MSCDUNet: A Deep Learning Framework for Built-Up Area Change Detection Integrating Multispectral, SAR, and VHR Data

Haoyang Li, Fangjie Zhu, Xiaoyu Zheng, Mengxi Liu, Student Member, IEEE, and Guangzhao Chen

Abstract—Built-up area change detection (CD) plays an important role in city management, which always uses very high spatial resolution (VHR) remote sensing data to extract refined spatial information. Recently, many CD models based on deep learning with VHR data have been proposed. However, due to the complex background information and natural landscape changes, VHR with optical RGB features is hard to extract changes exactly. To this end, we tend to explore the abundant channel information of multispectral and SAR data as a supplement to the refined spatial features of VHR images. We propose a new deep learning framework called multisource CD UNet++ (MSCDUNet), integrating multispectral, SAR, and VHR data for built-up area CD. First, we label and reform two new built-up area CD datasets containing multispectral, SAR, and VHR data: multisource built-up change (MSBC) and multisource OSCD (MSOSCD) datasets. Second, a feature selection method based on random forest is introduced to choose effective features from multispectral and SAR images. Finally, a multilevel heterogeneous feature fusion module is embedded in MSCDUNet to combine multifeatures for CD. Experiments are conducted on both the MSOSCD and the MSBC datasets. Compared to other CD methods based on VHR images, our proposal achieves the highest accuracy on both datasets and proves the effectiveness of multispectral, SAR, and VHR data fusion for CD. The dataset in the article will be available for download from the following link.1

Index Terms—Benchmark dataset, built-up, change detection (CD), deep learning (DL), multispectral data fusion, very high resolution (VHR).

I. INTRODUCTION

The ever-growing remote sensing data promote the development of remote sensing applications with the rapid increase of earth observation satellites [1]. As the basic data source of geography, remote sensing images with high spatial, spectral and temporal resolution improve the ability of earth surface observation and understanding [2]. In particular, very high spatial resolution (VHR) images increasing to the submeter level give more refined spatial details in mapping human activity spaces. In recent years, VHR images containing RGB channels have been widely used in earth observation including land use and land cover (LULC) classification [3], local climate zone classification [4], building extraction [5], change detection (CD) [6], etc. Built-up area CD is a specific CD task aiming at detecting changes from croplands, forest, and other natural land cover to buildings, impervious surface and semibuilt area or vice versa, which is essential for government to manage urban expansion and protect farm and natural resources [7].

CD is the operation of quantitative analysis and determination of land cover and land use changes from the same area over two or more distinct periods. Spectrum analysis based on multispectral images is difficult to characterize the details of building changes in urban areas due to the coarse spatial resolution and the mixed pixel problem. The increase in VHR images creates conditions for detecting complex and subtle changes in urban areas to meet the requirements of city management [8]. Recently, a large number of deep learning (DL) works have been proposed to solve the problem of urban CD with VHR images [9]. DL works, especially convolutional neural networks (CNNs), are able to extract complicated spatial information from VHR data through the robust extraction of texture features and spatial correlations by CNNs structures. Many dense prediction DL networks, such as UNet++ [10], STANet [11], DSAMnet [12], have been applied in CD tasks using bitemporal VHR images.

Compared with traditional CD tasks using VHR images, built-up area CD is more challenging under complex scene for the reason that built-up area CD focuses on the change between natural objects and built-up features, instead of direct vision changes [13]. Indeed, DL-based CNNs have done well in extracting spatial information from objects with homogeneous texture features and distinct boundary characteristics, such as a change from building to building. However, it is difficult for CNNs to distinguish changes between natural land cover and human activity related classes and extract the fine boundary, the RGB channel information of VHR images is susceptible to imaging quality, and not enough for high performance of CNNs based

Manuscript received March 31, 2022; revised May 17, 2022; accepted May 26, 2022. Date of publication June 8, 2022; date of current version July 4, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61976234, in part by the Guangdong Natural Science Foundation under Grant 2019A1515011057, and in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 22qndt2001.

(Corresponding author: Guangzhao Chen.)

Haoyang Li, Fangjie Zhu, Xiaoyu Zheng, and Mengxi Liu are with the Division of Landscape Architecture, Faculty of Architecture, The University of Hong Kong, Hong Kong SAR 999077, China (e-mail: lihy256@mail2.sysu.edu.cn; zhufj6@mail2.sysu.edu.cn; zhengxy75@mail2.sysu.edu.cn; liumx23@mail2.sysu.edu.cn).

Guangzhao Chen is with the Division of Landscape Architecture, Faculty of Architecture, The University of Hong Kong, Hong Kong SAR 999077, China (e-mail: chengzh7@hku.hk).

Digital Object Identifier 10.1109/JSTARS.2022.3181155

1https://github.com/Lihy256/MSCDUNet

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
CD algorithms such as changes from grassland to bare land and from bare land to semibuilt area. Meantime, bitemporal VHR data face the problem of low imaging quality, for example, the color of grassland may not be green, resulting in the confusion of visual presentation of natural objects. The limited channel information of VHR images may cause pseudochanges, and real changes might be easily ignored by visual interpretation and CNNs. In a word, the RGB channel information of VHR images is susceptible to imaging quality, and not enough for the high performance of CNNs-based CD algorithms.

Meanwhile, multispectral images can provide abundant and stable spectral information that helps to easily detect natural land cover changes using spectral differences [14]. SAR with the backscattering information also expresses different characteristics to help detect land cover changes [15]. Although the coarse spatial resolution of multispectral and SAR data restricts their use in urban CD, it is a potential way to integrate multispectral and SAR data with VHR data to improve the accuracy of detection with only VHR data as data source. Some studies have integrated multispectral and SAR features to improve VHR land cover classification and semantic segmentation [16]–[19], but there are few works of fusing multispectral, SAR, and VHR data for CD tasks. Most of the studies focus on combining multispectral and SAR data in coarse resolution CD [20]–[23]. However, fusing multispectral and SAR for CD is short of detailed spatial information and only using VHR faces the shortage of spectral and backscattering information. Therefore, the aim of the study is to integrate complementary features of multispectral, SAR, and VHR data to improve the accuracy of built-up area CD.

The main problems of constructing a multisource data integration framework for built-up area CD are listed as follows.

1) Few benchmark CD datasets can be used for the verification of built-up area CD algorithm. On the one hand, few datasets focus on built-up area CD. On the other hand, there is no CD dataset that contains multispectral, SAR, and VHR data simultaneously. Most of the datasets did not open the spatial coordinates, which restrict their reform for a specific task.

2) The multisource heterogeneous data with different spatial resolution may adversely affect the DL network. The relatively coarse resolution of multispectral and SAR data is possible to reduce the fine boundary details of VHR in the fusion process. Therefore, it is necessary to build an adaptive feature fusion module to fill the resolution gap and maintain the detailed spatial features and abundant channel features.

3) Quantities of features from multispectral and SAR data consume large amount of training time and lead to unbalanced inputs. Bands reduction and features selection of multispectral and SAR data can emphasize effective features and suppress useless information and noises. However, selecting optimal features from unbalanced inputs with CNNs testing is cumbersome and time-consuming.

To solve the above problems, in this article, we propose a DL-based multisource data fusion framework for built-up area CD. First, two benchmark datasets are made for built-up area CD. One is labeled based on GF-2 VHR images, and the other is reformed from an existing dataset—Onera Satellite CD (OSCD) dataset [24]. Second, a DL-based multisource data CD network named MSCDUNet is built for new built-up detection. A multilevel feature fusion module is proposed to adaptively fuse different spatial and spectral features. Third, to optimize the input features, the important features are extracted using random forest (RF) on Google Earth Engine (GEE).

The main contributions of this article are concluded as follows.

1) We label the first two multisource data built-up CD datasets MSBC and MSOSCD, filling the gap of built-up CD datasets including multispectral, SAR, and VHR.

2) Multisource CD UNet++ network (MSCDUNet) is proposed to extract and fuse diverse spatial and channel features with a multilevel feature fusion module, adaptively integrating different resolution features.

3) A simple and time-saving feature selection method is introduced to extract key features of multispectral and SAR data as a preprocess of MSCDUnet, which helps to reduce the amounts of data input and achieve higher accuracy than the original bands.

II. RELATED WORK

A. DL in CD

In the early stage, traditional CD methods based on image differencing and ratios [25], postclassification [26], feature learning and transformation [27] are widely used. Conventional hand-crafted feature-based methods have the inevitable difficulty in differentiating between relevant change and irrelevant change [28]. Recently, DL works, capable to learn discriminatory and representative characteristics from data in a hierarchical way, have been starting to catch on in CD [29].

Based on the structure and input form of the networks, DL-based CD can be divided into two main categories: single branch and double branches. Single branch usually means stacking bitemporal or multitemporal images together as a multiband image and feeding into a single-branch network to train and predict. FC-EF, a CD model based on UNet, was proposed to detect changes by concatenating bitemporal VHR images [30]. CD-UNet++ [31], an improved UNet++ network, was proposed to extract global and fine-grained information from the concatenated coregistered image pairs. UNetLSTM [32] was used to detect urban changes with multitemporal images.

Double-branch frameworks, particularly referring to Siamese-based networks, have been applied to extract deep features in each branch and detect the differences between two branches [33]–[37]. Siamese neural network [38] was first used to verify handwritten input board signatures. The main advantage of double-branch frameworks is that they can process two inputs by two subnetworks simultaneously. FC-Siam-conv and FC-Siam-diff [30] were proposed to train end-to-end from the origin bitemporal VHR images, showing an excellent accuracy and inference speed without postprocessing. ChangeNet [39] was also a parallel deep CNN, which took advantage of features from different levels to detect changes.
in the scene at a semantic level. To utilize spatial temporal information better, recurrent neural networks and attention modules were introduced to extract spatial and channel correlation. STANet [11], a spatial-temporal attention-based network, is integrated with a spatial and temporal attention mechanism to gain more spatial and temporal features from VHR images.

Recently, more and more DL-based CD works have been developed to solve specific applications. To map building changes, EGRCNN [40] appropriately incorporated the discriminative features and the estimated edge structures to enhance features of buildings. ChangeOS [41] used OBIA to extract building damage. To map detailed LULC changes from VHR images, a Siamese global learning (Siam-GL) framework [42] was applied in semantic CD. To decouple semantic and binary CD, ChangeMask [43] used encoder-transformer-decoder architecture with multitask loss function. In summary, most of the DL-based CD methods are concentrating on constructing effective CNN structure with VHR images.

B. Multisource Data Fusion for CD

With the development of remote sensing technology, various images can be acquired from different sensors. VHR images have abundant fine-grained spatial information. Multispectral and SAR data can reflect the physical properties of land objects in detail. However, VHR has few bands and unstable image quality [1]. The acquisition of multispectral images is easily affected by the weather. In addition, SAR data often contain speckle noise. These problems that cannot be ignored limit their application in the field of remote sensing. Multisource data fusion, as a bridge to complement the advantages of them, is emerging and developing.

Many studies on remote sensing classification have integrated multisource data to improve the performance, which can be divided into the following three levels [44].

1) Pixel-level fusion is the combination of raw data from multiple sources into single resolution data. The conventional method of it includes pixel algebra [45], autoencoder [46], and so on. Ferriaris et al. [18] used the idea of solving an inverse problem to fuse a low spatial resolution image and a low spectral resolution image, gaining a reference image with high spatial resolution and high temporal resolution for CD.

2) Feature-level fusion is the combination of features from raw data. All the useful features from the source image are extracted artificially or automatically. Li et al. [19] built an SVM-fusion classifier to fuse medium-resolution features extracted from Landsat data and high-resolution features from Google Earth imagery extracted by CNN. Yang et al. [47] built an attention-based fusion method to integrate VHR and auxiliary data to improve the performance of semantic segmentation.

3) Decision-level fusion combines the decisions of multiple classifiers into a common decision about the activity that occurred. Shao et al. [48] fused independent detection results from GF-1 and Sentinel-1A with the Dempster–Shafer theory at the decision level.

Research works have shown that the fusion of SAR and multispectral images has a high potential to be used for CD tasks [23], [49]. Most of these studies focus on the integration of multispectral and SAR data, which has a relatively small resolution gap. Yousif and Ban [50] utilized both the $k$-mean clustering algorithm and SVM on the concatenation between SAR and Landsat. Seo et al. [45] applied an RF regression to create a fused image for CD, where the fused image combined the surface roughness characteristics of the SAR image and the spectral characteristics of the MS image. With the development of DL, the pseudo-Siamese network is used for fusion features from SAR and multispectral, which is trained to regress the feature vector of bitemporal concatenated SAR-multispectral input data [51]. Hafner et al. [22] introduced a dual stream concept to process Sentinel-1 and Sentinel-2 separately before combining extracted features based on U-Net at a decision stage, whereas Zhang et al. [52] used the ResNet as a backbone for bipartite deep networks. To highlight the features from multispectral images, some research works used common multispectral indexes as one input. Benedetti et al. [53] fused three different features from SAR, Sentinel-2, and spectral difference by the weighted average block. Mishra and Susaki [21] developed an empirical relationship between change magnitude images and normalized difference ratio from SAR images and NDVI from multispectral images.

Several research works tried to fuse VHR with other images for CD, but without DL works. Touati and Mignotte [54] considered the CD as the estimation of overconstrained problem and fused different binary segmentation results, obtained from this similarity-feature map by different automatic thresholding algorithms. Solano-Correa et al. [55] aggregated the multisensory information in a homogeneous physical way based on linear transformations (tasseled caps and orthogonal equations) and Kwan et al. [56] presented a method named hybrid color map for fusion between Landsat and Worldview images.

To sum up, the fusion between SAR and multispectral has become a common technique for CD. However, fusion among VHR with SAR and multispectral images is rarely applied as far as we know even if any of the data source has a specific advantage for CD. Therefore, we tend to propose a multisource data fusion using VHR, SAR, and multispectral images together to detect built-up area changes and discuss the efficiency of each data source.

III. Dataset

In this section, we will introduce two datasets (MSBC and MSOSCD) that are prepared for experiments. Two datasets containing multispectral, SAR, and VHR multispectral data are built for experiments, which are named multisource OSCD (MSOSCD) dataset and multisource built-up change (MSBC) dataset. Both of the datasets be made openly available for all research needs.

A. MSBC Dataset

The study area of the MSBC dataset is located in Guigang City, Guangxi Province, China, as shown in Fig. 1. The original remote sensing image data contain: 1) bitemporal GF-2 VHR images with geographic registration of Guigang City obtained
in January 2018 and January 2019, with a spatial resolution of 2 m; 2) Sentinel-1 and Sentinel-2A satellite remote sensing images of Guigang City obtained in January 2018 and January 2019. The sentinel-1 radar image has two bands (VV and VH), and the sentinel-2A satellite image has 7 bands (B2, B3, B4, B8, B8A, B11, B12).

In this study area, we labeled the changed region vector files of the two parts of this city, dividing them into the training region and test region. In the training region, we used the VHR image of the changed area to make slice data, and clipped it with an overlap rate of 70%. Then, we clipped the radar data and multispectral data according to their geographical coordinate by slice data. The train and validation data were collected from the train region and test data were collected in the test region in Fig. 1. The train, validation, and test datasets contain 2079, 1164, and 526 pairs of data. The test dataset was without the overlap operation. Some images in this dataset are shown in Fig. 1. From the images of multispectral and SAR, we can determine that the change areas are apparent. Therefore, integrating multispectral and SAR with VHR data to improve CD is reasonable in vision.

B. MSOSCD Dataset

The second dataset is reformed from OSCD dataset. The origin dataset contains 24 pairs of multispectral images obtained from the Sentinel-2 satellites between 2015 and 2018. Locations of the images are worldwide including Brazil, USA, Europe, Middle-East, and Asia. For each location, OSCD provides pairs of 13-band images and annotations labeled based on Sentinel-2. The spatial resolution of OSCD ranges from 10 to 60 m, which does not satisfy the refined label of VHR requirements. The advantage of reforming OSCD is that the origin dataset offers time and location references to help collect corresponding VHR and SAR data, although test datasets from the origin dataset conceal the label resulting in less data usability for us.

We selected eight cities from OSCD and collected the corresponding VHR data. Considering that the resolution of the original labeled data is different from that of VHR data, the boundary of the dataset needed to be corrected and refined. First, we converted the original labels into vectors and manually corrected some labels to unchanged areas according to VHR data. These label errors were caused by the time difference between VHR and Sentinel-2 data. Second, we deleted labels not belonging to built-up area changes and refine the boundary of the changed area by comparing with the VHR images. Then, we converted these vectors into grid labels with the same resolution as the VHR images and clipped the images with an overlap rate of 50%. Finally, we collected Sentinel-1 data with the same obtained time as Sentinel-2 and formed the MSOSCD dataset. The train, validation, and test datasets are randomly split and contain 3805, 408, and 894 pairs of images, respectively. The origin datasets and some images in this dataset are shown in Fig. 2.

IV. METHODS

In this section, we will introduce three main methods. Important features are selected from multispectral and SAR data with RF on GEE in Section IV-A. In Sections IV-B and IV-C,
MSCDUNet is proposed for fusing multisource data to detect built-up changes, in which multilevel fusion and deep supervision are introduced to fuse multispatial and spectral information.

A. RF Feature Selection

Multispectral images and SAR data provide various and useful information to help detect land cover changes. However, the origin features provided may contain redundant information, which will not only slow down the operation speed of the DL network but may also bring invalidation information and reduce the accuracy of CD. Therefore, selecting appropriate features is necessary before training neural network. For keeping the physical meaning of bands and visualizing their importance, we choose RF as the feature filter, which is a stable method for feature selection. Besides, due to its high power and stability, it has been recommended for use in the field of remote sensing [45], [57].

RF [58] is an integrated algorithm composed of decision trees \{h(X, θ_k), k = 1, 2, \ldots, K\}, where \{θ_k\} is a random vector subject to independent and identically distribution and K represents the number of decision trees. For every node of each decision tree, extract m features from total N features randomly and calculate the amount of information contained in each feature, and then select the feature containing the most information to split the node. Each tree grows to the maximum without any cutting. 1) Randomly and repeatedly select K group samples from original training dataset for K decision trees. The samples that are not selected each time form the out of bag data. 2) For every node of each decision tree, extract m features from total N features randomly and calculate the amount of information contained in each feature, and then select the feature containing the most information to split the node. 3) Each tree grows to the maximum without any cutting. 4) The well-trained decision trees form an RF. The classification result depends on the number of votes created by each decision trees, which can be formulated as

\[
P_q = \frac{1}{K} \sum_{i=1}^{K} z^i_q (x) \tag{1}
\]

where K is the number of decision trees, q represents the category to which sample x belongs, and \(z^i_q (x)\) represents the probability that sample x is determined to belong to class q in the i\textsuperscript{th} decision tree. The most important key to build decision tree is the optimal segmentation criteria, obtaining the best classification result for each leaf. Breiman recommended using the Gini impurity index \(I_G\) as the optimal segmentation criteria, which is the probability of minimizing misclassification and can be formulated as

\[
I_G (f) = \sum_{i=1}^{c} f_i (1 - f_i) \tag{2}
\]

where \(f_i\) is the probability of class i (i \in \{0, 1\}) at node c. After training, the features for each tree are ranked the features by \(I_G\), giving the importance. Then, the average importance of features in RF is calculated to get the final feature importance value and ranking.

Based on the GEE, some indexes including NDVI [60], SAVI [61], LSWI [62], MNDWI, BSI, IBI are used to represent features of multispectral images. Gray-level co-occurrence matrix (GLCM) [63] is used to extract feature from SAR data. The above features are fed into RF, and then low ranked features are filtered until a balance is reached between accuracy and the number of features.

B. Multisource CD UNet++

Considering that UNet++ has only one input, which is not suitable for image pairs with different resolution, a Siamese version of UNet++ is employed to extract features from image pairs. The pseudo-Siamese model is chosen because two branches in this model do not share the same weights, which meet the requirements of feature extraction for heterogeneous input images.

Fig. 3 shows the designed pseudo-Siamese UNet++ model. In the network, each branch follows the architecture of UNet++, which consists of an encoder subnetwork and a decoder subnetwork that are connected through a series of nested dense convolution block. The most important and useful part for UNet++ is the redesigned skip pathway. Compared with UNet, the redesigned skip pathway bridges the semantic gap between the feature maps from the encoder and decoder and fuses them appropriately. Given node \(X_{i,j}\), its i indexes the downsampling layer along the encoder and j indexes the convolution layer along the skip pathway. Its output value \(x_{i,j}\) can be formulated as

\[
x_{i,j} = \begin{cases} 
\mathcal{H} (x_{i-1,j}), & j = 0 \\
\mathcal{U} (\left[ x_{i,k} \right]_{k=0}^{i-3}), & j > 0 
\end{cases} \tag{3}
\]

where \(\mathcal{H}(\cdot)\) is double convolution layer with ReLU and batch normalization layers, \(\mathcal{U}(\cdot)\) is an upsampling layer, and [\cdot] denotes the concatenation layer.

Take node \(X^{2,2}\) as an example, it receives the skip connection from all the previous convolution units at the same level and the corresponding up-sampled output of the lower level, instead of which just receives the value of \(X^{3,1}\) in UNet. In this way, the
encoder feature maps are closer to the corresponding decoder parts in semantic level, which is helpful for optimization.

In the proposed model, two parallel UNet++ are used to extract different features from different sensors, keeping characteristic of images as much as possible. And the final feature map is the concatenation from four precious fusion features. Notably, we retain periodic feature fusion results to facilitate the smooth implementation of deep supervision.

C. Multilevel Fusion and Deep Supervision

1) Multilevel Fusion: After fusing the multispectral and SAR data and VHR data with concatenation, we keep every fusion result for multilevel fusion and deep supervision. As shown in Fig. 4, the eight output nodes \( \{X_{0,1}, X_{0,2}, X_{0,3}, X_{0,4}, Y_{0,1}, Y_{0,2}, Y_{0,3}, Y_{0,4}\} \) are pairwise fused by concatenation, acquiring \( \{out1, out2, out3, out4\} \), a convolutional layer that reduces the 32-dimensional (32-D) feature image to a 2-D CD probability map, is followed to obtain subnetwork output results \( \{out1, out2, out3, out4\} \). Then, a new feature \( out5 \) is generated by concatenating the four feature fusion results

\[ out5 = out1 \otimes out2 \otimes out3 \otimes out4 \quad (4) \]

where \( \otimes \) represents the concatenation operation. Then, \( out5 \) is followed by two convolutional layers to get the final output. Therefore, five outputs \( \{out1, out2, out3, out4, out5\} \) are generated by concatenation and then features from different resolutions and levels are embedded in the final output.

Each convolution operation uses padding to ensure the consistent size of the feature images and the input images. The fused image contains abundant spatial details. Meanwhile, the integration of features from upper and lower levels is beneficial to the subsequent mining of multilevel fusion features.

Multilevel fusion could integrate heterogeneous from each subnet. The fusion module adaptively fuses the different resolution multilayer features after decoding that represents the changing information, which can maintain the spatial details and spectral features in the multilayer.

2) Deep Supervision and Loss Function: Inspired by UNet++, deep supervision is implemented through supervision side-output layers from subnetwork, which can not only help network learn more meaningful features from different levels but also overcome the vanish gradient problems. In our model, a weighted loss function with multilevel outputs is introduced to realize deep supervision, which can be formulated as

\[ L = w_D \ast L_{\text{dice}} + w_B \ast L_{\text{BCE}} \quad (5) \]

\[ L_{\text{dice}} = \frac{1}{N} \sum_{i=1}^{N} L_{\text{dice},i} \quad (6) \]

where \( L_{\text{dice},i} \) is the Dice coefficient loss from \( out_i \), \( L_{\text{BCE}} \) is the binary cross entropy loss from \( out5 \), and \( w \) refers to the weight that balances the two loss.

Dice coefficient loss is a common loss function for segmentation and CD tasks. It is based on Dice coefficient, which calculates the similarity of two sets. So, this loss can be formulated as

\[ L_{\text{dice},i} = 1 - \frac{2 \ast |\hat{Y}_i \cap Y|}{|\hat{Y}_i| + |Y|} \quad (7) \]

where \( Y \) and \( \hat{Y}_i \) represent the ground truth labels and the predicted probabilities from \( out_i \), respectively.

Dice loss performs well for the scene with seriously imbalanced positive and negative samples, and pays more attention to mining the foreground area in the process of training. Meanwhile, it is a region-related loss, which means loss and gradient of one pixel are related to its label and prediction as well as its neighborhoods. However, it is unstable because of the preference, especially for small targets. The BCE loss is to alleviate oscillation during training.

However, only using Dice loss always got a fluctuating convergence process. BCE loss often is used in binary classification and more stable than Dice loss. Therefore, we combine BCE loss and Dice loss to gain a stable train process. BCE loss can be formulated as

\[ L_{\text{BCE}} = Y \log (\hat{Y}) + (1 - Y) \log (1 - \hat{Y}) \quad (8) \]

where \( Y \) and \( \hat{Y} \) represent the ground truth labels and the predicted probabilities, respectively.

V. EXPERIMENTAL AND RESULTS ANALYSIS

A. Feature Selection by RF

Feature selection is carried out based on the accuracy of CD by RF with different feature inputs and GEE. The amount of our dataset involved in the research is large. We produced two datasets, one covering the whole area of Guigang, about 10 602 km² and the other covering eight representative cities around the world. We tend to use a platform with convenient data acquisition and fast operation. GEE, an ideal choice, is one of the most advanced cloud geographic information processing platforms with petabytes of geographic data. Meanwhile, it provides RFs and other classic machine learning algorithms of calls. Compared with other traditional platforms, it has a much faster running speed and more convenient data analysis methods. For these advantages, it has been widely used in the field of ecology [64], agriculture [65], [66], and urban extraction [67].

We select features on two datasets, respectively, with the same steps. First, select 5000 samples from changed area and 5000
samples from unchanged area randomly. Then, representative features are introduced to characterize images according to their importance in RF. We choose some common normalized indexes to generalize the feature of multispectral images, whereas GLCM is a method to calculate the texture features in SAR data. After that, use the idea of stepwise regression backward method \cite{68} to remove or retain features until the accuracy no longer changes or the number of features meets certain requirements.

In detail, GLCM could depict texture with 18 statistical properties of texture feature, which are presented in SUP Table I. And to present the difference between the two-phase images, we calculate the 18 statistical properties based on the difference and quotient images. Finally, we gain 36 statistical properties from them in total.

Then, the normalized indexes we choose for multispectral images are presented in SUP Table II. They can highlight typical ground objects such as vegetation, buildings, soil, and water in the image, which are more conductive to image feature extraction. Like SAR feature calculation, we calculate the normalized indexes for all multispectral images, and then subtract and divide the images in the same area according to the imaging time, reflecting the changing trend of the later phase compared with the previous phase. In the end, we gain 14 images to reflect the spectral change of land objects. To sum up, the class and the number of optional features are presented in Table I. After screening each band and alternative features, the final retained features are presented in the second column $S_1 \ast S_2$ in SUP Table III.

### B. Implementation Details

The preprocessing of VHR remote sensing images includes histogram matching and Laplacian filtering operation, so that the texture features and spectral features of the two images are consistent. The difference and quotient images of the two-time phases and the original images are input into the network. Then we resample multispectral and SAR data to $256 \times 256$ as VHR slice data. Building and preprocessing the datasets are completed by QGIS, GDAL-Python, and MATLAB.

For the model training, the optimizer is Adam, the initial learning rate is 0.01, the beta is 0.9 and 0.999, and the weight decay is 0.0001. The loss function is the combination of Dice loss and BCE loss. The learning rate decay strategy is StepLR with step size as 10, gamma as 0.8, and last epoch as 200. For the loss function, $w_D$ is 0.6 and $w_B$ is 0.4 based on experiments (shown in Section V). All models are trained for 250 epochs. The data augmentation for train data consists of random rotations of 90° and 180°, random horizontal and vertical flips, salt and pepper noise, and Gaussian noise. The validation and test datasets do not have an augmentation operation. Our models and experiments are based on the open-source DL framework PyTorch. The experimental environment is Centos 7.5.1804. The GPU is GeForce RTX 2080ti. The CPU is Intel(R) Xeon(R) CPU E5-2680.

### C. Comparisons With Baseline Methods

#### 1) Comparative Models:

1. Fully Convolutional-Early Fusion (FC-EF): FC-EF is a U-Net-based network, in which the bitemporal images are concatenated as a multiband image for input. Skip connections modules are used to transport the multiscale information from lower to deeper layers to maintain spatial information.
2. Fully Convolutional-Siamese-Concatenation (FC-Siam-conc): FC-Siam-conc is a FC-EF-based model, which uses a Siamese encoder to extract multifeature from bitemporal images in two branches input.
3) Fully Convolutional-Siamese-Difference (FC-Siam-diff): FC-Siam-diff is another FC_EF based model, in which the skip connections can enhance the differences between bitemporal features with two branches input.

4) Feature Constraint Change Detection Network (FCCDN): FCCDN, a dual encoder decoder network backbone for the CD task, constrains features both in bitemporal feature extraction and feature fusion. Features are extracted and fused by a nonlocal feature pyramid network.

5) Siamese NestedUnet Networks (SNUnet): SNUnet, a fully convolutional Siamese network for CD, is constructed from an improved UNet++ network. Features from different levels are combined to detect multilevel information.

6) Deeply Supervised Attention Metric-Based Network (DSAMnet): DSAMnet uses Siamese network and metric-based attention module. Dice loss and BCL loss are combined into deep supervision to improve the performance of CD.

2) Evaluation: In this article, we use the intersection over union (IoU) and F1 scores to evaluate the results. For the binary classification problem, the examples can be divided into true positive (TP), false positive (FP), true negative (TN), and false negative (FN) according to the combination of their real categories and learner prediction categories. The evaluation metrics are calculated based on these four indicators.

IoU, a concept usually used in object detection and semantic segmentation, is the overlap rate between the generated candidate bound and the original ground truth bound, which is the ratio of their intersection to union. The numerator part calculates the number of pixels to correctly predict in the foreground, and the denominator part calculates the number of pixels in the images and sets the real foreground and predicted foreground. The process of calculating a debit note can be expressed as

\[
\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}.
\]

F1 score is another evaluation metric in the statistical analysis of binary classification. Before calculating the F1 score, we need to calculate precision and recall. Precision refers to the proportion of samples with predicted value of positive and original value of positive in all samples with predicted value of positive. Recall refers to the proportion of samples with predicted value of positive and original value of positive in all samples with original value of positive. The calculation process of F1 score can be expressed as follows:

\[
\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

\[
\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}
\]

\[
\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.
\]

3) Comparison on MSOSCD Dataset: As can be seen from Table II, the proposed method MSCDUnet outperforms all baselines on the MSOSCD dataset, achieving the highest IoU and F1 scores of 83.01% and 91.85% on validation dataset and 84.25% and 92.81% on test datasets. The second-ranked SNUnet on validation dataset gets an IoU of 78.25% and an F1 of 89.47% and second-ranked DSAMnet on test dataset gets an IoU of 77.54% and an F1 of 88.89%, which proves the advancement of integrating multisource data to supply the RGB information of VHR images. The FC-Siam-conc shows an IoU of 61.25% and an F1 of 76.35% on test datasets, which is better than FC_EF, an IoU of 40.12% and an F1 of 59.58 and FC-Siam-diff, an IoU of 48.74% and an F1 of 65.85%. From the results, we determine that more deeper networks with attention module such as DSAMnet and SNUnet obtain better results than FC_EF based simple structure. However, our proposal with a relatively simple network but integrating multisource data can reach a better performance than complex DL structure.

Fig. 5 provides more intuitive pictures of each network’s performance on the MSOSCD datasets. FC-EF-based methods, which have the ability to detect relatively apparent change areas, not do well in extracting the whole and exact boundaries of built-up changes. Among the deep networks, DSAMnet and SNUnet can get more detailed detection results from VHR images. However, because of the vague boundary and texture of natural changes, the SNUnet and DSAMnet show noises spreading inner and boundaries of change areas. Our proposal gives more homogeneous results and refined boundaries of buildings, which demonstrates that the integration of multispectral, SAR, and VHR can improve the performance of built-up area detection.

4) Comparison on MSBC Dataset: As can be seen from Table III, the proposed method also achieves the highest results of all baselines on the MSBC dataset, achieving the highest IoU and F1 scores of 54.89% and 70.58% on validation dataset and 47.12% and 64.21% on test dataset. The second-ranked SNUnet obtains an IoU of 53.75% and an F1 of 69.77% on validation datasets and an IoU of 46.66% and an F1 of 60.87% on test dataset, which is much better than EF_FC based method that IoU score lower than 40% and F1 score lower than 50% on test dataset. Because the test dataset of MSBC is different place in Guigang City from the train and validation area, the performances of all the methods show the limits on test areas. Although CD faces the small data problem on MSBC dataset, our proposal reaches the best score that proves the fusing multisource data can relieve the sample problem.

Fig. 6 further demonstrates the behavior of different methods on the MSBC dataset. As can be seen from odd rows of columns 1 and 2 in Fig. 6, all the methods can focus on the two changed areas, but our proposal presents a relatively clear shape of the new construction. According to the ground truth (the odd row of the last column in Fig. 6), the change area shows a similar texture with the natural landcover nearby that confuses the models resulting in almost all approaches giving a vague boundary. The FC_EF based methods tend to miss more changes from VHR, which may result from the limit feature extraction ability of the architecture. DSAMnet tends to present more noises may result from not fully trained with the small dataset. Similar to the results of MSOSCD datasets, the MSCDUNet reaches the
best score of extracting information of changes among all the baselines on MSBC dataset.

D. Ablation Experiment

On the basis of multidata input, MSCDUnet integrates multispectral and SAR data to supply RGB information of VHR images. We therefore design ablation experiments to verify the improvement of multisource data input for CD. In Table IV, the “RGB” baseline denotes the basic model Unet++ with RGB bitemporal information input, which is separated from our architecture. The “RGB + S1S2” means MSCDUnet with input as the original Sentinel-1 and Sentinel-2 features without handcraft feature calculation and feature selection. The detailed S1S2 features are presented in the first column in Table IX in the Appendix. The experiment of “RGB + S1S2” tends to
compare with the selected features with the same model to prove the progress of RF feature selection. The “RGB + S1∗S2∗” model is the proposal of this article with the selected features as the second column in SUP Table III. Meanwhile, to verify the validity of multispectral and SAR data, respectively, we add four inputs that are “RGB + S1,” “RGB + S2,” “RGB + S1∗,” and “RGB + S2∗,” which means VHR combining with features from Sentinel-1 or Sentinel-2. Similarly, “S1” or “S2” means original features and “S1∗” and “S2∗” means selected features.

As can be seen from Table IV, the MSCDUnet with the selected features S1∗S2∗ integrated with VHR input obtains the best performance on two datasets. The RGB with original bands of Sentinel-1 and Sentinel-2 (RGB + S1S2) gets 86.12% and 47.68% IoU on two datasets, respectively, where F1 are 92.89% and 64.57%. This difference demonstrates that RF concentrates
the important features of the data and reduces noise of them. The Unet++ with RGB input (RGB) obtains a much lower score than “RGB + S1*S2” and “RGB + S1’S2,” which proves the effectiveness of data fusion module and deep supervision module.

Considering VHR images combined with only one auxiliary data, we could determine that “RGB + S1,” “RGB + S2,” “RGB + S1*,” and “RGB + S2*” all perform better than “RGB,” which only uses VHR data. The performances of VHR with one auxiliary data get different results of two datasets. On MSOSCD, VHR with selected features gets a higher accuracy and the experiments on MSBC get a contrary result. The secondary accuracy of test on MSOSCD is “RGB + S1*S2” and on MSBC is “RGB + S1.” Anyway, MSCDUnet with “RGB + S1*S2” input gets the highest accuracy, which demonstrates the great feature representation ability of integrating VHR and selected features of Sentinel-1 and Sentinel-2.

To sum up, we could determine that integrating RF-selected multispectral and SAR features with a simple fusion network will get a much higher score than the similar model with bitemporal VHR input or the combination of VHR and original multispectral or SAR features. The great improvement of the MSCDnet with “RGB + S1*S2” not only further proves the effectiveness of multisource data integration, but also shows the great potential of traditional hand-craft features combined with DL work in CD.

VI. DISCUSSION

A. Loss Function

As we explain in Section III, we propose a hybrid loss function with Dice and BCE loss to realize deep supervision and refine category imbalance. Because Dice loss always has an adverse impact on back propagation, we use BCE loss to balance the loss function. The selection of weight \( w_D \) and \( w_B \) is based on experiments. We define \( \lambda = \frac{w_D}{w_B} \) that is the variability in our experiments to select the best hyperparameter of the weight of loss function. \( \lambda = 0 \) means the experiment only uses BCE loss. In Table V, we can determine that most hybrid losses are prior than only using BCE loss unless when \( \lambda \) equals 0.5. Finally, we select \( \lambda = 1.5 \) based on the best IoU. Therefore, we appoint \( w_D = 0.6 \) and \( w_B = 0.4 \).

B. Efficiency and Time Analysis

To quantify the model size and time consumption, we compare our proposal with other methods with different perspectives, including the floating-point operations (FLOPs), the number of parameters (Params), and the time to predict a patch of images (Time). For DL-based algorithms, FLOPs and Params represent the computational complexity and model size that directly influence the time consumption of applications. In Table VI, we can determine that our proposal has the second-highest FLOPs and Params, also consuming the more time. However, considering that the input of our proposal is VHR, multispectral, and SAR data and others are only VHR data, the model size and time consumption of our method are acceptable. Compared with DSAMnet that has similar model size with us, our proposal gets much higher accuracies in Tables II and III, which determines that our proposal has a relatively high accuracy and appropriate time consumption.

Another time consumption in our proposal is RF on GEE. The training time of MSBC dataset is about 2 min and MSOSCD is about 3 min, which is also related to the network speed at that time. Compared with time of DL training, that is almost negligible. Considering feature selection time that may be dependent on the proficiency of operator, the time consumption on two datasets will not waste over half an hour. Also on GEE, we can simply preprocess and download images, which will also save our time. Therefore, our framework is time-saving and efficient even though we use VHR, multispectral, and SAR data all. Although our proposal is not an end-to-end method, the framework could integrate advantages of different features.
to improve the accuracy of built-up area CD without RF as a preprocessing of data input.

VII. CONCLUSION

In this article, a new DL-based framework called MSCDUnet is proposed for built-up area CD by integrating multispectral, SAR, and VHR data. We also provide two new CD datasets: MSOSCD and MSBC, focusing on built-up area CD consisting of multisource data. RF is used to select effective features from plenty of origin and hand-craft features. MSCDUnet uses a multilevel data fusion module to combine the different spatial and spectral data and maintain spatial details from VHR data at the same time. A deep supervision module after the fusion module is used to generate change maps with more spatial and multilevel information coming from different subnetworks. Experimental results demonstrate that the proposed MSCDUnet outperforms other models on both two datasets. The great performance of our framework not only demonstrates the effectiveness of integrating multispectral, SAR, and VHR data for built-up area detection but also proves that the prior knowledge of multispectral and SAR features could improve the DL-based CD with VHR. In the future, we will explore more stable architecture and advanced algorithms to maintain multisource features to improve the performance of CD in diversified scenarios.

APPENDIX

TABLE VII

| Texture Features from GLCM |
|----------------------------|
| **Abbreviation** | **Full name and or explanation** |
| ASM | Angular Second Moment; measures the number of repeated pairs |
| CONTRAST | Contrast; measures the local contrast of an image |
| CORR | Correlation; measures the correlation between pairs of pixels |
| VAR | Variance; measures how spread out the distribution of gray-levels is |
| IDM | Inverse Difference Moment; measures the homogeneity |
| SAVG | Sum Average |
| SVAR | Sum Variance |
| SENT | Sum Entropy |
| ENT | Entropy. Measures the randomness of a gray-level distribution |
| DVAR | Difference variance |
| DENT | Difference entropy |
| IMCCORR1 | Information Measure of Corr. 1 |
| IMCCORR2 | Information Measure of Corr. 2 |
| MAXCORR | Max Corr. Coefficient |
| DISS | Dissimilarity |
| INERTIA | Inertia |
| SHADE | Cluster Shade |
| PROM | Cluster prominence |

TABLE VIII

| Feature Index | Formula |
|---------------|---------|
| NDVI | NIR – Red |
| SAVI | NIR + Red + 0.5 (NIR – Red) + 1.5 |
| LSWI | NIR + SWIR1 |
| MNDWI | Green + SWIR1 |
| BSI | (SWIR1 + Red) + (NIR + Blue) |
| IBI | 2 + SWIR1 |

TABLE IX

| Class | Feature | Definition |
|-------|---------|------------|
| Original multispectral bands | B2m1 | Blue in the previous moment |
| B3m1 | Green in the previous moment |
| B4m1 | Red in the previous moment |
| B8m1 | NIR in the previous moment |
| B8A4m1 | Red Edge 4 (864.8nm) in the previous moment |
| B11m1 | SWIR 1 in the previous moment |
| B12m1 | SWIR 2 in the previous moment |
| B2m2 | Blue in the after moment |
| B3m2 | Green in the after moment |
| B4m2 | Red in the after moment |
| B8m2 | NIR in the after moment |
| B8A4m2 | Red Edge 4 (864.8nm) in the after moment |
| B11m2 | SWIR 1 in the after moment |
| B12m2 | SWIR 2 in the after moment |
| Vertical transmit/horizontal receive in the previous moment | vhm1 |
| Vertical transmit/horizontal receive in the previous moment | vhm2 |
| Single co-polarization, vertical transmit/vertical receive in the previous moment | vhm1 |
| Single co-polarization, vertical transmit/vertical receive in the after moment | vhm2 |
| CLASS | FEATURE | DEFINITION |
|-------|---------|------------|
|       | vh_divide | \( \frac{v h}{m_2} \) |
| Difference/divide operation in SAR | vh_minu | \( v h_{m_2} - v h_{m_1} \) |
|       | vv_minu | \( v v_{m_2} - v v_{m_1} \) |
|       | vvvh_minu | \( v v v h_{m_2} - v v v h_{m_1} \) |
|       | crossvh_divi_de | \( \frac{crossvh_{m_2}}{crossvh_{m_1}} \) |
| GLCM from difference/divide operation in SAR | crossvh_divi_de_CORR | \( \frac{CORR(crossvh_{m_2})}{crossvh_{m_1}} \) |
|       | crossvh_divi_de_SVAR | \( \frac{SVAR(crossvh_{m_2})}{crossvh_{m_1}} \) |
|       | crossvh_divi_de_SHADE | \( \frac{SHADE(crossvh_{m_2})}{crossvh_{m_1}} \) |
|       | BSI_divide | \( BSI_{m_2} \) |
|       | BSI_minu | \( BSI_{m_2} - BSI_{m_1} \) |
| Difference/divide operation in multispectral images | LSWI_divid e | \( \frac{LSWI_{m_2}}{LSWI_{m_1}} \) |
|       | LSWI_minu | \( LSWI_{m_2} - LSWI_{m_1} \) |
|       | MNDWI_divide | \( MNDWI_{m_2} \) |
|       | MNDWI_minu | \( MNDWI_{m_2} - MNDWI_{m_1} \) |
|       | NDVI_divide | \( NDVI_{m_2} \) |
|       | NDVI_minu | \( NDVI_{m_2} - NDVI_{m_1} \) |

\[ *v v h = v v * v h \]
\[ *c r o s s v h = v v / v h \]

REFERENCES

[1] H. Tamiminia, B. Salehi, M. Mahdianpari, L. Quackenbush, S. Adeli, and B. Brisco, “Google Earth engine for geo-big data applications: A meta-analysis and systematic review,” *ISPRS J. Photogramm.*, vol. 164, pp. 152–170, 2020.

[2] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, “Deep learning in remote sensing applications: A meta-analysis and review,” *ISPRS J. Photogramm.*, vol. 152, pp. 166–177, 2019.

[3] X. Tang, C. E. Woodcock, P. Olofsson, and L. R. Hutrya, “Spatiotemporal assessment of land use/land cover change and associated carbon emissions and uptake in the Mekong River Basin,” *Remote Sens. Environ.*, vol. 256, 2021, Art. no. 112336.

[4] M. Kim, D. Jeong, and Y. Kim, “Local climate zone classification using a multi-scale, multi-level attention network,” *ISPRS J. Photogramm.*, vol. 181, pp. 345–366, 2021.

[5] H. Guo, B. Du, L. Zhang, and X. Su, “A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery,” *ISPRS J. Photogramm.*, vol. 183, pp. 240–252, 2022.

[6] L. Mou, L. Bruzzone, and X. X. Zhu, “Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[7] Y. Tan, S. Xiong, and Y. Li, “Automatic extraction of built-up areas from panchromatic and multispectral remote sensing images using double-stream deep convolutional neural networks,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 3988–4004, Nov. 2018.

[8] D. Wen, X. Huang, H. Liu, W. Liao, and L. Zhang, “Semantic classification of urban trees using very high resolution satellite imagery,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1413–1424, Apr. 2017.

[9] N. Zang, Y. Cao, Y. Wang, B. Huang, L. Zhang, and P. T. Mathiopoulos, “Land-use mapping for high spatial resolution remote sensing image via deep learning: A review,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5372–5391, May 2021.

[10] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbaksh, and J. Liang, “Unet++: A nested U-Net architecture for medical image segmentation,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[11] H. Chen and Z. Shi, “A spatial-temporal attention-based method and a new dataset for remote sensing image change detection,” *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1662.

[12] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, “A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604816.

[13] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, “FCCDN: Feature constraint network for VHR image change detection,” *ISPRS J. Photogramm.*, vol. 187, pp. 101–119, 2022.

[14] S. Saha, F. Bovolo, and L. Bruzzone, “Building change detection in VHR SAR images via unsupervised deep transcodung,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 1917–1929, Mar. 2021.

[15] D. Brunner, G. Lemoine, and L. Bruzzone, “Earthquake damage assessment of buildings using VHR optical and SAR imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2403–2420, May 2010.

[16] L. Jing, T. Wang, M. Zhao, and P. Wang, “Adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox,” *Sensors*, vol. 17, no. 2, 2017, Art. no. 414.

[17] B. Bigdeli, P. Pahlavani, and H. A. Amirkolaee, “An ensemble deep learning method as data fusion system for remote sensing multisensor data fusion for urban change detection using a dual stream U-Net,” *IEEE Geosci. Remote Sens. Lett.*, vol. 265, 2021, Art. no. 112634, doi: 10.1109/LGRS.2021.3119856.

[18] B. Le Saux, A. Boulch, and Y. Gousseau, “Urban change detection for multispectral earth observation using convolutional neural networks,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2115–2118.
Q. Zhu et al., “Building damage assessment for an earthquake using a combination of SAR and optical data,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 4023–4034, Oct. 2018.

[45] D. K. Seo, Y. H. Kim, Y. D. Eo, M. H. Lee, and W. Y. Park, “Fusion of SAR and optical images using random forest regression for change detection,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, no. 10, pp. 4245–4257, Oct. 2021.

[46] M. Farahani and A. Mohammadzadeh, “Domain adaptation for unsupervised change detection of multisensor multitemporal remote-sensing images,” *Int. J. Remote Sens.*, vol. 41, no. 10, pp. 3902–3923, 2020, doi: 10.1080/01431161.2019.1711239.

[47] X. Yang et al., “An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery,” *ISPRS J. Photogramm.*, vol. 177, pp. 238–262, 2021, doi: 10.1016/j.isprsjprs.2021.05.004.

[48] Z. Shao, H. Fu, P. Fu, and L. Yin, “Mapping urban impervious surface by fusing optical and SAR data at the decision level,” *Remote Sens.*, vol. 8, no. 11, 2016, Art. no. 945, doi: 10.3390/rs8110945.

[49] A. Asokan and J. Andrade, “Change detection techniques for remote sensing applications: A survey,” *Earth Sci. Inform.*, vol. 12, no. 2, pp. 143–160, 2019, doi: 10.1007/s12145-019-00380-5.

[50] O. Yousif and Y. Ban, “Fusion of SAR and optical data for unsupervised change detection: A case study in Beijing,” in *Proc. Joint Urban Remote Sens. Event.*, 2017, pp. 1–4.

[51] Y. Chen and L. Bruzzone, “Self-supervised change detection by fusing SAR and optical multi-temporal images,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 3101–3104.

[52] H. Zhang, J. Liu, and L. Xiao, “Bipartite residual network for change detection in heterogeneous optical and radar data,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 352–335.

[53] A. Benedetti, M. Picchiani, and F. Del Frate, “Sentinel-1 and Sentinel-2 data fusion for urban change detection,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 1962–1965.

[54] R. Touati and M. Mignotte, “An energy-based model encoding nonlocal pairwise pixel interactions for multisensor change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1046–1058, Feb. 2018.

[55] Y. Solano-Correa, F. Bovolo, and L. Bruzzone, “An approach for unsupervised change detection in multitemporal VHR images acquired by different multисpectral sensors,” *Remote Sens.*, vol. 10, no. 4, 2018, Art. no. 533, doi: 10.3390/rs10040533.

[56] C. Kwan, J. Yang, and B. Chou, “Change detection using original and fused Landsat and Worldview images,” in *Proc. IEEE 10th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf.*, 2019, pp. 449–454.

[57] M. Belgiu and L. Dr˘uţ, “Random forest in remote sensing: A review of applications and future directions,” *ISPRS J. Photogramm.*, vol. 114, pp. 24–31, 2016.

[58] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[59] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[60] F. J. Kriegler, W. A. Malila, R. F. nalepka, and W. Richardson, “Preprocessing transformations and their effects on multispectral recognition,” in *Proc. Int. Symp. Remote Sens. Environ.*, vol. 6, 1969, pp. 97–131.

[61] A. R. Huete, “A soil-adjusted vegetation index (SAVI),” *Remote Sens. Environ.*, vol. 25, no. 3, pp. 295–309, 1988.

[62] K. Chandrasekar, M. Sesha Sai, P. S. Roy, and R. S. Dwevedi, “Land surface water index (LSWI) response to rainfall and NDVI using the MODIS vegetation index product,” *Int. J. Remote Sens.*, vol. 31, no. 15, pp. 3987–4005, 2010.

[63] P. Mohanaiha, P. Sathyarayananay, and L. GuruKumar, “Image texture feature extraction using GLCM approach,” *Int. J. Sci. Res. Publ.*, vol. 3, no. 1, pp. 1–5, 2013.

[64] L. Pipia, E. Amin, S. Belda, M. Salinero-Delgado, and J. Verrelst, “Green LAI mapping and cloud gap-filling using Gaussian process regression in Google Earth Engine,” *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 403, doi: 10.3390/rs13030403.

[65] Z. Jin et al., “Smallholder maize area and yield mapping at national scales with Google Earth Engine,” *Remote Sens. Environ.*, vol. 228, pp. 115–128, 2019, doi: 10.1016/j.rse.2019.106167.

[66] L. Zhu, X. Liu, L. Wu, Y. Tang, and Y. Meng, “Long-term monitoring of cropland change near Dongting Lake, China, using the LandTrendr algorithm with Landtrendr imagery,” *Remote Sens.*, vol. 11, no. 10, 2019, Art. no. 1234, doi: 10.3390/rs11101234.

[67] J. Liang, Y. Xie, Z. Sha, and A. Zhou, “Modeling urban growth sustainability in the cloud by augmenting Google Earth Engine (GEE),” *Comput. Environ. Urban Syst.*, vol. 84, 2020, Art. no. 101542, doi: 10.1016/j.compenvurbsys.2020.101542.

[68] C. Telmo, J. Lousada, and N. Moreira, “Proximate analysis, backwards stepwise regression between gross calorific value, ultimate and chemical analysis of wood,” *Bioresource Technol.*, vol. 101, no. 11, pp. 3808–3815, 2010.