BSCDNet: A Building Change Detection Network With Category Differentiation Using a Graph Attention Mechanism and Multitask Learning

Qian Shen, Shikang Tao, Rui Yang, Xin Zhang, and Min Wang

***Abstract*—Recent building-oriented change detection studies considered only morphological changes in buildings, and few publicly available change detection datasets further distinguish among building types. In this study, we propose a semantic change detection network named BSCDNet that considers both morphological and semantic changes in buildings. BSCDNet adopts a multitasking branch structure with object classification, change detection, and segmentation to simultaneously achieve object-level semantic classification and change analysis of buildings in bitemporal, high-spatial-resolution (HSR) imagery. In the object classification branch, a graph attention network is utilized to capture the spatial and semantic correlations among buildings during classification. The change detection branch applies both spatial and channel attention mechanisms to eliminate nonbuilding interference and enhance change features. Moreover, the segmentation branch adopts a distinctive instance segmentation procedure that improves the accuracy of object segmentation. We created a building change detection dataset with category differentiation based on HSR imagery to validate the proposed method. Ablation experiments verify the effectiveness and advantages of the above-mentioned task branches. Furthermore, in comparative experiments with several SOTA semantic change detection methods such as HRSCD, SCDNet, and MR_CD, BSCDNet reached the optimal level in terms of F1 and mIoU when evaluating the change detection performance, as well as kappa and score for evaluating the classification performance.**

***Index Terms*—Building, deep learning, graph attention network (GAT), semantic change detection (SCD), Siamese network.**

I. Introduction

In THE urban remote sensing field, change detection and analyses of urban elements, such as buildings, are vital in urban planning, land use management, illegal construction detection, and disaster damage assessment [1], [2], [3], [4], [5], [6]. Change detection can generally be categorized into two modes: binary change detection (BCD) and semantic change detection (SCD) [7], [8]. The former simply identifies changed/unchanged areas in binary mode, whereas the latter further obtains the semantics, i.e., categories of changed areas, by involving certain image classification processing techniques.

Currently, most research on building-oriented change detection focuses on morphological changes, such as the appearance, disappearance, and deformation of buildings [9], [10], [11], [12], [13]; this detection method is regarded as BCD or single-class SCD. Obtaining semantic types with morphological changes provides reasonable clues for analyzing the driving forces of urban changes. However, research on building change detection with category differentiation or building semantic change detection (BSCD) is scarce and urgently needs to be strengthened.

SCD, which can obtain the semantic types of changed buildings by certain image classification processes, offers a reasonable solution for the proposed building change detection issue. SCD can be categorized into four implementation modes [14], [15], [16]: training a land cover mapping network to compare the results for pixels in the image pair (Mode 1); utilizing two-branched Siamese networks for bitemporal image change analysis (Mode 2); decoupling the SCD task into independent BCD subtasks and semantic segmentation subtasks (Mode 3); and integrating the BCD and semantic segmentation into a single multitask network so that land cover information can be used for change detection (Mode 4). In these schemes, multitask networks can simultaneously extract building semantic types and morphological changes in an end-to-end manner. Therefore, a multitask network is a convenient and advanced strategy for BSCD. However, the collaborative training mode, in cases with multitasks, has high network design requirements.

Building distributions exhibit strong spatial correlations; i.e., buildings of the same type generally exhibit a clustered tendency, and different building types may have specific accompanying phenomena. Semantic segmentation, which is commonly utilized in SCD for semantic extraction, can hardly utilize these object-level classification clues due to its pixel-wise image classification process. A graph neural network (GNN) is a deep learning technique that has emerged in recent years for classifying graph-structured data. In image classification using GNNs,

Manuscript received 30 October 2023; revised 21 January 2024 and 11 March 2024; accepted 7 May 2024. Date of publication 17 May 2024; date of current version 13 September 2024. This work was supported in part by the National Key R&D Program of China under Grant 2021YFB3901300, and in part by the National Natural Science Foundation of China under Grant 42071301. *(Corresponding author: Min Wang.)*

Qian Shen, Shikang Tao, Rui Yang, and Min Wang are with the Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing 210023, China, also with the School of Geography, Nanjing Normal University, Nanjing 210023, China, also with the Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China, and also with the State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing 210023, China (e-mail: shenqian_gis@163.com; tsk2920722405@163.com; 1341655531@qq.com; sysj0918@126.com).

Xin Zhang is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China, and also with the Qilu Aerospace Information Research Institute, Jinan 250132, China (e-mail: zhangxin@radi.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3402431

graph nodes generally represent image objects, whereas graph edges represent specific relationships among these objects. Spatial positions and neighborhood relationships among objects are then encoded as an adjacency matrix. In this scenario, GNNs utilize edges to aggregate node features and generate new feature representations, which capture the interactions and dependencies among neighboring nodes and enable GNNs to effectively learn contextual information in classification. Compared with semantic segmentation technology, GNNs naturally capture and utilize the correlations among spatial objects, which should be beneficial for improving the accuracy of object classification, such as for buildings.

GNNs are divided into several variants, namely the graph convolutional network [17], GraphSAGE [18], and the graph attention network (GAT) [19]. GAT utilizes an attention mechanism to weigh different nodes according to their importance, which enables GAT to learn feature representations more accurately than other variants. In this study, we utilize the GAT for BSCD, which enables the simultaneous extraction of morphological and semantic changes of buildings. The characteristics and main contributions of this study are as follows.

1) A novel BSCD model, BSCDNet, is introduced. BSCDNet utilizes several deep learning techniques, such as instance segmentation, graph convolution, and multitask learning, that enable end-to-end classification and change detection for buildings via high-spatial-resolution imagery. To our knowledge, BSCDNet is the first network to consider building category differentiation within the SCD research field.
2) A new building classification strategy for BSCD is proposed. In contrast to the FCN-based semantic segmentation approach commonly used in SCD, BSCDNet employs a specific scheme similar to instance segmentation to obtain building objects, which are subsequently input into the GAT for object-level classification. This scheme effectively utilizes the aggregation and adjacency relationships specific to buildings and thus improves the accuracy of classification and change detection.
3) A new BSCD dataset. The current publicly available SCD/BCD datasets lack building category differentiation information. We thus created the first building semantic change detection dataset (BSCDD) with category differentiation by visual interpretation and object delineation for several scenes of GF-2 and unmanned aerial vehicle images. The BSCDD will be publicly available at https://github.com/SianGIS/building-semantic-change-detection-dataset.

The remainder of this article is organized as follows. Section II presents a review of the current SCD methods. Section III introduces the design of BSCDNet. Section IV presents experiments on different benchmark datasets and method performance analyses. Section V presents further discussion, and Section VI provides the conclusions and discussion of future work.

II. Related Work

SCD based on deep learning typically includes four modes: postclassification change detection [20], [21], [22], [23], direct change detection [24], [25], [26], separation of two subtasks [27], and integration of two subtasks (multitask change detection) [7], [28], [29], [30], [31], [32], [33], [34]. Postclassification change detection obtains change areas by overlaying and comparing classification results from bitemporal images. Its process is quite simple but relies heavily on accurate image registration and classification. In direct change detection, SCD is decomposed into a certain amount of BCD and requires output fusion, which is cumbersome when handling multiple semantic types. Methods with separate modes are completely independent without the mutual assistance capabilities of multitask networks, which could limit the accuracy of the method. Methods based on multitask SCD networks commonly take change detection as the main task and classification as an auxiliary task. Bitemporal image features are extracted through the encoder of a Siamese network. These features are fed to the classification branches of the decoder for semantic extraction and fused and fed to the change detection branch to obtain the binary-mode change results. In multitask mode, classification training and change detection tasks are simultaneously performed in a single network, offering a convenient solution for operators. However, the corresponding network design involves more techniques than other SCD modes. Subtle network design achieves mutual promotion among different network branches [15]. Due to its characteristics and advantages, multitask learning-based SCD is a popular research field. The most recent topics in this field include enhancing change features based on semantic constraints [34], using Siamese semantic-aware encoders to extract multiscale features [14], and guaranteeing temporal symmetry via a temporal-symmetric transformer model [31].

BSCD includes direct change detection [12], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48] and multitask change detection [10], [20], [49], [50], [51], [52]. In the former, only changed buildings are detected, whereas in the latter, all buildings and the corresponding changes are simultaneously extracted. In direct change detection, bitemporal features are extracted from Siamese encoders and fused as changed features in the decoder. Popular research topics in direct change detection include multiscale joint supervision [43], introducing a self-attention mechanism for context modeling in the spatiotemporal dimension [45], and introducing a cross-attention mechanism to more robustly extract feature differences between bitemporal images [46]. In multitask change detection methods, the semantic segmentation branch bitemporally extracts buildings, and the change detection branch extracts the changed buildings. The key issue lies in designing the network structure and controlling losses to ensure balance and consistency among different tasks. Strategies for this task include utilizing semantic feature constraints for change detection [50], improving loss functions to mitigate sample imbalance [52], and introducing an intrascale cross-interaction module to fuse the feature maps from different branches [51].

Currently, the publicly released SCD datasets include bitemporal images, changed regions, and their semantic class labels. Representative datasets include the SECOND [7], Hi-UCD [53], and HRSCD [15] datasets. These datasets cover several urban and rural areas and include multiple types of land use/land cover classes; however, buildings are commonly classified as one type

without further category differentiation. Building change detection datasets include datasets created for direct change detection and multitask change detection. The direct change detection datasets include LEVIR-CD [54], GDS [55], SYSU-CD [56], etc., in which only changed buildings are labeled. The multitask building change detection datasets include WHU-CD [57] and HRSCD [15], in which all the buildings and their changes are labeled. However, none of these building change detection datasets further subdivide the building categories.

III. Method Design

In pixel-level SCD methods, fragmented or conjoined building extraction results often exist in semantic segmentation/classification. However, the GAT needs morphologically complete building instances to learn the spatial and semantic correlations among buildings for classification. For this purpose, BSCDNet utilizes an object detection-based approach for BSCD, which consists of classification, change detection, and mask branch. Specifically, the building classification branch uses the buildings detected by object detection as graph nodes and defines the connection relationships among nodes based on the geometric and semantic features of buildings to extract categories. Then, the change detection and mask branches extract change information and perform segmentation based on individual building objects.

BSCDNet utilizes and modifies the structure of the two-stage instance segmentation network of Mask R-CNN [58] to conduct bitemporal building object extraction. Then, the spatial and semantic features of building objects are utilized to construct a dual-branch GAT, and object classification is implemented through attention mechanism-based spatial relationship reasoning among graph nodes. In this section, object classification implementation based on a GAT is introduced, and the network and corresponding branch design are described to facilitate understanding.

A. Building Classification

Given a graph $G = (V, E)$, where V represents the set of nodes and E represents the set of edges, the adjacency matrix $A \in \mathbb{R}^{N\times N}$ defines the adjacency relationships among the nodes, where N is the number of nodes. The feature F_i represents the feature vector of node i ($i \in V$). The GAT adopts an attention mechanism to allocate different weights to neighboring nodes. Given a node i and its neighbor node j, the attention coefficients e_{ij} are calculated as follows:

$$e_{ij} = {W_e}^T \left([WF_i || WF_j]\right), j \in N_i \tag{1}$$

where W is the weight matrix with shared parameters, vector W_e is a learnable weight vector, "$\|$" denotes the channel concatenation operation, and N_i is the neighborhood of node i. To make the coefficients easily comparable across different nodes, e_{ij} is normalized with the softmax function

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k\in N_i}\exp\left(e_{ik}\right)} \tag{2}$$

where α_{ij} represents the normalized attention coefficient. The aggregated feature of node i is then calculated as follows:

$$\hat{F_i} = LeakyReLU\left(\sum_{j\in N_i}\alpha_{ij}W_\alpha F_j\right) \tag{3}$$

where W_α represents a learnable weight matrix used for linear transformation of the input features.

Building objects are obtained by the region proposal network (RPN) [59], which forms the nodes of the GAT network. The spatial and semantic relationship graphs are constructed based on the geometric and semantic features of the building objects. To construct the spatial relationship graph, the geometric feature for a node pair (i, j) is defined as follows:

$$F_{s(i,j)} = \left[\log\left(\frac{|x_i - x_j|}{w_i}\right), \log\left(\frac{|y_i - y_j|}{h_i}\right), \log\left(\frac{w_j}{w_i}\right), \log\left(\frac{h_j}{h_i}\right)\right]^T \tag{4}$$

where $F_{s(i,j)}$ represents the 4-D geometric feature of the node pair; (x_i, y_i) and (x_j, y_j) are the center coordinates of nodes i and j, respectively; and (w_i, h_i) and (w_j, h_j) are the width and height of the proposed regions of nodes i and j, respectively. Then, the adjacency matrix $\mathrm{A_{spa}}$ of the spatial relationship graph is calculated as follows:

$$\mathrm{A_{spa(i,j)}} = \mathrm{ReLU}\left(\mathrm{ReLU}\left(\mathrm{W}_s F_{s(i,j)}\right) F_{s(i,j)}\right) \tag{5}$$

where W_s denotes the learnable weight matrix.

The adjacency matrix $\mathrm{A_{sem}}$ for the semantic relationship graph is calculated by evaluating the semantic similarity of every node pair

$$\mathrm{A_{sem(i,j)}} = \frac{\left(W_i F_i\right)^T \cdot W_j F_j}{\sqrt{d}} \tag{6}$$

where F_i and F_j represent the proposed region features of nodes i and j, d is the feature dimension, and W_i and W_j are the learnable weight matrices.

For building object classification, image features are first aggregated in the GAT layers based on $\mathrm{A_{spa(i,j)}}$ and $\mathrm{A_{sem(i,j)}}$ and are then fused by concatenation

$$F_i^{l+1} = \left[\left|\sum_{j\in N_i}\alpha_{i,j}^{spa}W_a^{spa}A_{spa(i,j)}F_j^l\right| \times \left|\sum_{j\in N_i}\alpha_{i,j}^{sem}W_a^{sem}A_{sem(i,j)}F_j^l\right|\right] \tag{7}$$

where W_a^{spa} and W_a^{sem} denote the learnable weight matrices and F_j^l represents the feature of node j in the lth layer.

BSCDNet includes a three-layer GAT for feature dimension reduction. The features are subsequently classified via a softmax function to obtain building categories. Through the above-mentioned operations, clustering and proximity relationships among buildings are obtained for image classification.

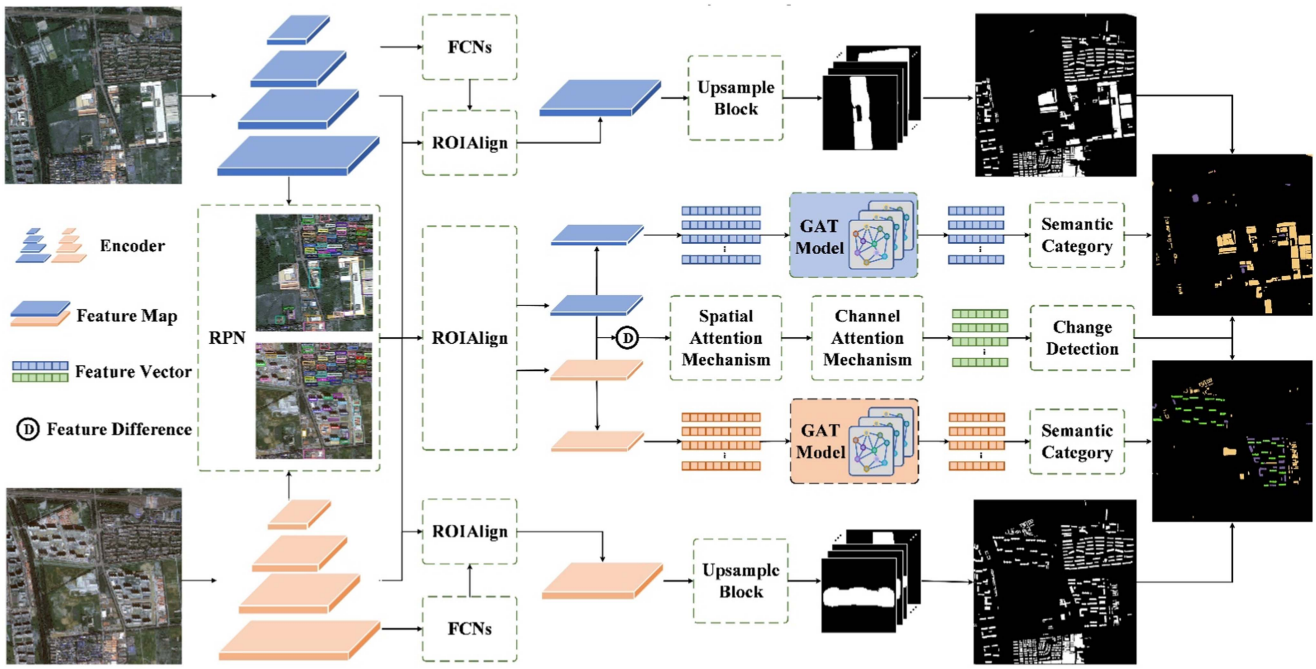

Fig. 1. Design of BSCDNet.

B. Model Structure

1) Overall Architecture: BSCDNet is an SCD method designed based on building objects. As shown in Fig. 1, BSCDNet adopts ResNet-50 to extract bitemporal features. The building object proposals are obtained via the RPN, and the object features are extracted via ROIAlign [58]. Based on the building object proposals, BSCDNet incorporates classification, change detection, and mask branches to extract the category, change information, and mask, respectively, for each building proposal in a bitemporal image. The classification branch forms three-layer dual-branch spatial and semantic graphs, which use the GAT to learn spatial correlations that buildings of the same type generally exhibit a tendency to cluster together, and different building types may have specific accompanying relationships. The change detection branch obtains change features through bitemporal object feature subtraction, in which spatial and channel attention mechanisms are utilized for feature enhancement. The mask branch uses FCNs to initially extract the binary building segmentation results and then concatenates the results with the object features using ROIAlign to instantly segment the buildings. The changed building proposals are filtered using the change detection branch. The classification and mask branches provide semantic categories and masks, respectively, for the changed building proposals.

2) Building Classification Branch: BSCDNet first utilizes an encoder to extract multistage image features F. Then, building proposals are generated through the RPN, and a high-confidence set of building proposals $P = \{(x_n, y_n, w_n, h_n), n \in \{1, 2, 3, \ldots, 512\}\}$ is used for subsequent classification, change detection, and mask extraction. In the building classification branch, F and P are processed with ROIAlign to obtain the semantic features of each building proposal. The object features are then flattened and passed through two fully connected layers to obtain semantic feature vectors with a length of 512 to form $F_{sem} \in R^{512\times512}$

$$F_{sem} = MLP\left(ROIALign\left(F, P\right)\right) \tag{8}$$

where MLP denotes two fully connected layers. To utilize the spatial relationships among different categories of buildings to assist in building classification, the classification branch uses the building proposals as nodes to construct spatial and semantic relationship graphs. The geometric features F_{spa} among the nodes are calculated based on the spatial information of the objects recorded in P using (4).

The connectivity between the nodes is defined using (5) and (6). On this basis, a dual-branch structure with parallel processing and multilayer spatial and semantic relationship graphs is designed for feature aggregation, as shown in Fig. 2. The attention matrices α_{sem} and α_{spa} associated with the semantic relationship graph and spatial relationship graph are computed using the GAT. Subsequently, attention-based aggregation is performed on each node's geometric and semantic features

$$\begin{aligned} \hat{F}_{spa} &= \sigma\left(\alpha_{spa} W_{spa} F_{spa}\right) \\ \hat{F}_{sem} &= \sigma\left(\alpha_{sem} W_{sem} F_{sem}\right) \end{aligned} \tag{9}$$

where $\hat{F}_{spa}$ and $\hat{F}_{sem}$ represent the weighted spatial and semantic features, σ is the LeakyReLU activation function and W_{spa} and W_{sem} are learnable weight matrices. To alleviate gradient vanishing during the stacking of multiple GNN layers, BSCDNet introduces residual connections in the dual-branch structure. The feature vectors before and after aggregation are fused via pixel-wise addition. The fused features of different branches are then concatenated

$$F_g^{l+1} = \left[\hat{F}_{spa}^{l} + F_{spa}^{l} \middle\| \hat{F}_{sem}^{l} + F_{spa}^{l}\right], \quad l = 0, 1, 2 \tag{10}$$

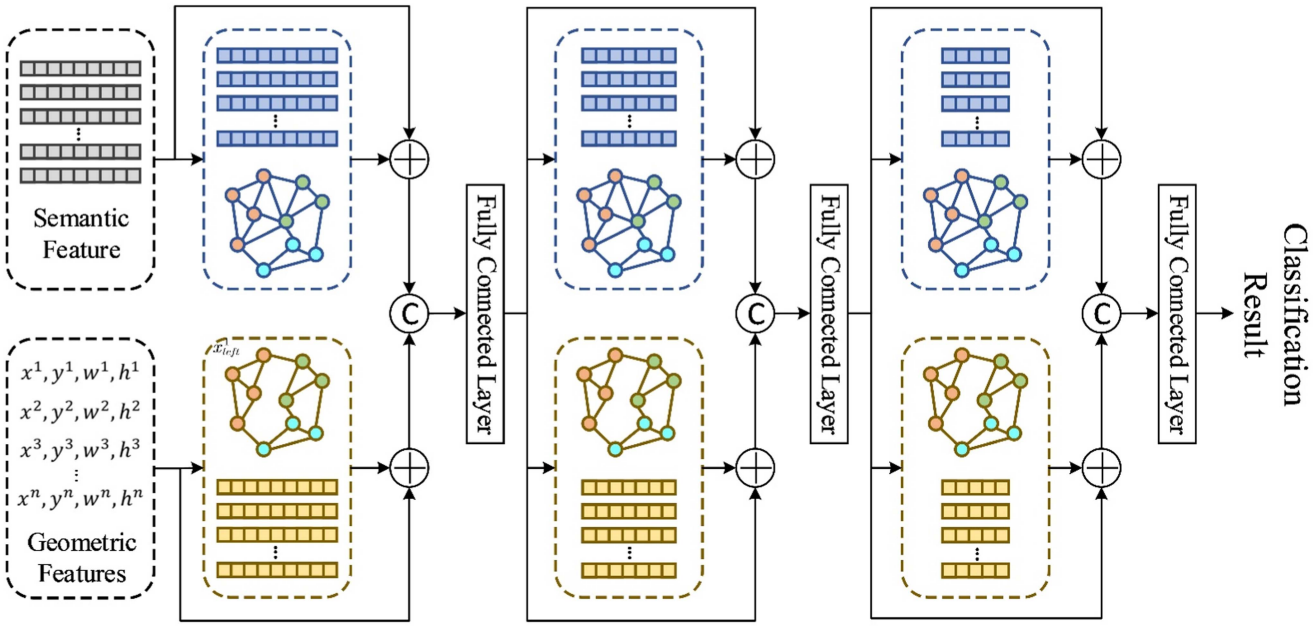

Fig. 2. Dual branches of spatial and semantic relationship graphs.

where F_g^{l+1} represents the output features of the lth dual-branch layer. The dimensions of F_g^{l+1} are compressed through the fully connected layer, and the features are subsequently fed to the next layer. A three-layer dual-branch structure is formed for the classification branch. The first two layers output feature vectors with lengths of 256 and 128, respectively, and the third layer conducts building classification through a fully connected layer and a softmax function.

3) Change Detection Branch: As shown in Fig. 3, bitemporal image features are processed via subtraction, and ROIAlign obtains the changed building features F_d. The object features are extracted based on the rectangular regions of the building proposals. When using F_d directly for change detection, nonbuilding regions may have a negative impact on the change detection results. The spatial attention mechanism (SAM) [60] adaptively learns the attention weights of features and focuses on the importance of different regions. To suppress nonbuilding interference, a SAM is introduced to enhance F_d

$$F_d^{sam} = \sigma\left(Conv_{1\times1}\left(\left[\max\left(F_d\right)\right| \left|\text{avg}\left(F_d\right)\right]\right)\right) F_d \tag{11}$$

where F_d^{sam} is the spatial attention weighted change feature, σ is the sigmoid function, and $max(\cdot)$ and $avg(\cdot)$ represent maximum pooling and average pooling, respectively. The channel attention mechanism (CAM) [61] then selects the key channels of the features. The features first undergo global max and average pooling, and the two output feature vectors are fused and processed with a sigmoid function to obtain the channel attention coefficient for feature enhancement

$$F_d^{cam} = \sigma\left(GMP\left(F_d^{sam}\right) + GAP\left(F_d^{sam}\right)\right) F_d^{sam} \tag{12}$$

where F_d^{cam} represents the feature after CAM enhancement and $GMP(\cdot)$ and $GAP(\cdot)$ represent the global max and average pooling, respectively. The enhanced features are subsequently flattened and input into a fully connected layer for final change detection. The change detection branch performs the aforementioned operations on both bitemporal buildings to extract the change information.

4) Mask Branch: The Mask RCNN, a representative method for instance segmentation, uses ROIAlign to obtain a low-resolution feature map (28×28), which is insufficient for capturing object details in FCN-based object segmentation. To address this issue, BSCDNet adopts a different scheme for accurate building instance segmentation [62]. The image features extracted by the encoder are first fed to the FCNs for the pixel-level semantic segmentation of buildings, as shown in Fig. 4. The image features and semantic segmentation results are processed by ROIAlign, and the outputs are concatenated to obtain image features that jointly represent object boundaries and masks. Then, the features are upsampled by three upsampling blocks consisting of 3×3 convolution and deconvolution layers. Finally, the object masks are obtained via 1×1 convolution and resampling.

5) Loss Function: Change detection, GAT classification, and RPN classification all use the multiclass cross-entropy loss function

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log p_{ij} \tag{13}$$

where N is the number of samples, K is the number of classes, and y_{ij} and p_{ij} represent the label and the predicted probability of the ith sample, respectively.

In the segmentation branch, the FCNs and object segmentation process both utilize the binary cross-entropy loss function

$$L = -y\log\left(\hat{y}\right) - \left(1-y\right)\log\left(1-\hat{y}\right) \tag{14}$$

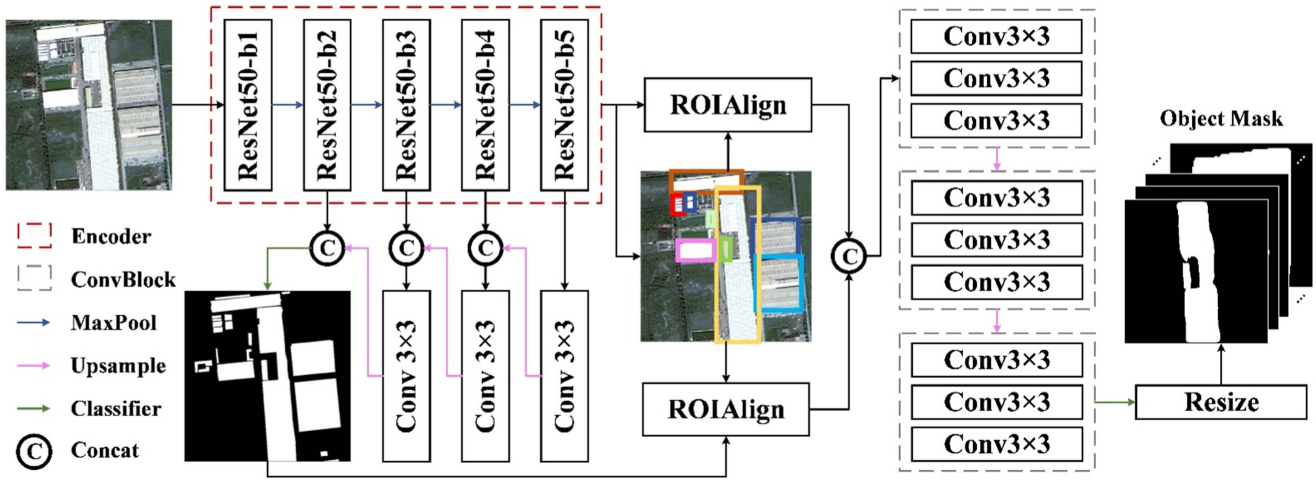

Fig. 4. Mask branch.

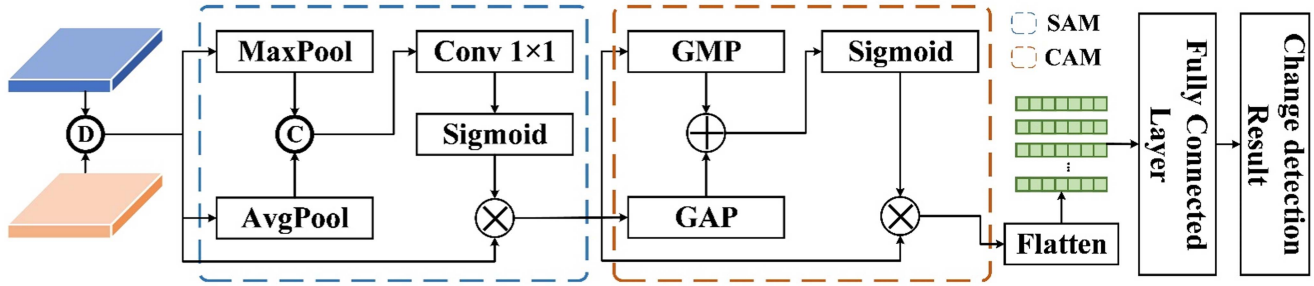

Fig. 3. Attention-based change detection branch.

where y represents the ground truth of the given sample (which is 0 or 1) and $\hat{y}$ is the network prediction. SmoothL1 was used in bounding box regression

$$L(y,\hat{y}) = \begin{cases} 0.5(\hat{y}-y)^2 & \text{if } |\hat{y}-y| < 1 \\ |\hat{y}-y| - 0.5 & \text{otherwise} \end{cases}. \quad (15)$$

The total loss of BSCDNet is finalized as follows:

$$L = L_{class1} + L_{class2} + L_{rpn1} + L_{rpn2} + L_{seg1} + L_{seg2} + L_{cd1} + L_{cd2} \quad (16)$$

where L_{class1} and L_{class2} denote the losses of the classification branches in phases 1 and 2, L_{rpn1} and L_{rpn2} represent the losses of RPN classification in phases 1 and 2, L_{seg1} and L_{seg2} denote the losses of the mask branches in phases 1 and 2, and L_{cd1} and L_{cd2} denote the losses of the change detection branches in phases 1 and 2.

IV. Experiments

A. Experimental Procedure

1) Datasets: The publicly released SCD datasets were primarily designed for multiclass SCD of urban scenes. None of the SCD datasets further subdivide the building categories, and it is impossible to verify the performance of BSCDNet in building SCD. Therefore, the research team collected several high-resolution aerial and satellite remote sensing images to create the BSCDD. The BSCDD encompasses partial regions of the Beijing urban area (BSCDD_BJ) and Yangzhou rural area (BSCDD_YZ). These two regions significantly differ in terms of data sources, building morphology, and distribution; these differences are used to comprehensively evaluate the proposed method.

a) BSCDD_BJ: For BSCDD_BJ, bitemporal images covering parts of Beijing in 2015 and 2019 were collected from GF-2 and were 9785 × 10061 in size with spatial resolutions of 1 m. The buildings in the dataset were divided into six categories: shantytowns, low-rise (LR) apartments, medium- and high-rise (M&HR) apartments, commercial and office (C&O) buildings, industrial and warehouse (I&W) buildings, and auxiliary buildings. The building objects were delineated using ArcGIS software, and their categories and change information were annotated at the object level. A total of 19 000 buildings were identified in the bitemporal images, as illustrated in Fig. 5. These images and labels were cropped to 512 × 512 nonoverlapped image patches, producing 400 image pairs. The pairs were further randomly divided into training (250 pairs), validation (50 pairs), and testing (100 pairs) sets. The training sets and validation sets were expanded using rotation, cropping, and copy-paste [63] for sample enhancement.

b) BSCDD_YZ: BSCDD_YZ consists of three pairs of aerial images covering Yangzhou. The data were captured in

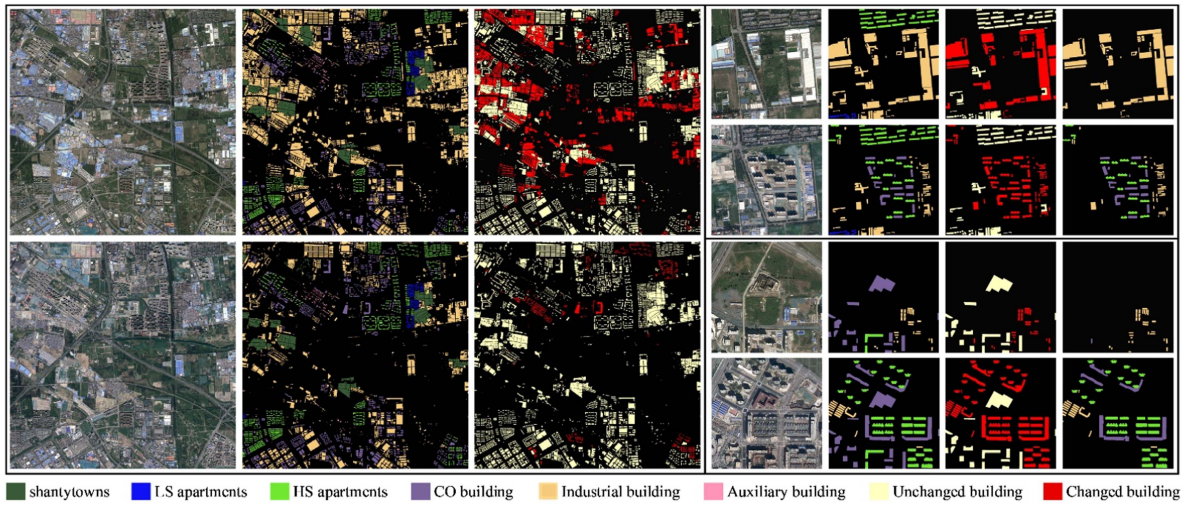

Fig. 5. BSCDD_BJ dataset.

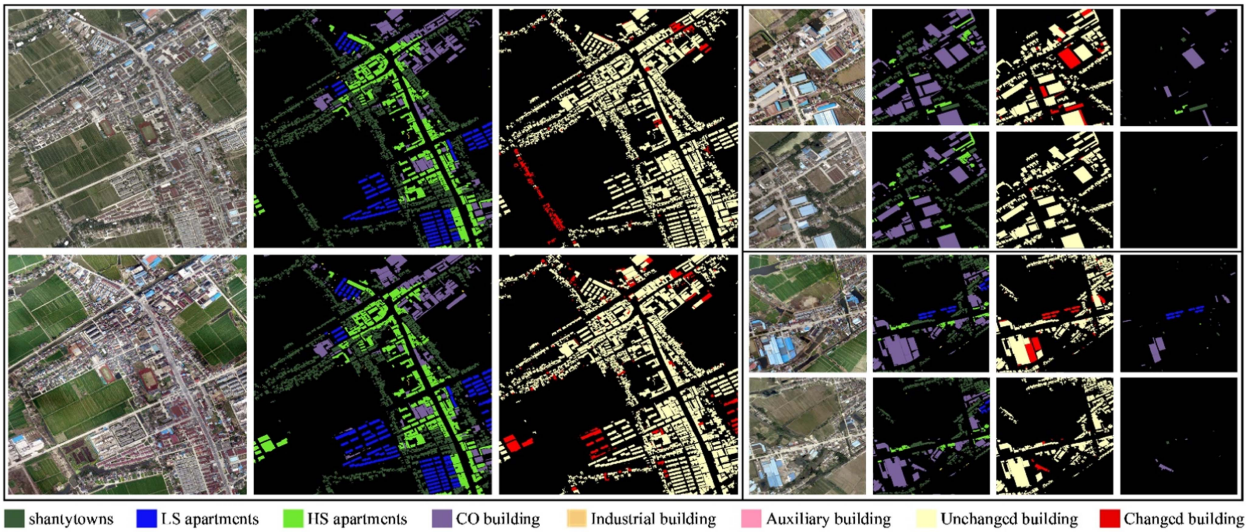

Fig. 6. BSCDD_YZ dataset.

2017 and 2020, with a spatial resolution of 0.1 m and image sizes of 33 774 × 23 946, 36 927 × 29 164, and 26 238 × 24 841. The buildings in these images were classified as low-rise residential buildings, M&HR apartments, C&O buildings, I&W buildings, or auxiliary buildings. A total of 20 544 buildings were delineated, as illustrated in Fig. 6. The images were cropped to 512 × 512 image patches without overlap, which produced 462 image pairs. The pairs were further randomly divided into training (300 pairs), validation (62 pairs), and testing (100 pairs) sets. The training sets and validation sets were augmented in the same way as the BSCDD_BJ.

2) Method Implementation: Several pixel-level and object-level SCD methods were selected and implemented for comparison. The pixel-level change detection methods include HRSCD.str1 [15], HRSCD.str4 [15], SCDNet [28], Bi-SRNet [30], and MTSCD [14]. The semantic segmentation and change detection branches adopted cross-entropy loss and binary cross-entropy loss with loss weights of 0.25 and 0.5, respectively.

HRSCD.str1: In this method, UNet is used to extract multi-class buildings. The extraction results are compared to conduct pixel-wise SCD.

HRSCD.str4: This method involves a multitask learning network with a semantic segmentation branch and a change detection branch. The category of the changed building is obtained by comparing the change detection and semantic segmentation results at the pixel scale.

SCDNet: This is a typical multitask SCD method. The semantic segmentation branch classifies an image into unchanged regions and changed regions with category information. The change detection branch utilizes multiscale difference features to achieve BCD. Similarly, the change detection and semantic segmentation results at the pixel scale are compared for SCD.

Bi-SRNet: In this method, the bitemporal features are merged through a deep CD unit to extract the changed areas. The segmentation branch uses Siamese and cross-temporal semantic reasoning blocks to improve the semantic representations and

model the semantic correlations of features to achieve the semantic segmentation of changed regions with temporal feature alignment.

MTSCD: This method adopts a two-level difference feature model to extract change information and generate a spatial attention weight map to provide prior information regarding change areas for the semantic segmentation branch. The semantic segmentation branch directly classifies the two temporal image features constrained by spatial attention weights via FCNs.

Considering that semantic building change detection studies are lacking and that BSCDNet is an object-level change detection method, two object-level SCDs were implemented based on several classic networks, i.e., the two-stage instance segmentation network MaskRCNN and the one-stage instance segmentation network SOLO [64]. The encoder for both of these comparative methods is ResNet-50.

MR_CD: In this approach, a dual RPN is used to extract bitemporal building proposals. Then, the semantic categories and masks of the building objects are extracted through the classification branch and mask branch of the Mask R-CNN. The change detection branch fuses the bitemporal image features of the objects and classifies the objects into changed and unchanged categories. The segmentation branch adopts binary cross-entropy loss, as do both the classification branch and the change detection branch.

SOLO_CD: This method uses Siamese encoders to extract bitemporal image features and divides them into a grid of fixed-size cells. The segmentation and classification branches extract the categories of objects in the grid. The change detection branch concatenates the bitemporal grid features to perform change detection.

All the models were implemented with PyTorch 1.7.1, which was powered by two NVIDIA 2080ti GPUs with 11 GB of RAM. The batch size was uniformly set to 4 for all model training, the same hyperparameters were used, and the number of training epochs in all experiments was uniformly set to 50. The Adam optimizer was used with a default learning rate of 1e-4. During the training process, the model with the lowest loss in the validation set was saved as the final training result.

3) Accuracy Measures: To comprehensively evaluate the performance of the models, accuracy metrics such as precision, recall, F1-score, and mIoU were selected to evaluate the accuracy of the changed areas. The kappa coefficient and score were used to evaluate the categories of the changed pixels. These measures are calculated as follows:

$$\begin{aligned}\text{Precision} &= \frac{TP}{TP+FP}\\ \text{Recall} &= \frac{TP}{TP+FN}\\ F1 &= 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}\\ mIOU &= \frac{1}{2}\left(\frac{TP}{TP+FP+FN}+\frac{TN}{TN+FN+FP}\right)\end{aligned} \tag{17}$$

where TP denotes the number of positive samples that were correctly predicted, TN denotes the number of negative samples that were correctly classified, FP denotes the number of positive samples that were incorrectly classified as negative, and FN denotes the number of negative samples that were incorrectly detected as positive.

Assuming that the number of categories in the dataset is C, where 0 represents the nonchanged class, the SCD confusion matrix is $S_{ij}(0 \le i \le C-1,\ 0 \le j \le C-1)$. The kappa coefficient is calculated as follows:

$$\begin{aligned}\text{Kappa} &= \frac{p_0 - p_e}{1-p_e}\\ p_0 &= \frac{\sum_{i=0}^{C-1} S_{ij}}{\sum_{i=0}^{C-1}\sum_{i=0}^{C-1} S_{ij}}\\ p_e &= \frac{\sum_{i=0}^{C-1}\left(S_{i+}\times S_{+i}\right)}{\left(\sum_{i=0}^{C-1}\sum_{i=0}^{C-1} S_{ij}\right)^2}\end{aligned} \tag{18}$$

where S_{i+} denotes the sum of the rows of confusion matrix S_{ij} and S_{+i} represents the sum of the columns of this matrix. Based on mIoU and kappa, the comprehensive score can be calculated

$$\text{Score} = 0.3\times mIOU + 0.7\times \text{Kappa}. \tag{19}$$

B. Results and Comparison

1) Quantitative Analysis:

a) BSCDD_BJ: As shown in Table I, the precision of MTSCD is lower than that of SCDNet and Bi-SRNet. However, other metrics for MTSCD were significantly higher than those of the other pixel-based change detection methods. The quantitative indicators of MR_CD were higher than those of the other comparative methods. The F1-score and mIoU, which are related to the changed area, of this method were 77.42% and 79.16%, respectively. The kappa and score, which evaluate the categories of change areas, were 68.19% and 71.48%, respectively. These indicators were significantly higher than those of the other comparative methods. The precision of BSCDNet was 4.63% higher than that of the second-ranked MR_CD, indicating that BSCDNet reduced the number of false alarms. BSCDNet's F1-score and mIoU were 2.59% and 2.1% greater than those of MR_CD, respectively. The kappa and score values of BSCDNet were also higher than those of all other comparative methods, which indicated that BSCDNet performed better in building classification.

b) BSCDD_YZ: The quantitative results of BSCDNET and the seven comparative methods on YZ are shown in Table II. Similar to the results on BSCDD_GF, in this case, object-based methods performed much better than pixel-based methods. The mIoU, kappa, and score of MR_CD were higher than those of MTSCD. BSCDNet performed the best in terms of all the accuracy metrics, among which kappa was the most improved; notably, the kappa coefficient of BSCDNet was 4.9% and 1.41% higher than those of SOLO_CD and MR_CD, respectively. In addition, the mIoU of BSCDNet was 4.32% and 0.8% higher

TABLE I
QUANTITATIVE RESULTS OF BCDNET AND THE SEVEN COMPARISON METHODS ON THE BSCDD_BJ DATASET

Method	Precision	Recall	F1-score	mIoU	Kappa	Score
HRSCD.str1	44.79	55.97	49.76	60.87	41.79	47.51
HRSCD.str4	66.27	37.46	47.86	61.69	36.45	44.02
SCDNet	91.2	42.63	58.1	67.42	53.76	57.86
Bi-SRNet	***83.78***	57.62	68.28	73.22	63.22	66.22
MTSCD	70.41	73.02	71.69	74.98	63.76	67.13
SOLOv2	75.58	30.39	43.35	59.94	38.54	44.96
MR_CD	74.92	**80.08**	**77.42**	**79.16**	**68.19**	**71.48**
BSCDNet	**79.55**	***80.48***	***80.01***	***81.26***	***71.16***	***74.19***

The best and second-best methods in each accuracy metric column are marked in bold and bold italics, respectively.

TABLE II
QUANTITATIVE RESULTS OF BCDNET AND THE SEVEN COMPARISON METHODS ON THE BSCDD_YZ DATASET

Method	Precision	Recall	F1-score	mIoU	Kappa	Score
HRSCD.str1	53.54	54.08	53.81	61.19	38.81	45.53
HRSCD.str4	40.16	38.91	39.52	53.28	21.46	31.01
SCDNet	74.19	43.13	54.55	63.26	35.93	35.93
Bi-SRNet	***87.33***	52.25	65.38	70.00	55.23	59.66
MTSCD	77.10	**80.51**	78.77	78.97	65.14	69.29
SOLOv2	83.81	71.23	77.01	77.91	67.32	70.5
MR_CD	85.36	77.89	**81.46**	**81.43**	**70.81**	**73.99**
BSCDNet	**86.53**	***85.63***	***86.08***	***85.51***	***75.01***	***78.16***

The best and second-best methods in each accuracy metric column are marked in bold and bold italics, respectively.

than those of SOLO_CD and MR_CD, respectively. This indicated that BSCDNet improved the change detection and classification abilities, leading to a 4.73% improvement in the score metric.

2) Qualitative Analysis:

a) BSCDD_BJ: Fig. 7 shows eight representative change detection results obtained on BSCDD_BJ. HRSCD.str1 yielded more regular boundaries for the extracted changed buildings than did the other methods. However, errors exist in the morphology and categories of the semantic segmentation results for bitemporal buildings, with numerous false alarms in various areas, as shown in Fig. 7. Among the pixel-based SCD methods, Bi-SRNet and MTSCD had the most complete results and displayed high accuracy but yielded obvious classification errors and missed alarms. Many missed alarms were observed for HRSCD.str4 and SCDNet, in which the extracted changed buildings were fragmented and the building classification was inaccurate, as illustrated in areas 2 and 4. SOLO_CD exhibited many missed alarms. MR_CD exhibited missed and false alarms, as illustrated in areas 3 and 4, respectively. In addition, C&O buildings exhibited classification errors, as illustrated in area 2. The changed buildings extracted by BSCDNet were most consistent with the morphology and category labels.

b) BSCDD_YZ: Fig. 8 shows eight representative change detection results obtained on the BSCDD_YZ dataset. HRSCD.str1 yielded many false alarms, as shown in areas 2 and 4. Other pixel-level comparative methods cause missed alarms and inaccurate classifications. For example, the results of HRSCD.str4 and SCDNet in areas 2 and 3 included the misidentification of I&W buildings as C&O buildings. Bi-SRNet had missed alarms and classification errors in areas 3 and 4, respectively, and MTSCD displayed relatively low segmentation and classification accuracy in area 1. The object-based change detection methods performed better than did the pixel-based methods, but defects still existed. Among all these comparative methods, BSCDNet performed best in changed building classification, exhibiting almost no missed alarms.

C. Ablation Experiments

We conducted several ablation experiments to evaluate the module performance on BSCDD_BJ dataset. The experiments were conducted with the combined attention-based change detection modules, segmentation branches, dual-branch GAT classifiers, and three branches.

The base network (first row in Table III) is a multitask change detection network based on object detection. On the basis of building object detection in bitemporal images, the base network includes a change detection branch, a classification branch, and an object segmentation branch with two ResNet-50 encoders. The change detection branch concatenates the bitemporal image features and extracts the change features of each building object via ROIAlign. The change features are input into the fully connected layer for change detection. The classification branch

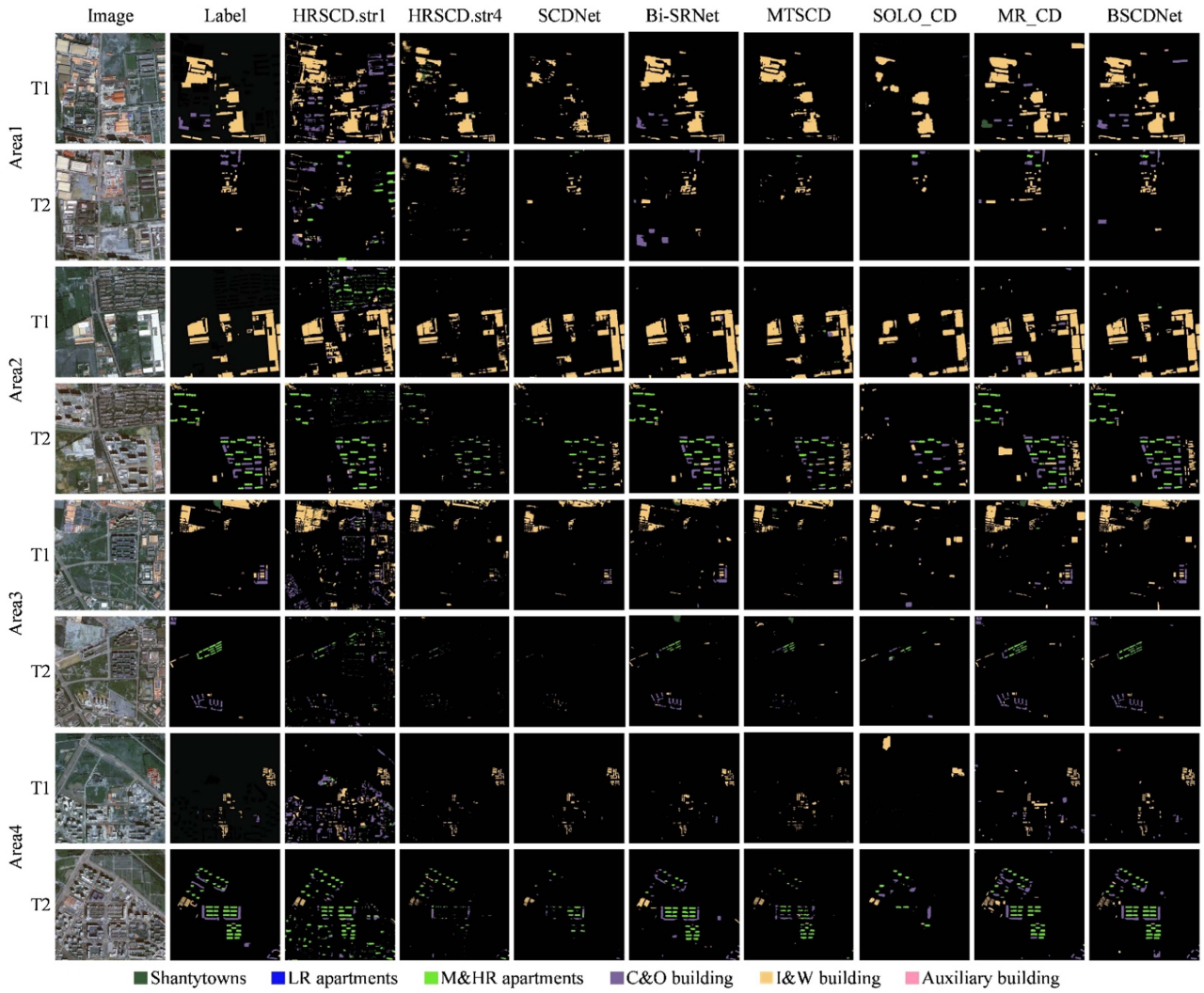

Fig. 7. Typical change detection results of BSCDNet and comparison on the BSCDD_BJ dataset.

TABLE III
QUANTITATIVE ANALYSIS RESULTS OF ABLATION EXPERIMENTS

Method	Precision	Recall	F1-score	mIoU	Kappa	Score
Base	74.92	80.08	77.42	79.16	68.19	71.48
Base+Attention	80.20	78.42	79.30	80.74	71.13	74.01
Base+Mask	**81.79**	**79.34**	**80.54**	**81.74**	**71.50**	**74.57**
Base+GAT	75.58	***79.96***	78.11	79.52	70.35	73.10
BSCDNet	***83.80***	78.74	***81.19***	***82.29***	***72.98***	***75.77***

The best and second-best methods in each accuracy metric column are marked in bold and bold italics, respectively.

uses a fully connected layer to classify objects. The object segmentation branch adopts three upsampling blocks consisting of 3×3 convolutional layers and deconvolution layers to extract the masks of building objects.

To verify the impact of the attention-based change detection branch on change detection, the change detection branch in the base network was replaced with that of BSCDNet (second row in Table III). Introducing the attention-based change detection branch significantly increased the precision of the method from 74.92% to 80.2% compared to that of the base network, indicating that introducing the attention mechanism significantly reduced false detection issues. Although the recall decreased, the F1-score, mIoU, and score increased by 1.88%, 1.58%, and 2.53%, respectively. The base network with the mask branch of BSCDNet performed better than the base network in all the metrics except for recall (third row in Table III); additionally, the two comprehensive indicators, F1-score and mIoU, increased by 3.12% and 2.58%, respectively. Improving

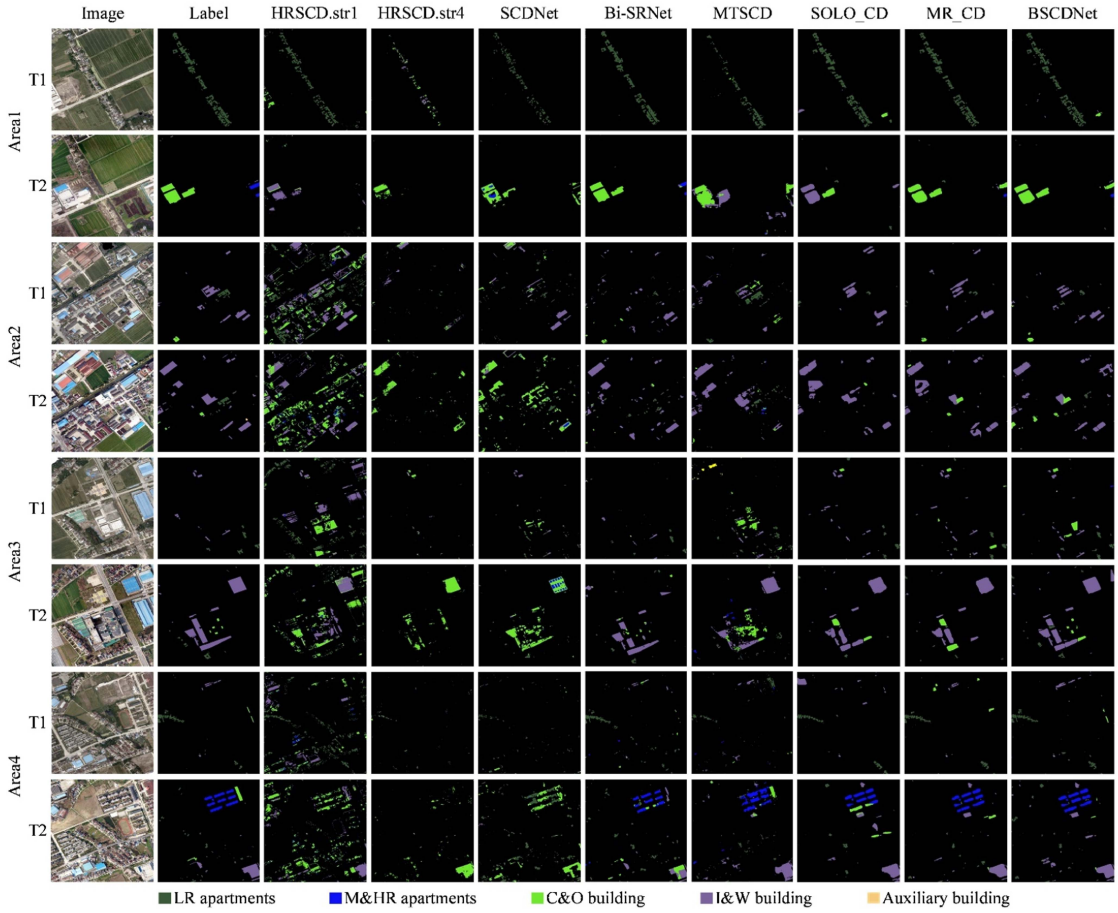

Fig. 8. Typical change detection results of BSCDNet and comparison on the BSCDD_YZ dataset.

the building segmentation accuracy also improved the kappa coefficient.

The dual-branch GAT classifier was introduced in the base network for comparisons among different classifiers (fourth row in Table III). The kappa coefficient of the classifier introduced in the dual-branch GAT was significantly higher than that of the CNN-based base network. The other indicators used to measure BCD performance were similar to those of the base network. The dual-branch GAT improved the building classification and change detection performance of the models.

To verify whether different modules can jointly improve network performance, all branches were added to the base network to form BSCDNet for comparative analyses by separately introducing different modules (fifth row in Table III). BSCDNet performed best on all the accuracy measures, among which the precision improved the most; the precision of BSCDNet was 8.88% higher than that of the method with the branch added to the base network separately. This suggested that jointly using different branches reduced the number of false alarms. An increased kappa coefficient indicates that the classification performance was improved.

Several typical sets of results of ablation experiments with different model branches are shown in Fig. 9. Introducing the attention-based change detection branch resulted in better performance and fewer false alarms than those for the base network. For example, in area 1, many false alarms existed in the result of the base network in the I&W building; the attention-based change detection branch performed better in this area. In the comparison of mask branches, the mask branch of BSCDNet provided the most accurate building masks. In area 1, the mask branch of BSCDNet yielded more complete extraction of large C&O buildings than did the base network. In area 2, the boundaries of the changed buildings from the mask branch of BSCDNet were more regular. In the comparison of classifiers, the base network identified M&HR apartments in area 1 and C&O buildings in area 2 as C&O buildings and I&W buildings, respectively. The GAT helps capture the spatial and semantic correlations among the building category during classification, and the model with the introduced dual-branch GAT correctly predicted these building categories. BSCDNet extracted the changed buildings and their categories more robustly than did the base network.

D. Building Classification Accuracy Analysis

In this section, SCDNet and MR_CD were selected as two representative methods for comparing and analyzing the performance of different SCD networks in the task of building classification. The semantic segmentation branch of SCDNet, a typical pixel-level SCD method, uses an attention mechanism and a

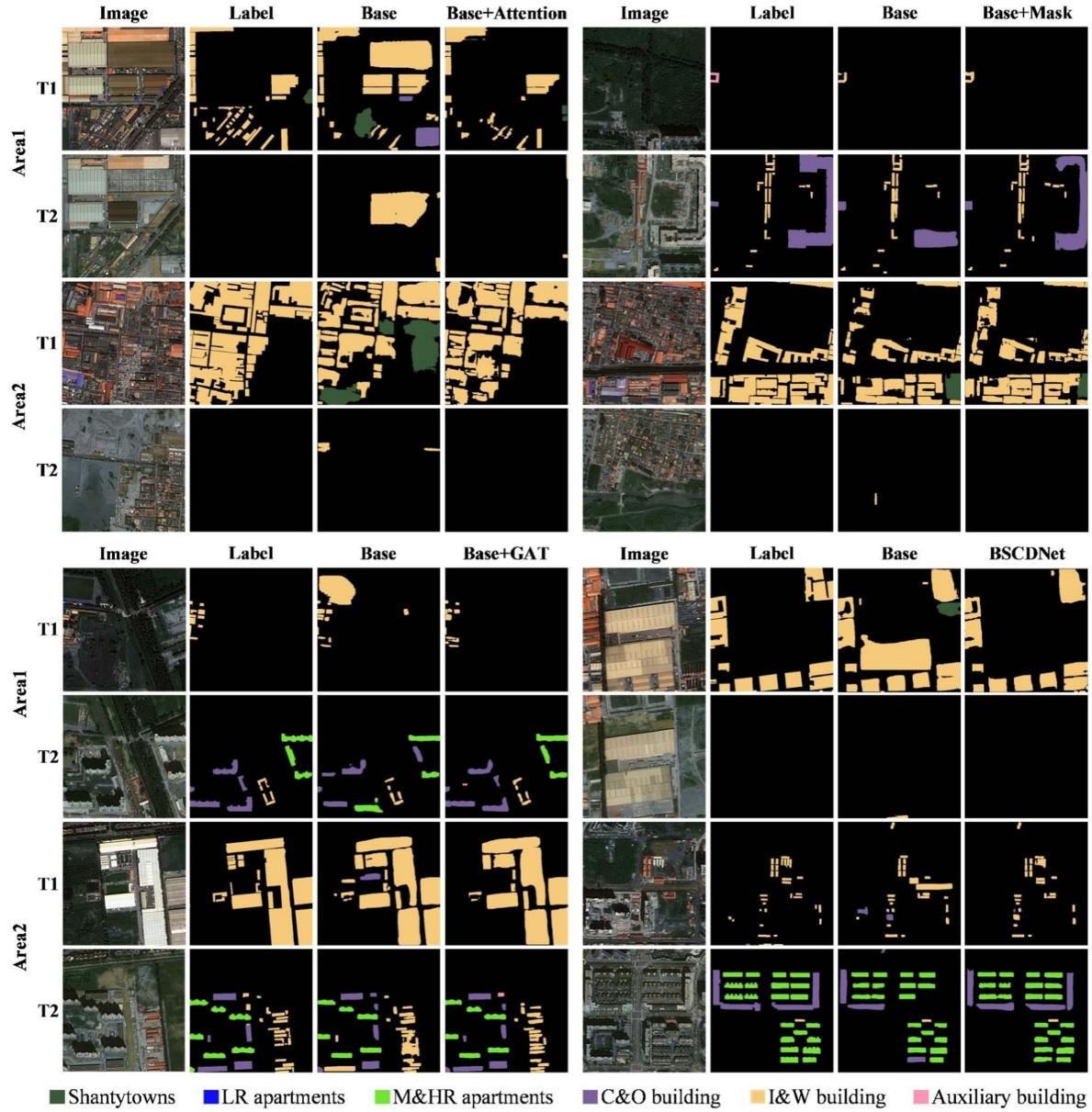

Fig. 9. Typical change detection results of different modules of the ablation experiment.

TABLE IV
QUANTITATIVE RESULTS OF BCDNET AND THE TWO COMPARISON METHODS ON THE BSCDD_BJ DATASET

Method	Precision	Recall	F1-score	mIoU	Kappa	Score
SCDNet	***89.50***	65.68	75.03	75.61	65.08	68.24
MR_CD	86.69	**77.39**	**81.77**	**80.57**	**73.61**	**75.70**
BSCDNet	**87.81**	***81.68***	***84.63***	***83.87***	***75.08***	***77.72***

The best methods in each accuracy metric column are marked in bold .

deep supervision strategy for image classification. Moreover, MR_CD is a typical object-based BSCD method that performs building classification via instance segmentation.

The qualitative and quantitative analyses of the building classification results are shown in Fig. 10 and Table IV. The recall and kappa coefficient of SCDNet were 65.68% and 65.08%, respectively, which were significantly lower than those of the other methods and were caused by missed detections. BSCDNet's mIoU and kappa were 3.3% and 1.47% higher than those of MR_CD, respectively, indicating that BSCDNet performed better than the latter in building classification. Fig. 10 shows three representative building extraction results obtained

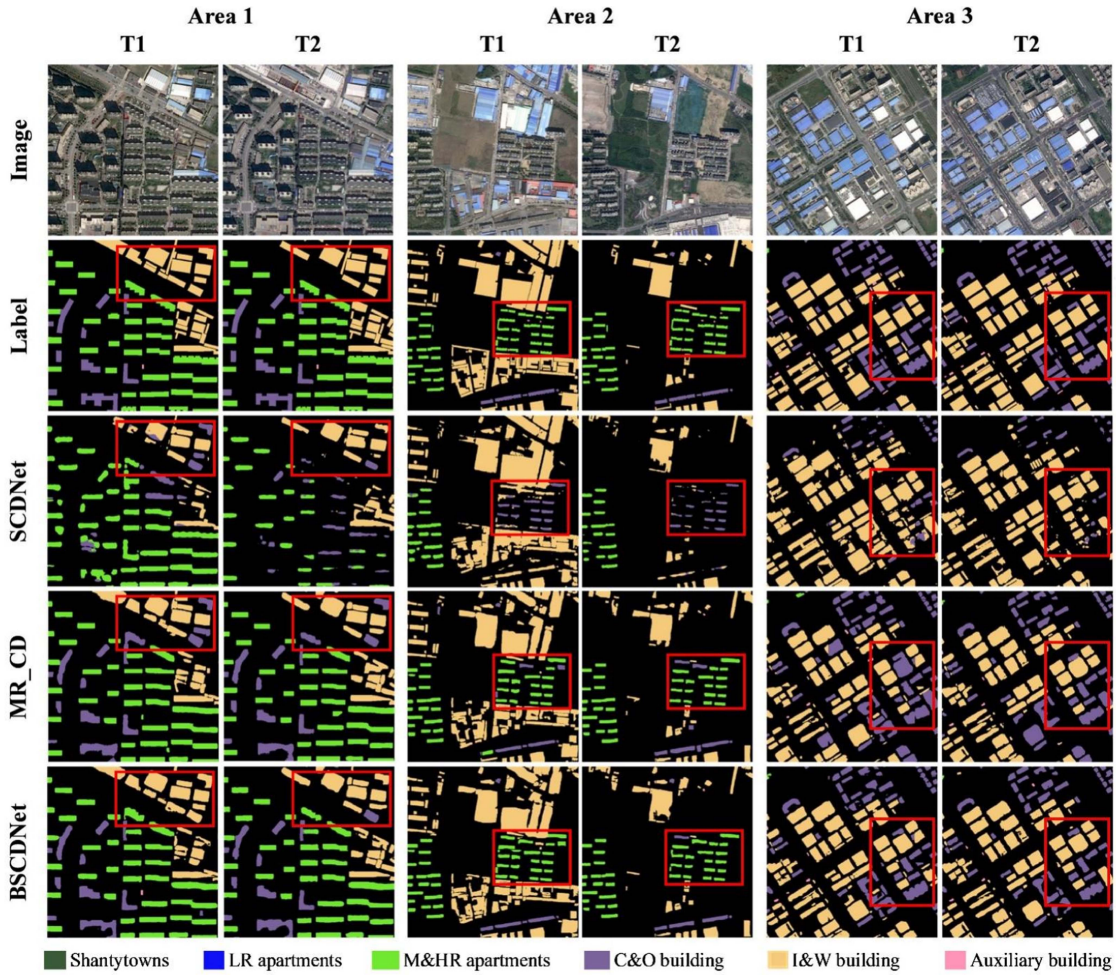

Fig. 10. Typical building semantic segmentation results of BSCDNet and comparative methods on the BSCDD_BJ dataset.

on BSCDD_BJ. Despite improvements in attention modules and deep supervision, the semantic segmentation branch of SCDNet still produced fragmented buildings, as shown in the buildings indicated with red boxes in area 2. MR_CD performs image classification without considering contextual information among buildings; this method occasionally classifies partial I&W and M&HR buildings as C&O buildings, as illustrated by the red boxes in areas 1 and 2. On the other hand, BSCDNet provides more accurate classification results than do the other two methods for clustered buildings, as illustrated in area 1. These results verify the superiority of the GAT-based semantic extraction mechanism of BSCDNet, as aggregation and adjacency relationships specific to buildings are utilized in classification.

V. Discussion

A. Further Analysis of Pixel-Level and Object-Level Change Detection Methods

The results of the comparative experiments show that the accuracies of the object-based building SCD methods, namely MR_CD and BSCDNet, are higher than those of the pixel-based SCD methods, namely Bi-SRNet and MTSCD. HRSCD.str1 is a postclassification change detection method that extracts bitemporal buildings through semantic segmentation and subsequently determines changes in pixels. The change detection performance depends on the semantic segmentation accuracy. Thus, classification errors and small bitemporal image positional deviations might cause false alarms. The pixel-level multitask SCD methods, HRSCD.str4, SCDNet, Bi-SRNet, and MTSCD, consist of change detection and semantic segmentation branches, and the latter provides semantic information for the former. Both the classification and change detection branches are pixel based, and errors in each branch easily lead to fragmented and inaccurate classifications. Fig. 11 shows the results for different branches of SCDNet and BSCDNet based on BSCDD_BJ. BSCDNet, as an object-based change detection method, results in more accurately extracted changed buildings and semantic classes than pixel-based methods.

B. Model Efficiency Analysis

The model parameters (params) and floating-point operations per second (FLOPs) were selected to evaluate the model efficiency. The params and FLOPs values of the different methods are shown in Table V. Among the methods, HRSCD. str1

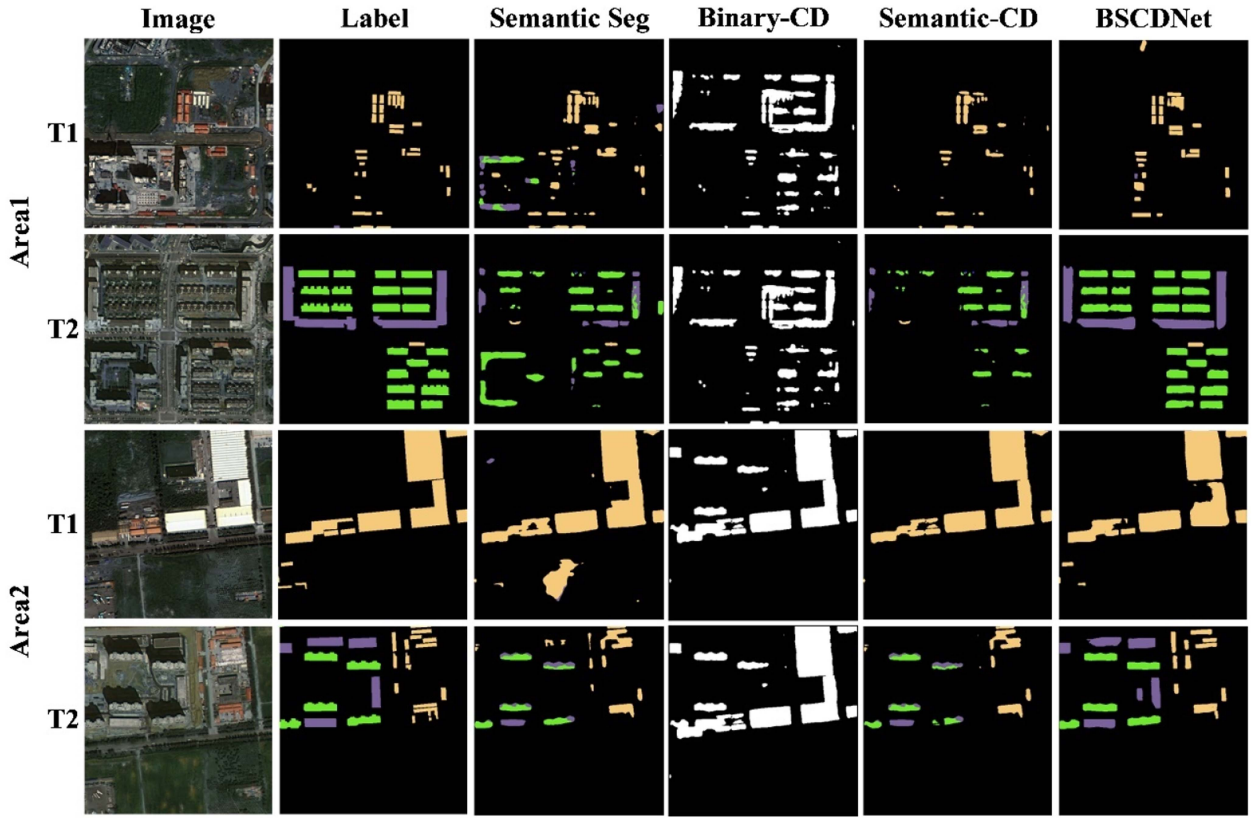

Fig. 11. Result comparison of pixel-based and object-based BSCD methods.

TABLE V
COMPARISON OF COMPUTATIONAL COMPLEXITY

Method	Params(MB)	FLOPs(G)
HRSCD.str1	59.45	471.95
HRSCD.str4	69.19	139.07
SCDNet	27.82	1088.27
BiSRNet	23.38	1522.37
MTSCD	94.55	1162.75
SOLOCD	95.44	476.82
MR_CD	68.97	825.25
BSCDNet	78.53	1111.04

and HRSCD. str4 yielded the lowest params values and computational complexity. BSCDNet is at a medium level in terms of parameters and FLOPs. Compared with SCDNet, MTSCD, and SOLO_CD, BSCDNet has fewer params. In terms of FLOPs, BSCDNet outperforms BiSRNet and MTSCD and is on the same level as SCDNet. BSCDNet has higher accuracy in change detection on the BSCDD_BJ and BSCDD_YZ datasets than does the other methods. This demonstrates that BSCDNet achieves a balance between efficiency and accuracy. The combination of low computational complexity and high accuracy validates the ability of BSCDNet to achieve the highest performance while maintaining reasonable computational utility.

VI. CONCLUSION

In this study, we propose a novel BSCD method, BSCDNet, that considers both morphological and semantic changes in buildings. As a multitask method, BSCDNet conducts object classification, change detection, and segmentation simultaneously and utilizes a GAT by adopting an attention mechanism to capture the spatial and semantic correlations among the building objects during classification. In future work, we will expand the BSCDD into a more generalized dataset by including more regions and building types for developing and testing BSCD methods. Moreover, we will adopt weakly supervised change detection technology to improve the performance of BSCDNet with unlabeled samples.

REFERENCES

[1] T. Bai et al., "Deep learning for change detection in remote sensing: A review," *Geo-Spatial Inf. Sci.*, vol. 26, no. 3, pp. 262–288, 2023.
[2] Z. Du, J. Yang, C. Ou, and T. Zhang, "Agricultural land abandonment and retirement mapping in the Northern China crop-pasture band using temporal consistency check and trajectory-based change detection approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 4406712.
[3] J. Ge, H. Tang, N. Yang, and Y. Hu, "Rapid identification of damaged buildings using incremental learning with transferred data from historical natural disaster cases," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 105–128, 2023.
[4] E. Hamidi, B. G. Peter, D. F. Muñoz, H. Moftakhari, and H. Moradkhani, "Fast flood extent monitoring with SAR change detection using Google Earth Engine," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 4201419.

[5] W. Li, P. Ma, H. Wang, and C. Fang, "SAR-TSCC: A novel approach for long time series SAR image change detection and pattern analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5203016.

[6] H. Yoon and S. Kim, "Detecting abandoned farmland using harmonic analysis and machine learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 201–212, 2020.

[7] K. Yang et al., "Asymmetric Siamese networks for semantic change detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5609818.

[8] P. Yuan, Q. Zhao, X. Zhao, X. Wang, X. Long, and Y. Zheng, "A transformer-based Siamese network and an open optical dataset for semantic change detection of remote sensing images," *Int. J. Digit. Earth*, vol. 15, no. 1, pp. 1506–1525, 2022.

[9] Z. Chen et al., "EGDE-Net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement," *ISPRS J. Photogrammetry Remote Sens.*, vol. 191, pp. 203–222, 2022.

[10] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 4410213.

[11] Q. Shu, J. Pan, Z. Zhang, and M. Wang, "MTCNet: Multitask consistency network with single temporal supervision for semi-supervised building change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 115, 2022, Art. no. 103110.

[12] J. Yuan, L. Wang, and S. Cheng, "STransUNet: A Siamese TransUNet-based remote sensing image change detection network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9241–9253, Nov. 2022.

[13] H. Zheng et al., "HFA-Net: High frequency attention Siamese network for building change detection in VHR remote sensing images," *Pattern Recognit.*, vol. 129, 2022, Art. no. 108717.

[14] F. Cui and J. Jiang, "MTSCD-Net: A network based on multi-task learning for semantic change detection of bitemporal remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 118, 2023, Art. no. 103294.

[15] 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.

[16] S. Tian, Y. Zhong, Z. Zheng, A. Ma, X. Tan, and L. Zhang, "Large-scale deep learning based binary and semantic change detection in ultra high resolution remote sensing imagery: From benchmark datasets to urban application," *ISPRS J. Photogrammetry Remote Sens.*, vol. 193, pp. 164–186, 2022.

[17] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, Feb. 2017.

[18] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 1025–1035.

[19] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, Feb. 2018.

[20] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7651–7668, Sep. 2021.

[21] J. Pyo, K.-J. Han, Y. Cho, D. Kim, and D. Jin, "Generalization of U-Net semantic segmentation for forest change detection in South Korea using airborne imagery," *Forests*, vol. 13, no. 12, 2022, Art. no. 2170.

[22] R. Su and R. Chen, "Land cover change detection via semantic segmentation," 2019, *arXiv:1911.12903*.

[23] H. Xia, Y. Tian, L. Zhang, and S. Li, "A deep Siamese postclassification fusion network for semantic change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5622716.

[24] Q. Zhu et al., "Land-use/land-cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 63–78, 2022.

[25] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[26] D. Wang, F. Zhao, C. Wang, H. Wang, F. Zheng, and X. Chen, "Y-Net: A multiclass change detection network for bi-temporal remote sensing images," *Int. J. Remote Sens.*, vol. 43, no. 2, pp. 565–592, 2022.

[27] F. Pacifici, F. Del Frate, C. Solimini, and W. J. Emery, "An innovative neural-net method to detect temporal changes in high-resolution optical satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 9, pp. 2940–2952, Sep. 2007.

[28] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, and P. He, "SCDNET: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, 2021, Art. no. 102465.

[29] D. Wang, X. Chen, N. Guo, H. Yi, and Y. Li, "STCD: Efficient Siamese transformers-based change detection method for remote sensing images," *Geo-Spatial Inf. Sci.*, pp. 1–20, 2023.

[30] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5620014.

[31] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 183, pp. 228–239, 2022.

[32] S. Tian, X. Tan, A. Ma, Z. Zheng, L. Zhang, and Y. Zhong, "Temporal-agnostic change region proposal for semantic change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 204, pp. 306–320, 2023.

[33] M. Zhao et al., "Spatially and semantically enhanced Siamese network for semantic change detection in high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2563–2573, Apr. 2022.

[34] Y. Deng et al., "Feature-guided multitask change detection network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9667–9679, Dec. 2022.

[35] A. Eftekhari, F. Samadzadegan, and F. D. Javan, "Building change detection using the parallel spatial-channel attention block and edge-guided deep network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 117, 2023, Art. no. 103180.

[36] W. Wang, X. Tan, P. Zhang, and X. Wang, "A CBAM based multiscale transformer fusion approach for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6817–6825, Sep. 2022.

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

[38] N. Shi, K. Chen, and G. Zhou, "A divided spatial and temporal context network for remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4897–4908, Jul. 2022.

[39] Q. Ding, Z. Shao, X. Huang, and O. Altan, "DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, 2021, Art. no. 102591.

[40] X. Wang et al., "A high-resolution feature difference attention network for the application of building change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102950.

[41] P. Zhu, H. Xu, and X. Luo, "MDAFormer: Multi-level difference aggregation transformer for change detection of VHR optical imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 118, 2023, Art. no. 103256.

[42] X. Song, Z. Hua, and J. Li, "Remote sensing image change detection transformer network based on dual-feature mixed attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5920416.

[43] Y. Dai, K. Zhao, L. Shen, S. Liu, X. Yan, and Z. Li, "A Siamese network combining multi-scale joint supervision and improved consistency regularization for weakly supervised building change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4963–4982, May 2023.

[44] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5622519.

[45] L. Yan and J. Jiang, "A hybrid Siamese network with spatiotemporal enhancement and two-level feature fusion for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 4403217.

[46] Y. Wu et al., "CSTSUNet: A cross swin transformer based Siamese u-shape network for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Dec. 2023, Art. no. 5623715.

[47] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5610111.
[48] Z. Li, C. Tang, L. Wang, and A. Y. Zomaya, "Remote sensing change detection via temporal feature interaction and guided refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5628711.
[49] Q. Shen, J. Huang, M. Wang, S. Tao, R. Yang, and X. Zhang, "Semantic feature-constrained multitask Siamese network for building change detection in high-spatial-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 189, pp. 78–94, 2022.
[50] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 101–119, 2022.
[51] Y. Sun, X. Zhang, J. Huang, H. Wang, and Q. Xin, "Fine-grained building change detection from very high-spatial-resolution remote sensing images based on deep multitask learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 8000605.
[52] 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.
[53] S. Tian, A. Ma, Z. Zheng, and Y. Zhong, "Hi-UCD: A large-scale dataset for urban semantic change detection in remote sensing imagery," *Neural Inf. Process. Syst.*, Dec. 2022.
[54] 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.
[55] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.
[56] 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, Jun. 2022, Art. no. 5604816.
[57] 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. 2019.
[58] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[59] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
[60] 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.
[61] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[62] X. Shen et al., "DCT-mask: Discrete cosine transform mask representation for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8716–8725.
[63] G. Ghiasi et al., "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2917–2927.
[64] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and fast instance segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 17721–17732.

Qian Shen is currently working toward the Ph.D. degree majoring in cartography and geographic information systems with the School of Geographical Sciences, Nanjing Normal University, Nanjing, China.

His research interests include remote sensing image processing and computer vision, deep learning, and their applications in remote sensing.

Shikang Tao received the B.S. degree in geographic information science in 2020 from the School of Geography, Nanjing Normal University, Nanjing, China, where he is currently working toward the Ph.D. degree.

His research interests include remote sensing and computer vision, deep learning, and their applications in remote sensing.

Rui Yang received the B.S. degree in computer science and technology from Changshu Institute of Technology, Suzhou, China, in 2018 and the M.S. degree in computer science and technology from Nanjing Normal University, Nanjing, China, in 2021 where he is currently working toward the Ph.D. degree with the School of Geographical Sciences.

His research interests include image processing, behavior recognition, and object detection.

Xin Zhang received the Ph.D. degree in cartography and geography information system from the Institute of Geographic Sciences and Natural Resources Research, Beijing, China, in 2004.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include service of remote sensing information and digital ocean system.

Min Wang received the Ph.D. degree in cartography and geographic information system from the Chinese Academy of Sciences, Beijing, China, in 2003.

He is currently a Professor with Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education Nanjing, China. His research interests include remote-sensing image processing, segmentation, feature extraction, classification, and remote-sensing image mining.

Dr. Wang has authored more than 70 peer-reviewed scientific manuscripts in journals including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and *Photogrammetric Engineering and Remote Sensing*.