A CNN-Transformer Network With Multiscale Context Aggregation for Fine-Grained Cropland Change Detection

Mengxi Liu, *Student Member, IEEE*, Zhuoqun Chai, Haojun Deng, and Rong Liu

***Abstract*—Nonagriculturalization incidents are serious threats to local agricultural ecosystem and global food security. Remote sensing change detection (CD) can provide an effective approach for in-time detection and prevention of such incidents. However, existing CD methods are difficult to deal with the large intraclass differences of cropland changes in high-resolution images. In addition, traditional CNN based models are plagued by the loss of long-range context information, and the high computational complexity brought by deep layers. Therefore, in this article, we propose a CNN-transformer network with multiscale context aggregation (MSCANet), which combines the merits of CNN and transformer to fulfill efficient and effective cropland CD. In the MSCANet, a CNN-based feature extractor is first utilized to capture hierarchical features, then a transformer-based MSCA is designed to encode and aggregate context information. Finally, a multibranch prediction head with three CNN classifiers is applied to obtain change maps, to enhance the supervision for deep layers. Besides, for the lack of CD dataset with fine-grained cropland change of interest, we also provide a new cropland change detection dataset, which contains 600 pairs of 512 × 512 bi-temporal images with the spatial resolution of 0.5–2m. Comparative experiments with several CD models prove the effectiveness of the MSCANet, with the highest F1 of 64.67% on the high-resolution semantic CD dataset, and of 71.29% on CLCD.**

***Index Terms*—Change detection (CD), cropland, deep learning (DL), remote sensing, transformer.**

I. Introduction

AGRICULTURAL production is the guarantee of worldwide food security [1]. However, affected by recent rapid population growth and dramatic climate change, cropland, as the basic unit of agricultural activities, have suffered many disadvantageous changes, including afforestation, lake digging, reserve expansion, and illegal building construction [2]. These nonagriculturalization events not only disturb local agricultural ecosystems, but also threaten global food supply [3]. Therefore, in order to obtain timely cropland information to ensure cropland production and food security [4], fast and dynamic change detection (CD) on cropland is extremely important [5].

As the cropland is widely distributed, it is labor- and time-consuming to acquire cropland dynamics through manual field investigation [6]. With the wide application of satellite images, remote sensing technology have been served as an effective and realistic approach to many aspects, such as terrain classification [7], building footprint extraction [8], as well as land cover CD [9]. Traditional CD methods are mainly based on multispectral images, which extract rich spectral, textural and structural features for rapid pixel- or object-wisely change results. For instance, change vector analysis (CVA) [10], principal component analysis [10], and multivariate alteration detection [12], [13] have been widely applied in CD researches for their advantages in succinct feature representation and rapid change extraction.

Nevertheless, since simple features are difficult to meet the needs of diverse and high-precision change extraction, thus the machine learning (ML) based methods with hand-craft feature engineering are applied to CD tasks [14], [15]. For example, Vries *et al.* [16] incorporated random forest-based postclassification and traditional CVA to the updating of annual cropland change mapping. However, these ML-based methods require prior expertise to construct and select features manually, which is of low generalization performance in different regions and datasets [17]. Moreover, it is difficult to acquire fine-grained dynamic results due to limited spatial resolution of multispectral images [18].

On account of rapid progress in artificial intelligence technology and remote sensing platforms, the focus of CD research has turned into deep learning (DL) models and high-resolution images (HRIs) [19], [20]. Based on convolutional neural networks (CNNs) structure, these DL models are capable to automatically learn multilevel change information from HRIs [21] by reconstructing classical models, such as UNet [22], [23], DeepLab [24], [25], and ResNet [26], [27]. While CD results are easily affected by seasonal, irradiant and atmospheric disturbances between images, many novel techniques have been introduced in recent CD networks to better perceive changes from bi-temporal images, including multiscale feature fusion [28], [29], attention mechanism [30], [31], recurrent neural networks [32]–[34], and so on, which have been proved to be effective in enhancing the

Manuscript received March 26, 2022; revised April 18, 2022; accepted May 18, 2022. Date of publication May 23, 2022; date of current version June 3, 2022. This work was supported in part by National Natural Science Foundation of China under Grant 61976234, and in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 22qntd2001. *(Corresponding author: Rong Liu.)*

The authors are with the Guangdong Provincial Key Laboratory for Urbanization and Geo-simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China (e-mail: liumx23@mail2.sysu.edu.cn; chaizhq@mail2.sysu.edu.cn; denghj5@mail2.sysu.edu.cn; liurong25@mail.sysu.edu.cn).

Code and dataset in the article will be available for download from the following link https://github.com/liumency/CropLand-CD.

Digital Object Identifier 10.1109/JSTARS.2022.3177235

feature extraction capability of the model. However, these traditional CNN-based methods still face two bottlenecks: one is the problem of information loss in the process of feature encoding and decoding, and the other is the problem of its exponentially growing computational consumptions with increasing layers and data size.

Recently, the transformer, which was initially designed for natural language processing tasks [35], has also received extensive attention in the field of computer vision, such as image classification [36], segmentation [37], object recognition [38] and image captioning [39], etc. In comparison to CNN, transformer has shown strong ability to model global dependencies to alleviate loss of long-range information [40]. Inspired by these works, Chen *et al.* [41] introduced transformer into CD tasks and implemented a bitemporal image transformer (BIT), which encode the input image into context-rich semantic tokens in a differencing-base CD framework.

Even though existing methods have made great achievements in remote sensing CD, there are still challenges to achieve fine-grained cropland CD. The performance of a DL model largely depends on the training dataset, and many previous works have provided well-annotated datasets for CD, such as High resolution semantic change detection dataset (HRSCD) [42] and SECOND [43] for semantic CD, SYSU-CD [44] and SVCD [45] for binary CD, and BCDD [46] and LEVIR-CD [31] for building CD, etc. So far, there is no dataset that specifically focuses on cropland changes, which greatly limits the development and application of cropland CD models. Therefore, how to efficiently and effectively model the multiscale information between bitemporal images is an urgent requirement in rapid cropland CD tasks.

To deal with the above problems, we propose a multiscale context aggregation network (MSCANet), and a high-resolution cropland change detection dataset (CLCD) in this article. The MSCANet first employs a CNN backbone to capture multiscale features from bitemporal images; then a multiscale context aggregator (MSCA) is utilized to model and aggregate the rich context information through transformer architecture; finally, a multibranch prediction head (MBPH) is applied to obtain change maps to further enhance feature extraction and learning of hidden layers. The MSCANet is constructed based on CNN-transformer structure, which can fully combine the advancements of both CNN and transformer to satisfy the urgent need of fast and accurate cropland CD. The CLCD consists of 600 pairs of bitemporal images annotated with various cropland changes, which can provide a benchmark for DL-based models on cropland CD tasks. The contributions of this article are summarized into three points.

1) An MSCANet with CNN-transformer hybrid architecture is proposed for cropland CD, in which a MSCA is designed to encode multiscale context information, and an MBPH is utilized to improve deep feature learning.
2) A high-resolution CLCD is provided for all research needed, which contains 600 pairs of 512 × 512 images with spatial resolutions of 0.5–2 m.
3) Comparative experiments with six state-of-the-art (SOTA) CD models on the HRSCD [42] and CLCD illustrate that the proposed MSCANet can obtain the highest F1 scores of 64.67% and 71.29%, respectively.

The rest of the article is organized as follows. Section II reveals detail structures of the methodology, while Section III gives the experimental settings. The experimental results will be demonstrated and analyzed in Section IV. Ablation study and model efficiency will be discussed in Section V. Finally, Section VI concludes this article.

II. Methodology

As shown in Fig. 1, the MSCANet contains three parts: a CNN feature extractor, an MSCA, and an MBPH. Detail information of each part will be introduced in the following.

A. Feature Extractor

The MSCANet employs a CNN backbone as the feature extractor, which is modified from ResNet-18 [47] by removing the initial fully connected layer. Therefore, the feature extractor contains a 7 × 7 convolutional layer, and four residual blocks (ResBlocks). The first convolutional layer with a stride of 2 is used to extract half-size shallow features. Then a 3×3 max-pool layer with stride 2 is further employed to capture features with quarter size of the original image, with the aim to filter important features and reduce the number of parameters.

Each ResBlock contains two 3 × 3 convolutional layers, a Batch normalization [48] layer and a rectified linear unit (ReLU) function [49]. The feature is fused with the original input feature by element-wise addition before being input into the ReLU layer. Since the first convolutional layer in ResBlock-1 and ResBlock-4 adopts a stride of 1, the size before and after feature input remains unchanged, while the first convolutional layer in ResBlock-2 and ResBlock-3 adopts a stride of 2, so the output feature size is halved. Finally, the size of output characteristics of ResBlock-4 is 1/16 of the original input image. The output channel of each ResBlock is 64, 128, 256, and 512, respectively.

To obtain multiscale cropland information, the multiscale output of ResBlock-1, 2, and 4 will be forwarded into subsequent modules. Before that, a 3 × 3 and a 1 × 1 convolutional layers will be applied to the selected features to unify their channel size into 32.

B. Multiscale Context Aggregator

In order to further model and fuse multiscale information from the feature extractor, an MSCA is designed in MSCANet. The MSCA uses three token encoders and three token decoders, which are built based on transformer architecture, to capture and aggregate multiscale context information from the three features with different sizes.

1) Token Encoder: The token encoder aims to encode global context information of feature through a spatial attention module and a transformer module. The spatial attention module is first adopted to convert the input feature into a target-size three-dimensional token embedding for subsequent transformer module, considering the limitations on calculation and storage. According to Fig. 2(a), given an input feature, referred as

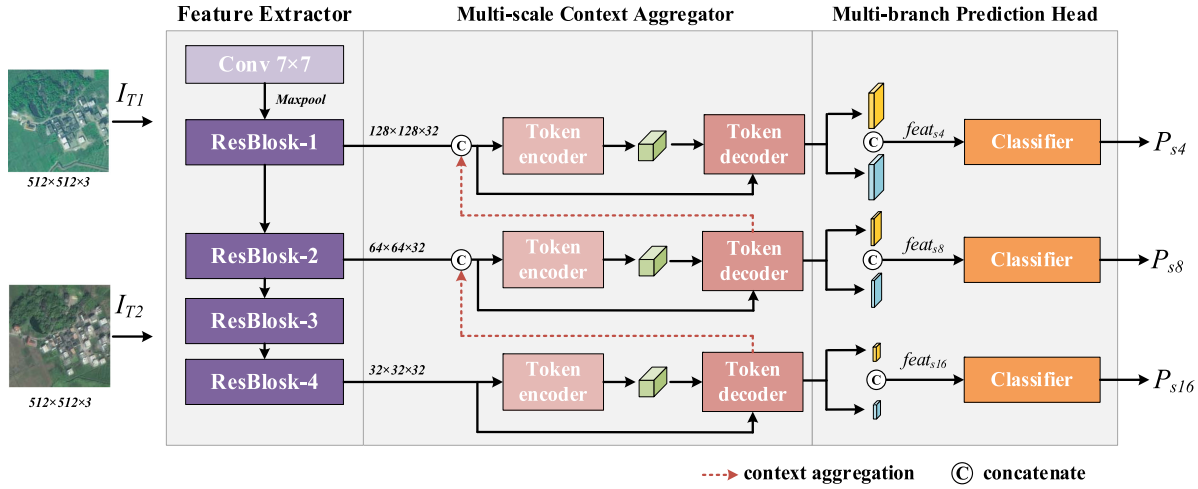


Fig. 1. Overview of the proposed MSCANet.

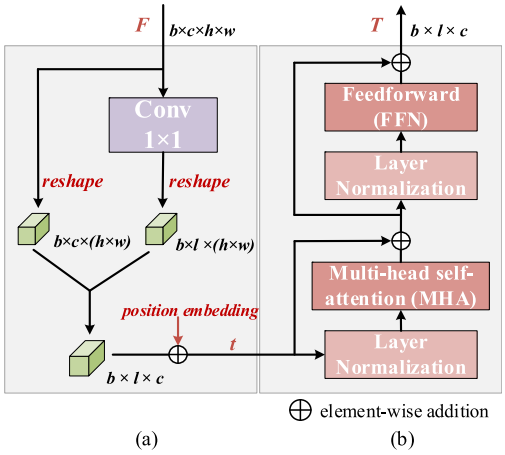


Fig. 2. Architecture of the token encoder. (a) Spatial attention module; (b) Transformer encoder.

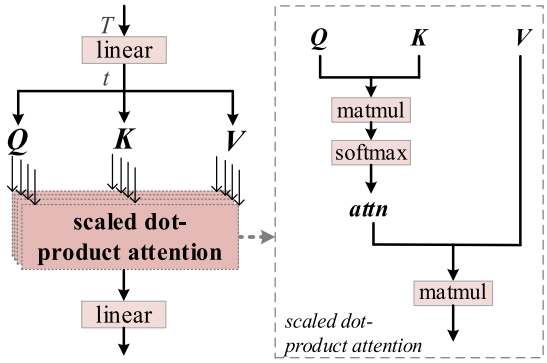


Fig. 3. Architecture of the MHA block.

$F \in \mathbb{R}^{b\times c\times h\times w}$, the spatial attention module adopts a 1×1 convolutional layer to obtain an intermediate feature, referred as $F' \in \mathbb{R}^{b\times l\times h\times w}$. Then, both F and F' will be reshaped into 3D tokens, referred as $f \in \mathbb{R}^{b\times c\times(h\times w)}$ and $f' \in \mathbb{R}^{b\times l\times(h\times w)}$, respectively. Finally, f and f' will be turn into a token embedding $t \in \mathbb{R}^{b\times l\times c}$ through *einsum* operation, which can be denoted as

$$t_{blc} = f'_{bl(hw)} f_{bc(hw)} \tag{1}$$

where b, c, h , w denote the batch size, number of channels, height, width of the input feature F, respectively; l is the token length, which is set to be 4 in the model.

Thereafter, the transformer encoder is utilized to model the context information in the token, in which a group of trainable parameters is first element-wisely added to the token t for position embedding (PE). The transformer encoder has a standard Transformer structure [40], which contains an MHA block and a feedforward (FFN) block, with a layer normalization (LN) layer applied before each block.

The architecture of the MHA block is shown in Fig. 3. The MHA first expands t into a new embedding t' by a linear layer, which can be denoted as

$$t' = tW^I,\ t' \in \mathbb{R}^{b\times l\times(n\times d\times 3)} \tag{2}$$

where W^I is the weight of the linear layer, n is the head number of MHA, d is the dimension for subsequent tensors. n and d are set for 8 and 64, respectively.

Then, the embedding t' will be forwarded into different heads of MHA. Parameters are not shared among heads. Each head contains two steps: linear transformation and scale dot-product attention (SDPA). Three linear layers are applied to map t' into query ($Q \in \mathbb{R}^{b\times n\times l\times d}$), key ($K \in \mathbb{R}^{b\times n\times l\times d}$), and value ($V \in \mathbb{R}^{b\times n\times l\times d}$), which can be denoted as

$$Q, K, V = t'W^Q, t'W^K, t'W^V \tag{3}$$

where W^Q, W^K, and W^V denotes the weights of the linear layers to map Q, K and V, respectively.

Thereafter, in SDPA, the correlation between Q and K is calculated through dot product operation and Softmax activation to generate an attention map, which will be used as the weight

of V. The process in SDPA can be expressed as

$$\text{SDPA}\left(Q,\ K, V\right) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{4}$$

The output of each head will be concatenated together before a linear layer is applied, then we will obtain the final output of the MHA, which can be expressed by the following formula:

$$\text{head}_i = \text{SDPA}\left(t'W_i^Q, t'W_i^K, t'W_i^V\right),\ i \in (0, n] \tag{5}$$

$$\text{MHA}\left(Q, K, V\right) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_n\right)W^O \tag{6}$$

where W_i^Q, W_i^K, and W_i^V denotes the weights of the linear layers of the ith head to map Q, K, and V, respectively; W^O is the weight of the last linear layer in MHA.

In FFN, two linear layers and a Gaussian error linear unit activation [50] are used to further transform the learning token of MHA.

2) Token Decoder: The token decoder receives two inputs, the convolutional feature F from feature extractor and the token embedding from token encoder, denoted as T. In the token decoder, first, a trainable parameter is added to F for PE, before F and T are input into a transformer decoder, with the aim to reproject the token embedding to the pixel space to and enhance the context information in F.

In the transformer decoder, a weight-shared LN layer is applied to F and T, before an MHA module is employed. The MHA is similar to that in the token encoder. The difference mainly lies in that the Query is mapped from F, while the Key and Value are mapped from T. Noted that mask mechanism is not applied here.

In the process of token encoding and decoding at different scales, context aggregate connection (see the dotted lines in Fig. 1) is introduced to aggregate the transformer features of higher-level into the convolutional features of lower-level.

C. Multibranch Prediction Head

To make better use of aforehand multiscale features, the MBPH adopts three CNN-based classifiers with the same architecture to generate change results to supervise feature learning of deep layers and help extract more useful features for CD.

After the multiscale features of inputs I_{T_1} and I_{T_2} are obtained by feature extractor and MHCA, features of the same scale will be fused together by concatenation and interpolated to the original image size. Then the classifiers will be applied to obtain three change maps from the multiscale features. Each classifier contains two 3×3 convolutional layers.

The MBPH outputs the multilevel prediction maps, referred as P_{s4}, P_{s8}, and P_{s16}, for in-depth supervision of the MSCANet, which provides auxiliary assistance to the model in capturing more effective features at multilevels for subsequent prediction. During training process, the MSCANet will be optimized through the sum of the cross-entropy loss between the three change maps and the ground truth Y. The formulation of the cross-entropy loss can be denoted as

$$L_{\text{CE}}\left(P, Y\right) = -\left[Y\log P + 1 - Y\log\left(1 - P\right)\right]. \tag{7}$$

Therefore, the total loss of the MSCANet can be expressed as

$$L = L_{CE}\left(P_{s4}, Y\right) + L_{CE}\left(P_{s8}, Y\right) + L_{CE}\left(P_{s16}, Y\right). \tag{8}$$

It can be seen from the objective function that deep supervision is implemented to the hidden layers in the MSCANet to generate more distinguish features. While for testing process, only P_{s4} will be used to obtain the final change result.

III. Experimental Settings

A. Datasets

1) High Resolution Semantic Change Detection Dataset: The HRSCD [42] is a semantic CD dataset, which contains 291 image pairs of 0.5-m RGB aerial images with size 10000 $\times$ 10000, with corresponding land cover information for each image, including five types of artificial surfaces, agricultural areas, forests, wetlands, and waters. All images were collected from urban and countryside areas in Rennes and Caen, French. In order to obtain fine-grained cropland change of interest, we reclassify the original labels of the bi-temporal images, by dividing the "agricultural areas" category into 1 and the rest categories into 0. Then the change annotation of cropland can be obtained by comparing the reclassified bitemporal labels.

For the convenience of model training, we tailor the original images in a nonoverlapping behavior and obtain 4398 pairs of 512 $\times$ 512 samples for cropland CD. These samples are separated for training, validation and test in the ratio of 6:2:2. Examples of in HRSCD are displayed in Fig. 4.

2) CropLand Change Detection: The CLCD dataset consists of 600 pairs image of cropland change samples, with 320 pairs for training, 120 pairs for validation and 120 pairs for testing. The bi-temporal images in CLCD were collected by Gaofen-2 in Guangdong Province, China, in 2017 and 2019, respectively, with spatial resolution ranged from 0.5 to 2 m. Each group of samples is composed of two images of 512 $\times$ 512 and a corresponding binary label of cropland change. As shown in Fig. 5, the main types of change annotated in CLCD include buildings, roads, lakes and bare soil lands, etc.

B. Comparative Methods

SOTA methods for bitemporal CD are employed in our experiments for comparison.

1) FC-EF [23] is an UNet-based CD method, which receives concatenation of bitemporal images as input, regarding them as separate channels.
2) FC-Siam-conc [23] is a variant of FC-EF, which applies the Siamese structure that shares weights to acquire multilevel features and concatenate them to coalesce change information.
3) DTCDSCN [51] is a Siamese FCN-based method with attention mechanism, which takes account of change information in both spatial and channel wise to extract more contextual features.
4) Multidirectional fusion pathway network (MFPNet) [29] is a multidirectional feature fusion method, which utilizes a multiscale fusion network with the multiway information

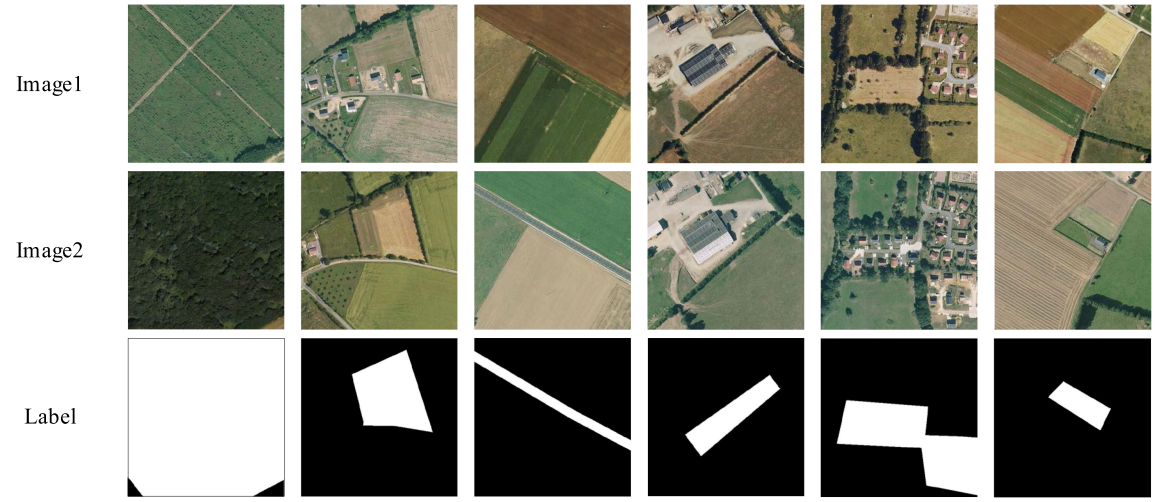


Fig. 4. Examples with size 512 × 512 in HRSCD dataset.

Fig. 5. Examples with size 512 × 512 in CLCD dataset. The main types of change annotated in CLCD include buildings, roads, lakes and bare lands.

flow for making data propagation easier while highlighting vital features.

5) Deeply supervised image fusion network (DSIFN) [28] uses the difference discriminant network for CD, and multilevel features are fused with the image difference map through the attention mechanism.
6) BiT [41] is a transformer-based feature fusion method, which integrates Siamese tokenizer and transformers encoder-decoder structure into the common CD network, thus performing capably to capture more meaningful and effective contextual concepts in global feature space.

C. Parameters and Metrics

The proposed model and all experiments involved are implemented in PyTorch. A batch size of 8 and a learning rate of 1^{-4} are adopted for all model training using an Adam optimizer. The training process lasts for 100 epochs, while data augmentation strategies are randomly applied to the training set to avoid over-fitting, including vertical and horizontal flip, and random rotation.

Four common metrics, precision (Pre), recall (Rec), F1-score and intersection over union (IoU), are selected for accuracy assessment. They can be defined as follows:

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

$$F1 = \frac{2\mathrm{Pre} \times \mathrm{Rec}}{\mathrm{Pre} + \mathrm{Rec}} \tag{11}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP} + \mathrm{FN}} \tag{12}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

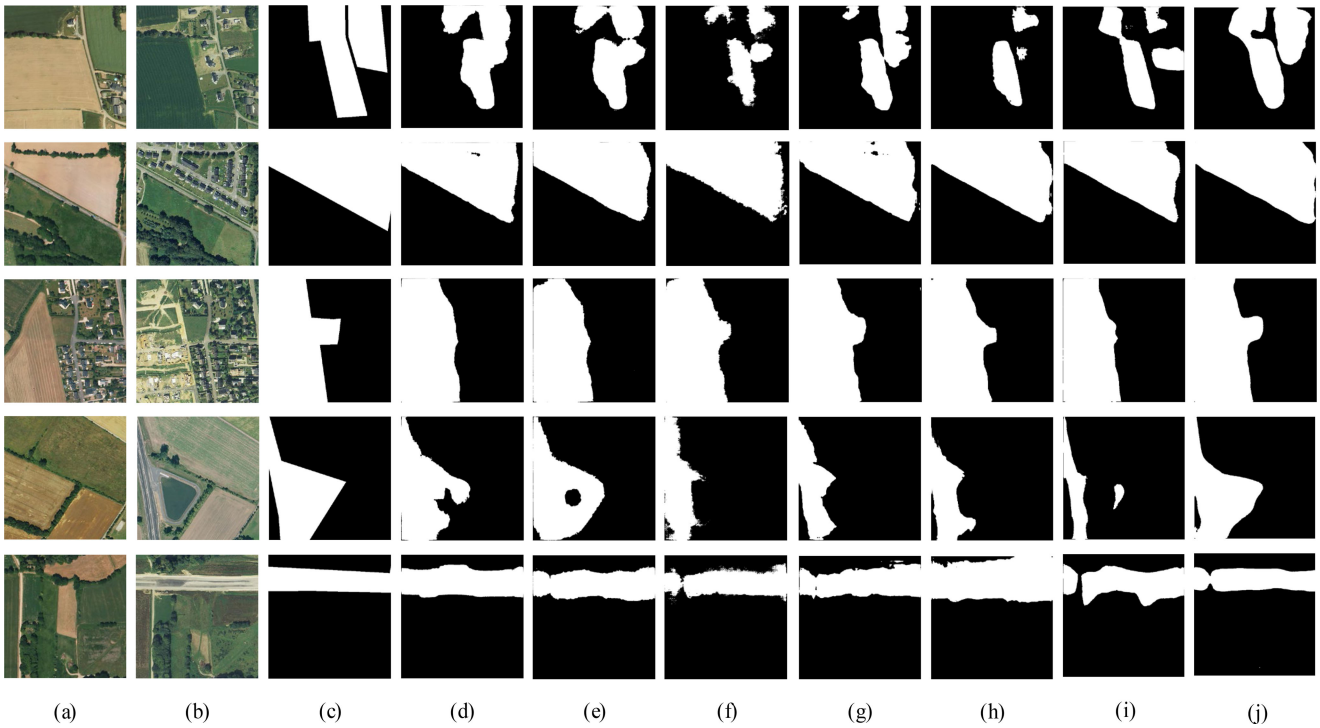


Fig. 6. Visualization of experimental results on HRSCD dataset. (a) Image1. (b) Image2. (c) Label. (d) FC-EF. (e) FC-Siam-conc. (f) DTCDSCN. (g) BiT. (h) MFPNet. (i) DSIFN. (j) MSCANet.

TABLE I
EXPERIMENTAL RESULTS ON HRSCD

Method	Pre(%)	Rec(%)	F1(%)	IoU(%)
FC-EF	72.75	50.30	59.48	42.33
FC-Siam-conc	72.23	47.53	57.34	40.19
DTCDSCN	75.79	48.83	59.39	42.24
BiT	71.30	52.23	60.30	43.16
MFPNet	76.42	54.98	63.95	47.01
DSIFN	**77.00**	54.27	63.66	46.70
MSCANet	70.17	**59.97**	**64.67**	**47.79**

TABLE II
EXPERIMENTAL RESULTS ON CLCD

Method	Pre(%)	Rec(%)	F1(%)	IoU(%)
FC-EF	71.70	47.60	57.22	40.07
FC-Siam-conc	73.27	52.91	61.45	44.35
DTCDSCN	54.49	66.23	59.79	42.64
BiT	61.42	62.75	62.08	45.01
MFPNet	**83.20**	60.74	70.22	54.11
DSIFN	79.07	63.79	70.61	54.58
MSCANet	75.36	**67.64**	**71.29**	**55.39**

IV. RESULTS AND ANALYSIS

A. Experiments on HRSCD

The Pre, Rec, F1, and IoU results on HRSCD are given in Table I. The F1 and IoU of FC-Siam-conc are the lowest among all methods, which are 57.34% and 40.19%, respectively, while those of fully convolutional–early fusion (FC-EF) are slightly higher, 59.48% and 42.33%. The performance of dual-task constrained deep siamese convolutional network (DTCDSCN) is between FC-Siam-conc and FC-EF, with F1 of 59.48%, followed by BiT, which obtain F1 of 60.30%. In general, MFPNet and DSIFN perform significantly better than aforementioned models, with F1 of 63.95% and 63.66%, respectively. The proposed MSCANet achieves the optimal recall, F1 and IoU values of 59.97%, 64.67%, and 47.79%, respectively, which are 4.99%, 0.72%, and 0.78% higher than the second-ranked MFPNet.

Fig. 6 visualizes experimental results of different methods in different scenarios on HRSCD dataset. In terms of cropland change into artificial surfaces, bare land, and roads, which are of distinct difference in appearance, most models can achieve relatively good recognition results. As can be seen in row 3 in Fig. 6, our proposed model can well extract the change of cropland to grassland when many methods fail. In addition, for the change of digging lakes (see row 4 in Fig. 6), the detection result of most methods is rather limited due to the relatively small number of relevant samples. In this case, the MSCANet can still completely identify such changes.

B. Experiments on CLCD

Quantitative results of all methods on CLCD are given in Table II. Different from results in HRSCD, the FC-Siam-conc with Siamese encoder and feature concatenation works better than FC-EF and DTCDSCN, with F1 of 61.45%. The following is BiT, showing the advancement of transformer structure than traditional UNet models. The performance of MFPNet and DSIFN are still bright in CLCD, which bumped F1 on CLCD to 70%. This can be attributed to the multiscale feature fusion strategy used in MFPNet and DSIFN, while the intraclass scale

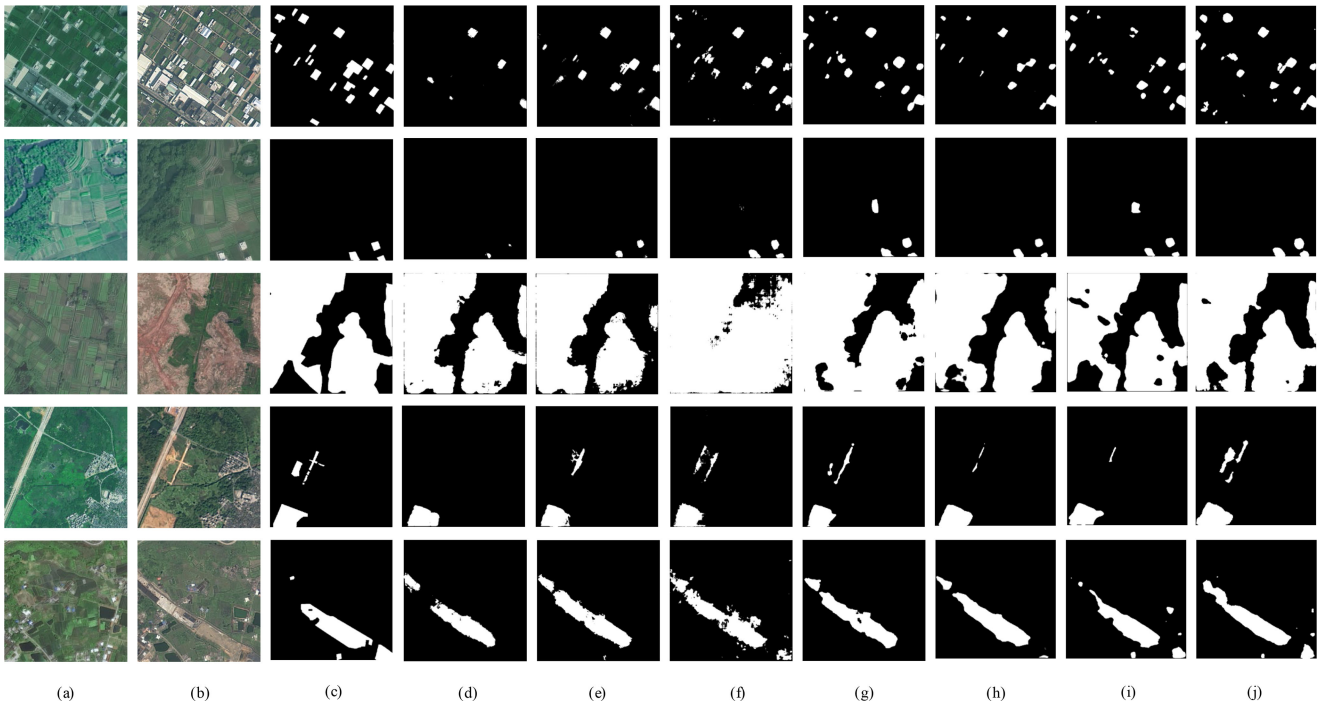


Fig. 7. Visualization of experimental results on CLCD dataset. (a) Image1. (b) Image2. (c) Label. (d) FC-EF. (e) FC-Siam-conc. (f) DTCDSCN. (g) BiT. (h) MFPNet. (i) DSIFN. (j) MSCANet.

difference in CLCD is much larger than HRSCD. Notably, the MSCANet gains the best results in Rec, F1, and IOU, reaching 67.64%, 71.29% and 55.39%, respectively, which are 1.41%, 0.68% and 0.81% higher than those of DSIFN.

Visualization comparison on CLCD is shown in Fig. 7. Compared to FC-EF, FC-Siam-conc can identify the change area more accurately. The change results by DTCDSCN are relatively fragmented, and suffer from severe misclassification caused by illuminant and phenological difference, which echoes the high recall and low precision of DTCDSCN, as given in Table II. Attributed to the multiscale feature fusion strategies, MFPNet and DSIFN can work well on cropland CD of various scales. Nonetheless, with the help of Transformer structure to encode semantic context information, BiT has better performance on cropland CD in complex scenes (such as row 1 in Fig. 7). On the whole, our MSCANet outperforms all comparative methods, which not only for better capability in edge preservation of large-scale changes, but also for more complete detection of small-scale changes, such as field roads and buildings, which is consistent with its highest recall, as given in Table II.

V. Discussion

A. Ablation Study

In this section, we conduct ablation study on CLCD to further verify the significance of MSCA and MBPH integrated in the MSCANet. The "base" model is the basic model for comparison without any tricks. "+MSCA" represent the "base" model with MSCA, while "MBPH" represent the "base" model with MBPH. Results of the ablation study are given in Table III. Compared with the "base" model with F1 of 68.71%, the F1 scores of "+MSCA" model and "+MBPH" model are improved by 0.64% and 2.24%, respectively, which preliminarily proves the validity of the MSCA and MBPH. The "+MSCA" obtains the highest recall rate of 72.71%, that is to say, the addition of MSCA is beneficial to reduce omission in CD, which is extremely important for cropland CD tasks. The MSCANet gains best results in the ablation experiments, which fully indicates the feasibility of the integration of MSCA and MBPH.

TABLE III
Ablation Study on CLCD

Method	Pre(%)	Rec(%)	F1(%)	IoU(%)
Base	70.34	67.15	68.71	52.15
+ MSCA	66.28	**72.71**	69.35	53.08
+ MBPH	73.67	68.42	70.95	54.98
MSCANet	**75.36**	67.64	**71.29**	**55.39**

Fig. 8 provides the visualized comparisons of the ablation results. From the example results, it can be seen that the edge of the CD result of "+MSCA" is closer to the original label, although there are some pseudo changes. It denotes that the MSCA module can effectively encode and aggregate multiscale context information between features, and thereby improving the semantic representation of the results. Compared with results by "base" model, the CD results by the "+MBPH" model have fewer false alarms. This shows that MBPH module is helpful to extract more discriminative features and reduce pseudo-changes by supervising the learning of deep hidden layers. Undoubtedly, the MSCANet, which combines the advantages of MSCA and MBPH, is superior in both boundary extraction and false alarms reduction.

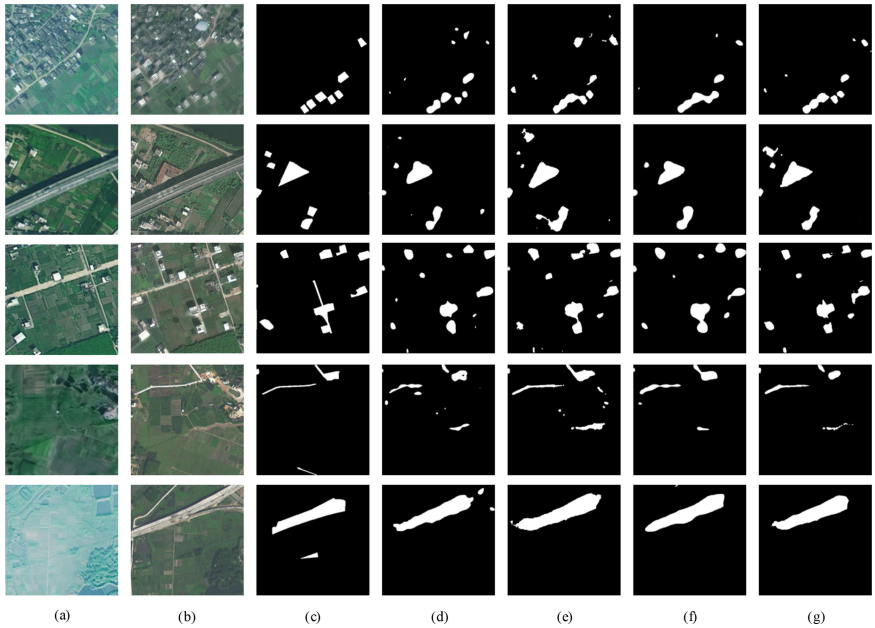


Fig. 8. Visualization of ablation study on CLCD dataset. (a) Image1. (b) Image2. (c) Label. (d) Base. (e) +MSCA. (f) +MBPH. (g) MSCANet.

TABLE IV
MODEL EFFICIENCY OF DIFFERENT METHODS

Method	FLOPs(G)	Params(M)
FC-EF	14.29	1.35
FC-Siam-conc	21.29	1.55
DTCDSCN	52.83	31.26
BiT	42.37	3.49
MFPNet	514.93	85.97
DSIFN	329.06	50.443
MSCANet	59.08	16.42

B. Model Efficiency

In order to get an in-depth understanding of different CD models in practical applications, we employ two metrics, floating points of operations (FLOPs) and number of parameters (Params), to further compare the model efficiency of all comparative methods. The FLOPs measures the computational complexity of the model by calculating the times of multiplication and addition operations, whose unit is 10^9 (G). The Params is the number of parameters that need to be learned during model training, corresponding to the space complexity of the model, in units of 10^6 (M).

Given two bitemporal inputs of size $1 \times 3 \times 512 \times 512$, the FLOPs and Params of all methods are given in Table IV. Combined with the previous analysis, it can be seen that in all models, FC-EF, FC-Siam-conc have the lowest FLOPs and Params. However, MFPNet and the DSIFN, which have excellent performance on both HRSCD and CLCD datasets, have the highest FLOPs and Params due to the use of complex multiscale feature fusion strategies. With the advantages of CNN-transformer hybrid architecture, MSCANet can achieve state-of-the-art CD performance under relatively lower FLOPs and Params, reflecting its feasibility and potential in rapid CD applications.

VI. CONCLUSION

In this article, an MSCANet and a new high-resolution dataset (CLCD) are proposed for cropland CD. The MSCANet employs a CNN-transformer structure, in which a pre-trained ResNet-18 is adopted to extract hierarchical features. Then, a transformer-based MSCA module is designed to encode and decode the context information in the multiscale features, with context aggregate connections applied to help feature fusion and aggregation across different levels. In the end, an MBPH is used to help enhance feature learning and capture more useful features.

Experiments on both HRSCD and CLCD proves the feasibility of the proposed MSCANet and the CLCD on cropland CD. The ablation study on CLCD further verify the effectiveness of the integrated MSCA and MBPH. More specifically, MSCA helps obtain the semantic properties of the change objects in terms of edge and morphology, while MBPH can reduce the pseudo changes in the results. Through the comparison of FLOPs and Params, the MSCANet further demonstrates its advantages in terms of space and computation complexity. All of the results have fully demonstrated the capability of the MSCANet in efficient and effective cropland CD.

REFERENCES

[1] H. C. J. Godfray *et al.*, "Food security: The challenge of feeding 9 billion people," *Science,* vol. 327, pp. 812–818, 2010.
[2] A. Molotoks, P. Smith, and T. P. Dawson, "Impacts of land use, population, and climate change on global food security," *Food Energy Secur.,* vol. 10, no. 1, 2021, Art. no. e261.
[3] L. See *et al.*, "Improved global cropland data as an essential ingredient for food security," *Glob. Food Secur., Agric. Policy Econ. Environ.,* vol. 4, pp. 37–45, 2015.
[4] T. Garnett *et al.*, "Sustainable intensification in agriculture: Premises and policies," *Science,* vol. 341, pp. 33–34, 2013.

[5] H. Mueller, P. Rufin, P. Griffiths, A. J. Barros Siqueira, and P. Hostert, "Mining dense Landsat time series for separating cropland and pasture in a heterogeneous Brazilian Savanna landscape," *Remote Sens. Environ.,* vol. 156, pp. 490–499, 2015.

[6] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sens.,* vol. 12, no. 10, May 2020, Art. no. 1688.

[7] T. J. Pingel, K. C. Clarke, and W. A. McBride, "An improved simple morphological filter for the terrain classification of airborne LIDAR data," *ISPRS J. Photogramm. Remote Sens.,* vol. 77, pp. 21–30, 2013.

[8] P. Liu *et al.*, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.,* vol. 11, no. 7, p. 830, Apr. 2019.

[9] M. Hu, C. Wu, L. Zhang, and B. Du, "Hyperspectral anomaly change detection based on autoencoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,* vol. 14, pp. 3750–3762, 2021.

[10] R. D. Johnson and E. S. Kasischke, "Change vector analysis: A technique for the multispectral monitoring of land cover and condition," *Int. J. Remote Sens.,* vol. 19, pp. 411–426, 1998.

[11] G. F. Byrne, P. F. Crapper, and K. K. Mayo, "Monitoring land-cover change by principal component analysis of multitemporal landsat data," *Remote Sens. Environ.,* vol. 10, pp. 175–184, 1980.

[12] A. A. Nielsen, K. Conr Ad Sen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.,* vol. 64, pp. 1–19, 1998.

[13] G. Doxani, K. Karantzalos, and M. Tsakiri-Strati, "Monitoring urban changes based on scale-space filtering and object-oriented classification," *Int. J. Appl. Earth Observ. Geoinformation,* vol. 15, pp. 38–48, 2012.

[14] T. Habib, J. Inglada, G. Mercier, and J. Chanussot, "Support vector reduction in SVM algorithm for abrupt change detection in remote sensing," *IEEE Geosci. Remote Sens. Lett.,* vol. 6, no. 3, pp. 606–610, Jul. 2009.

[15] K. J. Wessels *et al.*, "Rapid land cover map updates using change detection and robust random forest classifiers," *Remote Sens.*, vol. 8, no. 11, Oct. 2016, Art. no. 888.

[16] B. De Vries, M. Dec Uyper, J. Verbesselt, A. Zeileis, M. Herold, and S. Joseph, "Tracking disturbance-regrowth dynamics in tropical forests using structural change detection and landsat time series," *Remote Sens. Environ.,* vol. 169, pp. 320–334, 2015.

[17] B. De Vries, M. Dec Uyper, J. Verbesselt, A. Zeileis, M. Herold, and S. Joseph, "Tracking disturbance-regrowth dynamics in tropical forests using structural change detection and Landsat time series," *Remote Sens. Environ.,* vol. 169, pp. 320–334, 2015.

[18] F. Gao *et al.*, "Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery," *Remote Sens. Environ.,* vol. 188, pp. 9–25, 2017.

[19] L. ZhiYong, T. Liu, J. A. Benediktsson, and N. Falco, "Land cover change detection techniques: Very-high-resolution optical images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 44–63, Mar. 2022.

[20] K. Tan, J. Xiao, A. Plaza, X. Wang, X. Liang, and P. Du, "Automatic change detection in high-resolution remote sensing images by using a multiple classifier system and spectral–spatial features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,* vol. 9, no. 8, pp. 3439–3451, Aug. 2016.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, pp. 436–444, 2015.

[22] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet++," *Remote Sens.,* vol. 11, 2019.

[23] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *Proc. Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2115–2118.

[24] N. Venugopal, "Automatic semantic segmentation with deeplab dilated learning network for change detection in remote sensing images," *Neural Process. Lett.,* vol. 51, pp. 2355–2377, 2020.

[25] Y. Wang *et al.*, "Mask deeplab: End-to-end image segmentation for change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.,* vol. 104, 2021, Art. no. 102582.

[26] Q. Ke and P. Zhang, "CS-HSNet: A cross-siamese change detection network based on hierarchical-split attention," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,* vol. 14, pp. 9987–10002, 2021.

[27] N. Venugopal, "Sample selection based change detection with dilated network learning in remote sensing images," *Sens. Imag.,* vol. 20, pp. 1–22, 2019.

[28] A. Cz *et al.*, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, 2020.

[29] J. Xu, C. Luo, X. Chen, S. Wei, and Y. Luo, "Remote sensing change detection based on multidirectional adaptive feature fusion and perceptual similarity," *Remote Sens.,* vol. 13, no. 15, Aug. 2021, Art. no. 3053.

[30] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4403718.

[31] 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, May 2020, Art. no. 1662.

[32] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, "Detecting urban changes with recurrent neural networks from multitemporal sentinel-2 data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 214–217.

[33] 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.

[34] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.

[35] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst., Annu. Conf. Neural Inf. Process. Syst.*, I. Guyon *et al.*, Eds. Long Beach, CA, USA: Curran Associates, 2017, pp. 5998–6008.

[36] M. Liu, C. Zhang, H. Bai, R. Zhang, and Y. Zhao, "Cross-Part learning for fine-grained image classification," *IEEE Trans. Image Process.,* vol. 31, pp. 748–758, 2021.

[37] D. Guo and D. Terzopoulos, "A transformer-based network for anisotropic 3d medical image segmentation," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 8857–8861.

[38] D. Chen, H. Hsieh, and T. Liu, "Adaptive image transformer for one-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12247–12256.

[39] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10578–10587.

[40] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sens.,* vol. 13, 2021.

[41] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art. no. 5607514.

[42] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Understanding,* vol. 187, 2019, Art. no. 102783.

[43] K. Yang *et al.*, "Asymmetric siamese networks for semantic change detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022, Art. no. 5609818.

[44] 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, pp. 1–16, 2022, Art. no. 5604816.

[45] N. Bourdis, D. Marraud, and H. Sahbi, "Constrained optical flow for aerial image change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2011, pp. 4176–4179.

[46] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.,* vol. 57, no. 1, pp. 574–586, Jan. 2018.

[47] 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.

[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[49] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. JMLR Workshop Conf.*, 2011, pp. 315–323.

[50] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

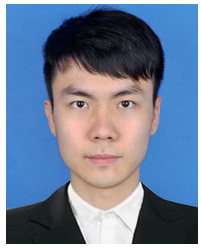
[51] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.,* vol. 18, no. 5, pp. 811–815, May 2021.

Mengxi Liu (Student Member, IEEE) received the B.S. degree in geographic information science in 2019 from Sun Yat-sen University, Guangzhou, China, where she is currently working toward the Ph.D. degree in cartography and geographic information system with the School of Geography and Planning.

Her research interests include intelligent understanding of remote sensing images, change detection, and domain adaptation.

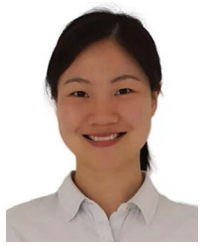
Zhuoqun Chai is currently working toward the B.S. degree with the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China.

His research interests include urban spatial analysis and image processing with deep learning.



Haojun Deng received the B.S. degree in geographic information science in 2020 from Sun Yat-sen University, Guangzhou, China, where he is currently working toward the M.S. degree in cartography and geographic information system with the School of Geography and Planning.

His research interests include machine learning and deep learning in phenology and urbanization.



Rong Liu received the B.S. degree in surveying and mapping engineering from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, in 2018.

She is currently an Associate Professor with the School of Geography and Planning, Sun Yat-sen University, China. She was a Postdoc Researcher or a Senior Researcher with the Remote Sensing Technology Institute, German Aerospace Center, Germany, and also with Signal Processing in Earth Observation, Technical University of Munich, from 2018 to 2021. Her research interest includes remote sensing image processing and evolutionary computation.