Spatially and Semantically Enhanced Siamese Network for Semantic Change Detection in High-Resolution Remote Sensing Images

Manqi Zhao, Zifei Zhao, Shuai Gong, Yunfei Liu, Jian Yang, Xiong Xiong, and Shengyang Li

***Abstract*—Given a pair of bitemporal very high resolution (VHR) remote sensing images, the semantic change detection task aims to locate land surface changes and identify their semantic classes. The existing algorithms use independent branches to locate and identify separately without considering the association between branches. In this article, we propose an end-to-end spatially and semantically enhanced Siamese network (SSESN) for semantic change detection. The SSESN aggregates the rich spatial and semantic information in the VHR image through a designed spatial and semantic feature aggregation module. Additionally, a change-aware module is proposed to decouple the aggregated features. Features in the binary branch are fused to the semantic branches as prior location information. This allows the spatially enhanced features to predict changed regions and the semantically enhanced features to refine the region categorizations. Experimental results show that our method provides comparable results with the state-of-the-art binary change detection and semantic change detection algorithms.**

***Index Terms*—Change aware (CA), change detection, remote sensing image, siamese network, spatial and semantic aggregation.**

I. Introduction

CHANGE detection in remote sensing images is essential in earth observation systems, distinguishing between images of the same geographic area taken at different times [1], [2]. Change detection is crucial in various applications, including ecosystem monitoring [3], urban planning [4], resource management [5], and damage assessment [6]. Change detection is commonly divided into binary change detection (BCD) and semantic change detection (SCD) tasks. BCD aims to identify the pixels corresponding to the changed regions to distinguish the changed regions from the unchanged ones. In contrast, SCD needs to further identify the change category based on the distinguished changed regions [7].

Manuscript received November 30, 2021; revised January 22, 2022 and February 18, 2022; accepted March 2, 2022. Date of publication March 16, 2022; date of current version April 6, 2022. This work was supported by the Space Science and Application of China Manned Space 392 Engineering DataBase in the National Basic Science Data Center under Grant NBSDC-DB-17. *(Corresponding author: Shengyang Li.)*

The authors are with the Key Laboratory of Space Utilization, Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhaomanqi19@csu.ac.cn; zhaozifei18@csu.ac.cn; gongshuai19@csu.ac.cn; liuyunfei@csu.ac.cn; yangjian20@csu.ac.cn; xiongxiong20@csu.ac.cn; shyli@csu.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3159528

Benefiting from the development of satellite imaging technology, many satellites, including Ikonos, QuickBird, Spot-5, and GaoFen, can acquire very high resolution (VHR) images at the meter or submeter resolutions. VHR images provide rich spatial distribution information and texture details of the surfaces, improving the image interpretation and context extraction at the pixel level [8]. These characteristics help the change detection task obtain more accurate changed regions and distinguish the features between different images more effectively [9].

Among the existing change detection algorithms, deep learning methods benefit from powerful feature representation capabilities, thereby efficiently using the rich spatial distribution information and semantic information in the VHR image. However, the features of different layers extracted by neural networks are often imbalanced between semantic and spatial: shallow features retain more spatial location information, while deep features contain more semantic information [10]. Moreover, the existing SCD methods handle BCD tasks simultaneously but use independent branches to process SCD and BCD separately [11]. These methods ignore the correlation between the two tasks: BCD can help SCD obtain prior location information of the changed regions.

To balance and aggregate the rich spatial and semantic information in VHR images, we propose a spatial and semantic feature aggregation (SSFA) module that integrates the feature pyramid structure and dense connections between multilevel features. Furthermore, to take advantage of the correlations between BCD and SCD, we design a change-aware (CA) module that uses binary features as prior information of semantic features.

Our main contributions in this article are as follows.

1) We propose a novel end-to-end spatially and semantically enhanced Siamese network (SSESN), which fuses and enhances the spatial and semantic information in VHR remote sensing image pairs. The SSESN achieves state-of-the-art performance on both the BCD and SCD datasets.
2) We design an SSFA module that aggregates the rich spatial and semantic information in the VHR image pairs. SSFA integrates the pyramid structure and establishes dense connections between multilevel features, making the extracted features more representative.
3) We devise a CA module that establishes the association between BCD and SCD tasks. CA uses spatial attention

and channel attention mechanisms to decouple the spatial and semantic information in the aggregated features. The features in the binary branch are then concatenated to the semantic branch as prior information, which further improves the accuracy of SCD.

The rest of this article is organized as follows. Section II summarizes the related work. Section III describes the proposed SSESN. Section IV shows our experimental results and discussion. Finally, Section V concludes this article.

II. Related Work

In this section, we propose a review of BCD and SCD algorithms. Sections II-A and II-B correspond to BCD and SCD, respectively. In addition, we also provide a specific discussion of the problems in the existing methods and present a brief synopsis of our corresponding solutions.

A. BCD Algorithms

Many BCD methods have been proposed to distinguish changed regions. Change vector analysis is usually applied to multispectral images acquired by various multispectral sensors [12], [13]. Formed by calculating the difference between the data of each image band at each time, the change vector provides the change intensity and direction, with the changed and unchanged regions identified by thresholds [14], [15]. Using canonical correlation analysis, multivariate change detection extracts change information from linear combinations of the raw data with maximal correlation [16]. Regions with little or no change in the image have absolute values close to zero, while significantly changed regions have large absolute values [17]. More recent methods take advantage of the enormous parameter space and powerful feature representation of deep learning networks to enhance the extraction of change information in remote sensing images [18]. Benefiting from the weight sharing structure and the dual-input mechanism, the Siamese network [19] has been widely incorporated into BCD methods [20], [21]. Deep feature extraction and enhanced representation further improve the performance of BCD algorithms [22]. Some methods use a pyramid structure to extract multiscale features to obtain more robust spatial features [23]. Attention mechanisms add the ability to enhance feature representation and discrimination [24], [25]. As illustrated in Section I, the rich spatial distribution and semantic information in VHR images indicate that it is necessary to explore some aggregation structures to take full advantage of the information. We propose an SSFA module to acquire well-balanced spatial and semantic aggregated features.

B. SCD Algorithms

Most research into BCD distinguishes between changed regions in fixed semantic scenes, such as buildings, land, and vegetation, without distinguishing specific semantic categories [26]. Therefore, there are few existing methods for SCD [27]. The existing algorithms treat SCD as a multicategory classification problem, classifying the specified semantic categories of two changing images separately [7]. Alternatively, SCD is treated as a semantic segmentation problem of changed and unchanged regions [11]. Yang *et al.* [28] present an asymmetric Siamese network with heterogeneous feature extraction to alleviate the categorical ambiguity in the semantic identification process. Most of the works treat the localization of changed regions and identify the semantic classes separately. Daudt *et al.* [7] provide a simple combination to add features from the semantic encoder to the binary decoder structure. However, we believe that each process emphasizes particular aspects while looking for overall changed regions. The localization process pays more attention to *where* changes occur, while the identification process pays more attention to *what* changes. This focus can be naturally linked to the attention mechanism [29]–[31]. Furthermore, the change position obtained from BCD can be used as *a priori* information to reduce the burden of semantic prediction. In this article, we propose a CA module to establish the connection between BCD and SCD tasks.

III. Methodology

An overview architecture of our proposed method is shown in Fig. 1. First, the Siamese weight-sharing backbone structure extracts multilevel features from the bitemporal VHR image pairs. The extracted multilevel features F_{L1}, F_{L2}, F_{L3}, and F_{L4} are fed into the SSFA module to integrate high-level semantic and low-level spatial information. Next, the aggregation features F_{T1} and F_{T2} and their feature difference map F_{diff} are sent into the CA module, corresponding to two semantic branches and a binary branch, respectively. Finally, both the binary and SCD results of the VHR image pairs are predicted by convolution and upsampling layers.

In this section, we first introduce two prominent components of our proposed architecture: the SSFA module for aggregating multilevel spatial and semantic features in Section III-A and the CA module for establishing the association between BCD and SCD tasks in Section III-B. Then, we describe the process of generating accurate binary and semantic probability maps along with the loss function in Section III-C.

A. SSFA Module

VHR images contain rich semantic information and precise spatial location information. It is well known that feature fusion is an effective way to enhance feature representation. Generally, there are different ways of fusion of semantic information and spatial information. Spatial fusion means aggregation between different resolutions and scales. For example, in [20] and [32]–[34], the pyramid structure is applied to gradually upsample high-level features from top to bottom and fuse same-level features. In contrast, semantic fusion requires aggregation between channels and depths. For example, in [35]–[38], densely connected network branches are organized between features with different depths and channels.

To take full advantage of semantic and spatial information in VHR images, we design the SSFA module, as shown in Fig. 2(a). SSFA takes the feature set extracted by the Siamese network as input. The set contains features with different depths F_{L1}, F_{L2}, F_{L3}, and F_{L4}, with atrous convolutions of rate 2 applied

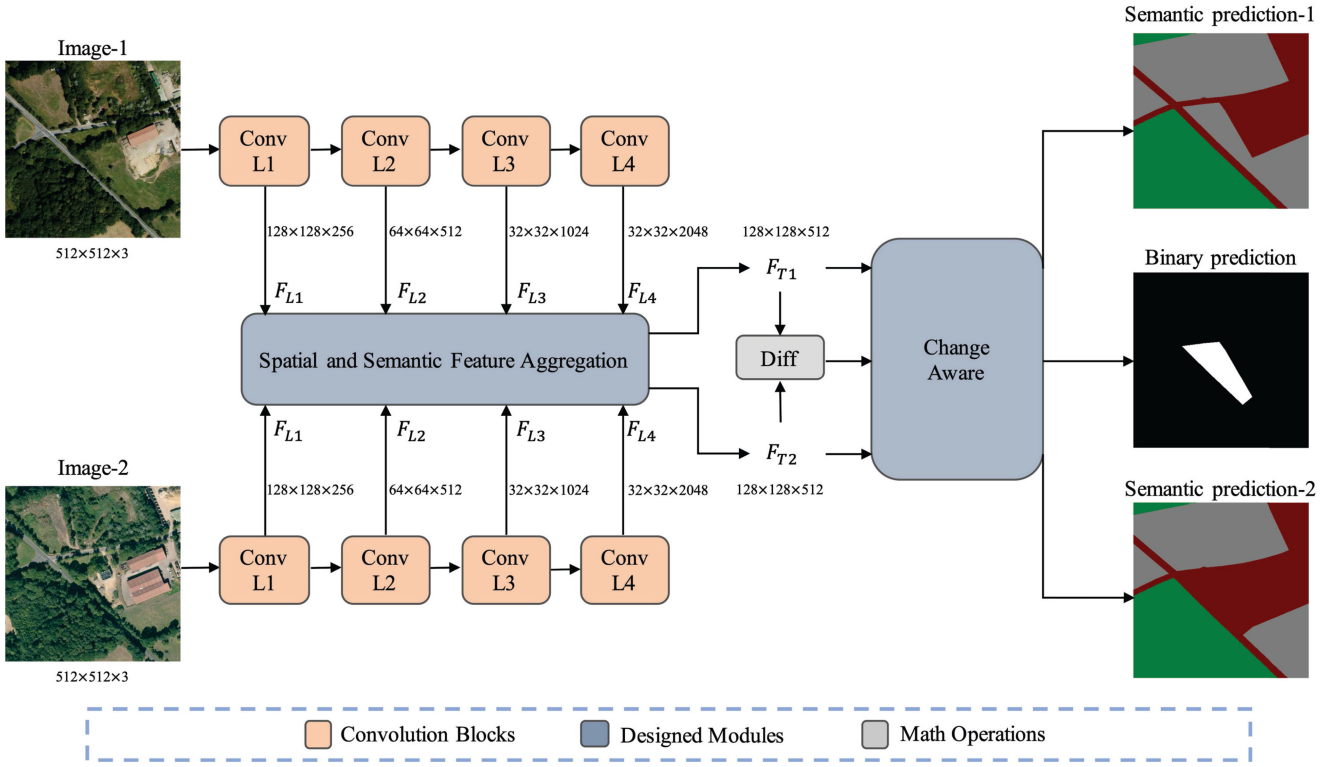


Fig. 1. Overview architecture of the SSESN. It consists of three prominent components: a Siamese structure to extract multilevel features from the bitemporal VHR image pairs, an SSFA module to integrate high-level semantic and low-level spatial information, and a CA module to decouple the fusion features and generate final prediction maps.

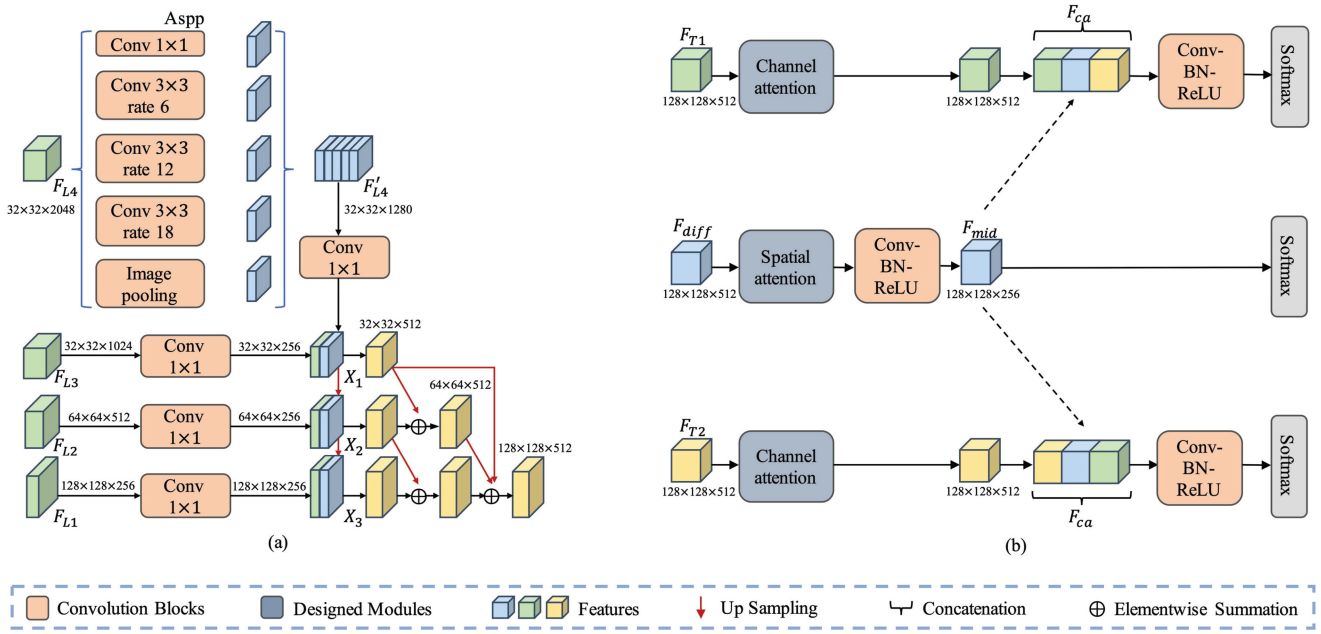


Fig. 2. Detailed structure of the designed SSFA and CA modules.

to the last layer to preserve the resolution of deep features. For convenience, the spatial image resolution is represented by *stride*, e.g., *stride 8* corresponds to 1/8 input image resolution.

First, the Atrous Spatial Pyramid Pooling (Aspp) module [39] is applied on the last layer feature, F_{L4}. Aspp consists of three 3×3 atrous convolutions with different rates, a 1×1 convolution, and an image pooling layer to adjust the field of view and capture enhanced features at multiple scales. The output features of each layer are concatenated to the fusion feature F'_{L4} at *stride 16*. Then, 1×1 convolution is

adopted to reduce the dimension of F'_{L4}. The reduced F'_{L4} is upsampled to *stride 8* and *stride 4*, and features with the same spatial resolution are dimensionally reduced and concatenated to obtain the fusion features X_1, X_2, and X_3. Although the fusion features have information at multiple scales, there is no interaction between high-level and low-level features, so it is necessary to fuse the multilevel features further.

A pyramid structure with dense connections is designed to aggregate the features X_1, X_2, and X_3. Each aggregation node is the summation of input features $Y_s = \Sigma_{i=1}^{3} a(X_i, s)$, where $X = \{X_1, X_2, X_3\}$ represents the input feature set, and Y_s represents the output feature at *stride* s. The function $a(X_i, s)$ consists of a bilinear sampling for upsampling X_i to *stride* s and a shortcut connection for maintaining resolution. Using this scheme, all features are aggregated to *stride 4*, which is then suitable for spatial and semantic balance [40]. Compared with the conventional down–up stream with a shortcut connection at the same resolution, dense connections in the pyramid structure further enhance the semantic information. The resulting SSFA module finally obtains the aggregated features with enhanced semantic and spatial information.

B. CA Module

Most of the existing methods focus only on changed regions as identified in the BCD task. Although some methods pay more attention to the semantic connotations of the changed regions as required by the SCD task, all of these approaches regard BCD and SCD as two separate tasks: the head of the network uses completely separate branches to predict the binary and semantic change maps. Nevertheless, there is a connection between the BCD and SCD tasks. More precisely, the binary branch focuses on using spatial information to distinguish *changed or unchanged*, while the semantic branch focuses on using semantic information to predict *what changed*. Therefore, using the spatial information of the binary branch as *a priori* knowledge alleviates the burden of the semantic branch to determine the spatial changed regions and results in better prediction of the semantic changes. Moreover, the channel and spatial attention serve as tools to enhance the semantic and spatial feature responses [30], [31], making the features more discriminative for each branch.

We design the CA module to establish the connection between the binary and semantic branches, as shown in Fig. 2(b). CA takes the semantic branch fusion features F_{T1}, F_{T2}, and the binary branch fusion feature F_{diff} as input. The channel and spatial attention modules are added at the beginning of CA to enhance the discriminative ability of the binary branch and semantic branch features, respectively. Fig. 3 shows the designed channel and spatial attention modules.

The spatial attention module is formed as an hourglass structure [41]. It first downsamples the features through convolution to highlight possible changed regions from the large receptive field and then uses transposed convolution to upsample them to restore the resolution. The learning process can be expressed by

$$F_{\text{spa}} = \sigma(H(F_{\text{diff}})) \otimes F_{\text{diff}} \tag{1}$$

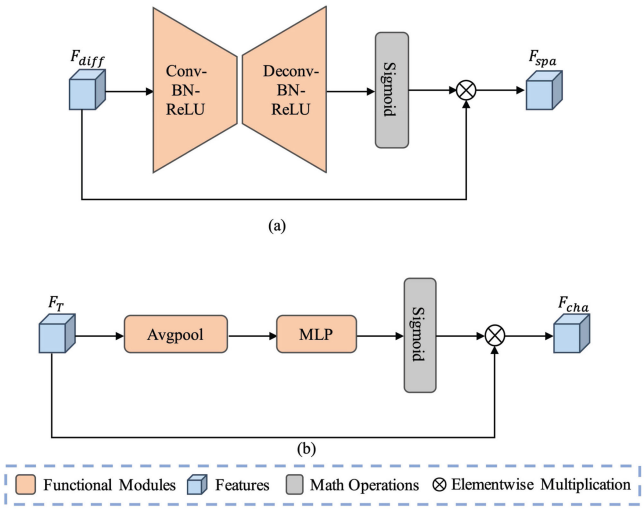


Fig. 3. Designed channel and spatial attention modules. (a) Spatial attention module. (b) Channel attention module.

TABLE I
DEFINITIONS OF EVALUATION METRICS

Metrics	Formula
P	$TP/(TP+FP)$
R	$TP/(TP+FN)$
F1	$2PR/(P+R)$
OA	$(TP+TN)/(TP+TN+FN+FP)$
mIOU	$TP/(FP+FN+TP)$
Kappa	$(p_o-p_e)/(1-p_e)$

where σ represents a sigmoid activation function, H denotes the hourglass structure, F_{diff} and F_{spa} correspond to the input feature and the output spatially enhanced feature, respectively, and $\otimes$ is elementwise multiplication.

The channel attention module takes the semantic branch fusion feature as input and obtains the average pooling feature F_{avg} through a global average pooling layer. A multilayer perceptron (MLP) consisting of two linear transformation layers and a ReLU nonlinearity layer in between is then applied to F_{avg} to form the channel weight, and the sigmoid activation function is used to normalize the weight. Finally, the channel weight and the input feature are multiplied to obtain the semantically enhanced output feature F_{cha}. The channel attention module can be expressed as

$$F_{\text{cha}} = \sigma(\text{MLP}(F_{\text{avg}})) \otimes F_{\text{avg}}. \tag{2}$$

After applying the spatial and channel attention modules, the binary branch feature can pay more attention to the location information of the changed regions, and the semantic branch feature has a more powerful semantic representation ability of the changed regions. Then, for the binary branch, we use several Conv-BN-ReLU blocks to refine F_{spa} into F_{mid}. Considering that feature in the binary branch has *a priori* location information of the changed regions, we concatenate the semantic branch

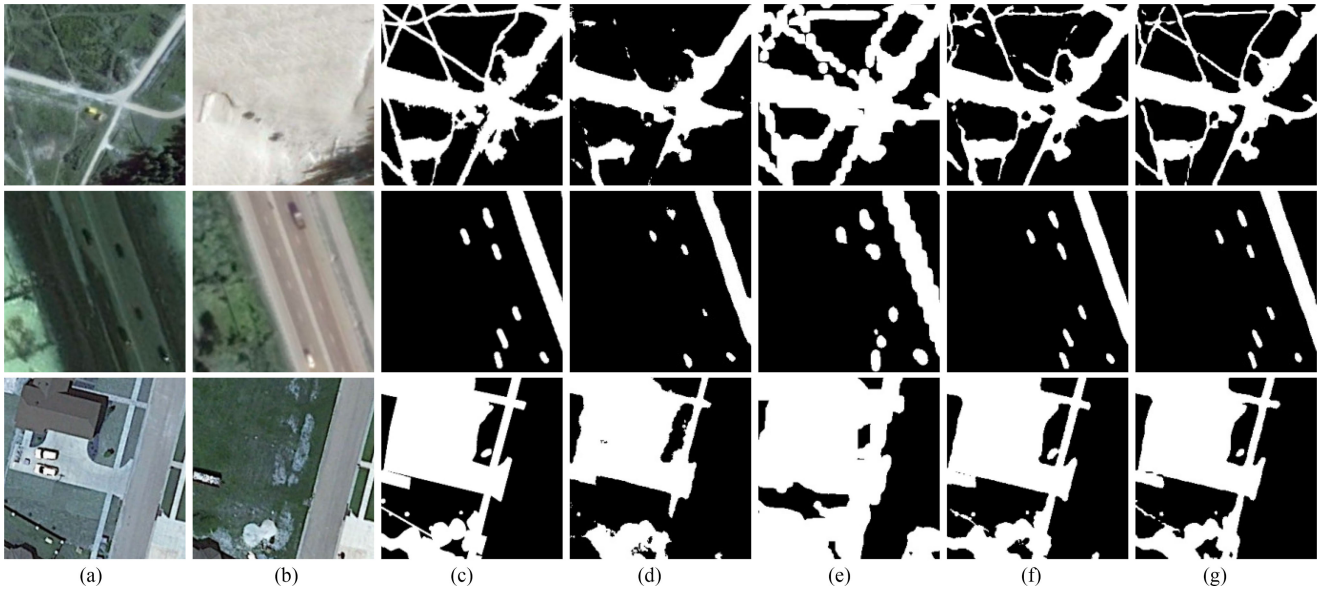


Fig. 4. Visualization results on the CDD dataset. (a) and (b) Original bitemporal image pairs. (c) Ground truth. (d) Result from FC-Siam-conc. (e) Result from DASNet. (f) Result from SNUNet-CD. (g) Result from our SSESN.

TABLE II
QUANTITATIVE RESULTS ON THE CDD DATASET

Method	P	R	F1
FC-Siam-Conc	0.710	0.666	0.687
FCN-PP	0.826	0.806	0.805
UNet++_MSOF	0.895	0.871	0.876
DASNet	0.914	0.925	0.919
SNUNet-CD	0.968	0.967	0.967
SSESN (Ours)	**0.973**	**0.962**	**0.967**

TABLE III
QUANTITATIVE RESULTS ON THE HRSCD AND SECOND DATASETS

Dataset	Method	OA	mIOU	Kappa
HRSCD	FC-EF	0.913	0.582	0.707
	FC-Siam-diff	0.907	0.557	0.647
	FC-Siam-conc	0.908	0.533	0.609
	HRSCD.str4	0.910	0.610	0.719
	SSESN (Ours)	**0.919**	**0.643**	**0.756**
SECOND	FC-EF	0.871	0.581	0.180
	FC-Siam-diff	0.879	0.619	0.268
	FC-Siam-conc	0.886	0.638	0.242
	HRSCD.str4	0.889	0.672	0.294
	SSESN (Ours)	**0.890**	**0.708**	**0.311**

features F_{cha} and the refined binary branch feature F_{mid} to F_{ca}. Similar Conv-BN-ReLU blocks are also used to refine F_{ca}. The CA module successfully establishes the connection between BCD and SCD.

C. Loss Function

Both the binary and semantic final features are obtained through a channel reduction convolutional layer and an upsampling layer. Then, a softmax layer is applied to the final features to generate the binary and semantic prediction maps. To optimize the proposed network and handle the imbalanced distribution of each class, we adopt the focal loss function [42] for both the binary and semantic branches. The loss function can be expressed as

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{M}\alpha_j y_{ij}(1-p_{ij})^{\gamma}\log(p_{ij}) \tag{3}$$

where N is the total number of training samples, M is the number of all classes, $y_{ij} \in \{0,1\}$ indicates whether a specific sample i belongs to the label of class j, $p_{ij} \in [0,1]$ is the probability that sample i is predicted to belong to class j, γ is the tunable parameter, and α_j is the balanced weight for class j.

The prediction semantic and binary maps can be denoted as $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_c$, separately. Specifically, given the ground truth for SCD $\mathcal{G}_1$ and $\mathcal{G}_2$ and for BCD $\mathcal{G}_c$, the joint loss function $\mathcal{L}_{\text{all}}$ is the linear combination of loss functions on separate branches

$$\mathcal{L}_{\text{all}} = \mathcal{L}(\mathcal{M}_1,\mathcal{G}_1) + \mathcal{L}(\mathcal{M}_2,\mathcal{G}_2) + 2\mathcal{L}(\mathcal{M}_c,\mathcal{G}_c). \tag{4}$$

IV. EXPERIMENTS AND DISCUSSIONS

A. Evaluation Datasets and Metrics

CDD [43] is one of the most common BCD evaluation datasets. The CDD dataset contains 11 pairs of multispectral

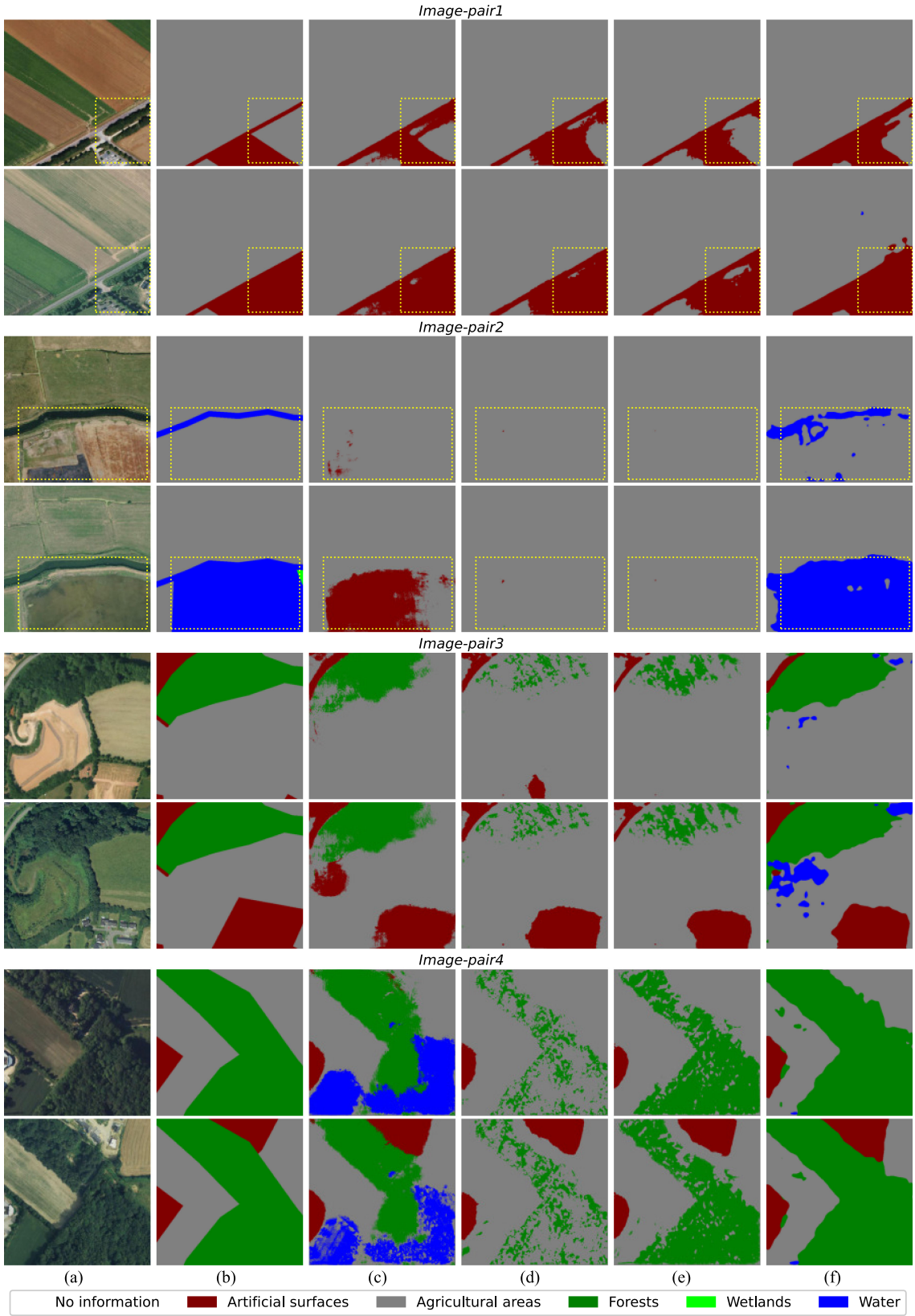


Fig. 5. Visualization results on the HRSCD dataset. (a) Original bitemporal image pairs. (b) Ground truth. (c) Result from HRSCD.str4. (d) Result from FC-Siam-diff. (e) Result from FC-EF. (f) Result from our SSESN.

TABLE IV
ABLATION STUDY RESULTS ON THE HRSCD AND CDD DATASETS

	HRSCD			CDD			FLOPS(G)
	OA	mIOU	Kappa	P	R	F1	
SSESN (baseline model)	0.903	0.570	0.696	0.926	0.916	0.919	
SSESN (w SSFA)	0.912	0.615	0.723	0.957	0.942	0.950	+ 24.77
SSESN (w CA)	0.916	0.621	0.736	0.949	0.934	0.942	+ 71.92
SSESN (w SSFA & CA)	0.919	0.643	0.756	0.973	0.962	0.967	+ 96.69

remote sensing images in different seasons acquired by Google Earth. The spatial resolution of these images ranges from 3 to 100 cm per pixel. 16 000 pairs of images with a size of 256×256 pixels are generated from the original images through cropping and rotation operations. Among the 16 000 image pairs, 10 000 are selected as the training set and 3000 for each of the validation and test sets. Precision rate (P), recall rate (R), and F1 score (F1) are adopted as BCD evaluation metrics.

HRSCD [7] is one of the few publicly available SCD datasets. The HRSCD dataset contains 291 pairs of aerial images with a resolution of 50 cm per pixel from the National Institute of Geographic and Forest Information's BD ORTHO database. There are six categories in the HRSCD dataset: artificial land, agricultural areas, forests, wetlands, water, and no change category. In the dataset, 99.232% of all pixels are labeled as no change category, and 0.653% stands for agricultural areas and artificial land, which brings extreme label imbalance. Considering the expensive training and testing costs brought by the original $10\,000 \times 10\,000$-pixel images, we cropped the original images to 512×512 pixels and chose images that focused on the changed regions. We selected 5713 pairs of images as the training set and 2854 pairs as the test set.

SECOND [28] is another SCD dataset containing 4662 pairs of aerial images with each image size 512×512. There are seven categories in the SECOND dataset: nonvegetated ground surface, tree, low vegetation, water, buildings, playgrounds, and nonchange category. 61.2% of all pixels correspond to the nonchange category, 29.9% stands for nonvegetated ground surface, while the remaining classes lower than 5%. Limited by open-source availability, we randomly split the original public training set into new training and testing sets with a ratio of 4:1, corresponding to 2374 and 594 pairs, respectively. Overall accuracy (OA), mean intersection over union (mIOU), and Kappa coefficient are adopted as SCD evaluation metrics.

The definition and the formula of the evaluation metrics are shown in Table I, where TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively. p_o corresponds to observed agreement between ground truth and predictions, and p_e is the expected agreement between ground truth and predictions.

B. Implementation Details

We implemented the proposed SSESN on the Pytorch framework and conducted all the experiments on two Nvidia Titan RTX GPUs. We augmented the data using random flips and rotations. The optimization process used the Adam optimizer [44], and the batch size was set to 16. The training process contained 50 epochs in total with an initial learning rate of 0.001. The learning rate decays by 0.1 at the epoch of 20. In addition, the weight decay and momentum were set to 0.0001 and 0.9, respectively. As for parameters in the loss function, we set γ to 2. For the balanced weight, we set α_j to 0.2 if class j occupies more than 50% pixels, 2.0 if less than 5% pixels, and 1.0 if in between.

C. Comparison and Analysis

For the BCD task, we selected several representative deep-learning-based methods for comparison. FC-EF, FC-Siam-conc, and FC-Siam-diff [20] integrate the UNet [33] structure and develop a baseline model for change detection based on the Siamese network structure. Similarly, FCN-PP [45] utilizes a U-shaped network structure and introduces pyramid pooling to enlarge the receptive field to overcome the limitations of traditional global pooling. The variants of UNet structure, i.e., UNet ++ [37], was also introduced into change detection. UNet+ + _ MSOF [46] inputs image pairs into the UNet++ backbone network and uses multiside output fusion for hierarchical supervision. SNUNet-CD [47] integrates the Siamese network, UNet++, and channel attention mechanisms. The dense skip connection reduces the uncertainty of the edge pixels of the changed regions. DASNet [24] applies the spatial and channel attention mechanisms to describe local features of the changed regions and recognize the pseudo-changes.

As shown in Table II, our method achieved state-of-the-art performance on the CDD dataset. More precisely, the F1 score of our method reached 0.967, matching the best result from SNUNet-CD but with a higher precision rate of 0.973. In order to evaluate and compare more intuitively, we visualized the experimental results, as shown in Fig. 4.

Fig. 4 shows three pairs of representative images divided into three rows to comprehensively display the change detection results in different scenarios. The first row shows our model's accurate recognition of complex road network changes; our method yielded the most coherent lines. The color and background in the image pairs in the second row varied greatly, and our method successfully captured the changed vehicles and roads. The third row of images shows typical architectural changes in change

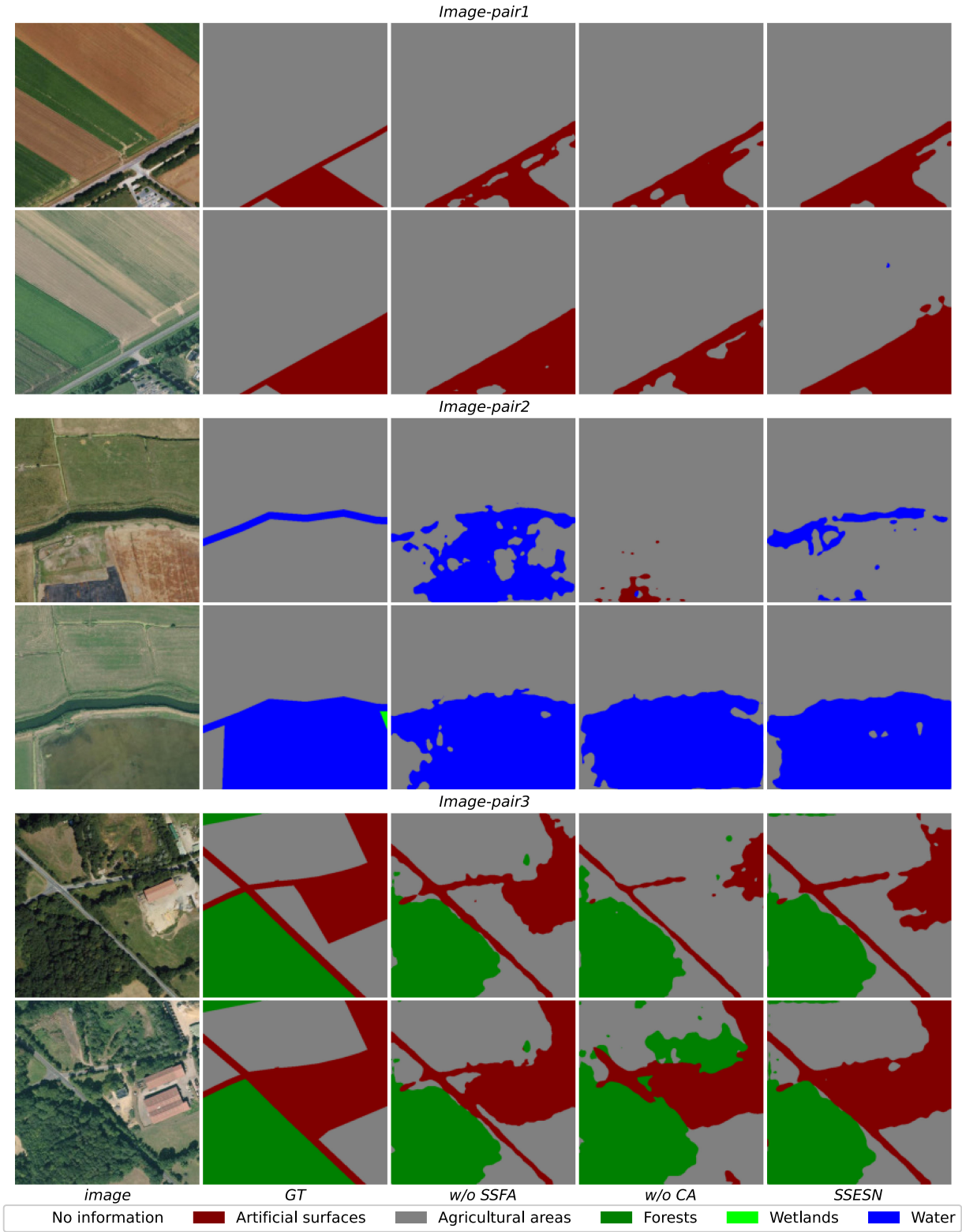


Fig. 6. Ablation results on the HRSCD dataset.

detection. Our method retained most of the changed details in the extraction of building contours.

For the SCD task, we used FC-EF, FC-Siam-conc, and FC-Siam-diff [20] as baseline models by changing output category numbers. The fourth strategy proposed in [7] explores the combination of SCD and BCD based on the U-shaped structure: the encoder of the semantic branch is merged with the decoder of the binary branch through skip connections and finally obtains the

BCD and SCD maps from the change and semantic branches, respectively. This method is denoted by HRSCD.str4 for short.

The quantitative results in Table III show that our method achieved the best performance on both the HRSCD and SECOND datasets. Our approach gets the highest OA (0.919, 0,890), mIOU (0.643, 0.708), and Kappa coefficient (0.756, 0.311) on the two datasets. Taking the method HRSCD.str4 as an example, although HRSCD.str4 combines BCD and SCD, the results show that this combination was suboptimal, with mIOU scores of 0.610 and 0.672, and Kappa coefficient values of 0.719 and 0.294. Our method exceeds the performance of HRSCD.str4 by 5.4% on the mIOU metric and 5.8% on the Kappa metric. Similarly, we have a more intuitive visual comparison and analysis of the results on the HRSCD dataset in Fig. 5.

We selected four bitemporal image pairs containing the most comprehensive semantic change categories as representative in Fig. 5. *Image-pair1* and *Image-pair2* focus on comparing the accuracy of changed regions, while *Image-pair3* and *Image-pair4* perform comprehensive comparisons of changed regions and semantic categories. Compared with other methods, the SSESN best identified changed regions and categories. More specifically, *Image-pair2* shows large-scale changes between water and agricultural areas. Only SSESN correctly predicted the changed regions and categories; all other methods failed. In addition, other image pairs reflect small-scale changes between artificial surfaces and agricultural areas. Owing to the interference of some details (such as trees in agricultural areas), the prediction results generally contained noise. Although small-scale changes were more susceptible to effects such as being misidentified as noise, the SSESN reliably provided more accurate contour details. Benefiting from more effective aggregation and extraction of semantic information, the SSESN determined more accurate semantic categories in all image pairs.

D. Ablation Study

In order to further validate the effectiveness of the proposed SSFA and CA, we performed an ablation study on the HRSCD and CDD datasets. Without changing the experimental parameter settings, SSFA and CA were added to the baseline model to perform the ablation study. The baseline model corresponds to the method without SSFA and CA modules. As shown in Table IV, on the HRSCD dataset, adding SSFA increased the mIOU by 7.9% and the Kappa coefficient by 3.9%; adding CA increased the mIOU and the Kappa coefficient by 8.9% and 5.7%, respectively. The complete SSESN model exceeds the baseline model by 12.8% on the mIOU score and 8.6% on the Kappa score. Results on the CDD dataset reveal a similar conclusion. Compared with the baseline model, the complete SSESN model has better performance on P, R, and F1 scores with the improvement of 5.0%, 5.1%, and 5.1%, respectively. In summary, SSFA and CA improved the model performance of our method for BCD and SCD tasks, and the quantitative results confirmed the effectiveness of our proposed modules.

In addition, Fig. 6 shows the visualization results of the ablation study. With CA existed alone, the model extracted the correct change region from the three image pairs. However, the lack of semantic aggregation information in SSFA led to the model producing errors in change category prediction (*Image-pair2*). With SSFA presented only, the model produced incorrect results when predicting the changed regions. Therefore, semantic and spatial information fused by SSFA cannot be appropriately decoupled. The visualization results further demonstrate the value of the proposed SSFA and CA structures.

V. Conclusion

In this article, we propose an end-to-end SSESN for the SCD of VHR images called SSESN. The SSESN retains and integrates rich spatial and semantic information simultaneously and gives more accurate predictions of changed regions and semantic categories through our designed SSFA and CA modules. Experimental results show that the proposed model achieves competitive results with the existing state-of-the-art methods on the CDD and HRSCD datasets. In the future, we will further explore the fusion of spectral information for SCD and develop a method suitable for hyperspectral SCD.

Acknowledgment

The authors would like to thank Dr. R. Caye Daudt for the open-source HRSCD dataset.

References

[1] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.

[2] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, 2004.

[3] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin, "Digital change detection methods in ecosystem monitoring: A review," *Int. J. Remote Sens.*, vol. 25, no. 9, pp. 1565–1596, 2004.

[4] H. Luo, C. Liu, C. Wu, and X. Guo, "Urban change detection based on Dempster-Shafer theory for multitemporal very high-resolution imagery," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 980.

[5] X.-P. Song *et al.*, "Global land change from 1982 to 2016," *Nature*, vol. 560, no. 7720, pp. 639–643, 2018.

[6] 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.

[7] 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.

[8] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2021.

[9] L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 609–630, Mar. 2013.

[10] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[11] 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.

[12] W. A. Malila, "Change vector analysis: An approach for detecting forest changes with landsat," in *Proc. LARS Symp.*, 1980, p. 385.

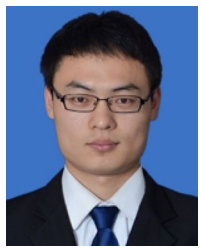
[13] E. F. Lambin and A. H. Strahlers, "Change-vector analysis in multitemporal space: A tool to detect and categorize land-cover change processes using high temporal-resolution satellite data," *Remote Sens. Environ.*, vol. 48, no. 2, pp. 231–244, 1994.

[14] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 1, pp. 218–236, Jan. 2007.

[15] F. Bovolo, S. Marchesi, and L. Bruzzone, "A framework for automatic and unsupervised detection of multiple changes in multitemporal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2196–2212, Jun. 2012.
[16] A. A. Nielsen, K. Conradsen, 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, no. 1, pp. 1–19, 1998.
[17] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Feb. 2007.
[18] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Auton. Robots*, vol. 42, no. 7, pp. 1301–1322, 2018.
[19] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 539–546.
[20] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 4063–4067.
[21] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.
[22] 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.
[23] C. Zhang *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.
[24] J. Chen *et al.*, "DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, Nov. 2020.
[25] 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.
[26] K. Sakurada, M. Shibuya, and W. Wang, "Weakly supervised silhouette-based semantic scene change detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 6861–6867.
[27] S. Tian, Z. Zheng, A. Ma, and Y. Zhong, "Hi-UCD: A large-scale dataset for urban semantic change detection in remote sensing imagery," *CoRR*, vol. abs/2011.03247, pp. 1–6, Dec. 2020.
[28] 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.
[29] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[32] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interve.*, 2015, pp. 234–241.
[34] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[36] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2403–2412.
[37] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, 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*. Berlin, Germany: Springer, 2018, pp. 3–11.
[38] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.
[39] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, pp. 1–14, Jun. 2017, *arXiv:1706.05587*.
[40] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 3069–3087, Sep. 2021.
[41] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.
[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[43] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 42, no. 2, pp. 565–571, 2018.
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[45] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, and A. K. Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 982–986, Jun. 2019.
[46] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet++," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1382.
[47] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Feb. 2021.

Manqi Zhao received the B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2019. He is currently working toward the Ph.D. degree in computer applied technology with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing.

His research interests include remote sensing change detection, satellite, unmanned aerial vehicles, and conventional video object tracking.



Zifei Zhao received the M.Eng. and B.Eng. degrees in photogrammetry and remote sensing from the Shandong University of Science and Technology, Qingdao, China, 2015 and 2018. He is currently working toward the Ph.D. degree with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China.

His research interests include remote sensing image processing and satellite video analysis, such as object detection.

Shuai Gong received the B.Eng. degree in computer science and technology from Hunan University, Changsha, China, in 2019. He is currently working toward the M.S. degree in computer technology with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China.

His research interests include remote sensing change detection, satellite video analysis, and knowledge graph.

Yunfei Liu received the B.Eng. and M.S. degrees in computer science and technology from Beijing Forestry University, Beijing, China, in 2013 and 2016 respectively.

He is currently an Engineer with the Technology and Engineering Center for Space Utilization, Chinese Academy of Science, Beijing. His research interests include big data mining and analysis.

Jian Yang received the B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2020. He is currently working toward the M.S. degree in signal and information processing with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing.

His research interests include object tracking for unmanned aerial vehicles and intelligent video analysis.

Xiong Xiong received the B.Eng. degree in robot science and engineering from Northeastern University, Shenyang, China, in 2020. He is currently working toward the M.S. degree in computer technology with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China.

His research interests include knowledge graph and information extraction.

Shengyang Li received the M.S. degree in computer science and technology from the Shandong University of Science and Technology, Qingdao, China, in 2003, and the Ph.D. degree in remote sensing image processing and analysis from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2006.

He is currently a Professor with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences. His research interests include ground data processing system technology, target detection and tracking in video satellite, and machine learning in remote sensing image classification and recognition.