution land cover classes in the 2013 imagery (see Table I).

The predictions are evaluated with the mean intersection-over-union (mIoU) metric over a subset of the overall 2250 tiles using held out high-resolution labels—50 in the *validation* phase of the competition and 57 in the *test* phase of the competition.

[1]EPSG:3857

TABLE I
MAPPING FROM THE LOW-RESOLUTION NLCD CLASS LABELS TO THE HIGH-RESOLUTION CLASS LABELS

		Approximate class frequencies			
NLCD class name	**Class labels**	Water	TC	LV	Imperv.
Open Water	Water	98%	2%	0%	0%
Developed, Open Space	(mixed)	0%	39%	49%	12%
Developed, Low Intensity	(mixed)	0%	31%	34%	35%
Developed, Medium Intensity	Impervious	1%	13%	22%	64%
Developed High Intensity	Impervious	0%	3%	7%	90%
Barren Land (Rock/Sand/Clay)	Low Vegetation	5%	13%	43%	40%
Deciduous Forest	Tree Canopy	0%	93%	5%	0%
Evergreen Forest	Tree Canopy	0%	95%	4%	0%
Mixed Forest	Tree Canopy	0%	92%	7%	0%
Shrub/Scrub	Tree Canopy	0%	58%	38%	4%
Grassland/Herbaceous	Low Vegetation	1%	23%	54%	22%
Pasture/Hay	Low Vegetation	0%	12%	83%	3%
Cultivated Crops	Low Vegetation	0%	5%	92%	1%
Woody Wetlands	Tree Canopy	0%	94%	5%	0%
Emergent Herbaceous Wetlands	Tree Canopy	8%	86%	5%	0%

High-resolution labels were never given to the participants to encourage them to focus on how to best exploit weak supervision. Noisy low-resolution land cover labels are globally available, for example, from the MODIS MCD12Q1.006 product or the Copernicus Global Land Cover Layer [20], and sparse high-resolution labels can easily be collected to calculate label statistics similar to those shown in Table I. Therefore, any innovations in this competition setup have immediate implications for similar high-resolution land cover mapping around the globe. Proposed methods will have to deal with label noise, mismatched land cover class definitions, super-resolution of labels, and the challenges of combining multiple types of data sources in a model.

A. Baseline

We provided several naive baseline approaches in a public GitHub repository prior to the start of the competition.[2] The first approach, `NLCD difference`, simply calculates the low-resolution change from the pair of NLCD layers and assigns the low-resolution class labels to high-resolution class labels according to the mapping given in Table I. The other methods—`U-Net both`, `U-net separate`, `FCN both`, and `FCN separate`—train U-Net and fully convolutional network (FCN) models with NAIP inputs and NLCD labels, then infer pseudo-NLCD labels over the imagery, and finally compute the high-resolution change by the same differencing and assignment step used in `NLCD difference`. The methods named "both" consist of a single model trained on pairs of NAIP imagery from 2013 with NLCD labels from 2013 and NAIP imagery from 2017 with NLCD labels from 2016, while the methods named "separate" consist of two models: one trained on the 2013 imagery and labels, and one trained on the 2017 imagery and labels. For model architectures, the FCN model is a five-layer fully convolutional model with 64 3×3 filters in each layer and ReLU activations, while the U-Net uses a ResNet-18 backbone and uses the implementation from the Segmentation Models PyTorch library [21]. In all cases, we train with all available data (2250 tiles).

The baseline results on the validation phase are shown in Table II. As expected, the `NLCD difference` method gives the worse performance as many changes in the landscape are only identifiable at higher resolutions. We observe that, under both architectures, training separately in the different years of data gives improved performance and that the FCN architecture gives the overall better performance. Qualitatively, we observe in Fig. 2 that the U-Net architectures are able to fit to the low-resolution labels and make low-resolution predictions despite using high-resolution imagery as input. In contrast, the FCN architecture is limited (by construction) with a receptive field of 11 pixels. Here, the model is unable to fit to the low-resolution labels and produces higher fidelity output as a result. While the NLCD classes themselves are not meaningful at a 1-m resolution (e.g., the distinction between "Developed, High Intensity" and "Developed, Low Intensity" cannot be defined at 1-m resolution), this gives a better result when the class labels are remapped into the high-resolution class labels. The best performing baseline model achieves a mIoU score of 0.453 on the validation set.

TABLE II
BASELINE APPROACH RESULTS ON THE VALIDATION SET

Method	**Mean IoU**
NLCD difference	0.1389
U-Net both	0.3374
U-Net separate	0.3610
FCN both	0.4188
FCN separate	0.4530

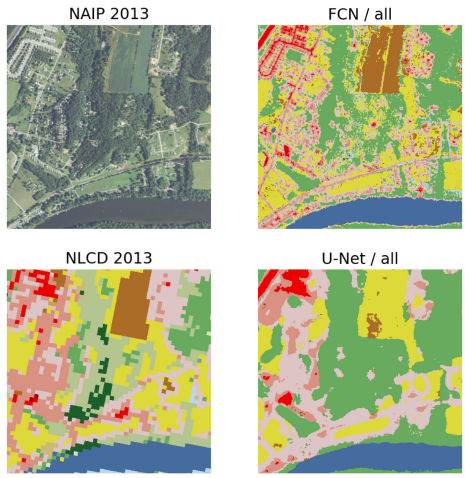

Fig. 2. Example predictions (**right column**) from the baseline FCN and baseline U-Net models that are trained to predict low-resolution labels from high-resolution inputs (**left column**). We observe that the U-Net and FCN models produce qualitatively different predictions despite using the same inputs and labels. The U-Net model (with a full receptive field) attempts to reproduce the low-resolution labels and doesn't identify small features in the imagery, while the FCN model (with an 11-pixel receptive field) does identify small features in the imagery.

III. ORGANIZATION, SUBMISSIONS, AND RESULTS

There were 139 unique registrations at the CodaLab competition website[3] during the development phase and 20 teams entered the test phase after screening the descriptions of their

[2][Online]. Available: https://github.com/calebrob6/dfc2021-msd-baseline/

[3][Online]. Available: https://competitions.codalab.org/competitions/27956

TABLE III
TOP RANKED TEAMS AND THEIR APPROACHES

Rank	Team	mIoU	Data		Approach					
			L8	Other	Ensemble	Refine LR labels	Pseudo labels	Multi-task	Specific class	Filtering, thresholding
1	*AsheLee*	0.6772	✓		✓	✓	✓	✓		✓
2	*tulilin*	0.6657	✓	✓	✓	✓	✓			✓
3	*baoqianyue*	0.6445			✓		✓		✓	
4	*EVER*	0.6435	✓				✓	✓		✓

approaches submitted by the end of the development phase. With active participation from all registered teams, 2033 submissions were received during the development phase. During the test phase, the maximum number of submissions per team was limited to ten, and 115 submissions were received. The final ranking was determined based on the mIoU averaged over eight types of change.

The first to fourth ranked teams were awarded as winners of the DFC2021 Track MSD and presented their solutions during the 2021 IEEE International Geoscience and Remote Sensing Symposium. The four winning teams are the following.

1) **1st place**: *AsheLee* team; Zhuohong Li, Fangxiao Lu, Hongyan Zhang, Guangyi Yang, and Liangpei Zhang from Wuhan University, China [22].
2) **2nd place**: *tulilin* team; Lilin Tu, Jiayi Li, and Xin Huang from Wuhan University, China [23].
3) **3rd place**: *baoqianyue* team; Qianyue Bao, Yang Liu, Zixiao Zhang, Dafan Chen, Yuting Yang, Licheng Jiao, and Fang Liu from Xidian University, China [24].
4) **4th place**: *EVER* team; Zhuo Zheng, Yinhe Liu, Shiqi Tian, Junjue Wang, Ailong Ma, and Yanfei Zhong from Wuhan University, China [25].

Table III summarizes the characteristics of the methods of the top four teams. We can see that these methods are diverse, but there are some common features. All teams used high-resolution labels, which are predicted by segmentation models trained with high-resolution images and low-resolution labels, as pseudolabels for retraining segmentation models. Similar to previous editions of DFC, many teams used ensembling (or model fusion) of different neural networks and refinement of high-resolution classification maps by morphological filtering and thresholding for further improvement.

The top two teams were unique in which they incorporated low-resolution label refinement, but the way they did it varied greatly. *AsheLee* built up a low-resolution segmentation model using Landsat-8 data for label refinement, while *tulilin* used global land cover products other than the data distributed in the contest to make corrections based on human knowledge. The teams differed in their use of Landsat-8 data and how they used it. *tulilin* and *EVER* used spectral indices extracted from Landsat-8 data as hand-crafted features in intermediate and postprocessing, respectively. *AsheLee* and *EVER* proposed neural networks for a multitask problem that solves land cover classification and change detection simultaneously, which is an ingenious way to deal with the problem setting. The changes related to water is particularly challenging, and *baoqianyue* built a classifier specialized only for water, while *tulilin* and *EVER* used water indices as features for accuracy improvement. The following sections detail the top two winning solutions.

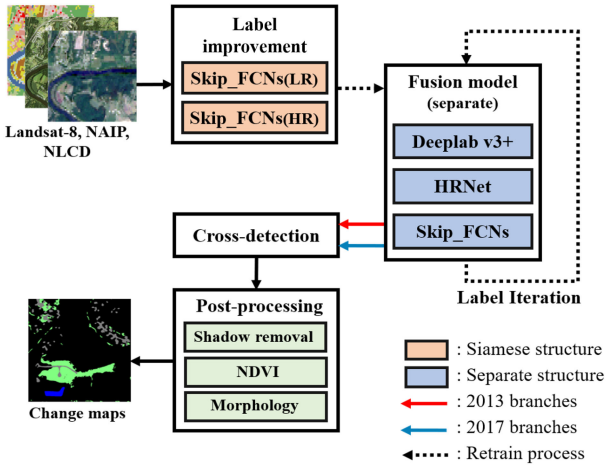

Fig. 3. Flowchart of the overall scheme.

IV. FIRST PLACE TEAM OF TRACK MSD

The method of the first place team in Track MSD includes four stages: label improvement, multimodel fusion, cross-detection, and postprocessing. The flowchart of the scheme is shown in Fig. 3.

A. Label Improvement

1) Designing of the Overall Scheme: In this section, we aim at improving the original low-resolution labels to generate pseudolabels for the subsequent parts, which mainly deals with the following two problems. 1) The acquisition times between the NLCD labels and NAIP imagery are not completely corresponding, which may lead to the mismatch of training pairs. 2) The spatial resolution between the NLCD labels and NAIP imagery are not matched, which may restrict the effect of advanced networks with large receptive field or deep encoder–decoder structure, e.g., UNet [26], DeepLab v3+ [27], etc.

With the consideration of the first problem, the acquisition times between pseudolabels and images should be aligned. Therefore, we first use the Landsat-8 images that have a wide time-span and the NLCD to generate time-aligned pseudolabels. Because only the 2013 and 2016 NLCD labels are provided, we use the 2017 Landsat-8 images and the 2016 NLCD labels to train a siamese SkipFCN, which is done for two considerations: 1) When the acquisition time of images is relatively close, their

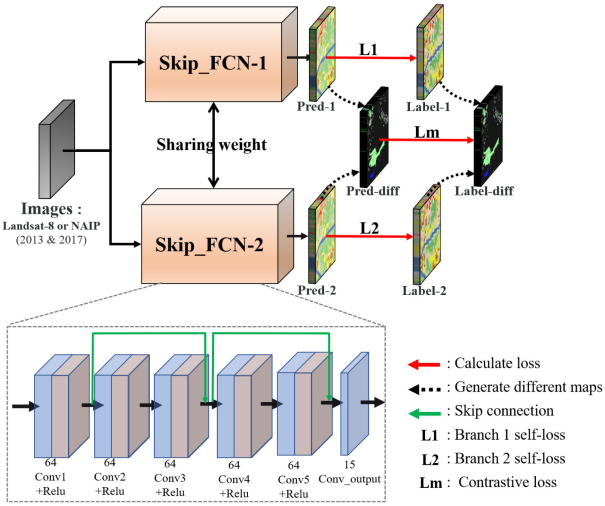

Fig. 4. Siamese SkipFCNs network.

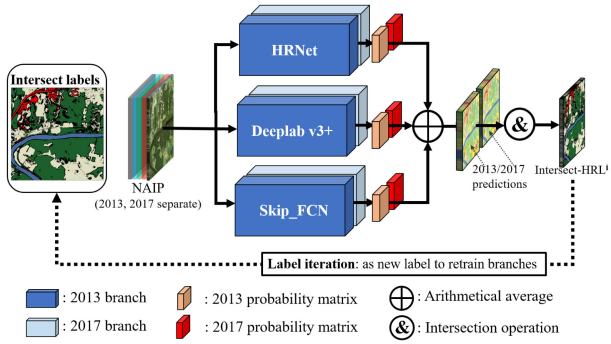

Fig. 5. Multimodel fusion and labels iteration.

feature differences are negligible; 2) since the pseudolabels will be used for training along with the NAIP images from 2013 and 2017, the pseudolabels generated by Landsat-8 images from 2017 can be used to describe the land cover features in 2017 more truly. Based on the above discussion, two Landsat-8/NLCD pairs (i.e., 2013/2013 and 2017/2016) are used at the very beginning of label improvement process and generate pseudolabels which have the same resolution as the Landsat-8 image. Compared with the original NLCD, its details are richer, and the edges are smoother. At the same time, the alignment of the acquisition time between the produced pseudolabel and the NAIP imagery have been completed. The labels generated in this phase are called Landsat-Low Resolution Labels (Landsat-LRL), which represents that they are generated with low-resolution Landsat-8 images.

After the first phase, we replace the input images to NAIP and train a new SkipFCN along with Landsat-LRL labels. Then a set of updated pseudolabels, which have the same resolution with the input NAIP image, would be generated, which are called NAIP-High Resolution Labels (NAIP-HRL). So far, we have solved both problems that we listed above in this stage.

2) Designing of Networks: Since the NLCD is mosaic and low-resolution, advanced networks with large receptive fields may overfit the noisy labels and make fuzzy predictions. Based on that, a Siamese SkipFCNs network is designed for labels resolution improvement, as shown in Fig. 4. The network contains two SkipFCNs with weight-sharing. Each SkipFCN has five Conv-ReLU layers that maintain small receptive fields [28] and skip connections in the intermediate layers that preserve shallow information [29]. These convolutional layers are all with 3×3 kernels and 1 pixel padding at each side, leading to the fact that the resolution of feature maps stays the same during the network forward process. This is the reason why the generated labels have the same resolution as the input images. During training, as Chopra *et al.* did in [30], a loss function including three parts (L_1, L_2, and L_m) is proposed. The L_1 and L_2 parts are two supervised losses between predictions and labels, which are the regular losses of semantic segmentation

$$L_1 = L_{\text{CE}}(\theta(x_1), y_1); L_2 = L_{\text{CE}}(\theta(x_2), y_2). \tag{1}$$

In (1), L_{CE} denotes the cross-entropy loss, $\theta(\cdot)$ denotes the composition of operations (convolution, activation, and batch normalization) used in the networks, and x_i and y_i denote the input and corresponding label of the ith branch. The L_m part is a contrastive loss

$$L_m = L_{\text{MSE}}(\tanh(\phi(x_1, x_2)), \text{XOR}(y_1, y_2)) \tag{2}$$

$$\phi(x_1, x_2) = \sqrt{\sum_{i=1}^{K} (\theta(x_1)_i - \theta(x_2)_i)^2}. \tag{3}$$

In (2) and (3), L_{MSE} denotes the mean square loss, and $\phi(\cdot, \cdot)$ denotes the Euclidean distance between two inputs, K denotes the number of prediction channels, and i denotes the ith channel of prediction. The tanh activation function maps the output to the range of [0,1) since the Euclidean distance is always positive. The XOR operation generates the binary original change map, in which 1 denotes changed pixel and 0 denotes unchanged pixel.

This loss function makes the network focus on not only the segmentation task but also the change detection task. During training, we randomly select two images as the input of the network. The two images may come from the same year or come from different years, so do the labels, which makes the network not sensitive to the order of inputs.

B. Multimodel Fusion

The previously generated NAIP-HRLs still contain errors, which renders them as not accurate enough for being used in training the advanced models. To reduce label error, as shown in Fig. 5, multiple models are trained separately in years, and their outputs are fused as new pseudolabels for further retraining. Specifically, by considering that the HRNet and Deeplab v3+ have reported state-of-the-art performance in general segmentation tasks [27], [31], and SkipFCN is a shallow structure that we proposed to maintain mapping details, the NAIP images and NAIP-HRL labels are used to train the years 2013 and 2017 branches of these three models for integrating the advantages of multimodels. Then, arithmetical average assignment is applied to integrate their outputs and intersection operation is implemented, which only preserves the identical parts in two years' predictions, to maintain the common high confidence

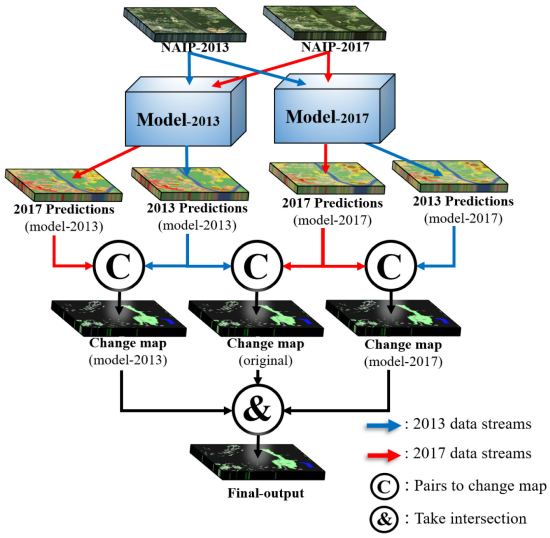

Fig. 6. Cross-detection structure.

parts in the bitemporal predictions. The intersected outputs, called "Intersected High-Resolution Labels (Intersect-HRL)," become the new training labels in the next iteration. This process is repeated until the results are stable and with high precision.

C. Cross-Detection Structure

At the final stage of our scheme, the predicted land cover maps of 2013 and 2017 need to be intersected into final change maps. By considering the "single year overfitting" issue that is brought by separate training process in the multimodel fusion phase. A practical structure, called "cross-detection," is applied to reduce the redundant errors and implement more interactions between branches. As shown in Fig. 6, it is performed by using the 2013 branch of the fusion model to predict 2017 validation set and repeating the process in the 2017 branches. Then two cross-change maps are taken as a restriction to the original change map.

D. Postprocessing

After the final outputs of cross-detection, shadow-removal, NDVI restriction, and morphological methods are implemented in the postprocessing step. For removing the uncertain area covered by shadow, the intensity channel of hue-saturation-intensity color model and NIR channel of images are used to detect and remove false alarms in the shadow. To discriminate low vegetation and impervious surface better, NDVIs are performed to restrict change maps. Finally, several morphology methods, including erosion, dilation, and small object removal, are used to remove remaining slight errors.

E. Results and Discussion

To demonstrate our attempts at each stage more clearly, the experimental results of the applied method on the development phase are reported. For the purpose of evaluating the results

TABLE IV
IoU OF DIFFERENT ATTEMPTS ON DEVELOPMENT PHASE (TEAM *ASHELEE*)

Algorithms	Labels	Extra	mIoU
UNet	NLCD	-	0.3610
Siam-Skip_FCNs	NLCD	-	0.4827
Siam-Skip_FCNs	Landsat-LRL	-	0.5380
Skip_FCN	Inersect-HRL[1]	-	0.5994
HRNet	Inersect-HRL[1]	-	0.5994
Deeplab v3+	Inersect-HRL[1]	-	0.6541
Fusion(H+D+U)	Inersect-HRL[2]	-	0.6534
Fusion(H+D+S)	Inersect-HRL[2]	-	0.6689
Fusion(H+D+S)	Inersect-HRL[3]	-	0.6794
Fusion(H+D+S)	Inersect-HRL[3]	CD	0.6906
Fusion(H+D+S)	Inersect-HRL[3]	CD+PP	**0.7025**

Note: H: HRNet, D: Deeplab v3+, U: UNet, S: Skip_FCN. CD: Cross-detection, PP: Post-processing. Intersect-HRL[i]: i-th label iteration

Note: H: HRNet; D: Deeplab v3+; U: UNet; S: SkipFCN.
CD: Cross-detection; PP: Postprocessing.
Intersect-HRL ^{i}: i th label iteration.

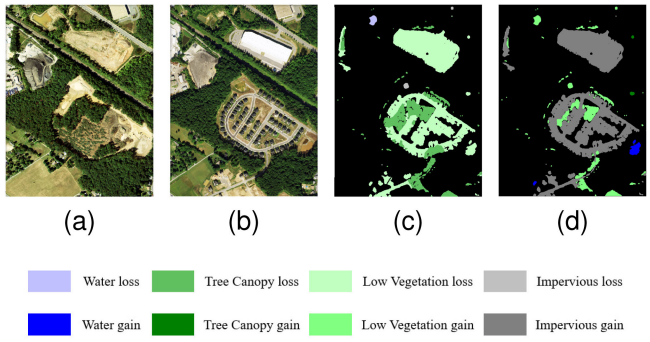

Fig. 7. Example results of the method of Team *AsheLee*. (a) and (b) NAIP image (2013/2017). (c) and (d) Land cover change (loss/gain).

quantitatively and qualitatively, we compare the four stages IoU score of our method in Table IV and illustrate the visual results in Fig. 7. By comparing the quantitative results between the baseline method trained on UNet and our approach, we can see that by applying the label improvement in the first stage, the score increases from 0.3610 to 0.5380. This suggests that the preliminary label-resolution improvement greatly improves the detection accuracy. Due to the continued label iteration in the second stage, the maximum score reaches 0.6541 for the single model. By adopting the multimodel fusion, we find that the fusion strategy conducting on HRNet, Deeplab v3+, and UNet performs even worse with a mIoU of 0.6534. However, the score of the fusion model that includes HRNet, Deeplab v3+, and SkipFCN increases to 0.6794, which is because these networks are quite different in structure, and, thus, their results can well complement each other. This sign also indicates that the multimodel fusion strategy can perform much better than every included networks when their structures have complementary advantages. Finally, due to the effect of cross-detection and postprocessing in removing false detections, the score reaches 0.7025 in development phase and reaches 0.6772 on the test dataset. In the future, based on the widely existing weak labels and high-resolution images, we will apply this method to detect

landscape or urban changes that occur in a wider coverage and further verify the approach effectiveness on more variable remotely sensed data at a larger scale.

V. Second Place Team of Track MSD

The algorithm proposed by the second-place team, a semisupervised deep learning approach for high-resolution classification and change detection using low-resolution NLCD labels, is described in this section. The NLCD labels were refined using global land cover products as a preprocessing step. For generating high-resolution classification and change maps, an FCN [32] was trained in two stages, i.e., using low-resolution NLCD labels and high-resolution pseudolabels, respectively. Modified normalized difference water index (MNDWI) [33], ensemble training, and decision-level fusion were used to improve the performance. Furthermore, a series of postprocessing steps from the pixel level to the scene level were implemented on the change maps in order to reduce commission errors in the change detection results.

A. Preprocessing

The data preprocessing includes the following steps.

1) Label Reclassification: The 16-class NLCD labels were reclassified into four target classes: *Water*, *Tree Canopy*, *Low Vegetation*, and *Impervious* according to the approximate correspondence between NLCD and target classes [34]. For pixels with NLCD classes which may correspond to more than one target classes (e.g., *Developed Open Space*, *Developed Low Intensity*, and *Barren Land*), they were removed from the NLCD labels and were not used as training samples.

2) Label Refinement: NLCD labels were refined using five global land cover products: FROM-GLC10 [35], GLCFCS30 [36], Globeland30 [37], Global Forest Change [38], and Global Surface Water [39]. Similarly, the labels of these land cover products were reclassified to the target classes. Only pixels with the same label in NLCD, FROM-GLC10, GLCFCS30, and Globeland30 were preserved. For Global Forest Change and Global Surface Water, they were used to remove the erroneous *Tree Canopy* and *Water* labels, respectively.

3) Training Sample Generation: The NAIP images and refined NLCD labels were cropped into a series of patches with the size of 512×512 as training samples. The NAIP images of the year 2013 were assigned with the NLCD labels of the year 2013, and the NAIP images of the year 2017 were assigned with the NLCD labels of the year 2016.

B. Network Training, Classification, and Change Detection

The overall framework for generating classification and change maps was shown in Fig. 8, which mainly composed of three steps: training with low-resolution NLCD labels, training with high-resolution pseudolabels, and decision-level fusion. Algorithm details are described as follows.

1) FCN Network Model: As shown in Fig. 8, FCN has five 3×3 convolutional layers for feature extraction and one

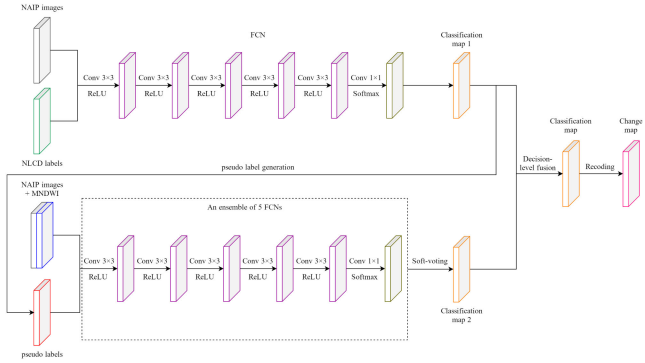

Fig. 8. Overall framework for generating classification and change maps.

1×1 convolutional layer for classification. Without any downsampling layers (i.e., pooling layers), the network has small receptive field and, hence, would not tend to overfit the low-resolution and noisy labels. In this way, high-resolution classification maps (i.e., classification maps with the same resolution as NAIP images) can be obtained after training with low-resolution NLCD labels.

2) Network Training and Classification: First, bi-temporal NAIP images along with the NLCD labels were all fed into one FCN for training. Subsequently, classification maps of all NAIP image tiles were predicted. High-resolution pseudolabels were then generated from the classification maps. Specifically, the class probability of the four target classes was calculated for each pixel, and only pixels with the maximum class probability larger than a threshold were preserved and fed into the pseudolabels. In the next training period, MNDWI, which can better discriminate water from impervious layer than normalized difference water index (NDWI), was extracted from the Landsat-8 images

$$\text{NDWI} = \frac{G - \text{NIR}}{G + \text{NIR}} \tag{4}$$

$$\text{MNDWI} = \frac{G - \text{SWIR}}{G + \text{SWIR}} \tag{5}$$

where G, NIR, and SWIR represent the pixel value of the green band, NIR band, and short-wave infrared band, respectively.

MNDWI was concatenated with the four-band NAIP images as the input features and high-resolution pseudolabels were used as new training samples during the next training period. In addition, the sample set was divided into five parts and an ensemble of five FCNs was trained. The classification maps of this training period were generated via the soft-voting of the five FCNs.

3) Decision-Level Fusion and Change Map Generation: After the network training, the classification maps of the two training periods were fused in the decision level. Specifically, for each pixel, the maximum class probability of each classification map was compared and the label corresponding to the higher probability was assigned to the pixel. Change maps were generated from the fused bi-temporal classification maps according to the encoding rules of the contest.

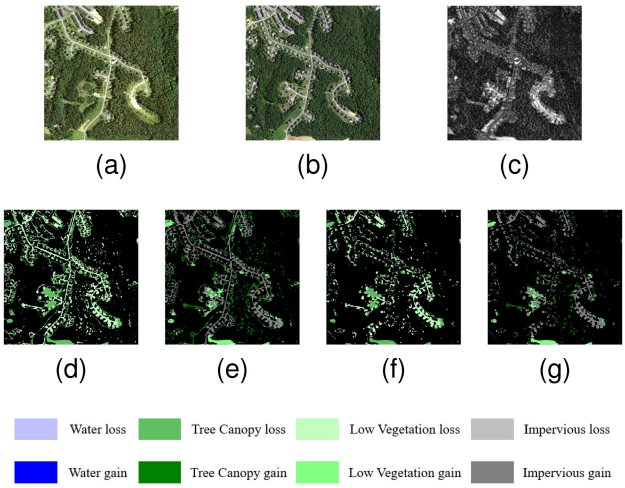

Fig. 9. Illustrations of the effect of CVA. (a) NAIP image of the year 2013. (b) NAIP image of the year 2017. (c) Feature difference map generated by CVA. (d) Change map before postprocessing (loss map). (e) Change map before postprocessing (gain map). (f) Change map after postprocessing based on CVA (loss map). (g) Change map after postprocessing based on CVA (gain map).

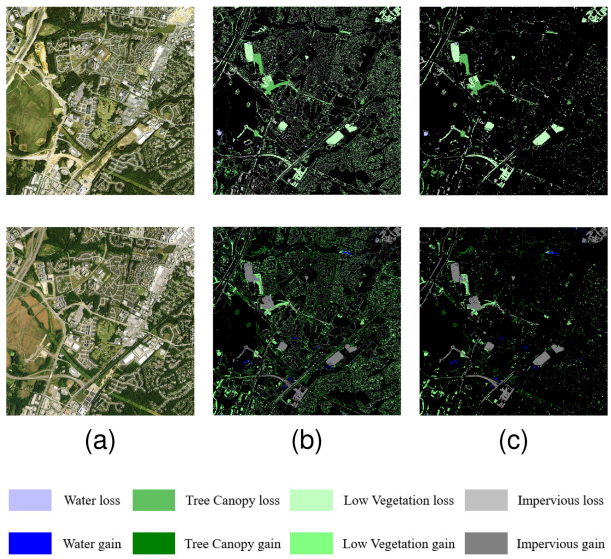

Fig. 10. Illustrations of the effect of spatial postprocessing. (a) Bi-temporal NAIP images (top: NAIP image of the year 2013; bottom: NAIP image of the year 2017). (c) Change map before postprocessing (Top: loss map; bottom: gain map). (d) Change map after spatial postprocessing (top: loss map; bottom: gain map).

C. Postprocessing

Four postprocessing steps from the pixel level to the scene level were performed on the change maps in order to reduce the commission errors.

1) Postprocessing Based on Change Probability: For each change pixel, the change probability was calculated with the following equation:

$$\text{prob}_{\text{change}} = \text{prob}_{2013} \times \text{prob}_{2017} \tag{6}$$

where prob_{2013} and prob_{2017} represent the maximum class probability of the years 2013 and 2017, respectively. For change pixels with change probability smaller than a threshold, the change labels were modified according to the class labels with the second maximum class probability. In addition, two kinds of changes, *Impervious* to *Water* and *Impervious* to *Tree Canopy* were nearly impossible according to the NLCD statistics [34]. For the pixels with these two kinds of changes, the change labels were modified to *Low Vegetation* to *Water* and *No Change*, respectively.

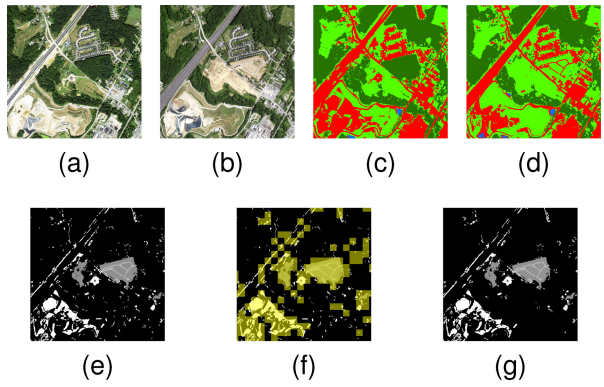

Fig. 11. Illustrations of the effect of the scene-level postprocessing. (a) NAIP image of the year 2013. (b) NAIP image of the year 2017. (c) Classification map of the year 2013 (blue: *Water*; dark green: *Tree Canopy*; green: *Low Vegetation*; red: *Impervious*). (d) Classification map of the year 2017. (e) Change map the before scene-level postprocessing. (f) Blockwise change map generated from the bi-temporal classification maps (marked in yellow). (g) Change map after the scene-level postprocessing.

2) Postprocessing Based on Change Vector Analysis: Change vector analysis (CVA) [40] was conducted on the bi-temporal feature maps generated by the last feature extraction layer of the FCN to obtain the feature difference map

$$\text{diff} = \sqrt{\sum_{i=1}^{n} (\text{feat}_{2013}^{i} - \text{feat}_{2017}^{i})^2} \tag{7}$$

$$\text{diff}_{\text{norm}} = \frac{\text{diff} - \text{diff}_{\text{min}}}{\text{diff}_{\text{max}} - \text{diff}_{\text{min}}} \tag{8}$$

where feat_{2013}^{i} and feat_{2017}^{i} represent the ith-dimensional feature in 2013 and 2017, respectively, and n is the total number of the features.

For each change pixel, if the feature difference is smaller than a threshold, the pixel was removed from the change map. A representative example for the CVA postprocessing is demonstrated in Fig. 9.

3) Spatial Postprocessing: Morphological opening and closing operators were performed on the change maps to deal with the noise in the edges of the objects that were caused by the difference of imaging angle between the bi-temporal NAIP images. Area and length-and-width ratio were used to further eliminate the small and long-narrow errors in the change maps. The effect of spatial postprocessing is shown in Fig. 10.

4) Postprocessing in the Scene Level: Motivated by [41], each image tile was divided into a series of blocks. For each block, the class distribution histogram was counted according to the classification maps. Each column of the histograms was

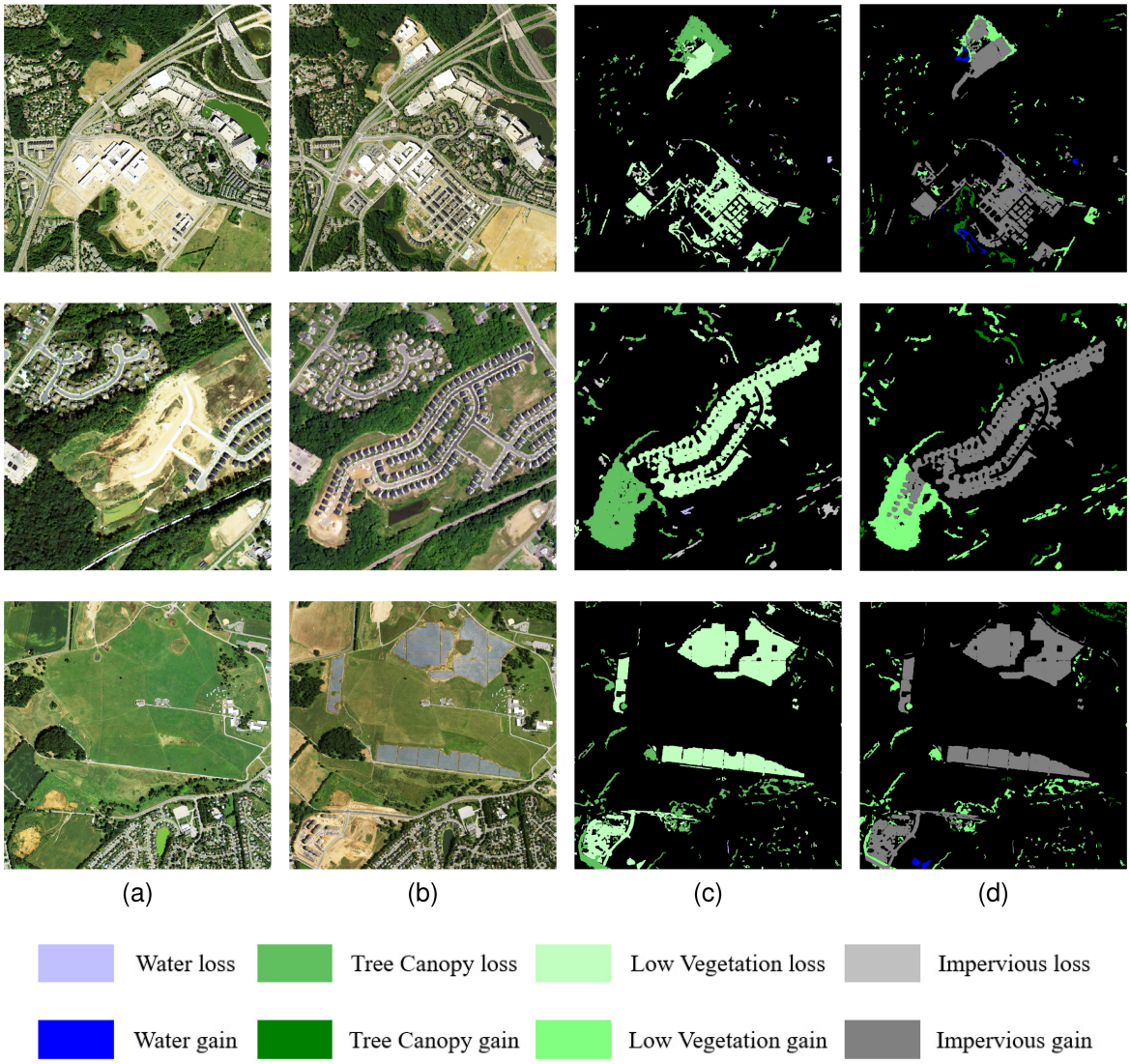

Fig. 12. Some example visual results generated by the proposed method. (a) NAIP image of the year 2013. (b) NAIP image of the year 2017. (c) Loss map. (d) Change map.

regarded as a feature of the corresponding block and the bi-temporal feature difference of each block was calculated. For each block, if the feature difference was smaller than a threshold, this block was considered no change and all the change objects in the block were removed. The effect of the scene-level postprocessing is shown in Fig. 11.

D. Results and Discussion

In this section, experimental results of the proposed method on the test dataset of the contest are reported. In order to investigate the effect of various components in the proposed method, six modules were tested.

1) NLCD-FCN: The FCN model was trained with NAIP images and low-resolution NLCD labels.

2) Pseudo-FCN: The FCN model was trained with NAIP images and high-resolution pseudolabels.

3) NLCD-MNDWI-FCN: NAIP images and MNDWI extracted from Landsat-8 images were used as input features of the FCN, and NLCD labels were used as training samples.

4) Pseudo-MNDWI-FCN: NAIP images and MNDWI extracted from Landsat-8 images were used as input features of the FCN, and high-resolution pseudolabels were used as training samples.

5) FCN-Fusion: Decision-level fusion was conducted on the classification maps generated by NLCD-FCN and Pseudo-MNDWI-FCN.

6) FCN-Fusion-Post: The proposed method. Postprocessing was performed on the change maps generated by FCN-fusion.

Table V presents the change detection accuracies on the test dataset for different modules. When training with NAIP images and NLCD labels (NLCD-FCN), the mean IoU of change detection was 0.5468. However, when training with NAIP images and pseudolabels (Pseudo-FCN), the mean IoU decreased by 2.47%,

TABLE V
CHANGE DETECTION ACCURACIES OF THE PROPOSED METHOD ON THE TEST DATASET OF THE CONTEST (EVALUATED BY INTERSECTION OVER UNION OF EACH CHANGE CLASSES, WITH BEST PERFORMANCE SHOWN IN BOLD)

Change Class	NLCD-FCN	Pseudo-FCN	NLCD-MNDWI-FCN	Pseudo-MNDWI-FCN	FCN-fusion	FCN-fusion-post
Water loss	0.7779	0.3764	0.6613	0.8278	**0.9183**	0.9132
Tree Canopy loss	0.7252	0.7799	0.6834	0.7822	0.8136	**0.8300**
Low Vegetation loss	0.6149	0.7221	0.5390	0.7098	0.7226	**0.7276**
Impervious loss	0.3830	0.3899	0.1904	0.3706	0.4474	**0.5553**
Water gain	0.2463	0.2198	0.0855	0.2663	0.2806	**0.3027**
Tree Canopy gain	0.3268	**0.5145**	0.3389	0.4813	0.4976	0.4984
Low Vegetation gain	0.6359	0.6764	0.4829	0.6276	0.6932	**0.7192**
Impervious gain	0.6645	0.4979	0.5621	0.7477	0.7748	**0.7790**
Mean IoU	0.5468	0.5221	0.4429	0.6017	0.6435	**0.6657**

which indicates that merely four-band NAIP images were insufficient to fit the pseudolabels that have higher resolution and larger data volume compared with the NLCD labels. After the MNDWI was added to the input features (Pseudo-MNDWI-FCN), the change detection accuracy was significantly raised (with the mean IoU 5.49% higher than that of NLCD-FCN) especially for the IoU of *Water loss* and *Impervious gain* which were highly influenced by the misclassification of *Water* and *Impervious*. Note that MNDWI can improve the change detection accuracy only when training with high-resolution labels. The result of NLCD-MNDWI-FCN shows a sharp decrease of the mean IoU to 0.4429 when MNDWI was used for training with low-resolution NLCD labels. Decision-level fusion of the results of NLCD-FCN and Pseudo-MNDWI-FCN (FCN-fusion) combined the advantages of both low-resolution and high-resolution labels and hence improved the mean IoU by 4.18% and increased the IoU of all the classes. By reducing commission errors in the change maps through postprocessing, FCN-fusion-post, the proposed method, further improved the accuracy and yielded the highest mean IoU of 0.6657. It is worth mentioning that when the pseudolabels were iterated, e.g., the method of the first-place team, the change detection accuracies can be further boosted. Some visual results generated by the proposed method are shown in Fig. 12. For future work, we will improve the transferability of the network models and promote the approach to a wider range of applications.

VI. CONCLUSION

During the last decades, Earth observation and remote sensing were dominantly used for deriving the current state of the Earth, i.e., answering questions such as "What can be observed where?" If temporal information was used then usually in the form of time series, i.e., exploiting the idea that the properties of a geo-/bio-physical process and therefore the signal might change (e.g., due to plant growth), but the underlying process remains fixed (e.g., same type of crop during a period of time). Only in the last years has the dynamic nature of the system Earth been moving into the focus of the remote sensing and computer vision communities and been the direct goal of corresponding methods to detect, monitor, and predict. On the one hand, the reason is better data availability. The increasing number of airborne and, more importantly, spaceborne sensors provides high-quality images with a high temporal resolution. On the other hand, methods to automatically interpret remote sensing imagery have evolved considerably. Together with the currently available required compute power, modern approaches have capabilities that go well beyond the production of static semantic maps based on a single image per scene. Instead, they allow to detect significant and meaningful changes. Early approaches aimed for simple binary maps, i.e., asking a Yes/No question whether a change had occurred in a given scene. Modern methods aim for a more fine-grained information and predict the type of change, i.e., "What has changed into what?" This task is of high importance as it not only characterizes the dynamic system Earth much better than static maps, but it also allows to assign different levels of relevance to different types of change (e.g., in the context of deforestation or the spread of impervious surfaces in urban areas). From a scientific point of view, this offers a multitude of interesting challenges. One of them is that fully supervised learning reaches its boundaries in the context of semantic change detection: Manually creating large-scale semantic annotations is already a tedious and costly task which scales very poorly to the amount of data required by modern deep learning methods. Creating training labels for change maps requires to analyze not only one but to carefully compare at least two images per scene requires an even larger workload. Requiring semantic classes of change only increases this difficulty. Thus, the availability of large-scale multitemporal image data with highly accurate annotations of semantic change is very unlikely. Instead, the corresponding approaches have to be able to handle either small datasets with accurate labels or large datasets with less accurate labels (weak supervision).

In this article, we summarize the Track MSD of the 2021 IEEE GRSS Data Fusion Contest, organized by the IEEE GRSS IADF TC, that was dedicated to exactly this real-world environmental challenge: to detect change within specific semantic classes. To this aim, Track MSD provided nine different layers of data: high-resolution aerial imagery from the NAIP for 2013 and 2017, five layers of low-resolution Landsat 8 multispectral imagery for each year from 2013 to 2017, and noisy low-resolution land cover labels from the NLCD for 2013 and 2016. These data not only cover the wide temporal range from 2013 to 2017 but also spread over the entire U.S. state of Maryland spanning. To increase realism, only low-resolution labeled samples were provided for training corresponding machine learning approaches (weak supervision).

The challenge of low-resolution labels was one of the main issues addressed by all winning teams in a similar manner: Semantic segmentation models trained on the high-resolution images but with low-resolution labels were used to predict high-resolution semantic maps which were used as pseudolabels in subsequent processing steps. Low-level image operations such as morphology and thresholding were employed to stabilize predictions, while fusing the output of an ensemble of neural networks was used to address the remaining variance in the estimates. Interestingly, traditional approaches, e.g., exploiting domain knowledge in the form of spectral indices, were used as well as modern approaches such as casting the task into a multiobjective learning problem where land cover classification and change detection are solved simultaneously.

The four top ranked solutions of this track presented their methods at IGARSS 2021, while the two top ranking solutions are described in this article in more detail. As in previous years, the DFC2021 attracted global attention with participants well distributed over the world, different affiliations, and career stages. This clearly illustrates the interest of the remote sensing and earth observation community to use the available tools and expertise to contribute to the social good. Furthermore, many of the contest participants were students, which shows that the DFC is introduced to early career scientists and used for educational purposes.

The data remain accessible after the DFC2021 on Azure in a read-only blob container[4] to allow further research and contributions. The CodaLab evaluation server and its public leaderboard[5] was reopened and made accessible from the contest website.[6] Thus, anyone can submit prediction results, obtain performance statistics, compare to other users, and, hopefully, improve on the results presented in this article.

Apart from the obvious societal impact, the DFC2021 Track MSD states also a very interesting challenge from a scientific point of view. The need of modern machine learning approaches to be trained on large-scale datasets inevitably leads to a decrease in label quality as it becomes infeasible to carefully curate thousands and millions of annotated images. Instead, existing products have to be leveraged which are (at least potentially) misaligned, at a lower resolution, partially outdated, or simply erroneous. Increasing the manual workload during annotation and quality assessment simply does not scale well enough. Modern methods of machine learning—in particular when applied to remote sensing and earth observation problems—need to be able to cope with these issues. Questions like if and how different sensor data should be fused to improve predictions, how to process different spatial and spectral resolutions within a common framework, and how to mitigate issues due to label noise or otherwise degraded reference data (e.g., lower resolutions) are far from being solved. All of these questions represent very active research directions where the future promises significant advancements. In this regard, the data of the DFC2021 provide a valuable benchmark dataset that can be used to evaluate all or only some of these aspects.

[4][Online]. Available: https://www.grss-ieee.org/community/technical-committees/2021-ieee-grss-data-fusion-contest-track-msd/ for download instructions

[5][Online]. Available: https://competitions.codalab.org/competitions/27956

[6][Online]. Available: https://www.grss-ieee.org/community/technical-committees/2021-ieee-grss-data-fusion-contest-track-msd/

Acknowledgment

The IADF TC chairs would like to thank the IEEE GRSS for continuously supporting the annual Data Fusion Contest through funding and resources.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] C. Robinson *et al.*, "Global land-cover mapping with weak supervision: Outcome of the 2020 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3185–3199, 2021.

[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] S. Kunwar *et al.*, "Large-scale semantic 3-D reconstruction: Outcome of the 2019 IEEE GRSS data fusion contest–Part A," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 922–935, 2021.

[18] Y. Lian *et al.*, "Large-scale semantic 3-d reconstruction: Outcome of the 2019 IEEE GRSS data fusion contest–Part B," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1158–1170, 2021.
[19] N. Yokoya *et al.*, "2021 data fusion contest: Geospatial artificial intelligence for social good [Technical Committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 287–C3, Mar. 2021.
[20] M. Buchhorn *et al.*, "Copernicus global land service: Land cover 100m: Collection 3: Epoch 2015: Globe (V3.0.1)," [Data set], Zenodo, 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3939038
[21] P. Yakubovskiy, "Segmentation models," 2019. [Online]. Available: https://github.com/qubvel/segmentation_models
[22] Z. Li, F. Lu, H. Zhang, G. Yang, and L. Zhang, "Change cross-detection based on label improvements and multi-model fusion for multi-temporal remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2054–2057.
[23] L. Tu, J. Li, and X. Huang, "High-resolution land cover change detection using low-resolution labels via a semi-supervised deep learning approach - 2021 IEEE data fusion contest track MSD," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2058–2061.
[24] Q. Bao *et al.*, "MRTA: Multi-resolution training algorithm for multitemporal semantic change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2062–2065.
[25] Z. Zheng, Y. Liu, S. Tian, J. Wang, A. Ma, and Y. Zhong, "Weakly supervised semantic change detection via label refinement framework," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2066–2069.
[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.
[27] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[30] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 539–546.
[31] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[33] H. Xu, "A study on information extraction of water body with the modified normalized difference water index (MNDWI)," *J. Remote Sens.*, vol. 9, no. 5, pp. 589–595, 2005.
[34] K. Malkin, C. Robinson, and N. Jojic, "High-resolution land cover change from low-resolution labels: Simple baselines for the 2021 IEEE GRSS Data Fusion Contest," 2021, *arXiv:2101.01154v1*.
[35] 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, pp. 370–373, 2019.
[36] X. Zhang, L. Liu, X. Chen, S. Xie, and Y. Gao, "Fine land-cover mapping in China using landsat datacube and an operational SPECLib-based approach," *Remote Sens.*, vol. 11, no. 9, pp. 1056–1073, 2019.
[37] J. Chen, A. Liao, J. Chen, S. Peng, L. Chen, and H. Zhang, "30-meter global land cover data product - GlobeLand30," *Geomatics World*, vol. 24, no. 1, pp. 1–8, 2017.
[38] M. C. Hansen *et al.*, "High-resolution global maps of 21st-century forest cover change," *Science*, vol. 342, no. 6160, pp. 850–853, 2013.
[39] J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, pp. 418–436, 2016.
[40] W. A. Malila, "Change vector analysis: An approach for detecting forest changes with landsat,," in *LARS Symposia*, pp. 326–335, 1980.
[41] D. Wen, X. Huang, L. Zhang, and J.-A. Benediktsson, "A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 609–625, Jan. 2016.

Zhuohong Li received the B.S. degree in communication engineering in 2020 from Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS).

His research interests include large-scale land cover mapping, semantic segmentation, and deep learning.

Fangxiao Lu received the B.S. degree in surveying and mapping engineering in 2020 from Wuhan University, Wuhan, China, where he is currently working toward the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS).

His research interests include road network extraction, semantic segmentation, and deep learning.

Hongyan Zhang (Senior Member, IEEE) received the B.S. degree in geographic information system and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2005 and 2010, respectively.

He has been a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University, since 2016. He is a Young Chang-Jiang Scholar appointed by the Ministry of Education of China, Beijing, China.

Dr. Zhang is currently a Reviewer for more than 30 international academic journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *IEEE Geoscience and Remote Sensing Letters*, and so on.

Lilin Tu received the B.S. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2020. He is currently pursuing the Ph.D. degree in photogrammetry and remote sensing in the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include hyperspectral image change detection, semantic segmentation, and deep learning.

Jiayi Li (Senior Member, IEEE) received the B.S. degree in surveying and mapping engineering from Central South University, Changsha, China, in 2011, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016.

She is an Assistant Professor with the School of Remote Sensing and Information Engineering, Wuhan University. She has authored more than 30 peer-reviewed articles (Science Citation Index (SCI) articles) in international journals. Her research interests include hyperspectral imagery, sparse representation, computation vision and pattern recognition, and remote sensing images.

Dr. Li is a Reviewer for more than ten international journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, IEEE SIGNAL PROCESSING LETTERS, and *International Journal of Remote Sensing*. She is the Guest Editor for the Special Issue on *Change Detection Using MultiSource Remotely Sensed Imagery for the Remote Sensing* (an open access journal from MDPI).

Xin Huang (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing in 2009 from Wuhan University, Wuhan, China, working with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS).

He is currently a Full Professor with Wuhan University, where he teaches remote sensing, image interpretation, etc. He is the Head of the Institute of Remote Sensing Information Processing (IRSIP), School of Remote Sensing and Information Engineering, Wuhan University. He has authored or coauthored more than 170 peer-reviewed articles (SCI papers) in international journals. His research interests include remote sensing image processing methods and applications. He has been supported by the National Program for Support of Top-notch Young Professionals (2017), the China National Science Fund for Excellent Young Scholars (2015), and the New Century Excellent Talents in University from the Ministry of Education of China (2011).

Dr. Huang was the recipient of the Boeing Award for the Best Paper in Image Analysis and Interpretation from the American Society for Photogrammetry and Remote Sensing (ASPRS) in 2010, the John I. Davidson President's Award from ASPRS in 2018, and the National Excellent Doctoral Dissertation Award of China in 2012. In 2011, he was recognized by the IEEE Geoscience and Remote Sensing Society (GRSS) as the Best Reviewer for *IEEE Geoscience and Remote Sensing Letters*. He was the winner of the IEEE GRSS Data Fusion Contest in the years of 2014 and 2021. He was the Lead Guest Editor of the special issue for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, *Journal of Applied Remote Sensing, Photogrammetric Engineering and Remote Sensing*, and *Remote Sensing*. He was an Associate Editor for the *Photogrammetric Engineering and Remote Sensing* from 2016 to 2019 and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS from 2014 to 2020 and has been serving as an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING since 2018. He has also been an Editorial Board Member of the *Remote Sensing of Environment* since 2019.

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 Computer Science 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 Microsoft's AI for Good Research Lab, 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.

Nikolay Malkin received the B.S. degree in mathematics from the University of Washington, Seattle, WA, USA, in 2015, and the Ph.D. degree from Yale University, New Haven, CT, USA, in 2021.

He is currently a Postdoctoral Researcher with Mila–Québec Artificial Intelligence Institute, Montreal, QC, Canada. His research interests include deep learning-based reasoning and induction of compositional structure in deep generative models, as well as applications to natural language processing and computer vision, including 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.

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

Pedram Ghamisi (Senior Member, IEEE) received the M.Sc. (Hons.) degree in remote sensing from 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 the Head of the Machine Learning Group, Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology (HZDR-HIF), Freiberg, Germany; Cofounder of VasoGnosis Inc., with two branches in the USA; and a Visiting Professor and Leader of AI4RS, Institute of Advanced Research in Artificial Intelligence (IARAI), Vienna, Austria.

His research focuses on ensemble methods for image analysis. Dr. Ghamisi was the Vice-Chair of the IEEE Geoscience and Remote Sensing Society Image Analysis and Data Fusion Technical Committee from 2019 to 2021.

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

His current research interests include ensemble methods for image analysis.

Dr. Hänsch was the Co-Chair (2017–2021) and is the current Chair (2021–2023) of the IEEE Geoscience and Remote Sensing Society Image Analysis and Data Fusion Technical Committee and the Co-Chair of the International Society for Photogrammetry and Remote Sensing Working Group II/1 (Image Orientation). He is currently an Associate Editor for *Geosience and Remote Sensing Letters* and an Editor-in-Chief for the *GRSS eNewsletter*.

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, Japan, where he leads the Geoinformatics Unit.

Dr. Yokoya was the Chair of the IEEE Geoscience and Remote Sensing Society Image Analysis and Data Fusion Technical Committee from 2019 to 2021. He is currently an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.