Semantic Information Collaboration Network for Semantic Change Detection in Remote Sensing Images

Xiaogang Ning, You He, Hanchao Zhang, Ruiqian Zhang, *Member, IEEE*, Dong Chang, and Minghui Hao

***Abstract*—Semantic change detection (SCD) extends the traditional change detection (CD) task to simultaneously identify the change areas and their corresponding land cover categories in bi-temporal images. This "from-to" change information holds significant value in numerous practical applications and is increasingly garnering attention in the remote sensing domain. However, prevalent challenges such as loss of detail and class imbalance significantly hinder the efficacy of SCD applications. To address this challenge, we propose a novel semantic information collaboration network (SIC-Net). This network incorporates a detail capture path and a spatial-temporal semantic coordination module aiming to effectively execute SCD by fusing detailed information with contextual features and harnessing synergies between binary CD and semantic segmentation. Additionally, we introduce a pseudo-label growth algorithm to mitigate the substantial loss of sample category information. The experimental results on two widely used SCD datasets demonstrate that SIC-Net outperforms other methods across various evaluation metrics, achieving SeK of 23.96% and 61.29%, respectively. These findings not only validate the effectiveness of the SIC-Net but also provide new ideas and directions for future research in the field of remote sensing, particularly in SCD.**

***Index Terms*—Change detection, pseudo-label growth algorithm (PLGA), semantic change detection (SCD), semantic coordination, semantic segmentation, spatial detail.**

I. Introduction

CHANGE detection (CD) is the process of identifying differences in the state of an object or phenomenon by observing it at different times [1]. Precise identification of alterations in Earth's surface attributes is essential for comprehending the interplay between humans and natural phenomena [2]. Hence, this task has consistently captivated the interest of the remote sensing community over an extended period. Remote sensing images CD plays an important role in updating geographic data, evaluating disasters, predicting disaster development trends, land use monitoring, and other tasks.

According to the type of semantic label information desired in the output change map, CD tasks can be divided into two categories: binary CD and semantic change detection (SCD) [3], [4]. Binary CD can only tell us where changes have taken place. However, in practical applications, we are interested not only in the location of changes but also in the type of changes. In order to address this deficiency, SCD methods have emerged [5], [6]. SCD needs to further identify the change category based on the changed regions to provide detailed "from-to" change information in practical application [7], [8]. Therefore, SCD provides detailed and information-rich perspective for monitoring land cover changes in the remote sensing context.

CD methods can be roughly divided into two categories: traditional and artificial intelligence based [9]. With the increasing application of automatic CD in remote sensing images, the limitations of traditional methods are becoming more obvious. The increase in spatial resolution of remote sensing images has resulted in richer image details. However, traditional methods face limitations in feature extraction, leading to a decrease in the accuracy of CD. Moreover, these methods are also prone to the influence of factors such as seasonal changes and lighting conditions [9]. Recently, deep learning-based (DL-based) methods have aroused the interest of many researchers due to the explosive development of artificial intelligence technology. Due to their excellent nonlinear feature extraction and learning ability, DL-based methods can better understand complex scenes, and their performance is far superior to traditional methods [10], [11]. For DL-based SCD algorithms, leveraging large volume remote sensing data, especially high-resolution data, to detect finer changes [12], has become an urgent problem that needs to be addressed.

SCD is an intricate task, which includes two subtasks: 1) semantic segmentation, and 2) binary CD. In the existing research, some approaches have tackled semantic segmentation and binary CD separately using independent branches [13], while others have relied on post-classification CD methods [14]. However, the inherent interdependency between these two subtasks has not been fully exploited. Specifically, the information gained from the land cover semantic segmentation tasks can potentially enhance the accuracy of CD [5]. Recognizing this, recent studies have proposed the use of Siamese networks to

Manuscript received 21 March 2024; revised 13 May 2024 and 5 June 2024; accepted 20 June 2024. Date of publication 27 June 2024; date of current version 24 July 2024. This work was supported by the National Key Research and Development Program of China under Grant 2023YFB3907602. *(Corresponding author: You He.)*

Xiaogang Ning, You He, Hanchao Zhang, Ruiqian Zhang, and Minghui Hao are with the Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping, Beijing 100036, China (e-mail: ningxg@casm.ac.cn; heyou19990208@163.com; zhanghc@casm.ac.cn; zhangrq@casm.ac.cn; haomh@casm.ac.cn).

Dong Chang is with the College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China (e-mail: dongchang@sdust.edu.cn).

The source code is available at https://github.com/hy-pn/SICNet.

Digital Object Identifier 10.1109/JSTARS.2024.3418632

extract image features and feed them into different task heads in order to improve the accuracy of SCD [15], [16], [17]. However, these networks have yet to fully exploit the inherent correlations between subtasks, leading to inconsistent detection results.

Despite some progress, existing research on SCD methods still faces the following challenges:

1) *Impact of Missing Details:* Existing methods may result in false positives and negatives due to the lack of detailed spatial information.
2) *Contradictory Results from Different Subtasks:* The relationship between binary CD and semantic segmentation often leads to conflicting outcomes, where detected change areas do not align with segmentation results.
3) *Class Imbalance Challenge:* The scarcity of positive samples in datasets containing both changed and unchanged regions hampers the accuracy of semantic segmentation techniques.

To address the deficiencies in SCD research in the field of remote sensing, this article proposes a semantic information collaboration network (SIC-Net), that effectively integrates detailed information and maximizes the synergistic relationship between the two subtasks. Moreover, we exploit a pseudo-label growth algorithm (PLGA) to increase the number of annotation pixels, alleviating class imbalance issues and thereby improving the accuracy of SCD. The major contributions in this article are as follows.

1) We propose a novel SIC-Net to improve SCD performance of remote sensing images. The network incorporates a dual-branch backbone (DBB), seamlessly integrating spatial details with contextual information, achieving adaptive alignment and significantly enhancing semantic awareness. The experimental results indicate that our SIC-Net achieves state-of-the-art performance on benchmark SCD datasets.
2) We design a spatial-temporal semantic coordination module (STSCM) in SIC-Net. The STSCM employs an attention mechanism to facilitate information exchange between the two subtasks, thereby harnessing the collaborative potential between them and further enhancing the network's robustness.
3) We develop a PLGA, which generates high-quality pseudo-labels based on semantic segmentation network to alleviate the issue of class imbalance between positive and negative samples.

The rest of this article is organized as follows. Section II reviews the related work. Section III introduces our methodology in detail. Section IV describes the datasets and the experimental settings. Section V presents the results of our experiments. The details extraction strategy and model efficiency are discussed in Section VI. Finally, we conclude our work in Section VII.

II. Related Work

A. Binary Change Detection

In recent years, binary CD techniques has gradually become an important means of acquiring dynamic land cover change information in the field of remote sensing. Traditional methods can be divided into visual analysis, algebra-based methods, transformation-based methods, classification-based methods, advanced models, and other hybrid approaches [18]. Algebra-based methods comprise image differing methods, image regression method [19], [20], image rationing method [21], and change vector analysis [22]. Transformation-based methods involve principal component analysis [23], Tasseled Cap [24], Gramm–Schmidt, and others. Classification-based methods primarily employ post-classification comparison techniques to identify changes.

However, the emergence of deep learning-based CD networks has captured significant attention. These networks exhibit robust feature learning capabilities and flexible model architectures, greatly enhancing the performance of binary CD [25], [26], [27]. In recent years, many binary CD networks based on deep learning have been proposed. Deep learning binary CD architectures can be roughly divided into single-stream networks and double-stream networks.

Single-stream networks typically refer to semantic segmentation networks. They utilize various data fusion techniques to integrate multiple temporal cycles of remote sensing images, generating intermediate data for input [5], [28], [29], [30].

Double-stream networks are commonly composed of two weight-sharing feature extraction streams that directly take bi-temporal images as the input [31]. In recent years, researchers have proposed various innovative network architectures and methods in the field of CD. Daudt et al. [32] introduced a fully convolutional network that includes a part for extracting features from dual-temporal remote sensing images using Siamese networks. Zhang et al. [33] designed a super-pixel sampling network for feature extraction and super-pixel segmentation in dual-temporal images. Additionally, there are methods like the super-resolution-based change detection network method [34] and a local–global pyramid network for building change detection [35]. The dual-task constrained deep Siamese convolutional network [36] and the semantic feature-constrained change detection network [37] both utilize two Siamese networks to constrain the binary CD network. Other innovative approaches include the SNUNet-CD network [38], dual-attention fully convolutional Siamese networks with weighted bilateral edge contrast loss [39], methods combining transformer [40], [41], [42], [43], [44], [45], and the SAM-CD method [46] based on the segment anything model [47]. While these methods have shown gradual improvements in tasks such as identifying change areas, addressing pseudo changes, small target CD, and change area boundary recognition, traditional binary CD methods often provide insufficient information. Using these methods to accurately identify change types still requires further improvement in practical applications.

B. Semantic Change Detection

SCD stands as an informative pixel-level CD method that concurrently identifies change regions in dual-temporal images and their corresponding land cover categories. It offers semantic information overlooked by binary CD methods and finds widespread applications in diverse domains, including urban planning, farmland conversion, and disaster monitoring.

In the realm of early SCD methods, several approaches have been explored. These encompass the most intuitive post-classification change detection [7], [48], as well as direct classification methods, among others. Due to the application and advancement of deep learning in remote sensing research, tasks related to SCD based on deep learning have been gradually investigated.

Existing deep learning methods for SCD can be categorized based on the number of encoders into three types: single-encoder methods, Siamese-encoder methods, and triple-encoder methods.

Single-encoder methods: These methods treat SCD as a multiclass semantic segmentation task. In such approaches, a single encoder is employed to extract features from fused imagery of two different periods, and the network is trained to classify semantic changes. For instance, HRSCD str2 [5] is a notable example.

Siamese-encoder methods: Siamese-encoder methods encompass three variations. The first variation involves Siamese encoders combined with a single decoder, treating SCD as a classification-after-change task. The second variation utilizes Siamese encoders with dual decoders. For example, Peng et al. [50], based on the Siamese U-Net network, introduced an SCD method. They further incorporated metric learning and deep supervision strategies to enhance the network's performance. Xia et al. [14] proposed a deep Siamese classification-after-fusion network.

The third and widely adopted variation employs Siamese encoders with three decoders. This architecture is designed to better capture the spatial and temporal features in the data, making it suitable for complex SCD tasks. This work highlighted the synergy between binary CD and semantic segmentation tasks, showcasing the potential of multitask deep learning in SCD. Therefore, most of the existing network structures for SCD use Siamese networks for feature extraction and then employ different heads to handle subtasks, addressing the issue of temporal correlation neglect in previous methods [5]. For instance, Zhao et al.[4] proposed a spatially and semantically enhanced Siamese network, which aggregates the rich spatial and semantic information in the remote sensing images through a designed spatial and semantic feature aggregation module. Zhu et al. [49] proposed the Siamese global learning framework, which alleviates the issue of class imbalance by improving the sample sampling mechanism. Zheng et al. [16] proposed a multitask architecture named ChangeMask, which decouples the SCD into a temporal-wise semantic segmentation and a binary CD, and designs a temporal-symmetric transformer to guarantee temporal symmetry. Chen et al. [51] proposed a feature constraint change detection network and proved that bi-temporal semantic segmentation branches can improve the precision of CD task. Ding et al. [15] compared several base architectures for SCD and proposed a bi-temporal semantic reasoning network (Bi-SRNet) on this basis. [52] introduces MTSCD-Net, which simultaneously extracts multiscale features from dual-temporal data and leverages the spatial attention weight map from the binary CD subtask to enhance the semantic segmentation subtask. Jiang et al. [53] developed a temporal-transform network, which captures temporal changes across dimensions through a designed temporal-transform module. Chen et al. [54] proposed MambaSCD based on the Mamba structure, comprehensively learning global spatial context information from input images to achieve spatiotemporal interaction of multitemporal features.

Triple-encoder methods: Triple-encoder methods involve networks with three independent encoders paired with separate decoders. These methods separate the processes of semantic segmentation and binary CD, offering greater flexibility in modeling and training. They overlook the temporal correlation between the two time-period images and the intrinsic correlation between the two subtasks. Representative architectures of this kind include HRSCD-sr.3 and HRSCD-sr.4 as described in [5]. Building upon this foundation, Ding et al. [55] introduced SCanNet, where comprehensive learning of spatiotemporal dependencies is conducted at both encoding and decoding stages. Wang et al. [56] proposed DESNet, which effectively improves the robustness to large-scale changes and the integrity of change objects. Chang et al. [57] proposed JFRNet, which transforms independent learning of multiple tasks into joint refinement of dual temporal features. These approaches significantly improved the performance of the triple-encoder methods.

However, there are still problems in the current research of SCD. On one hand, the existing SCD models pay more attention to the extraction of semantic information and ignore the importance of spatial details. On the other hand, the number of positive samples of land cover category labels in the existing SCD dataset is too small, resulting in poor semantic segmentation accuracy, which is also a common problem in SCD datasets. To alleviate the above issues, this article proposes a SIC-Net method for SCD tasks.

III. Methodology

In this section, we offer a detailed description of the proposed SIC-Net. The overall architecture is illustrated in Fig. 1(a), consisting of two shared-weight dual-branch backbones and a spatial–temporal semantic coordination module. The integration of two shared-weight DBB provide the network with robust feature extraction capabilities. These backbones, specifically designed with the detail capture path (DCP) and the semantic context path (SCP), collaboratively capture information at different hierarchical levels within the image. To enhance information exchange and fusion between these two paths, we introduce a detail guidance module (DGM). During the decoding phase, SIC-Net incorporates the STSCM to facilitate information exchange between the two subtasks, thereby enabling collaborative training for both tasks. The inputs of CD_head1 and CD_head2 are derived from different nodes in STSCM's change features, achieving the goal of deep supervision. The output of CD_head1 serves as the final binary CD result. Furthermore, before training, we leverage PLGA to predict semantic class labels for unchanged regions, subsequently generating pseudo-labels.

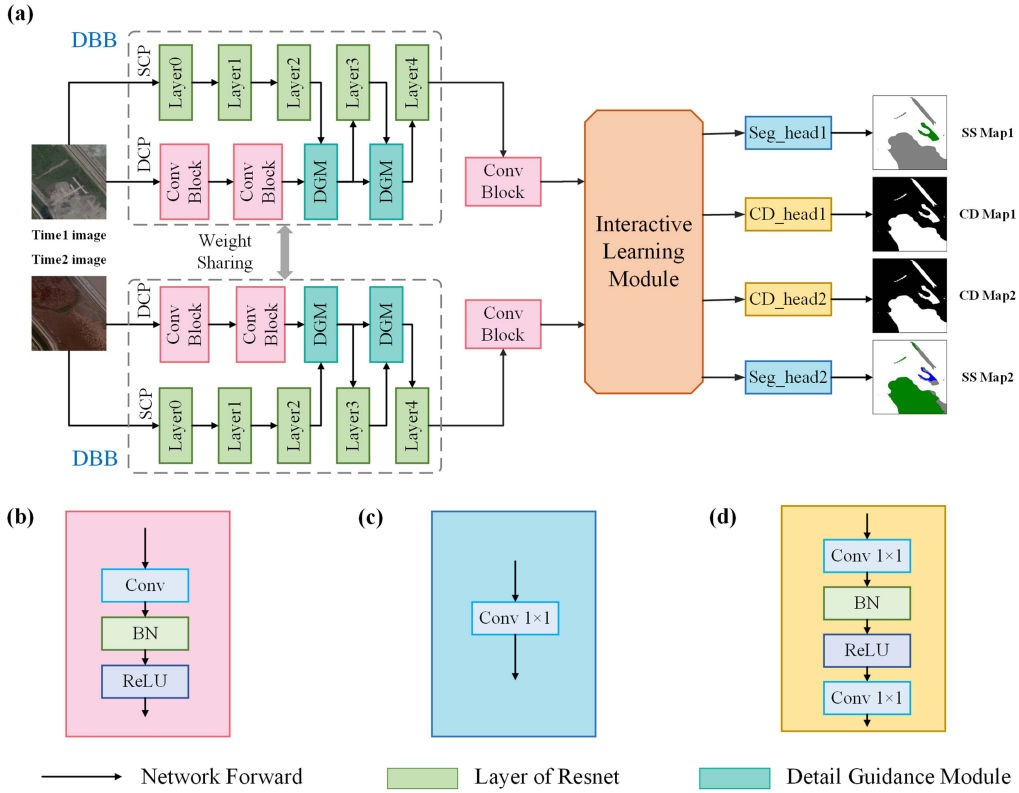


Fig. 1. Architecture of the proposed SIC-Net. (a) Overview of the proposed sic-net. DGM denotes detail guidance module. SCP denotes semantic context path. DCP denotes detail capture path. DBB refers to a dual-branch backbone composed of DCP and SCP. (b), (c), and (d) respectively denote the structures of conv block, seg_head1, and seg_head2, cd_head1, and cd_head2.

A. *Dual-Branch Backbones*

To alleviate the omission and misclassification of small patches caused by the lack of detailed information, it is crucial to balance the requirements for spatial details and a larger receptive field. Both factors are vital for achieving high segmentation accuracy [58], [59], [60], and they can mitigate the challenge of frequent false positives near changing boundaries [61], [62].

Therefore, building upon the Siamese encoder structure, SIC-Net employs DBB for extracting and aligning spatial details with contextual information. In this architecture, ResNet [63] serves as the SCP for extracting deep contextual information. The DCP architecture designed in this article is depicted in Fig. 1(a). It comprises four layers with spatial resolutions of 1/2, 1/4, 1/4, and 1/4 of the input image, respectively. After the initial extraction of fine-grained spatial details by the first two convolutional layers, the input is fed into the DGM, which adaptively integrate detailed information X_1 with contextual features X_2, dynamically adjusting the weighting between these two feature types. The fused feature X is then fed into the corresponding layer of the ResNet, encouraging the network to focus on finegrained details while extracting deep semantic information, thereby reducing the loss of detailed information. Through the DGM, the DCP, and the SCP can mutually influence each other, establishing a stronger collaborative relationship within the network.

The designed DGM is illustrated in Fig. 2. In the channel dimension, the finegrained feature X_1 and the contextual feature X_2 extract global information through adaptive average pooling, followed by concatenation to generate the attention matrix W_c. Then, utilize this weight matrix to adjust the channel-wise feature distribution of X_2. In the spatial dimension, the fine-grained feature X_1 generates two direction-specific attention maps, namely W_h and W_w, guiding the network to concentrate more on learning the target region. Particularly in handling directional features such as edges or textures, enhancing the model's perception of details and its ability to identify specific areas is achieved by guiding the network's attention. X_2 undergoes a 1x1 convolutional layers to adjust its channel dimension to match that of X_1

$$X'_2 = \text{Conv1} \times 1\,(X_2)\,. \tag{1}$$

Subsequently, W_c is applied as channel attention, while W_h and W_w serve as spatial attention features to recover missing finegrained features, ensuring that the network adequately focuses on spatial details and receives accurate guidance from contextual information

$$\bar{X}_2 = X'_2 \ * W_C \tag{2}$$

$$\tilde{X} = \bar{X}_2 \ * W_h + \bar{X}_2 * W_h + X_1. \tag{3}$$

Finally, following the residual structure paradigm, the fused features $\tilde{X}$ are added to the input contextual features, effectively integrating spatial details and enabling the exploration of more comprehensive semantic information.

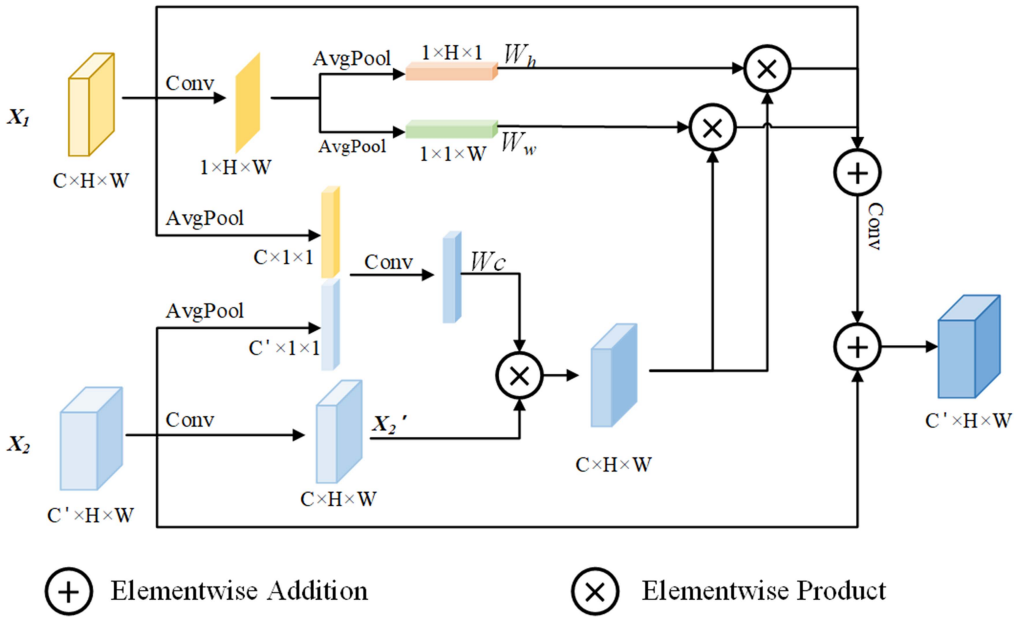


Fig. 2. Structures of the DGM. AvgPool denotes adaptive average pooling.

B. Spatial–Temporal Semantic Coordination Module

To alleviate conflicts between the results of the two subtasks, we introduce an attention-driven STSCM in this article. This module aims to delve into the intrinsic correlations between the two subtasks and leverage this correlation to optimize the overall model performance. By incorporating an attention mechanism, this module can selectively focus on specific regions of interest for the two subtasks, promoting more effective information exchange between them. In this way, the STSCM can fully exploit the potential causal relationships between spatial semantic information and temporal change features, leveraging the collaborative effects of the two subtasks and effectively alleviating conflicts in the results.

As shown in Fig. 3, STSCM concatenates the features f_1 and f_2, and then generates temporal feature f_{cd} through 4 Resblocks (as illustrated in Fig. 4) to accurately model the temporal changes in features. Simultaneously, a convolutional block composed of two residual blocks, a convolutional layer, batch normalization, and ReLU activation, is employed to process f_1 and f_2, resulting in new features f_1' and f_2'. To precisely measure the similarity between f_1' and f_2', we adopt the method of calculating the Euclidean distance, providing a quantitative metric for their similarity

$$W = \mathrm{Sigmoid}\left(\mathrm{dist}\left(f'_1 + f'_2\right)\right) \tag{4}$$

Where dist denotes a function for calculating the L2 distance, providing a specific numerical basis for measuring the similarity between features. The similarity attention matrix W generated by (4) is cleverly introduced into the binary CD branch to dynamically adjust the weights of feature distributions. This approach aims to significantly enhance the binary CD branch's perception of semantic changes. Specifically, W allows the network to weight these features based on the similarity between spatial semantic features at two different time points, thereby more effectively capturing and responding to changes between features. Subsequently, we concatenate the optimized change features obtained using W with the differential features computed from f_1 and f_2, and generate the final change features through convolutional blocks. We simultaneously output the binary CD feature after and before using W, and generate CD Map1 and CD Map2 as shown in Fig. 1, respectively, to achieve deep supervision. This mechanism enables the network to focus more on features that undergo changes at different time points, significantly enhancing the network's sensitivity to temporal information.

Furthermore, after generating the feature f_{cd} in the binary CD branch, spatial attention is employed to provide prior positional information about changing regions. This information is subsequently propagated to the semantic segmentation branch, aiding in further optimizing semantic segmentation features. This design allows the segmentation network to focus more on the changed regions, significantly enhancing its sensitivity to change information in category recognition tasks. Overall, the STSCM strengthens the network's focus on changed regions, fully leveraging the inherent connection between temporal features and spatial semantics, thereby significantly improving the network's performance in modeling spatiotemporal information.

C. Pseudo-Label Growth Algorithm

The existing binary CD datasets lack sufficient descriptions of land cover categories, making them inadequate for SCD requirements. Additionally, in the available SCD datasets, there is a shortage of positive samples for land cover category labels (SECOND [6]: 19.87%, Landsat-SCD [3]: 18.98%, MSSCD [65]: 2.7%), resulting in poor semantic segmentation accuracy. To address this issue, this study proposes a novel method for

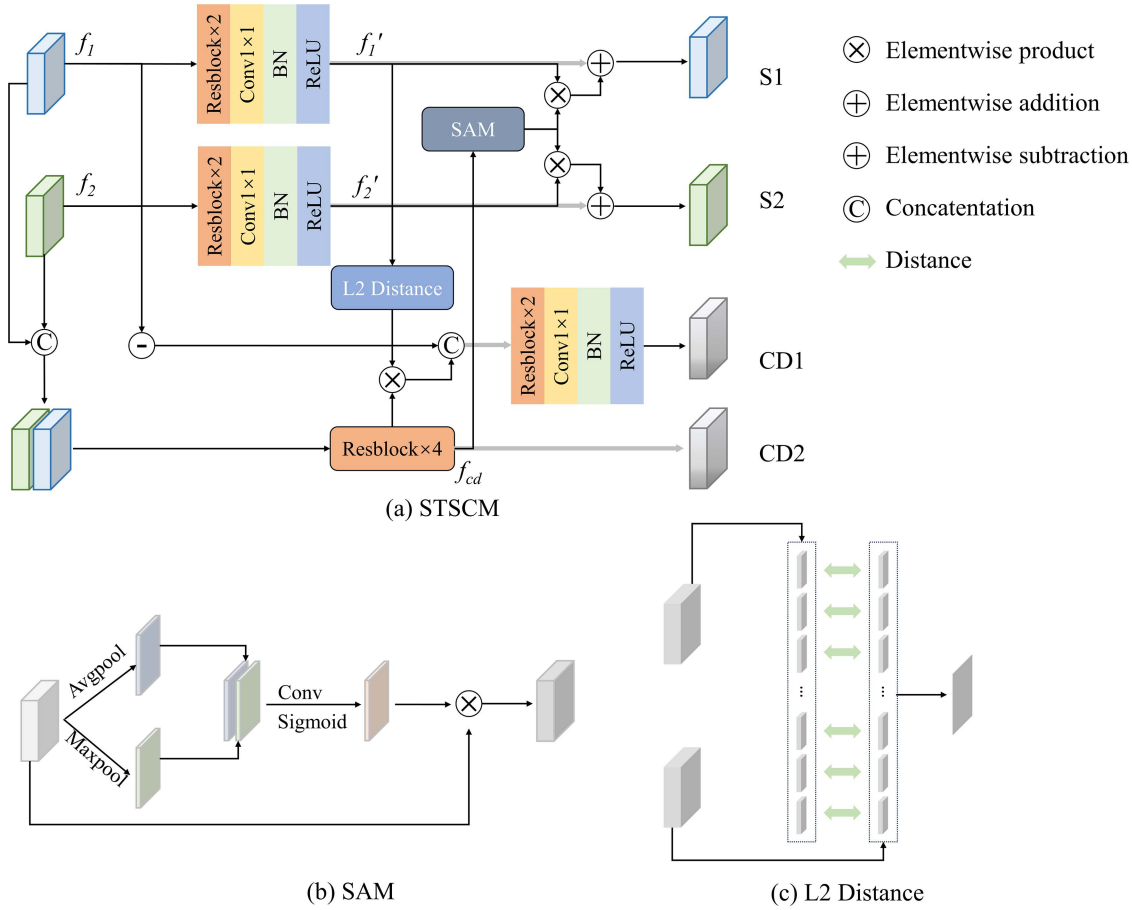


Fig. 3. Structures of the (a) STSCM, (b) SAM, (c) L2 distance. SAM denotes spatial attention modules in [64].

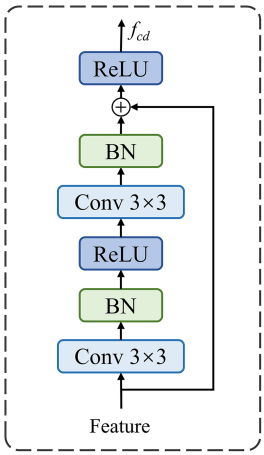


Fig. 4. Structures of the Resblock.

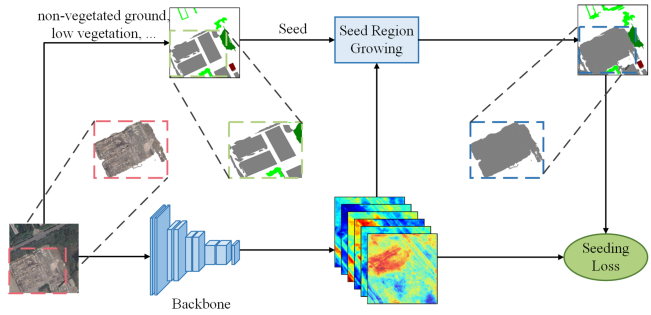


Fig. 5. Pseudo-label growth algorithm.

generating high-quality pseudo-labels, aiming to significantly enhance the performance of SCD.

Drawing inspiration from the article [66] and considering our specific requirements, we introduced a PLGA, as illustrated in Fig. 5. We use a small number of land cover categories from existing labels as seed clues, and employ the predicted probability map from the HRNet [67] semantic segmentation network as the judgment basis. Given the complexity of land cover in remote sensing images, we set fairly strict criteria for seed region growth. Specifically, we set the growth threshold to 0.99. This ensures that only regions with a very high likelihood of being representative of the given land cover category are considered for further expansion. To enhance the reliability of the generated labels, an additional constraint is introduced. This constraint is based on the ratio of the maximum probability to the second maximum probability in the probability distribution map

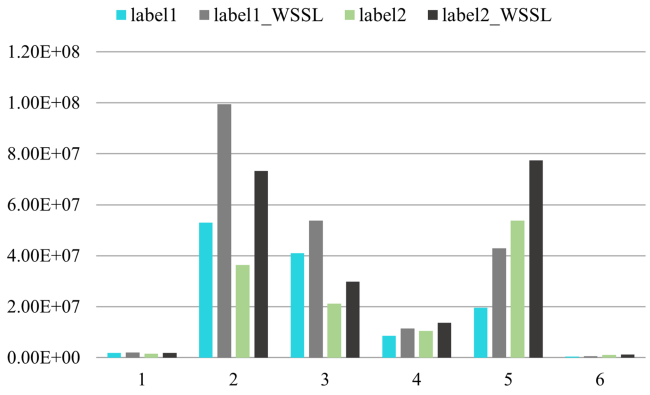


Fig. 6. Comparison of pixel count in training samples of SECOND dataset.

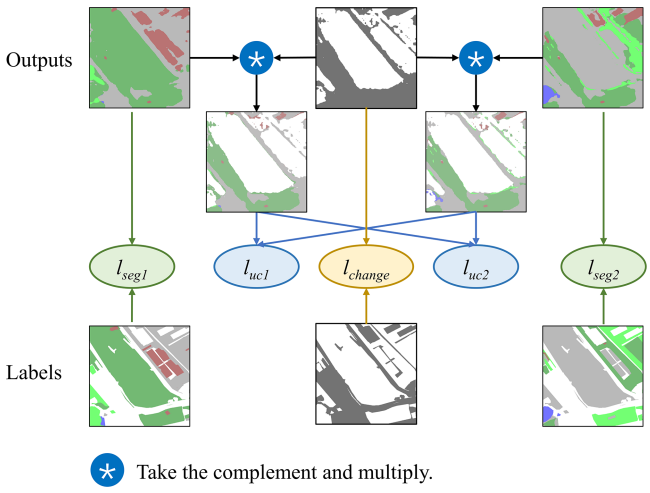


Fig. 7. Calculation of multitask losses.

derived from the prediction results. We set this ratio as another threshold, denoted as threshold2, with a substantial value of 5. The primary objective of this constraint is to prioritize regions where the predicted land cover category is significantly more dominant than alternative possibilities. By doing so, we aim to minimize potential misclassifications and enhance the overall accuracy of the generated labels. During the training process, the computation of the seed loss is derived from the calculation results of the labeled regions' loss.

The pseudocode of the PLGA is shown in Algorithm 1.

The pixel annotation counts for each category in the training samples of the SECOND dataset exhibit varying degrees of improvement before and after PLGA, as illustrated in Fig. 6.

Furthermore, we finetuned the loss function to incorporate consistency constraints when integrating pseudo-labeled samples into the training process. In this refinement, loss calculation occurs only when the land cover types of pseudo-labeled samples (representing unchanged areas) in both temporal images exhibit consistency. This strategic adjustment aims to expand the category annotation pool without compromising accuracy, ultimately preserving the authenticity of the network's performance.

Algorithm 1: PLGA.

Input: Image **I**, seed clues **S**, threshold1 $\mathbf{th}_1$, threshold2 $\mathbf{th}_2$.

Step 1: Predict the probability map **P**, **P** = HRNet(**I**).

Step 2: Set predicted probabilities of types not present in the **S** to 0.

Step 3: Obtain predicted result **R**, **R** = argmax(**P**).

Step 4: Filter the set of pixels **A** that satisfy the conditions based on the predicted probabilities.

$$A_1 = \{(i,j) \,|P\,(i,j)\rangle\, th_1\}$$

$$A_2 = \left\{(i,j) \left| \frac{\max\,(P\,(i,j))}{\max_2\,(P\,(i,j))} \right\rangle th_2 \right\}$$

$$A = A_1 \cap A_2$$

Step 5: Neighborhood growth is an iterative process for each class. For the iteration process of class c:

- Obtain the predicted region R_c and its corresponding seed S_c.
- Identify the categories within the growth range $R_c \times A$, then group adjacent pixels with the same color into the same region to generate connected domains. Afterward, filter out the connected domains **RA** that contain seed clues.
- Iterate over each element μ in R_c :
- If μ belongs to a connected domain in **RA** and is predicted as class c, assign a pseudo label **PL** (μ) = c.

Output: Pseudo label **PL**.

D. Loss Function

In order to ensure that each sub-task of SIC-Net receives proper optimization, as illustrated in Fig. 7, we designed a multitask loss function L as follows:

$$L = l_{\text{change}} + l_{\text{seg}} + l_{\text{uc}} \tag{5}$$

where l_{change} is the binary CD loss, l_{seg} is the semantic segmentation loss, and l_{uc} is unchanged consistency loss based on self-supervised learning strategy in the unchanged area [51].

The l_{change} uses weighted values of the binary cross-entropy function. Its specific form is shown in (6):

$$l_{\text{change}} = -y_c \log\,(p_c) - (1-y_c) \log\,(1-p_c) \tag{6}$$

where y_c and p_c denote the ground truth (GT) label and the predicted probability of change, respectively.

The l_{seg} is the cumulative loss derived from the semantic segmentation results of two periods.

$$l_{\text{seg}} = 0.5 * l_{\text{seg1}} + 0.5 * l_{\text{seg2}} \tag{7}$$

where l_{seg1} and l_{seg2} represent the cross-entropy loss.

l_{change} and l_{seg} are designed to learn binary CD and semantic segmentation of regions with semantic category information in labels, respectively. The regions without category information in the labels haven benefit for model training, so we further employed an unchanged consistency loss l_{uc}. The specific steps to calculate the loss: 1) Identifying unchanged regions. 2) Utilize

the semantic segmentation results one branch is used as the label of the other branch within these unchanged regions. The l_{uc} is the average of the losses computed from two time periods, namely l_{uc1} and l_{uc2}. The calculation for l_{uc1} or l_{uc2} uses the conventional cross-entropy functions

$$l_{\mathrm{uc1}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(p_i\right) \tag{8}$$

$$l_{\mathrm{uc}} = 0.5 * l_{\mathrm{uc1}} + 0.5 * l_{\mathrm{uc2}} \tag{9}$$

where N is the number of semantic classes, while y_i and p_i represent the GT label and the predicted probability of the ith class, respectively.

By employing the aforementioned trio of loss functions, we can directly train two semantic segmentation subtasks and binary CD subtasks. This comprehensive approach ensures a robust learning process for all targeted tasks.

IV. Dataset Description and Experiment Setting

A. Datasets

1) SECOND Dataset: The SECOND dataset is an SCD Dataset [6], collects 4662 pairs of aerial images from several platforms and sensors. Each image has size of 512 × 512, contains RGB channels, and is annotated at the pixel level in changed regions. The spatial resolution varies from 0.5 m to 3 m (per pixel). The changed regions account for 19.87% of the total image. The land cover categories of change regions in the previous and subsequent images are provided, including no-change, nonvegetated ground surface, tree, low vegetation, water, buildings, and playgrounds.

Among the 4662 pairs of temporal images, 2968 ones are openly available. There does not exist a standard splitting for this dataset, so we randomly split the dataset into a training set, and testing set based on the ratio train: test = 4:1.

2) Landsat-SCD Dataset: Landsat-SCD dataset is a recently published well-annotated dataset for SCD [3], the images in Landsat SCD dataset were collected from Landsat-like images captured between the years 1990 and 2020 in Tumushuke (39°39′N – 40°4′N, 78°53′E – 79°19′E), Xinjiang. The dataset, consisting of 8468 image pairs, and each image has size of 416 × 416, contains RGB channels. The spatial resolution is 30m per pixel. The changed regions account for 18.89% of the total image. The Landsat-SCD dataset provides 10 change types, each of which is a separate class representing land-cover transitions. To align with the SECOND dataset, we establish land cover categories of change regions in the previous and subsequent images. These categories encompass: no change, farmland, desert, building, and water. We adhere to the data split proposed by the authors [3] with 6053 pairs allocated for training, 1729 pairs for validation, and 686 pairs for testing, randomly sampled.

B. Evaluation Metrics

In this work, we use three types of evaluation metrics to evaluate the accuracy of the total task of SCD and two subtasks (binary CD and semantic segmentation). These include binary CD metrics: mean intersection over union (mIoU) and $F1$ score, semantic segmentation metrics: Kappa coefficient, and SCD metrics: overall accuracy (OA), Separated Kappa (SeK) coefficient [6] and F_{scd} [15]. We calculate the confusion matrix $Q = \{q_{i,j}\}$ through the prediction results and labels, where $q_{i,j}$ represents the number of pixels that are classified into class i while their GT index is j ($i, j \in \{0, 1, \ldots, N\}$) (unchanged class is set as the class 0).

mIoU and $F1$ are standard metrics commonly used in binary CD, with their calculation formulas as follows:

$$\mathrm{IoU}_1 = q_{00} \Big/ \left(\sum_{i=0}^{N} q_{i0} + \sum_{i=0}^{N} q_{0j} - q_{00}\right) \tag{10}$$

$$\mathrm{IoU}_2 = \sum_{i=1}^{N}\sum_{i=1}^{N} q_{ij} \Big/ \left(\sum_{i=0}^{N}\sum_{j=0}^{N} q_{ij} - q_{00}\right) \tag{11}$$

$$\mathrm{mIoU} = (\mathrm{IoU}_1 + \mathrm{IoU}_2) \Big/ 2 \tag{12}$$

$$R = \sum_{i=1}^{N}\sum_{j=1}^{N} q_{ij} \Big/ \sum_{i=0}^{N}\sum_{j=1}^{N} q_{ij} \tag{13}$$

$$P = \sum_{i=1}^{N}\sum_{j=1}^{N} q_{ij} \Big/ \sum_{i=1}^{N}\sum_{j=0}^{N} q_{ij} \tag{14}$$

$$F1 = \frac{2 * \mathrm{P} * \mathrm{R}}{\mathrm{P} + \mathrm{R}}\,. \tag{15}$$

$Kappa$ is a commonly used metric in semantic segmentation, with its calculation formula as follows:

$$p_0 = \sum_{i=1}^{N} q_{ii} \Big/ \sum_{i=1}^{N}\sum_{j=1}^{N} q_{ij} \tag{16}$$

$$p_e = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} q_{ij} * \sum_{j=1}^{N} q_{ji}\right) \Big/ \left(\sum_{i=1}^{N}\sum_{j=1}^{N} q_{ij}\right)^2 \tag{17}$$

$$\mathrm{Kappa} = \frac{p_0 - p_e}{1 - p_e}\,. \tag{18}$$

The SeK metric is used to evaluate change patterns, but due to label imbalance in SCD, pixels that are truly zero and also predicted as zero are ignored and assigned a zero result in SeK calculation. Its calculation formula is as follows:

$$\rho = \sum_{i=1}^{N} q_{ii} \Big/ \left(\sum_{i=0}^{N}\sum_{j=1}^{N} q_{ij} - q_{11}\right) \tag{19}$$

$$\eta = \left(\sum_{i=1}^{N} \left(\sum_{j=0}^{N} q_{ij} * \sum_{j=0}^{N} q_{ji}\right) + \sum_{j=1}^{N} q_{0j} * \sum_{j=1}^{N} q_{j0}\right) \Big/ \left(\sum_{i=0}^{N}\sum_{j=0}^{N} q_{ij} - q_{00}\right)^2 \tag{20}$$

$$\mathrm{SeK} = e^{\mathrm{IoU}_1 - 1} * (\rho - \eta) \Big/ (1 - \eta)\,. \tag{21}$$

F_{scd} is a metric derived from the F1 score, used to evaluate the accuracy of land cover classification within change areas. Its calculation formula is as follows:

$$P_{\text{scd}} = \sum_{i=1}^{N} q_{ii} \Big/ \sum_{i=1}^{N} \sum_{j=0}^{N} q_{ij} \tag{22}$$

$$R_{\text{scd}} = \sum_{i=1}^{N} q_{ii} \Big/ \sum_{i=0}^{N} \sum_{j=1}^{N} q_{ij} \tag{23}$$

$$F_{\text{scd}} = \frac{2 * P_{\text{scd}} * R_{\text{scd}}}{P_{\text{scd}} + R_{\text{scd}}}. \tag{24}$$

Through the above three types of evaluation metrics, the accuracy of SCD tasks can be comprehensively evaluated.

Finally, to comprehensively evaluate the computational efficiency of the model, we adopted two key metrics: parameter count (Params) and floating-point operations (FLOPs). FLOPs represent the number of floating-point operations required for the model to perform a single forward pass. A higher FLOPs value and a larger Params indicate a more complex model that requires more computational resources for inference. To compute these two metrics, we provide two input images of size $1\times3\times512\times512$ each, taken at two different times.

C. Experimental Settings

The experiments in this paper were run on a desktop workstation equipped with an NVIDIA GeForce RTX 3090 GPU boasting 24 G memory. All programs are implemented based on the PyTorch platform.

During data preprocessing and augmentation for each dataset, we applied normalization to images and employed random flipping and rotating techniques to process both image and label data. For the proposed SIC-Net, consistent experimental parameters are employed on different datasets, including batch size $= 8$, running epochs $= 50$, and initial learning rate $= 0.1$. Additionally, we adopted the stochastic gradient descent method to optimize the weights.

D. Comparative Methods

To comprehensively evaluate the performance of the proposed SIC-Net, we further compare it with several state-of-the-art methods in SCD tasks. The compared methods include the following.

1) The SSCD-l [15]: This network uses two Siamese ResNet to extract semantic information, and then input it into the binary CD head and two semantic segmentation heads.
2) The L-UNet [68]: This method is a UNet-like network, which can simultaneously handle binary CD and semantic segmentation by using fully convolutional long short-term memory networks.
3) The HRNet: This is the champion scheme of the SCD competition hosted by Sense Time in 2020. The network uses Siamese HRNet as the backbone to extract multiscale feature. Then, two semantic segmentation heads and one detection head are used. Its solution open source address.[1]
4) The SCDNet [50]: This method is based on a Siamese U-Net architecture, utilizing an attention mechanism and deep supervision strategy to improve performance.
5) The Bi-SRNet [15]: This method is an improvement based on SSCD-l network, using two types of semantic reasoning blocks to reason both single-temporal and cross-temporal semantic correlations.
6) The MTSCD-Net [52]: This method employs a Siamese encoder based on the Swin transformer along with a feature aggregation module to extract multiscale features. Subsequently, it explores the correlations between subtasks through designed modules.
7) MambaSCD [54]: This method, based on the Mamba architecture, aims to learn global spatial context information from input images and to learn spatiotemporal features through a mechanism for modeling spatiotemporal relationships.
8) SCanNet [55]:This network proposed a semantic change Transformer (SCanFormer) to explicitly model the change information between dual-temporal images, while utilizing dual-time consistency as additional supervision to guide the learning of semantic changes.
9) DEFO-MTLSCD [17]: This method facilitates feature interaction between the binary CD and semantic segmentation subtasks through the design of two modules, thus generating more representative encoding features.

V. Results

To comprehensively evaluate the performance of the proposed SIC-Net, we conducted ablation experiments on the SECOND dataset. This aimed to scrutinize the influence of each component of the model on the overall performance. Additionally, we compared SIC-Net with other state-of-the-art methods on two distinct datasets, thereby validating its superiority.

A. Ablation Study

To investigate the impact of DCP, STSCM, and PLGA in our proposed SIC-Net performance, we conducted a series of ablation experiments. These experiments encompassed the baseline, baseline-PLGA, baseline-PLGA-DCP, baseline-PLGA-STSCM, and SIC-Net. These experiments aimed to delve into the functionalities of these components and their contributions to overall performance. The quantitative results are shown in Table I.

First, we test the effectiveness of the PLGA by adding it to train baseline. The PLGA significantly improves the performance of baseline and increases the accuracy by around 0.78% in SeK, 0.97% in F_{scd}, and 1.07% in Kappa. This result indicates that training with pseudo-labels generated by PLGA contributes to enhancing the network's capability in land cover classification. The improvement in land cover classification accuracy also

1.[Online]. Available: https://github.com/LiheYoung/SenseEarth2020-ChangeDetection.

TABLE I
QUANTITATIVE RESULTS OF THE ABLATION STUDY ON THE SECOND DATASET

Method	OA (%)	SeK (%)	F_{scd} (%)	mIoU (%)	F1 (%)	Kappa (%)
Baseline	87.11	21.64	60.85	72.55	72.92	76.72
Baseline-PLGA	87.36	22.42	61.82	73.14	73.30	77.79
Baseline-PLGA-DCP	87.68	23.12	62.50	73.26	73.55	78.67
Baseline-PLGA-STSCM	87.66	23.46	62.89	73.39	73.96	78.39
SIC-Net	**87.80**	**23.96**	**63.26**	**73.68**	**74.30**	**79.04**

The best values are highlighted in bold.

assists in enhancing the precision of the binary CD task, with an increase of approximately 0.38% in F1. Therefore, it contributes to enhancing the overall accuracy of SCD.

Building upon the baseline-PLGA, we further validate the effectiveness of DCP and STSCM. With the introduction of DCP, we observed an increase of 0.70% in SeK, 0.68% in F_{scd}, and 0.88% in Kappa coefficient. Following the incorporation of STSCM, we saw an improvement of 1.04% in SeK, 1.07% in F_{scd}, and 0.66% in F1. The results indicate that both detailed information and collaborative training contribute to improving the accuracy of the SCD task. Finally, we evaluated the SIC-Net incorporating all these auxiliary designs. Compared to the baseline, the improvements were approximately SeK: 2.32%, F_{scd}: 2.41%. In summary, the integration of DCP, STSCM, and other auxiliary designs significantly improves the performance of SIC-Net, demonstrating its effectiveness in SCD tasks.

We present several inference results on SECOND dataset in Fig. 8. It can be observed that there are a considerable number of misclassifications in the baseline results. However, training with pseudo-labels generated by PLGA significantly improves this issue. With the inclusion of DCP, the experimental results of Baseline-PLGA-DCP gradually approach GT in identifying change regions (indicated by red dashed boxes in Fig. 8) with more accurate boundaries. The incorporation of STSCM and DCP methods further improves the prediction of land cover categories (yellow dashed box in Fig. 8). Furthermore, it effectively improves the issue of incomplete detection of change detection target areas [as shown in Fig. 8(c1), (c2)]. In summary, compared to the baseline method, by progressively integrating different design components, the changed regions and land cover categories determined by SIC-Net are much closer to the GT (as shown by the blue dashed box in Fig. 8).

To further understand the impact of spatial details on the SCD task, we selected two pairs of images as test samples and input them into the network. We then performed a visualization analysis of the three branches' features [i.e., f_1, f_2, and f_{cd} in Fig. 3(a)] before and after adding DCP. The features were visualized by calculating the mean value of the feature maps, and the results are shown in Fig. 9. From the visual results, it is evident that with the introduction of DCP, the features of the semantic segmentation branch become more complete, and the contours are clearer (black dashed box in Fig. 9). Additionally, the features of the binary CD branch exhibit significant differences. After integrating the detailed information, the boundaries of the change areas become more accurate (red dashed box in Fig. 9). This demonstrates the importance of detailed information in the SCD task, significantly influencing the accurate detection of both subtasks.

B. SECOND Dataset

In this section, we present the comparison results of SIC-Net and other SCD methods on the SECOND dataset. The quantitative results are reported in Table II. L-UNet was originally designed for binary CD and exhibits inferior performance in SCD tasks. The Kappa metric exhibited a noticeable decline relative to other methods. In terms of quantitative results, recently published methods such as SCanNet, DEFO-MTLSCD, and our proposed SIC-Net all demonstrate significant advantages. It is worth mentioning that the SIC-Net method achieves the highest SCD accuracy (Sek: 23.96%, F_{scd}: 63.26%). Compared with the performance-suboptimal DEFO-MTLSCD, the SeK metric improves by nearly 0.5%, while semantic segmentation accuracy (Kappa) increases by nearly 1.23%, but the improvement in binary CD accuracy (F1, mIoU) is relatively low. Compared with the MambaSCD method based on the Mamba structure, SIC-Net improves SeK by 1.79%, F1 by 1.56%, and Kappa by 0.46%. This result indicates that SIC-Net has achieved significant success in optimizing the balance between the two subtasks, resulting in higher overall performance.

Fig. 10 illustrates the comparison of prediction results between SIC-Net and other methods on the SECOND dataset. The first two columns of the figure show the dual-temporal images and their corresponding ground truth land cover labels. In the comparison, HRNet, MTSCD-Net, MambaSCD, and our SIC-Net demonstrate more comprehensive and accurate recognition capabilities, especially in identifying the "playground" class, which has fewer training samples, as indicated by the yellow dashed boxes in the figure. In contrast, other methods exhibit more noticeable misclassifications when facing such uncommon classes. Despite integrating spatial details in their unique ways, SCDNet and HRNet demonstrate instability across multiple scenarios, as depicted by the green dashed boxes, where they fail to precisely detect subtle changes. In this regard, our SIC-Net outperforms the DEFO-MTLSCD method. Moreover, although DEFO-MTLSCD shows certain capabilities in change detection, it encounters inconsistencies in some cases, where the semantic segmentation results of the two periods are consistent, as shown in Fig. 10(b) and (c). This further demonstrates the superior performance of SIC-Net in recognizing land cover categories and maintaining consistency in subtask results compared to the

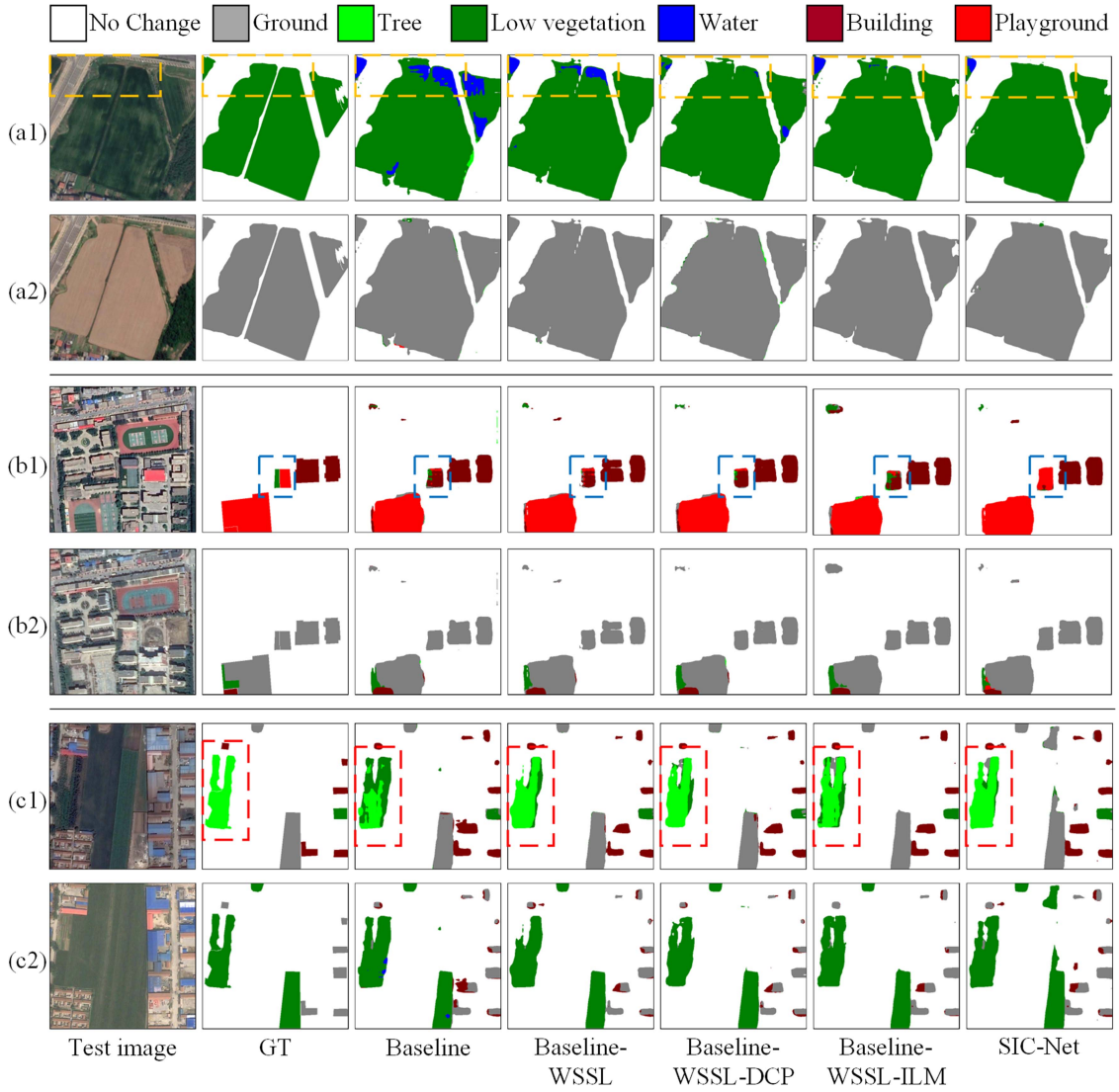


Fig. 8. Example of results in the ablation study on SECOND dataset.

suboptimal method DEFO-MTLSCD. Furthermore, compared to SCanNet and MambaSCD, SIC-Net exhibits fewer false detections and provides more complete representations of land objects in its prediction results. This convincingly demonstrates the advantages of our method in identifying small target changes and variations in land cover categories. These results not only highlight the effectiveness of our approach in handling SCD tasks but also underscore the significant benefits of effectively integrating contextual information with detailed information and coordinating training among subtasks.

C. Landsat-SCD Dataset

In this section, we compared the accuracy of SIC-Net and other SCD methods on the Landsat-SCD dataset, which consists of medium-resolution images. This evaluation validates the performance of our network and its compatibility with images of different resolutions. Additionally, to ensure fairness in the comparative experiments and solely validate the superiority of our proposed network architecture, we did not employ the PLGA in this scenario.

We use the same comparison methods as SECOND. As reflected in Table III, SIC-Net has achieved the best accuracy on the test set of Landsat-SCD, with an SeK of 61.29% and F_{scd} of 86.18%. Compared to the results on the SECOND dataset, the accuracy of detection results for various methods on the Landsat-SCD dataset is significantly higher. Bi-SRNet and MTSCD-Net, which performed well on the previous dataset, lag far behind our method, with lower binary CD accuracy (F1) at 85.98% and lower semantic segmentation accuracy (Kappa) at 86.37%. Although SCDNet achieves a relatively high Kappa coefficient, its binary CD accuracy is significantly lower compared to all other comparison methods when faced with medium-resolution images. Compared to these, DEFO-MTLSCD, MambaSCD, SCanNet, and SIC-Net demonstrate notably higher accuracy. Notably, since the Kappa in this paper evaluates the classification accuracy of the overlapping area between the predicted and actual change regions, a significantly high F1 score results in

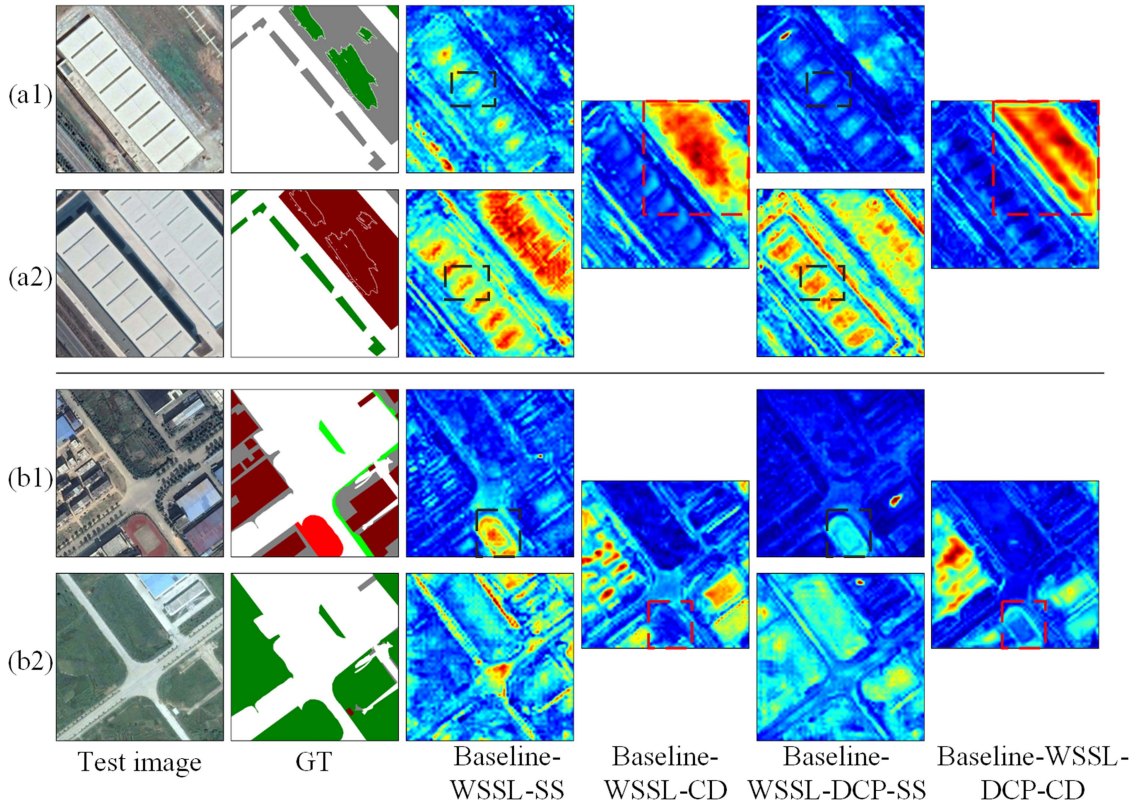


Fig. 9. Visualization of features in different branches.

TABLE II
COMPARISON RESULTS ON THE SECOND DATASET

Method	OA (%)	SeK (%)	F_{scd} (%)	mIoU (%)	F1 (%)	Kappa (%)
SSCD-l	87.05	21.65	60.99	72.43	72.82	77.15
L-Unet	86.24	17.99	57.24	70.63	70.35	73.88
HRNet	87.27	20.27	60.18	71.53	71.25	77.96
SCDNet	86.48	20.49	60.35	71.44	71.79	77.65
Bi-SRNet	87.45	22.20	61.63	72.83	73.16	77.71
MTSCD-Net	86.67	22.11	61.46	72.40	73.16	77.41
MambaSCD	87.71	22.17	61.70	72.66	72.74	78.58
SCanNet	87.66	22.93	62.37	73.10	73.48	78.67
DEFO-MTLSCD	87.75	23.47	62.61	73.58	74.25	77.81
SIC-Net	**87.80**	**23.96**	**63.26**	**73.68**	**74.30**	**79.04**

The best values are highlighted in bold.

TABLE III
COMPARISON RESULTS ON THE LANDSAT-SCD DATASET

Method	OA (%)	SeK (%)	F_{scd} (%)	mIoU (%)	F1 (%)	Kappa (%)
SSCD-l	91.07	44.18	79.73	82.74	85.69	88.98
L-UNet	89.59	39.04	76.23	81.03	84.21	84.85
HRNet	91.91	47.55	81.27	84.37	87.14	89.30
SCDNet	90.56	41.91	79.21	81.20	84.25	90.46
Bi-SRNet	91.46	45.41	80.59	83.12	85.98	90.07
MTSCD-Net	90.49	45.48	79.33	83.43	86.80	86.37
MambaSCD	92.29	50.04	82.14	85.57	88.30	88.93
SCanNet	93.05	53.18	84.00	86.51	89.10	90.91
DEFO-MTLSCD	94.08	58.34	85.95	88.67	90.94	**91.29**
SIC-Net	**94.45**	**61.29**	**86.18**	**90.44**	**92.48**	90.16

The best values are highlighted in bold.

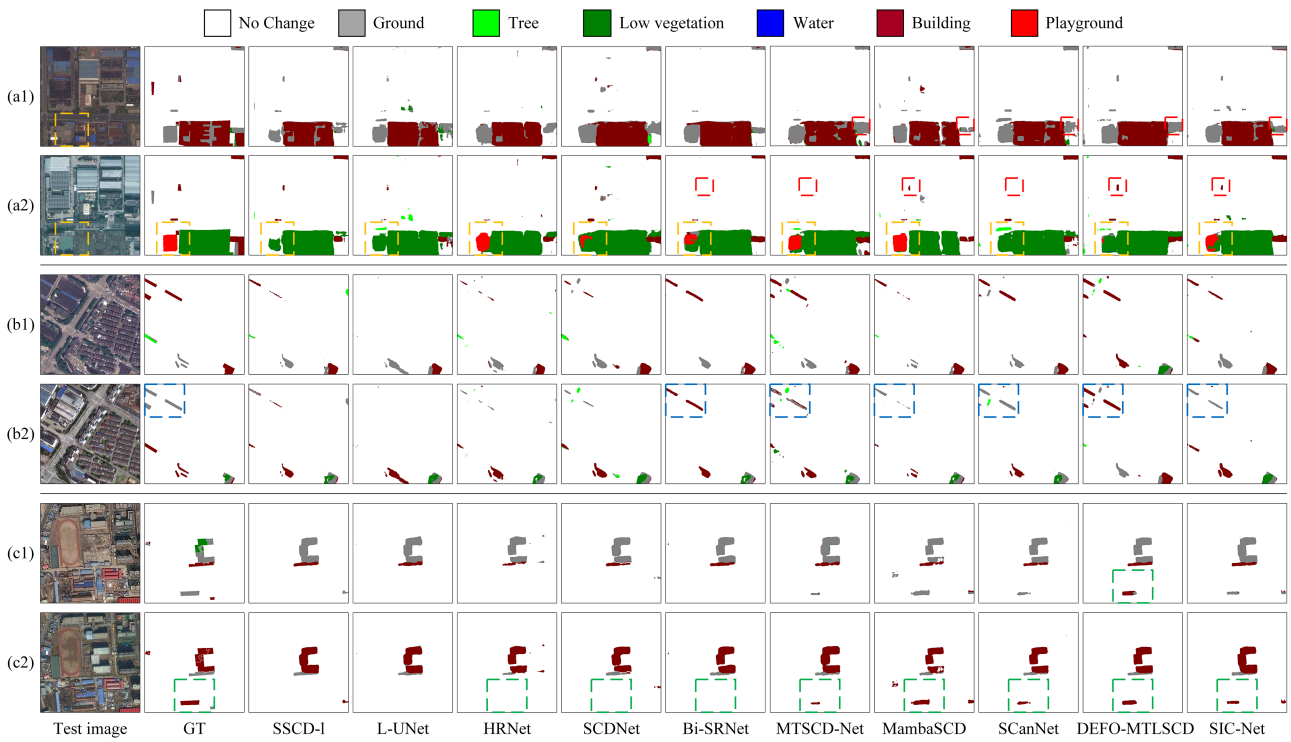


Fig. 10. Comparison of experimental results on SECOND dataset.

more pixels being involved in the Kappa calculation. This partly explains why SIC-Net slightly lags behind SCanNet and DEFO-MTLSCD in terms of semantic segmentation accuracy (Kappa). However, its significant advantage in binary CD accuracy with an F1 score of 92.48% leads to a notable overall accuracy improvement, with the SeK metric increasing by approximately 3%. The experimental results further demonstrate that SIC-Net, through the design of STSCM, effectively facilitates information exchange between the two subtasks, achieving a good balance and improving SCD accuracy. Additionally, the results indicate the robustness of our method in handling images of different resolutions.

Fig. 11 visually displays the SCD results of SIC-Net and the comparison methods on the Landsat-SCD dataset. Different scenarios of the Landsat-SCD dataset are selected. By observation, it can be noticed that the performance of most methods is quite satisfactory, with no significant differences. However, due to the limitations of medium-resolution images, unlike in the GT labels in SECOND dataset, there are numerous linear features (one pixel wide). Most comparative methods struggle to accurately predict this scenario. As shown in Fig. 11, SIC-Net outperforms other methods in many aspects, while the results of SSCD-l, SCDNet, and L-UNet are notably inferior to the rest. Notably, it excels in accurately identifying fine-scale features, as highlighted by the red dashed box in Fig. 11. SIC-Net can more comprehensively detect this kind of small targets that are easily overlooked by other methods. Additionally, SIC-Net demonstrates outstanding performance in accurately identifying change boundaries (as shown by the black dashed box in Fig. 11), while also outperforming comparative methods in identifying land cover categories (as indicated by the purple dashed box in the same figure). In the case of Fig. 11(a1) and (a2), other methods either produce false positives or exhibit partial omissions, resulting in incomplete target shapes, whereas SIC-Net demonstrates good stability under such circumstances. In contrast to DEFO-MTLSCD, which has suboptimal accuracy performance, SIC-Net can be observed from the prediction results that the misidentification of land cover types by SIC-Net is significantly alleviated (as indicated by the blue dashed box in Fig. 11). Experimental results demonstrate that our method performs better in identifying the types and extent of difficult samples in complex environments. Although inaccurate identification of change domains still exists, achieving precise recognition in such scenarios remains challenging even for humans.

VI. Discussion

A. Comparison of Detail Extraction Strategies

To further validate the superior capability of the DBB composed of DCP and SCP in capturing detailed information, this study conducted a series of comprehensive comparative experiments based on the SSCD-l architecture. In this process, we selected HRNet and U-Net as comparison objects, as they are both renowned for their excellent capture of complex features. To ensure the fairness of the experiments, similar to other methods, U-Net initially down sampled the input image to 1/4 size through convolution and pooling. Through this series of comparative experiments, we aimed to clearly elucidate the differences between DCP, multiscale learning, and skip connections, and

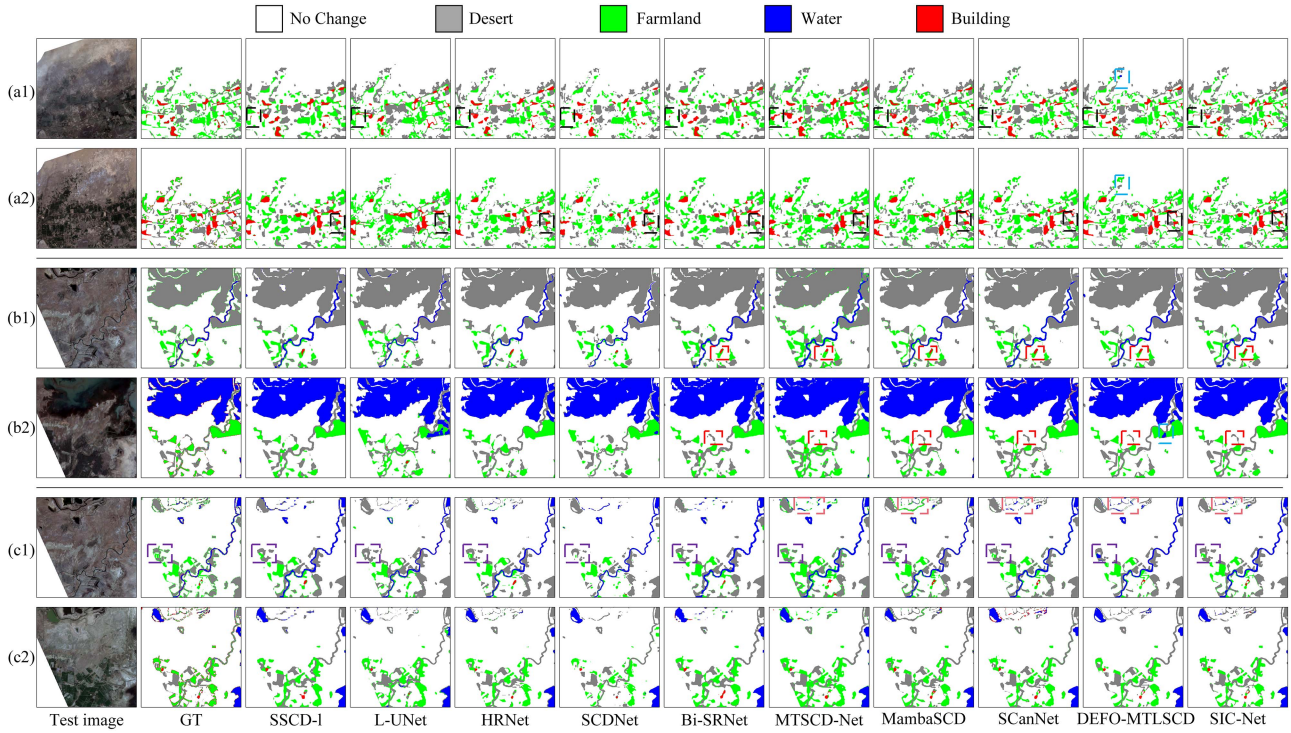


Fig. 11. Comparison of experimental results on Landsat-SCD dataset.

TABLE IV
COMPARISON RESULTS FROM DIFFERENT BACKBONES ON THE SECOND DATASET

Backbone	OA (%)	SeK (%)	F_{scd} (%)	mIoU (%)	F1 (%)	Kappa (%)
U-Net	86.04	17.95	57.57	70.40	70.24	74.50
HRNet	87.09	22.06	61.20	72.66	73.14	77.13
SSCD-l	87.05	21.65	60.99	72.43	72.82	77.15
SSCD-l-DCP	**87.37**	**22.25**	**61.61**	**72.85**	**73.25**	**77.56**

The best values are highlighted in bold.

more accurately assess the ability of DBB in capturing image details.

As illustrated in Table IV, the experimental results unequivocally indicate that SSCD-l-DCP achieves the highest accuracy in both binary CD (F1) and semantic segmentation (Kappa). This not only highlights the applicability of features extracted through the proposed DBB in the SCD task but also corroborates the effectiveness of integrating detailed information extracted through DCP with contextual features in enhancing the accuracy of both subtasks.

In Fig. 12, we randomly selected images from the SECOND validation set, showcasing features obtained using different backbone networks. The visual results indicate significant differences in features obtained with the addition of DCP, further confirming its capability to improve feature extraction. Comparatively, HRNet exhibits more noise points in its features. U-Net demonstrates lower stability and offers poorer feature quality in specific scenarios (first row of Fig. 12). Our method combines detailed information with contextual features, resulting in clearer object outlines extracted by SSCD-l-DCP. The results indicate that the designed DBB composed of SCP and DCP achieves a good balance between detailed information and contextual features, demonstrating the superiority and wide applicability of DCP in the SCD task.

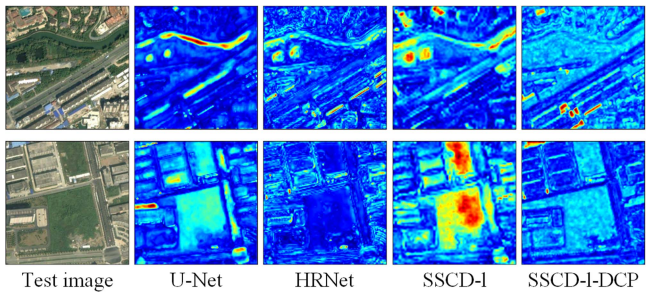


Fig. 12. Visualization of features in different backbones.

B. Model Efficiency

To comprehensively assess the performance of our proposed algorithm, we conducted a detailed comparison in terms of

TABLE V
COMPARISON RESULTS COMPUTATIONAL COMPLEXITY

Method	Params (M)	FLOPs(G)
SSCD-l	23.31	189.76
L-Unet	9.43	93.76
HRNet	34.90	168.35
SCDNet	37.25	147.32
Bi-SRNet	23.38	190.30
MTSCD-Net	94.55	290.69
MambaSCD	54.28	146.80
SCanNet	27.90	264.95
DEFO-MTLSCD	26.02	401.09
Baseline	25.73	210.59
Baseline-DCP	26.07	216.12
SIC-Net	33.83	280.74

computational complexity. Table V presents a comparison of SIC-Net with several benchmark methods regarding computational complexity. This comparison focuses on two metrics: Params and FLOPs.

Although the DBB structure used in this article inevitably leads to an increase in network parameters, the parameter count of SIC-Net is still relatively lower compared to methods like MTSCD-Net and MambaSCD. Moreover, compared to advanced methods such as DEFO-MTLSCD and MTSCD-Net, SIC-Net also exhibits certain advantages in terms of FLOPs, indicating its relatively lower computational complexity. This result demonstrates that the proposed method maintains high accuracy without introducing excessive computational complexity, achieving a good balance between accuracy and efficiency, thus showing significant performance advantages overall.

VII. CONCLUSION

In this article, we propose a novel SIC-Net for remote sensing images SCD. By deeply integrating the DBB with the STSCM, the network significantly enhances the model's capability in finegrained feature extraction, promotes synergistic learning between the two subtasks, and alleviates issues such as detail loss and result inconsistencies. Furthermore, we introduce the PLGA to address the issue of class imbalance in SCD tasks by increasing the annotated pixel count in the labels.

Experimental results suggest that SIC-Net achieved an improvement of over 2.32% compared to the baseline methods and obtained the highest accuracy on both SCD datasets. This demonstrates the outstanding performance of SIC-Net in the task of SCD in remote sensing image processing.

Additionally, we believe that DBB and STSCM have not fully exploited the potential of spatial detail and spatial–temporal dependency modeling. Therefore, we encourage exploring different architectures. Simultaneously, considering the complexity of SCD datasets, future research should consider incorporating semi-supervised or unsupervised learning to alleviate the limitations of feature extraction caused by limited data and extreme class imbalances in real-world remote sensing scenarios.

REFERENCES

[1] F. Bovolo and L. Bruzzone, "The time variable in data fusion: A change detection perspective," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 8–26, Sep. 2015.

[2] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2407, Jun. 2004.

[3] 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.

[4] 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, 2022.

[5] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Understanding*, vol. 1, no. 187, 2019, Art. no. 10278.

[6] K. Yang et al., "Asymmetric siamese networks for semantic change detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5609818.

[7] C. Wu, B. Du, X. Cui, and L. Zhang, "A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion," *Remote Sens. Environ.*, vol. 199, pp. 241–255, 2017.

[8] Q. Zhu, X. Guo, Z. Li, and D. Li, "A review of multi-class change detection for satellite remote sensing imagery," *Geo-Spatial Inf. Sci.*, vol. 27, no. 1, pp. 1–15, 2022.

[9] H. Jiang et al., "A survey on deep learning-based change detection from high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 7, 2022, Art. no. 1552.

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2018, vol. 11211, pp. 833–851.

[11] 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.

[12] J. Pan et al., "MapsNet: Multi-level feature constraint and fusion network for change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, 2022, Art. no. 102676.

[13] 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.

[14] 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, 2022, Art. no. 5622716.

[15] 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, 2022, Art. no. 5620014.

[16] 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.

[17] Z. Li, X. Wang, S. Fang, J. Zhao, S. Yang, and W. Li, "A decoder-focused multitask network for semantic change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5609115.

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

[19] A. Fernandez, P. Illera, and J. L. Casanova, "Automatic mapping of surfaces affected by forest fires in Spain using AVHRR NDVI composite image data," *Remote Sens. Environ.*, vol. 60, no. 2, pp. 153–162, May 1997.

[20] D. Seo, Y. Kim, Y. Eo, W. Park, and H. Park, "Generation of radiometric, phenological normalized image based on random forest regression for change detection," *Remote Sens.*, vol. 9, no. 11, 2017, Art. no. 1163.

[21] I.-H. Holobâcă, "Glacier mapper–A new method designed to assess change in mountain glaciers," *Int. J. Remote Sens.*, vol. 34, no. 23, pp. 8475–8490, 2013.

[22] H. Zhuang, K. Deng, H. Fan, and M. Yu, "Strategies combining spectral angle mapper and change vector analysis to unsupervised change detection in multispectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 5, pp. 681–685, May 2016.

[23] Y. Moisan, M. Bernier, and J. M. M. Dubois, "Detection of changes in a series of multitemporal ERS-1 images by principal components analysis," *Int. J. Remote Sens.*, vol. 20, no. 6, pp. 1149–1167, Apr. 1999.

[24] S. M. Jin and S. A. Sader, "Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances," *Remote Sens. Environ.*, vol. 94, no. 3, pp. 364–372, Feb. 2005.

[25] M. E. A. Arabi, M. S. Karoui, and K. Djerriri, "Optical remote sensing change detection through deep Siamese network," in *Proc. 38th IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 5041–5044.

[26] W. Zhang and X. Lu, "The spectral-spatial joint learning for change detection in multispectral imagery," *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 240.

[27] L. Li, C. Wang, H. Zhang, B. Zhang, and F. Wu, "Urban building change detection in SAR images using combined differential image and residual U-Net network," *Remote Sens.*, vol. 11, no. 9, 2019, Art. no. 1091.

[28] A. Ghosh, B. N. Subudhi, and L. Bruzzone, "Integration of Gibbs Markov random field and Hopfield-type neural networks for unsupervised change detection in remotely sensed multitemporal images," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3087–3096, Aug. 2013.

[29] J. Liu, M. Gong, J. Zhao, H. Li, and L. Jiao, "Difference representation learning using stacked restricted Boltzmann machines for change detection in SAR images," *Soft Comput.*, vol. 20, no. 12, pp. 4645–4657, 2014.

[30] H. Aghababaee, J. Amini, and Y. C. Tzeng, "Improving change detection methods of SAR images using fractals," *Scientia Iranica*, vol. 20, no. 1, pp. 15–22, 2013.

[31] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, "ChangeNet: A deep learning architecture for visual change detection," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2019, vol. 11130, pp. 129–145.

[32] 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.

[33] H. Zhang, M. Lin, G. Yang, and L. Zhang, "ESCNet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 28–42, Jan. 2023.

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

[35] T. Liu et al., "Building change detection for VHR remote sensing images via local–global pyramid network and cross-task transfer learning strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4704817.

[36] 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.

[37] 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.

[38] 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, 2022, Art. no. 8007805.

[39] 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, 2021.

[40] 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, 2022.

[41] G. Ailimujiang, Y. Jiaermuhamaiti, H. Jumahong, H. Wang, S. Zhu, and P. Nurmamaiti, "A transformer-based network for change detection in remote sensing using multiscale difference-enhancement," *Comput. Intell. Neurosci.*, vol. 2022, 2022, Art. no. 2189176.

[42] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.

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

[44] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.

[45] 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, 2022, Art. no. 5622519.

[46] L. Ding, K. Zhu, D. Peng, H. Tang, K. Yang, and L. Bruzzone, "Adapting segment anything model for change detection in VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5611711.

[47] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[48] T. Suzuki et al., "Semantic change detection," in *Proc. 15th Int. Conf. Control, Automat., Robot. Vis.*, 2018, pp. 1785–1790.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] L. Jiang, F. Li, L. Huang, F. Peng, and L. Hu, "TTNet: A temporal-transform network for semantic change detection based on bi-temporal remote sensing images," *Remote Sens.*, vol. 15, no. 18, 2023, Art. no. 4555.

[54] H. Chen, J. Song, C. Han, J. Xia, and N. Yokoya, "Changemamba: Remote sensing change detection with spatio-temporal state space model," *IEEE Trans. Geosci. Remote Sens.*, p. 1, 2024.

[55] L. Ding, J. Zhang, H. Guo, K. Zhang, B. Liu, and L. Bruzzone, "Joint spatio-temporal modeling for semantic change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5610814.

[56] R. Wang, H. Wu, H. Qiu, F. Wang, X. Liu, and X. Cheng, "A difference enhanced neural network for semantic change detection of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5509205.

[57] H. Chang, P. Wang, W. Diao, G. Xu, and X. Sun, "A triple-branch hybrid attention network with bitemporal feature joint refinement for remote sensing image semantic change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5613816.

[58] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 181, pp. 84–98, Nov. 2021.

[59] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2018, vol. 11217, pp. 334–349.

[60] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, Nov. 2021.

[61] G. Cheng, G. Wang, and J. Han, "ISNet: Towards improving separability for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5623811.

[62] M. Yang, L. Jiao, F. Liu, B. Hou, S. Yang, and M. Jian, "DPFL-nets: Deep pyramid feature learning networks for multiscale change detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6402–6416, Nov. 2022.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[64] 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.

[65] Y. He, H. Zhang, X. Ning, R. Zhang, D. Chang, and M. Hao, "Spatial-temporal semantic perception network for remote sensing image semantic change detection," *Remote Sens.*, vol. 15, no. 16, Aug. 2023, Art. no. 4095.

[66] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Proc. 31st IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7014–7023.

[67] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

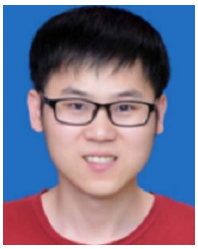
[68] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7651–7668, Sep. 2021.

Xiaogang Ning received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006.

He is a Full Professor and the Director of the Institute of Photogrammetry and Remote Sensing. Since 2006, he has been continuously working with the Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping, Beijing, China. He has authored or coauthored more than 50 peer-reviewed articles published in international journals. His research interests include high-resolution image processing, pattern recognition, and urban applications of remote sensing.

You He received the B.S degree in remote sensing science and technology from Central South University, Changsha, China, in 2021. He is currently working toward the M.S. degree in photogrammetry and remote sensing with the Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping, Beijing, China.

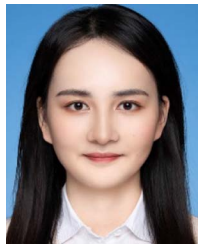
His research interests include change detection and semantic segmentation of remote sensing images.



Hanchao Zhang received the Ph.D. degree in photogrammetry and remote sensing from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2019.

He is currently an Assistant Research Fellow with the Chinese Academy of Surveying and Mapping, Beijing, China. His primary research interests include remote sensing interpretation, change detection, and deep learning.

Dr. Zhang has also been selected for the 8th Young Talent Support Project by the China Association for Science and Technology.



Ruiqian Zhang (Member, IEEE) received the B.S. degree in remote sensing science and technology and the Ph.D. degree in photogrammetry and remote sensing from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2015 and 2021, respectively.

She completed her postdoctoral research in June 2024 and is currently working with the Chinese Academy of Surveying and Mapping, Beijing, China. She has authored or coauthored more than 30 peer-reviewed articles published in international journals. Her research interests include image processing, change detection, and object detection in the realm of remote sensing and computer vision.

Dong Chang received the professional master's degree in surveying and mapping engineering from the School of Surveying and Spatial Information, Shandong University of Science and Technology, Qingdao, China, in 2023.

He is currently working with Beijing DeepBlueSpatial Remote Sensing Technology, Co., Ltd., Shandong University of Science and Technology. His research interests include intelligent understanding of remote sensing images, label optimization, and change detection.

Minghui Hao received the first master's degree in land administration from the University of Twente, The Netherlands, in 2011, and the second master's degree in photogrammetry and remote sensing from the Chinese Academy of Surveying and Mapping (CASM), Beijing, China, in 2011.

She is currently a Senior Engineer with CASM. Her research interests include the development and application promotion of the natural resource remote sensing image change monitoring technology.