remote sensing

MDPI

Article

SMNet: Symmetric Multi-Task Network for Semantic Change Detection in Remote Sensing Images Based on CNN and Transformer

Yiting Niu [1], Haitao Guo [1,*], Jun Lu [1], Lei Ding [1] and Donghang Yu [2]

[1] Institute of Geospatial Information, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China
[2] Naval Research Institute, Beijing 100070, China
* Correspondence: ghtgjp2002@163.com

Abstract: Deep learning has achieved great success in remote sensing image change detection (CD). However, most methods focus only on the changed regions of images and cannot accurately identify their detailed semantic categories. In addition, most CD methods using convolutional neural networks (CNN) have difficulty capturing sufficient global information from images. To address the above issues, we propose a novel symmetric multi-task network (SMNet) that integrates global and local information for semantic change detection (SCD) in this paper. Specifically, we employ a hybrid unit consisting of pre-activated residual blocks (PR) and transformation blocks (TB) to construct the (PRTB) backbone, which obtains more abundant semantic features with local and global information from bi-temporal images. To accurately capture fine-grained changes, the multi-content fusion module (MCFM) is introduced, which effectively enhances change features by distinguishing foreground and background information in complex scenes. In the meantime, the multi-task prediction branches are adopted, and the multi-task loss function is used to jointly supervise model training to improve the performance of the network. Extensive experimental results on the challenging SECOND and Landsat-SCD datasets, demonstrate that our SMNet obtains 71.95% and 85.65% at mean Intersection over Union (mIoU), respectively. In addition, the proposed SMNet achieves 20.29% and 51.14% at Separated Kappa coefficient (Sek) on the SECOND and Landsat-SCD datasets, respectively. All of the above proves the effectiveness and superiority of the proposed method.

Keywords: remote sensing images; semantic change detection; deep learning

Citation: Niu, Y.; Guo, H.; Lu, J.; Ding, L.; Yu, D. SMNet: Symmetric Multi-Task Network for Semantic Change Detection in Remote Sensing Images Based on CNN and Transformer. *Remote Sens.* **2023**, *15*, 949. https://doi.org/10.3390/rs15040949

Academic Editors: Lizhe Wang, Xiaodong Zhang, Jining Yan and Guanzhou Chen

Received: 17 December 2022
Revised: 2 February 2023
Accepted: 7 February 2023
Published: 9 February 2023
Corrected: 8 June 2023

1. Introduction

Accurate access to dynamics and change information on the land surface is important for understanding and studying the natural environment, human activities, and their correlations [1,2]. With the rapid development of remote sensing sensors, it is possible to obtain massive high-resolution remote sensing images, which further provides reliable data sources for studying CD. Remote sensing image CD is used to identify the differences between two images, which are located at the same position but are obtained at different times. Therefore, remote sensing image CD is of great signification in many fields including urban planning, ecosystem assessment, and natural resource management [3–5].

Over the years, deep learning has been successfully applied to image classification [6,7], object detection [8,9], and semantic segmentation [10,11], which contributes new ideas and methods for remote sensing image CD. A simple and commonly used method is stacking the bi-temporal images and then feeding them into a CNN to extract change information [12–14]. However, this method does not make full use of the information from the bi-temporal images. Therefore, more researchers adopted an architecture with two parallel CNN and some effective techniques (e.g., multi-scale feature fusion methods,

attention mechanism, deep supervision) to improve the performance of the architecture. For example, Zhang et al. [15] proposed a CD framework with a hierarchical fusion strategy and introduced the dynamic convolutional module for adaptive learning, to improve results integrity. Fang et al. [16] used a Siamese network based on U-Net++ to extract bi-temporal images' features and adopted an integrated channel attention module (CAM) for deep supervision. Ling et al. [17] proposed an integrated residual attention multi-scale Siamese network, which effectively obtains multi-scale semantic information to alleviate the lack of unpredictable change details and global semantic information. Chen et al. [18] designed a dual-attention mechanism to capture more discriminative features than compared methods, which improves CD accuracy. Peng et al. [19] proposed a difference-enhancement dense-attention CNN, which uses upsampling attention and difference-enhancement units to, respectively, extract and select change information. Guo et al. [20] proposed CosimNet for scene CD using thresholded contrastive loss to learn more discriminative metrics. Zhu et al. [21] adopted a global hierarchical mechanism to enhance unbalanced samples and improve SCD accuracy. Although CNN significantly improves CD accuracy for remote sensing images with its powerful feature extraction capabilities, the existing methods still suffer from the following problems. First, the CNN's receptive field is much smaller than the theoretical maximum, which makes it difficult to establish long-range dependencies in space and time. Hence, the global and contextual information of remote sensing images is often ignored by the CNN. Second, the CNN's methods of dealing with changed region boundaries are unsatisfactory. Last but not least, most algorithms for CD can predict whether image pixels have changed, but they cannot detect their semantic categories on the bi-temporal images.

To overcome the difficulty faced by CNN in handling bi-temporal features and to better leverage its feature extraction powers for SCD, we propose a method based on CNN and Transformer called symmetric multi-task network (SMNet) that not only locates changed areas but also identifies the type of change. First, to extract the local and global information from bi-temporal remote sensing images, a hybrid PRTB backbone composed of PR and TB is built [22], which extracts different hierarchical features. Then, a novel multi-content fusion module (MCFM) [23] is used to strengthen the change-related features obtained by subtracting the extracted corresponding hierarchical features. Finally, we use the multi-task prediction branches (i.e., two semantic and one change branches) to obtain SCD results. To improve overall network performance, the multi-task loss function is jointly used to supervise training. Extensive experiments on the SECOND and Landsat-SCD dataset demonstrate that our proposed method obtains better accuracy than compared methods. The main contributions of this research are as follows:

(1) We propose a novel multi-task model (SMNet) for remote sensing image SCD, which reduces confusion between semantic and change information. To sufficiently capture the local and global information from the bi-temporal images, we exploit the multi-scale feature extraction encoder that integrates the PR and TB to obtain more abundant semantic information. For enhancing the extraction of change-related features in complex scenes, the MCFM is introduced to mitigates false detection of fine-grained changes.

(2) Extensive experiments on two public datasets demonstrate the effectiveness of the proposed SMNet. The proposed method outperforms compared methods, yielding the highest SeK of 20.29% and 50.14%, respectively.

The rest of this paper is organized as follows. Related works are reviewed in Section 2. A detailed description of the proposed method is provided in Section 3. The experimental data, evaluation indicators, and training details are given in Section 4. The experimental results and discussion are reported in Section 5, and the conclusions are provided in Section 6.

2. Related Work

2.1. Binary Change Detection

Numerous scholars have conducted extensive research on BCD. To suppress background noise in remote sensing images, Chen et al. [24] designed a feature constraint CD

network that uses a self-supervised learning strategy to constrain feature extraction and feature fusion. Zhi et al. [25] proposed a novel neural network with a spatial-spectral attention mechanism and multi-scale dilation convolution modules, effectively alleviating the pseudo-changes caused by solar height and soil moisture in land cover CD for remote sensing images. To address the problems of feature diversity and scale-change flexibility, Lei et al. [26] employed scale-adaptive attention to establish relationships between feature maps and convolution kernel scales, utilizing a multi-layer perceptron (MLP) that fuses low-level details and high-level semantics to improve feature discrimination. Wei et al. [27] designed a location guidance module that accurately identifies changed regions. To handle the varying resolutions of bi-temporal images, a super-resolution module containing a generator and discriminator [28] is introduced to directly learn super-resolution features. Meanwhile, a stacked attention module is used to capture the more useful channel and spatial information, which effectively improved the accuracy of multi-resolution remote sensing image CD. The challenges of CD for remote sensing images are a small number of datasets and a huge work of data labeling. To alleviate the challenges, a semi-supervised convolutional network based on a generative adversarial network [29] is proposed, which uses two discriminators to enforce the feature distribution consistency of segmentation maps and entropy maps between the labeled and unlabeled data, to achieve high-precision CD on a small number of labeled datasets.

2.2. Semantic Change Detection

Despite the great success of BCD, it is difficult to meet the needs of practical production applications due to the lack of semantic information. Therefore, some scholars have researched SCD. To address a coarse boundary, Tsutsui et al. [30] developed a multi-task SCD method based on U-Net, which improves the accuracy of boundary predictions by simultaneously performing CD and semantic segmentation through a shared feature extraction network. A convolutional network for large-scale SCD [31] utilizes the multi-scale atrous convolution unit to enlarge the receptive field as well as capturing multi-scale information. Additionally, an attention mechanism and deep supervision strategy are further introduced to improve network performance. To overcome scale variation and class imbalance problems, a dual-task constrained deep Siamese convolutional network [32] is proposed, which introduces a dual-attention module to obtain discriminant features and results in a good performance on the WHU dataset. The temporal correlation of bi-temporal images is worth considering. To this end, an end-to-end network that combines CNN and recurrent neural network [33] is proposed to extract spatial and temporal information and directly applies changed labels for model training. A multi-task learning framework [34] is constructed, which uses a fully convolutional long short-term memory to capture the temporal relationship among spatial feature vectors and to boost the overall performance. Daudt et al. [35] applied multi-task learning to extract temporal correlations and feature-categorized information to improve SCD performance. To extract asymmetric change information from multi-temporal images, an asymmetric Siamese network [36] is designed to extract depth features through the asymmetric gating unit, which effectively distinguished changed features in complex scenes. Zheng et al. [37] constructed a multitask encoder–converter–decoder network with a reduced number of encoder branches and explored the relationships between semantic change and time symmetry.

3. Methodology

3.1. Overview

The network architecture of the proposed SMNet is shown in Figure 1. SMNet adopts an encoder–decoder structure, which is mainly composed of three parts: a multi-scale feature extraction encoder, multi-content fusion enhancement, and the multi-task prediction decoder. First, the feature extraction encoder contains two symmetric branches composed of the PR and TB, which are used to extract semantic features at different levels from bi-temporal images ($T1$ and $T2$). Two paired sets of feature maps, [$\boldsymbol{E}^1{}_{T1}$, $\boldsymbol{E}^2{}_{T1}$, $\boldsymbol{E}^3{}_{T1}$, $\boldsymbol{E}^4{}_{T1}$,

E^5_{T1}] and [E^1_{T2}, E^2_{T2}, E^3_{T2}, E^4_{T2}, E^5_{T2}], are obtained from the feature extraction encoder. Then, five difference feature maps [D_1, D_2, D_3, D_4, D_5] at each scale (E^i_{T1}, E^i_{T2}; i = 1, 2, 3, 4, 5) are generated for change analysis. To improve the ability of our network to capture useful information from complex scenes, we introduce the MCFM to aggregate foreground, background, and global features. The enhanced change features [F_1, F_2, F_3, F_4, F_5] are then generated by multi-scale feature enhancement composed of MCFM. Finally, the multi-task prediction decoder is used to make synchronous predictions on semantic and enhanced change features, which generate two semantic maps (S_1, S_2) and a binary change map (B). Further, semantic change maps (P_1, P_2) are generated by masking S_1 and S_2 with B.

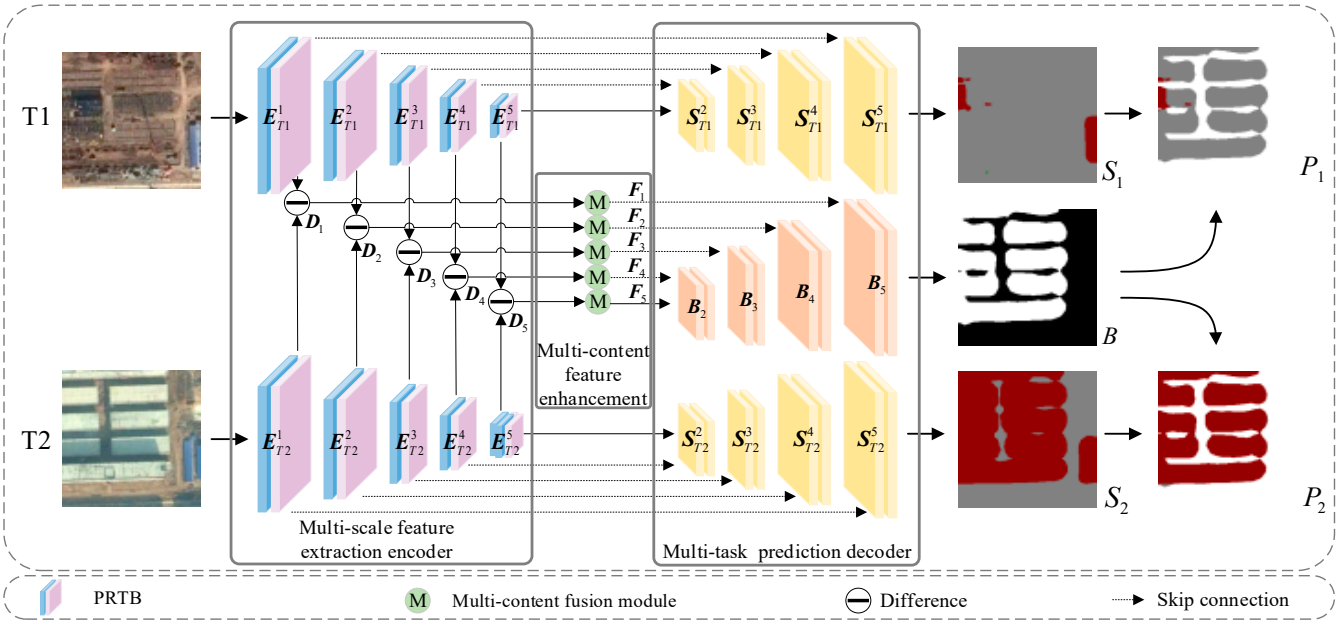

Figure 1. The architecture of the proposed SMNet.

3.2. Multi-Scale Feature Extraction Encoder

The remote sensing image SCD faces the challenge that pure CNN poorly captures global contextual information. The transformer compensates for this shortcoming by exploiting multi-head self-attention (MHSA) to establish a strong global dependency. This enables the receptive field to flexibly learn more powerful features. With the addition of the transformer, both local and global features can be extracted to facilitate semantic understanding. The PRTB backbone is thus constructed to extract multi-scale semantic features at different spatial levels.

Figure 2 illustrates the detailed structure of the PRTB. Compared with the residual blocks of ResNet, the PR block (see Figure 2a) sees the activation functions (batch normalization (BN) and rectified linear unit (ReLU)) as "pre-activation" of the weight layers. The PR block makes the information able to be directly transmitted from one layer to another, in both forward and backward passes [38], which makes SMNet easier to train and improves its generalizability. The input feature, X_{l-1}, in the $l-1$ layer is first fed to BN and ReLU layers. Then, 3×3 convolutional layers are used to capture the local information. Then, X_l in the l layer is generated by the PR block. The TB structure is illustrated in Figure 2b. The MHSA module is used to extract global information, and layer normalization (*LN*), *MLP*, and residual connection modules are introduced to improve representability. Finally, X_l is transformed into X_{l+1} through the n-layer TB. The computing formula is as follows:

$$X_l = BRC(BRC(X_{l-1})) + X_{l-1}, \tag{1}$$

$$\bar{X}_l = MHSA(LN(X_l)) + X_l, \tag{2}$$

$$\boldsymbol{X}_{l+1} = MLP(LN(\bar{\boldsymbol{X}}_l)) + \bar{\boldsymbol{X}}_l, \tag{3}$$

where $\boldsymbol{X}_i$ is the features extracted from the i-th layer after the images are input to the network, and $BRC(\cdot)$ indicates that the feature map passes through BN, ReLU, and 3×3 convolutional layers. $\bar{\boldsymbol{X}}_l$ is the output of the MHSA module. As shown in the left part of Figure 1, the feature extraction encoder consists of five stages. After the first stage, the characteristic resolution of the images remains unchanged, while the channel dimension is increased to 32. At the following stage, the feature resolution is halved, while the channel dimension is doubled. For example, input images $T1$ and $T2$ are both $256 \times 256 \times 3$, and the sizes of the semantic feature maps, $\boldsymbol{E}^i{}_{T1}$ and $\boldsymbol{E}^i{}_{T2}$ ($i = 1, 2, 3, 4, 5$), in the five stages are $256 \times 256 \times 32$, $128 \times 128 \times 64$, $64 \times 64 \times 128$, $32 \times 32 \times 256$, and $16 \times 16 \times 512$, respectively.

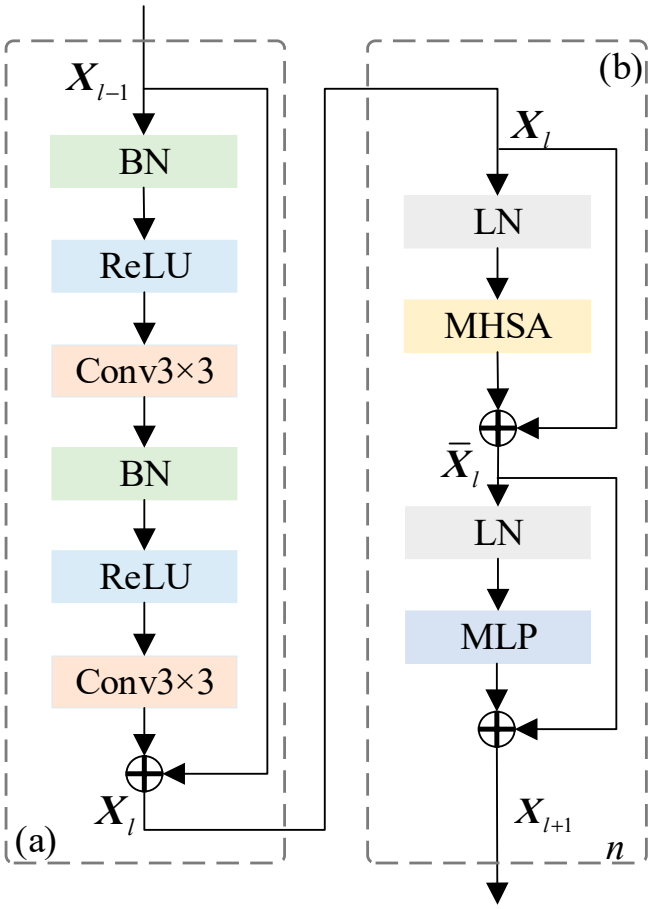

Figure 2. Preactivation-residual block (PR) and transformer block (TB): (**a**) the PR and (**b**) the TB.

3.3. Multi-Content Fusion Enhancement

Problems of false alarm, missed detection, and the loss of details are common in SCD methods and are usually caused by the insufficient ability of the model to extract foreground (region-of-interest) information. Therefore, we introduce the MCFM to handle feature extraction in the foreground and background, as well as globally. While capturing the foreground information, background and global information are supplemented to improve the discrimination of foreground features and obtain more effective edges. As shown in the middle part of Figure 1, we generate difference feature map $\boldsymbol{D}_i$ with correspondent semantic feature maps ($\boldsymbol{E}^i{}_{T1}$, $\boldsymbol{E}^i{}_{T2}$) for change analysis. Then, $\boldsymbol{D}_i$ is enhanced by the MCFM to obtain the enhanced feature, $\boldsymbol{F}_i$. The specific process is as follows:

$$\boldsymbol{D}_i = \boldsymbol{E}^i_{T1} - \boldsymbol{E}^i_{T2}, \tag{4}$$

$$\boldsymbol{F}_i = MCFM(\boldsymbol{D}_i), \tag{5}$$

where i = 1, 2, 3, 4, 5. The specific MCFM architecture is illustrated in Figure 3.

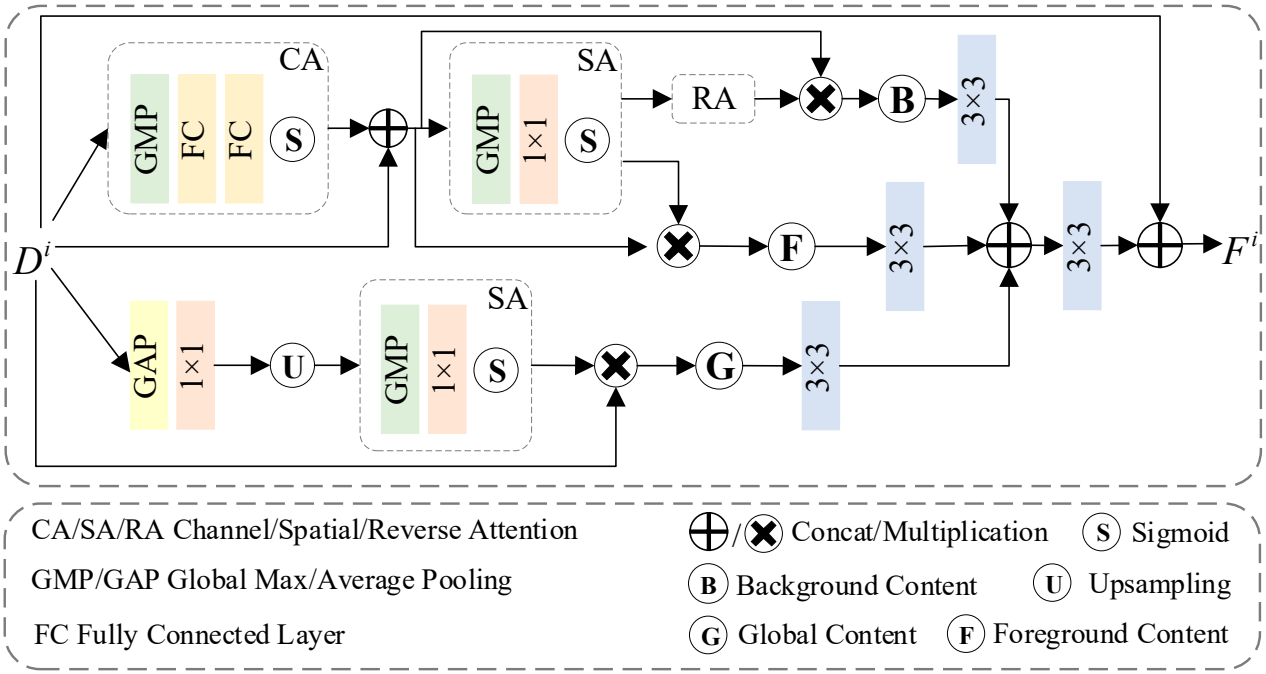

Figure 3. Multi-content fusion module.

3.3.1. Foreground and Background Branches

First, the difference feature, $\boldsymbol{D}_i \in \mathbb{R}^{C\times H\times W}\mathbb{R}^{C\times H\times W}$ (C, H, and W are the channel dimension, height, and width of the feature), is fed into the CA to reduce redundant information. The specific formula is as follows:

$$f_{ca} = CA(\boldsymbol{D}_i) \oplus \boldsymbol{D}_i, \tag{6}$$

where f_{ca} denotes the CA features, and $\oplus$ indicates concatenation in the channel dimension.

Then, to embed the CA information, we obtain the spatial attention (SA) feature, f_{sa}, through the SA. This is computed as

$$f_{sa} = SA(f_{ca}), \tag{7}$$

where f_{sa} denotes the SA features.

On the one hand, f_{sa} is used as an input to the foreground information alongside the CA feature. On the other hand, f_{sa} enters the background branch. Because f_{sa} indicates high-level features that omit spatial details, it leads to the lack of edge features in the continuous convolution process. Therefore, in the background branch, we introduce reverse attention to mine valid edge information. This process is written as follows:

$$f_{fore} = f_{sa} \otimes f_{ca}, \tag{8}$$

$$f_{back} = (1 - f_{sa}) \otimes f_{ca}, \tag{9}$$

where f_{fore} denotes foreground features, and f_{back} denotes background features. $\otimes$ represents element-wise multiplication.

3.3.2. Global Branch

The foreground and background branches pay more attention to the local features of images; thus, we add a global branch to obtain more global features. Specifically, we

apply global average pooling (*GAP*) on $\boldsymbol{D}_i$ to integrate the global information and reduce redundant parameters. It also utilizes a 1×1 convolution for feature smoothing. Then, we reconstruct the global feature to the same size as the difference, $\boldsymbol{D}_i$, by bilinear interpolation upsampling. This rough operation loses detailed information, but the reconstructed features reflect the overall characteristics of original features. Next, the reconstructed features are fed into the SA and residually multiplied with $\boldsymbol{D}_i$ to obtain the final global feature, f_{global}. The entire process is formulated as follows:

$$f_{global} = [SA(up(Conv_{1\times1}(GAP(\boldsymbol{D}_i))))] \otimes \boldsymbol{D}_i, \tag{10}$$

where $Conv_{1\times1}(\cdot)$ is the 1×1 convolution layer, and $up(\cdot)$ is the upsampling operation.

3.3.3. Feature Fusion

We obtain three kinds of features (i.e., f_{fore}, f_{back}, and f_{global}) and further reshape them using a 3×3 convolution layer. Then, we aggregate the reshaped features by concatenating along the channel dimension. We also apply a skip connection to retain the original features and generate the output features of $\boldsymbol{F}_i \in \mathbb{R}^{C\times H\times W}$. The entire process is written as follows:

$$\boldsymbol{F}_i = \left[Conv_{3\times3}\left(Conv_{3\times3}\left(f_{fore}\right) \oplus Conv_{3\times3}(f_{back}) \oplus Conv_{3\times3}\left(f_{global}\right)\right)\right] \oplus \boldsymbol{D}_i, \tag{11}$$

where $Conv_{3\times3}(\cdot)$ is 3×3 convolutional layer.

3.4. Multi-Task Prediction Decoder

The single-task methods use two decoders to directly generate two semantic change maps, which cause confusion between the change and semantic information. This normally reduces CD detection accuracy. Therefore, we design the decoder with multi-task prediction branches that include two semantic decoders for predicting the semantic maps of bi-temporal images and a change decoder for predicting binary change maps, as shown on the right side of Figure 1. Each of the three decoders contains four stages that receive features from the previous stages, and the semantic decoders integrate semantic feature maps $\boldsymbol{E}^i{}_{T1}$ and $\boldsymbol{E}^i{}_{T2}$ (i = 1, 2, 3, 4, 5). The change decoder integrates the multi-scale enhanced features $\boldsymbol{F}_i$ (i = 2, 3, 4, 5) through skip connections. Each stage consists of a bilinear upsampling layer and a 3×3 convolutional layer. At each stage of the decoder, the channel dimension is halved, while the feature resolution is doubled until the original image resolution is restored. For example, input images $T1$ and $T2$ are both $256 \times 256 \times 3$, and the feature maps in the decoder are $32 \times 32 \times 256$, $64 \times 64 \times 128$, $128 \times 128 \times 64$, and $256 \times 256 \times 32$. Then, two semantic maps (S_1 and S_2) and a BCD map (B) are generated by the classification layer. Finally, P_1 and P_2 are obtained by masking S_1 and S_2 with B. The calculations are as follows:

$$P_1, P_2 = B \cdot (S_1, S_2), \tag{12}$$

$$\boldsymbol{S}^i_{T1} = up(Conv_{3\times3}(\boldsymbol{E}^{i-1}_{T1})) + \boldsymbol{E}^i_{T1}, \tag{13}$$

$$\boldsymbol{S}^i_{T2} = up(Conv_{3\times3}(\boldsymbol{E}^{i-1}_{T2})) + \boldsymbol{E}^i_{T2}, \tag{14}$$

$$\boldsymbol{B}_i = up(Conv_{3\times3}(\boldsymbol{F}_{i-1})) + \boldsymbol{F}_i, \tag{15}$$

where $\boldsymbol{S}^i_{T1}$ and $S^i_{\boldsymbol{T2}}$ represent the feature maps of each stage of the two semantic decoders; $\boldsymbol{B}_i$ represents the feature map of each stage of the change decoder, i = 2, 3, 4, 5.

Only the change loss is used to train multi-task prediction; hence, semantic information is lacking for training supervision, which negatively affects network performance. We use the multi-task loss function ($\mathcal{L}_{mul}$), including the semantic losses, $\mathcal{L}_{sem1}$ and $\mathcal{L}_{sem2}$, and binary change loss, $\mathcal{L}_c$, to jointly guide model training. The calculation is as follows:

$$\mathcal{L}_{mul} = \alpha(\mathcal{L}_{sem1} + \mathcal{L}_{sem2}) + (1 - \alpha)\mathcal{L}_c, \tag{16}$$

where α is used to adjust the effect of different loss functions for the network. The semantic losses, $\mathcal{L}_{sem1}$ and $\mathcal{L}_{sem2}$, are the multi-class cross-entropy loss between the semantic segmentation results, S_1 and S_2, and the ground truth (GT) semantic change maps, $\mathcal{L}_1$ and $\mathcal{L}_2$. The no-change class (index "0") is excluded from the loss calculation. The calculation of $\mathcal{L}_{sem}$ on each pixel is as follows:

$$\mathcal{L}_{sem} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i), \tag{17}$$

where N is the number of semantic classes, and y_i and p_i denote the GT label and the predicted probability of the i-th class, respectively.

A severe class imbalance problem occurs with BCD, in which the number of unchanged pixels is much larger than the number of changed pixels. To alleviate this, binary change loss $\mathcal{L}_c$ combines binary cross-entropy and dice loss to jointly supervise the BCD. The formula is as follows:

$$\mathcal{L}_c = -(y_b \log(\hat{y}_b) + (1 - y_b)\log(1 - \hat{y}_b)) + 1 - \frac{2 \times y_b \times softmax(\hat{y}_b)}{y_b + softmax(\hat{y}_b)}, \tag{18}$$

where y_b is the GT label, which is obtained by replacing non-zero labels with "1" as a change label. $\hat{y}_b$ is the predicted probability of binary change.

4. Experimental Data and Evaluation Indices

4.1. Datasets

The SECOND [36] is a high-resolution dataset collected by several aerial platforms and sensors for remote sensing image SCD. Among the 4662 pairs of temporal images, 2968 are openly available, covering several cities (e.g., Hangzhou, Chengdu, and Shanghai). Each pair of data provides original images and corresponding semantic change labels. Each image has a fixed size of 512 × 512 pixels with the spatial resolution varying from 0.3 to 5 m. The dataset holds six land-cover classes, including non-vegetated ground surface, tree, low vegetation, water, building, and playground, and involves 30 common change categories. Figure 4 presents sample images from the SECOND. As can be seen, no-changed pixels account for more than 80% of the total, whereas the 30 change categories only take up small proportions, which poses a huge imbalance challenge to the SCD method.

The Landsat-SCD dataset [39] is made up of Landsat images collected between 1990 and 2020. The observation area is Tumshuk, Xinjiang, China. The dataset consists of 8468 pairs of images, each having a fixed size of 416 × 416 pixels with a resolution of 30 m. The dataset relates a no-change class and four land-cover classes, including farmland, desert, buildings, and water. Figure 5 shows sample images from the Landsat-SCD dataset. The dataset contains many complex detection scenes, where the buildings are small and scattered. Changed pixels account for about 19% of the total, which provides a realistic evaluation dataset for SCD methods.

4.2. Evaluation Metrics

In this paper, four evaluation metrics are utilized to assess the performance of different methods, including overall accuracy (OA), mean Intersection over Union (mIoU), separated Kappa (κ) coefficient (SeK), and a comprehensive score (Score). OA reflects the proportion of correctly classified samples to all samples, defined as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} \tag{19}$$

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

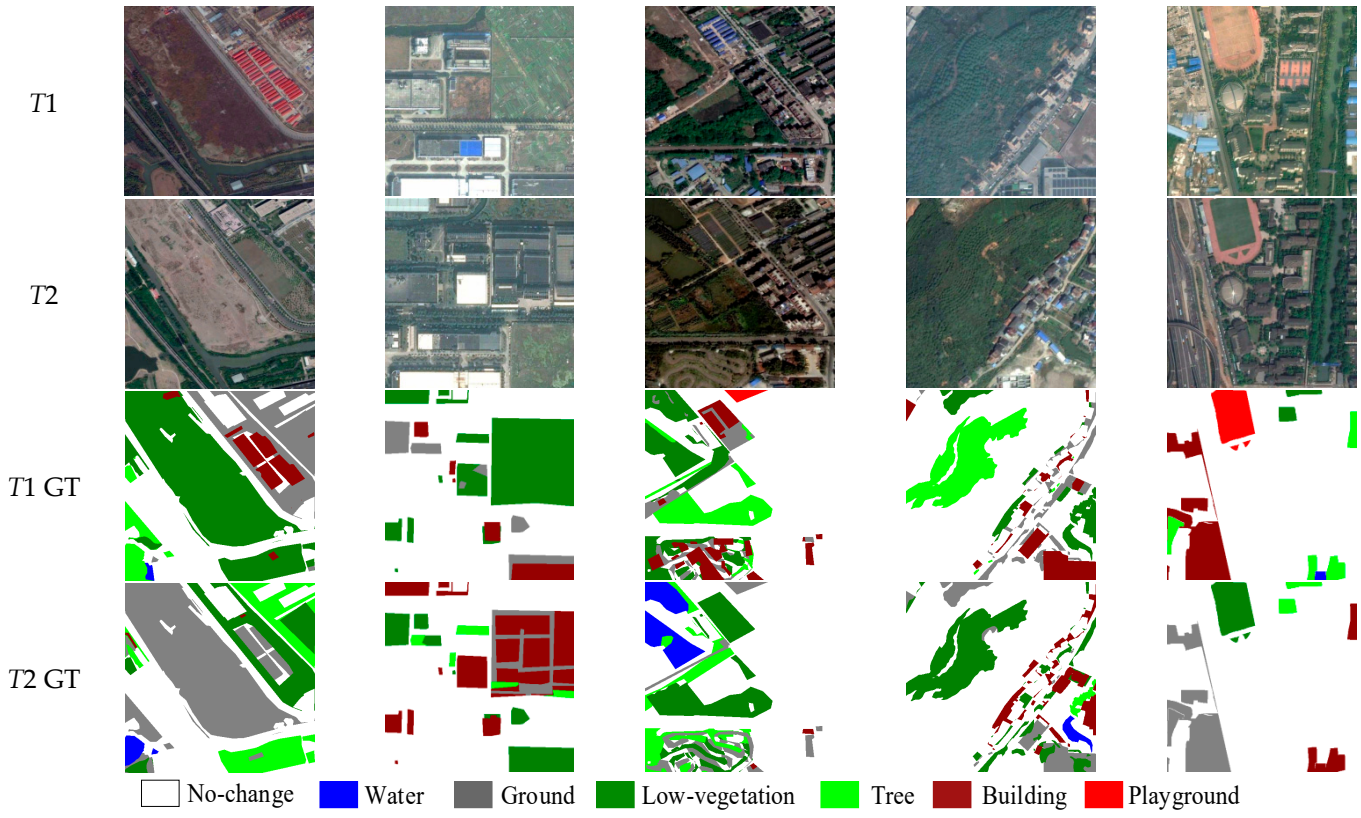

Figure 4. Example images from the SECOND Dataset.

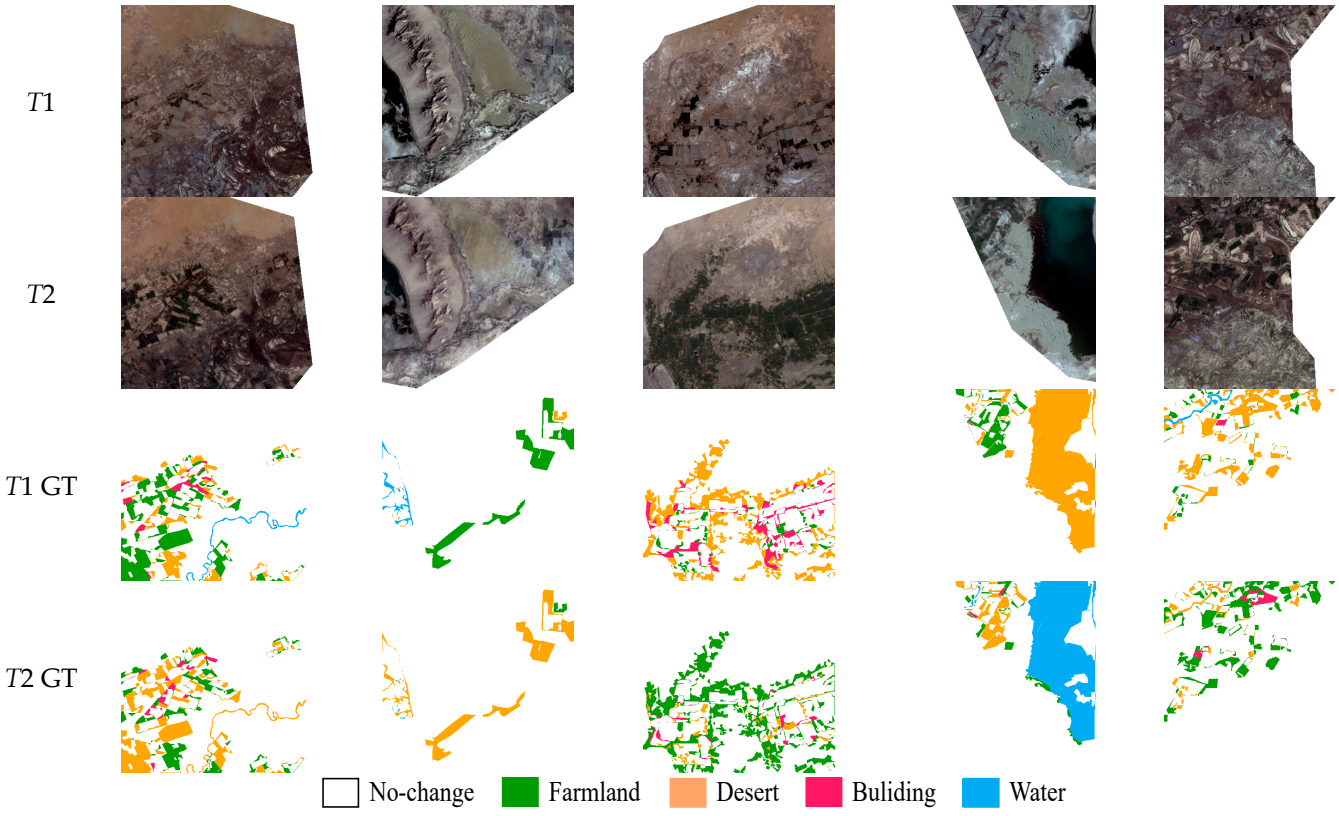

Figure 5. Example images from the Landsat-SCD Dataset.

Owing to the obvious imbalance of positive and negative samples, the OA calculation can easily be dominated by negative samples; hence, it will fail to provide a reasonable perspective of full-task accuracy. Thus, we turn to mIoU and SeK. The former is used to evaluate SCD results from the BCD perspective, and the latter takes the SCD perspective. mIoU is the mean value of the IoU of no-change pixels (IoU_1) and changed pixels (IoU_2):

$$IoU_1 = \frac{TN}{TN + FP + FN}, \tag{20}$$

$$IoU_2 = \frac{TP}{TP + FN + FP}, \tag{21}$$

$$mIoU = \frac{IoU_1 + IoU_2}{2}, \tag{22}$$

SeK is the combination of IoU_2 and the new κ after unconsidered true predictions of non-changed pixels. SeK is calculated as follows:

$$SeK = \kappa \times e^{IoU_2 - 1}, \tag{23}$$

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{24}$$

$$p_0 = \frac{\sum_{i=0}^{N} q_{ii}}{\sum_{i=0}^{N} \sum_{j=0}^{N} q_{ij}}, \tag{25}$$

$$p_e = \frac{\sum_{i=0}^{N} (q_{i+} \times q_{+i})}{\left(\sum_{i=0}^{N} \sum_{j=0}^{N} q_{ij}\right)^2}, \tag{26}$$

where $Q = \{q_{ij}, 0 \le i \le N, 0 \le j \le N\}$ is the confusion matrix, in which "0" represents the unchanged class, and N represents the number of categories. q_{i+} denotes the row sum of Q, and q_{+i} denotes the column sum. Based on mIoU and Sek, Score can be calculated as [37]:

$$Score = 0.3 \times mIoU + 0.7 \times SeK. \tag{27}$$

4.3. Training Details

All methods are implemented with PyTorch on Linux and trained using two NVIDIA RTX 2080Ti GPUs. During training, the same experimental parameters are used in all experiments. We use stochastic gradient descent to train the network with an initial learning rate of 0.01 and a weight decay of 0.9. The batch size and α are set to 4 and 0.5, respectively. We further split each dataset into training, validation, and testing sets randomly at a ratio of 7:1:2. To make better use of the GPU for training, we uniformly resize image patches to 256 × 256 and train the framework with 50 epochs. To improve robustness, data augmentation is performed via random flipping and rotating of the input images.

5. Results

5.1. Comparison Experiments

To verify the superiority of SMNet on SCD tasks, several excellent remote sensing image CD models were compared:

- FC-Siam-conc [12]: A fully convolutional Siamese network that fuses bi-temporal features through skip-connections for CD.
- FC-Siam-diff [12]: A fully convolutional Siamese network that utilizes multi-layer difference features to fuse bi-temporal information.
- DSIFN [40]: A deeply supervised differential network that generates change maps using multi-scale feature fusion.

- HRSCD-str3 [35]: A network that introduces temporal correlation information by constructing a BCD branch.
- HRSCD-str4 [35]: A Siamese network that designs a skip operation to connect Siamese encoders with the decoder of the CD branch.
- BiSRNet [41]: A bi-temporal semantic reasoning (SR) network that applies Siamese and cross-temporal SR to enhance information exchange between temporal and change branches.
- FCCDN [24]: A feature constraint CD network based on a dual encoder–decoder that uses a non-local feature pyramid network to extract and fuse multi-scale features and proposes a densely connected feature fusion module to enhance robustness.
- BIT [42]: A network that combines a CNN and transformer learns a compact set of tokens to represent high-level concepts that reveal change of interest in bi-temporal images. The transformer finds the relationship between semantic concepts in the token-based space-time.

The quantitative analysis of experimental results on the SECOND is shown in Table 1, indicating that our method delivers excellent performance. Specifically, SMNet achieves the best mIoU, SeK, and OA values of 71.95%, 20.29%, and 86.68%, respectively. Compared with the BIT, SMNet shows consistent improvements for all evaluation metrics, demonstrating that the PR and TB combination greatly assists feature extraction, resulting in the acquisition of much richer global information. Our multi-task method shares some of its obvious advantages with BiSRNet and FCCDN over the single-task methods, such as FC-Siam-conc, FC-Siam-diff, and DSIFN. The main reason may be that semantic and change information interferes with each other in the single-task method. The experimental results on the Landsat-SCD are shown in Table 2, where the proposed method achieves the best results on each evaluation metric. In particular, our method is 7.16% higher compared to BIT in SeK.

Table 1. Comparison results on SECOND dataset.

Method	mIoU/%	SeK/%	Score/%	OA/%
FC-Siam-conc	66.38	10.60	27.33	84.11
FC-Siam-diff	66.65	9.57	26.69	84.01
DSIFN	66.28	9.37	26.44	83.20
HRSCD-str3	65.85	11.89	28.08	82.85
HRSCD-str4	69.69	16.91	32.74	84.81
BiSRNet	69.13	15.08	31.30	83.90
FCCDN	69.38	17.06	32.75	86.44
BIT	70.17	18.06	33.69	86.01
SMNet (ours)	71.95	20.29	35.79	86.68

Table 2. Comparison results on Landsat-SCD dataset.

Method	mIoU/%	SeK/%	Score/%	OA/%
FC-Siam-conc	64.66	5.70	23.39	80.09
FC-Siam-diff	69.22	10.25	27.94	84.25
DSIFN	70.37	20.67	35.58	85.66
HRSCD-str3	78.23	32.70	46.36	90.65
HRSCD-str4	80.33	36.68	49.78	91.79
BiSRNet	80.44	37.65	50.49	92.16
FCCDN	77.10	29.72	43.94	90.29
BIT	82.60	43.98	55.56	93.45
SMNet (ours)	85.65	51.14	61.49	94.53

Figure 6 visualizes partial results on the SECOND. Since the multi-task methods decouple the SCD task into two sample sub-tasks of se. Since the multi-task methods decouple the SCD task into two sample sub-tasks of semantic segmentation and BCD, which largely reduces the output space. As a result, multi-task methods are more accurate and

complete than single-task methods (e.g., FC-Siam-conc). In complex scenes, as shown in the first and second groups in Figure 5, there are a large number of false and missed detection in compared methods due to the relatively similar spectral characteristics between low vegetation and trees in the images. However, the proposed method effectively establishes context information and clearly distinguishes low vegetation and trees by capturing richer global information. In the case of irregular changes at multiple scales, as shown in the third and fourth groups in Figure 5, FC-Siam-conc, BiSRNet, FCCDN, and BIT all have missed detection at different degrees. However, SMNet demonstrated higher adaptability to multi-scale change regions and produced smoother edges by distinguishing foreground and background information. Since the resolution of the Landsat-SCD dataset is relatively low, it poses a greater challenge to the performance of the model. In this dataset, the proposed model accurately identifies fine-grained changes that are easily missed by compared models, such as the drying of rivers and the change of small farmland and buildings in Figure 7.

Figure 8 presents the intermediate results of the BIT and the proposed method on the SECOND. It can be seen that BIT shows obvious misjudgment in the recognition of semantic types, which may be caused by its transformer receiving only the lowest semantic features from the CNN, resulting in inadequate semantic information. The proposed method not only effectively reduces the response to unrelated information and improves the classification ability of different objects via the PRTB backbone, but it also successfully leveraged the MCFM module to accurately detect change regions and improve boundary expressability. It is worth noting that in the case of extreme label imbalances, water is detected more accurately by SMNet, which demonstrates that the proposed method achieves excellent performance for remote sensing image SCD.

5.2. Ablation Experiments

To validate the contributions of the key components of our framework, four sets of ablation experiments were performed on the SECOND. In this paper, we use the backbone network composed of the P-Res block as the baseline. Each module is joined to the network framework individually for performance evaluation on basis of the baseline. The results are listed in Table 2.

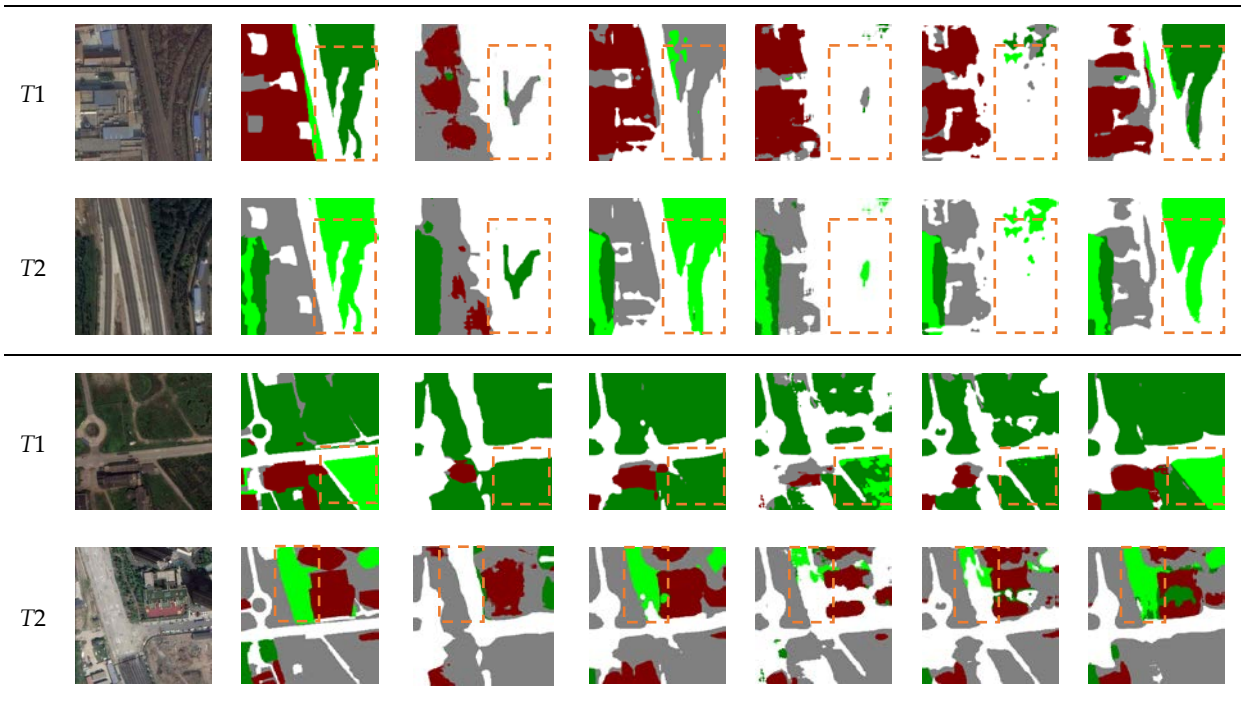

Figure 6. *Cont.*

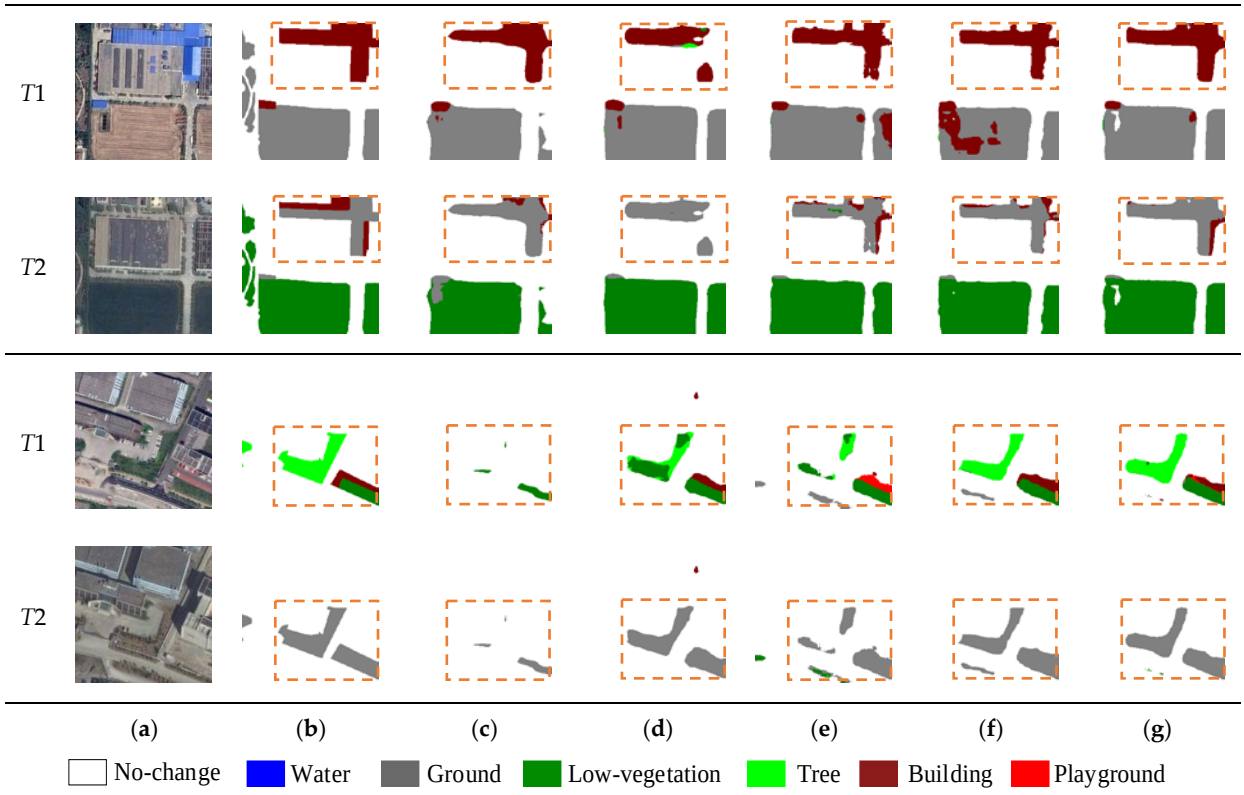

Figure 6. Comparisons of the change maps obtained by different methods on the SECOND dataset. (**a**) images, (**b**) GT, (**c**) FC-Siam-conc, (**d**) BiSRNet, (**e**) FCCDN, (**f**) BIT, and (**g**) SMNet (ours). We highlight interesting regions with orange boxes.

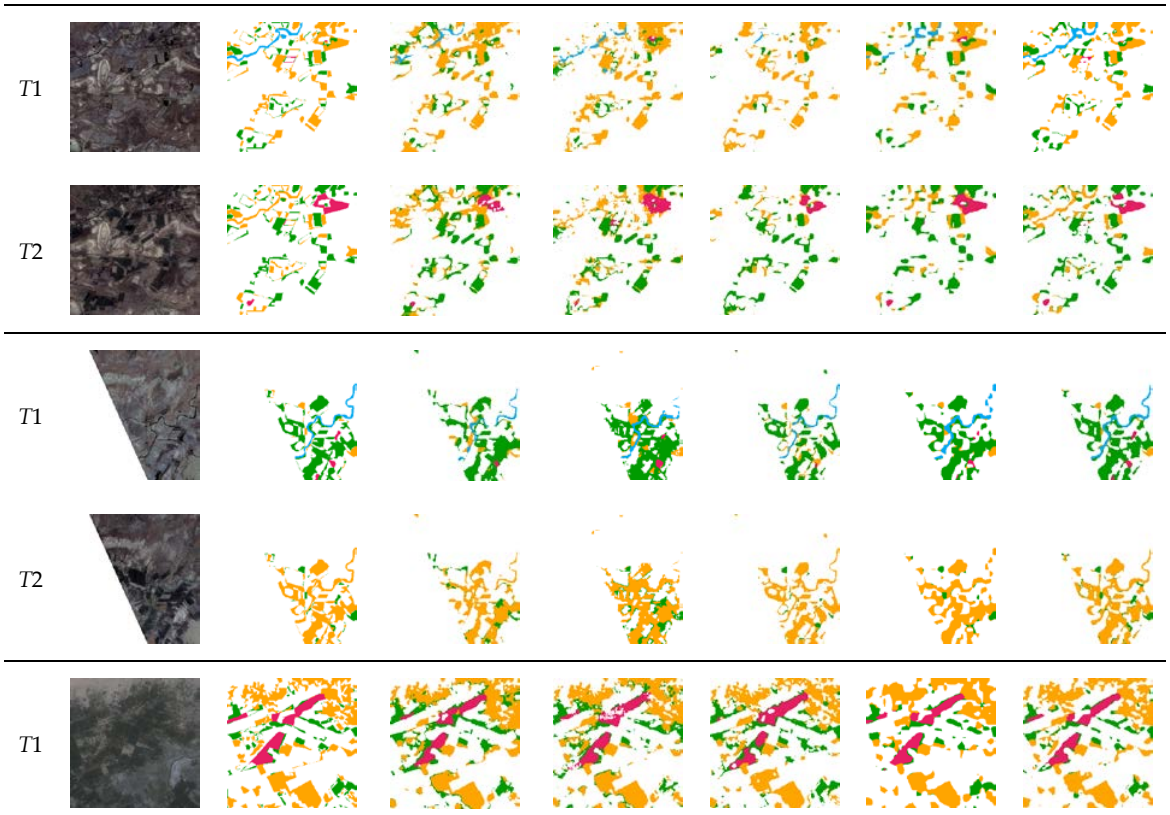

Figure 7. *Cont.*

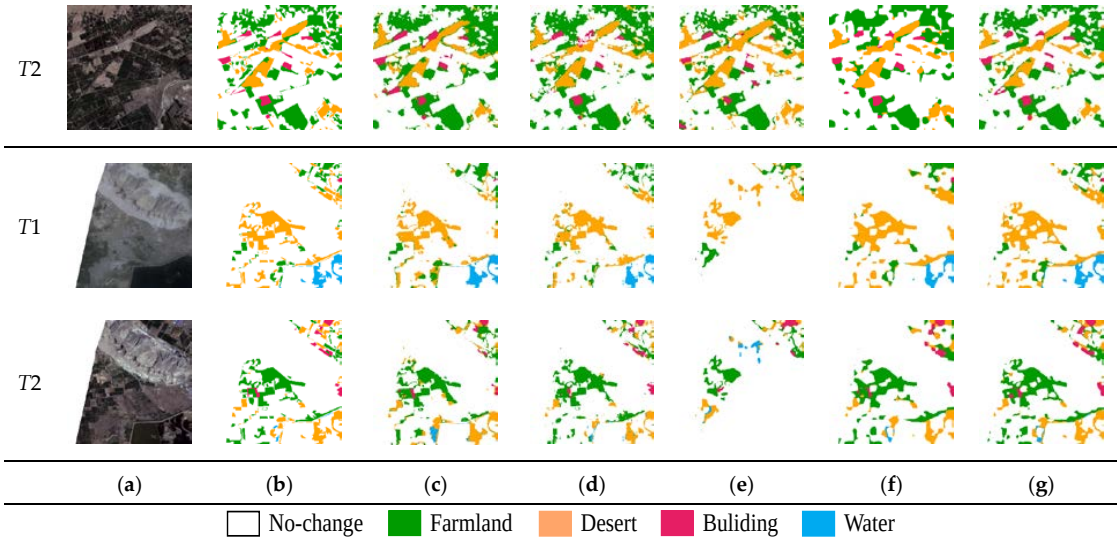

Figure 7. Comparisons of the change maps obtained by different methods on the Landsat-SCD dataset. (**a**) images, (**b**) GT, (**c**) HRSCD-str4, (**d**) BiSRNet, (**e**) FCCDN, (**f**) BIT, and (**g**) SMNet (ours).

As shown in rows 1 and 2 of Table 3, compared with the baseline, the addition of the multi-task loss achieves a gain of 0.34% and 0.73% for mIoU and Sek, respectively. This indicates that the multi-task loss improves semantic expressability via co-supervised model training. The third row of Table 2 shows that the addition of the PRTB backbone achieves significant improvements in all metrics, which is due to that our PRTB backbone considers more global information and context information of the model. From the last row of Table 2, introducing the MCFM increases the mIoU value from 71.50% to 71.95% and the Sek value from 19.45% to 20.29%, respectively. This illustrates that the MCFM is effective in highlighting the boundaries of change areas. Some of the visualization results are shown in Figure 9, which illustrates the improvements in terms of fewer missed and false detection. In particular, the ability to distinguish ground, low vegetation, and tree categories have been significantly increased.

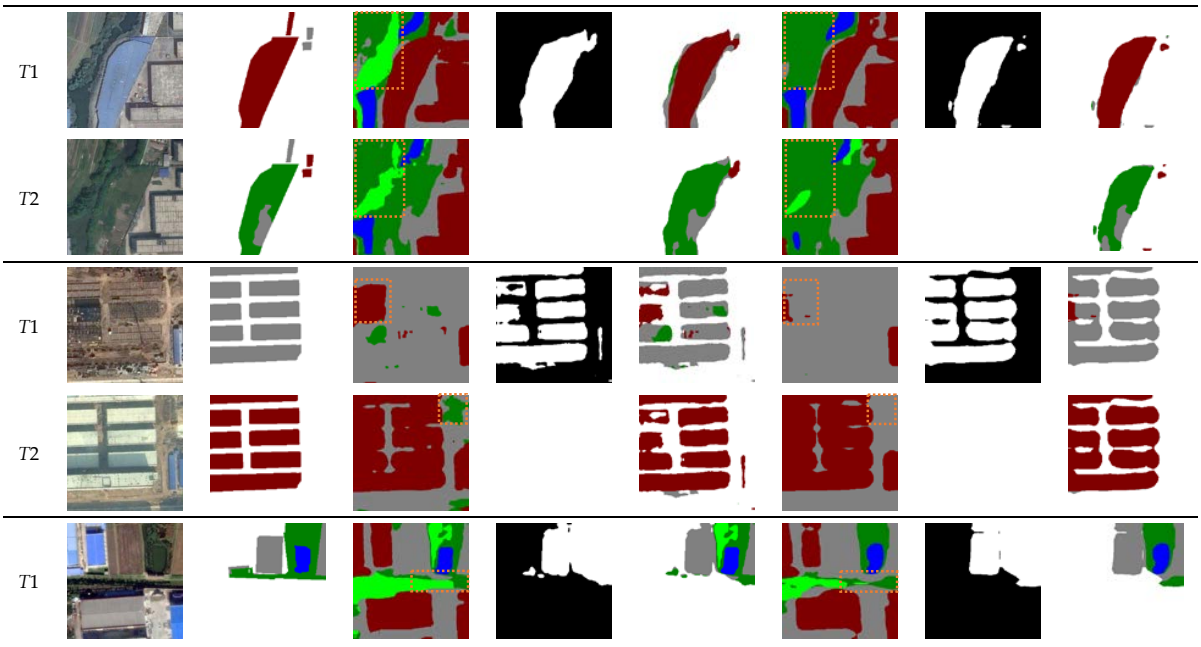

Figure 8. *Cont.*

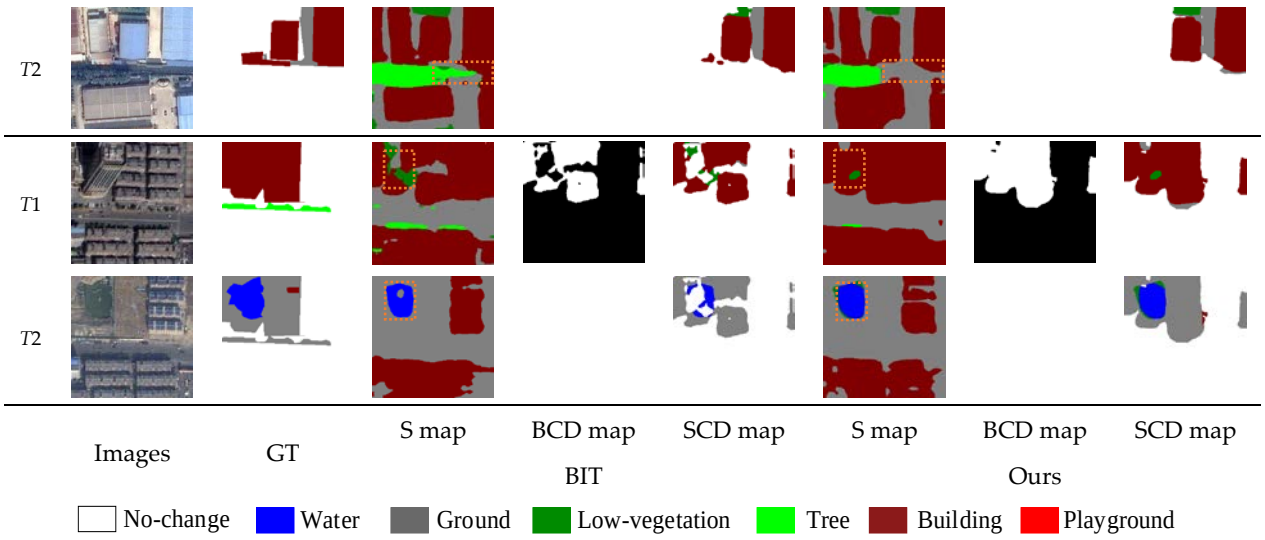

Figure 8. Comparisons of the results provided by the BIT and Ours. We highlight interesting regions with orange boxes. *Definitions*: S-map, semantic map.

Table 3. Ablation study on the SECOND dataset.

Model	mIoU/%	SeK/%	Score/%	OA/%
Base	70.28	17.10	33.06	85.83
Base + $\mathcal{L}_{mul}$	70.62	17.83	33.67	85.38
Base + PRTB	70.64	18.74	34.31	86.25
Base + $\mathcal{L}_{mul}$ + PRTB	71.50	19.45	35.10	86.44
Base + $\mathcal{L}_{mul}$ + PRTB + MCFM	71.95	20.29	35.79	86.68

In addition, the performance exhibited by different layer transformer blocks on the SECOND dataset is discussed in this paper, and the results are shown in Table 4. It can be seen that too many transformer blocks decrease the detection effectiveness of the network. When the number of n-layers is set to (1, 2, 4, 2, 1), the best detection results are obtained at mIoU and SeK, which reduces redundant computation.

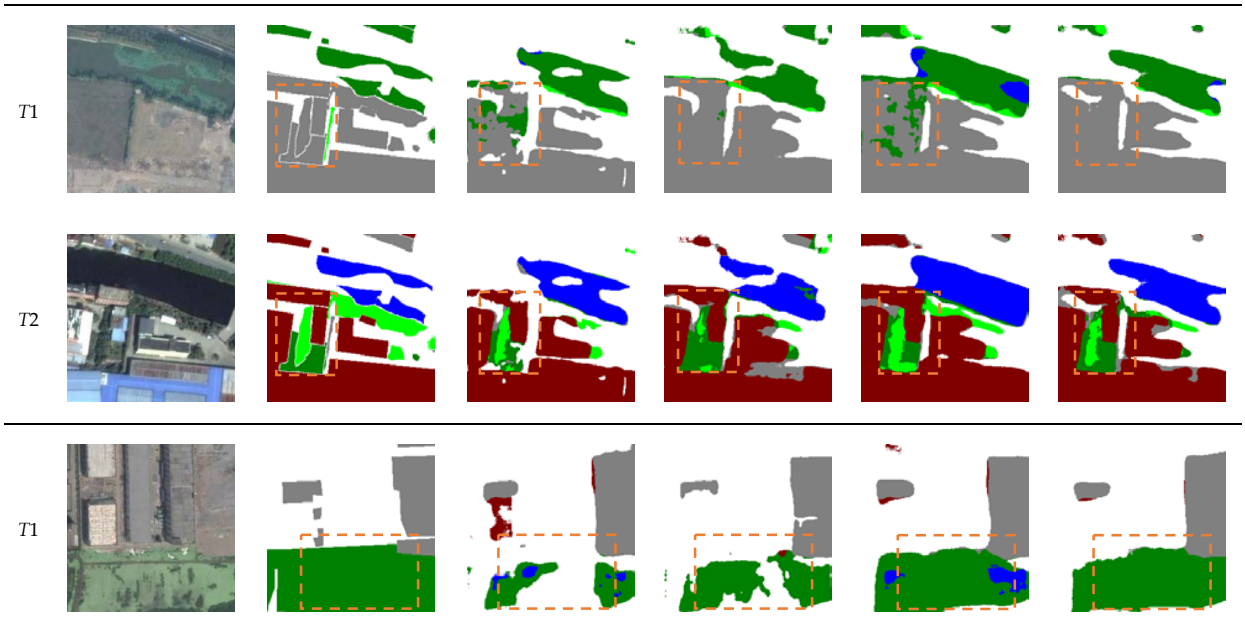

Figure 9. *Cont*.

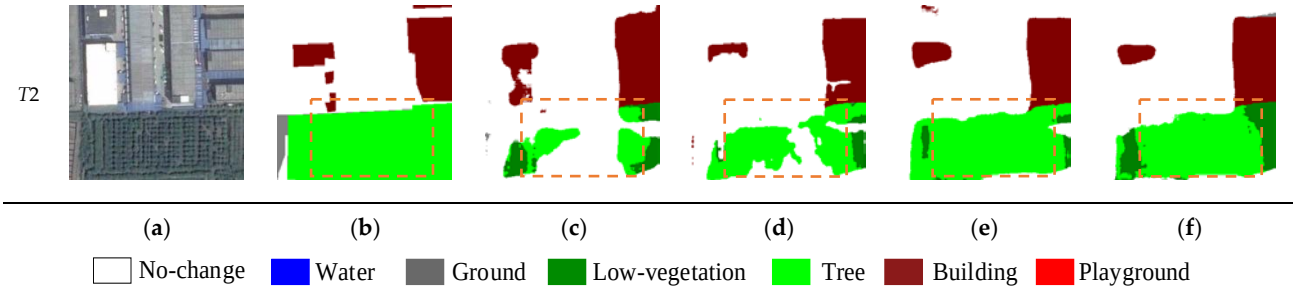

Figure 9. Visual comparisons of ablation experiment on the SECOND dataset. (**a**) images, (**b**) GT, (**c**) baseline model, (**d**) base + $\mathcal{L}_{mul}$, (**e**) base + $\mathcal{L}_{mul}$ + PRTB, and (**f**) base + $\mathcal{L}_{mul}$ + PRTB + MCFM. We highlight interesting regions with orange boxes.

Table 4. The effect of the *n*-layer transformer block on the SECOND dataset.

Layers	mIoU/%	SeK/%	Score/%	OA/%	Flops/G	Params/M
(1,1,1,1,1)	70.54	18.17	33.91	85.86	67.63	36.87
(2,2,2,2,2)	70.62	18.59	34.17	86.30	70.35	38.63
(1,2,4,2,1)	71.95	20.29	35.79	86.68	71.05	37.48
(4,4,4,4,4)	70.72	18.66	34.28	86.50	75.79	42.16

The loss function provides effective supervision information for network training, which has a significant impact on network performance. In this paper, α is introduced in the multi-task loss to balance the loss values between the semantic and change branches. In the process of training, the larger value of α indicates more supervision information from semantic branches to the model. To analyze the sensitivity of α, we chose four sets of values from 0 to 0.8 for our experiments. Figure 10 shows the effect of different α values in the multi-task loss on each evaluation metric. It can be seen that all curves of the SECOND dataset reach the best value when α = 0.5, which indicates that semantic loss and binary change loss play equal roles in the co-supervision network training. However, all evaluation metrics have the lowest value when α = 0.8, which illustrates that semantic supervision is greatly important for remote sensing image SCD.

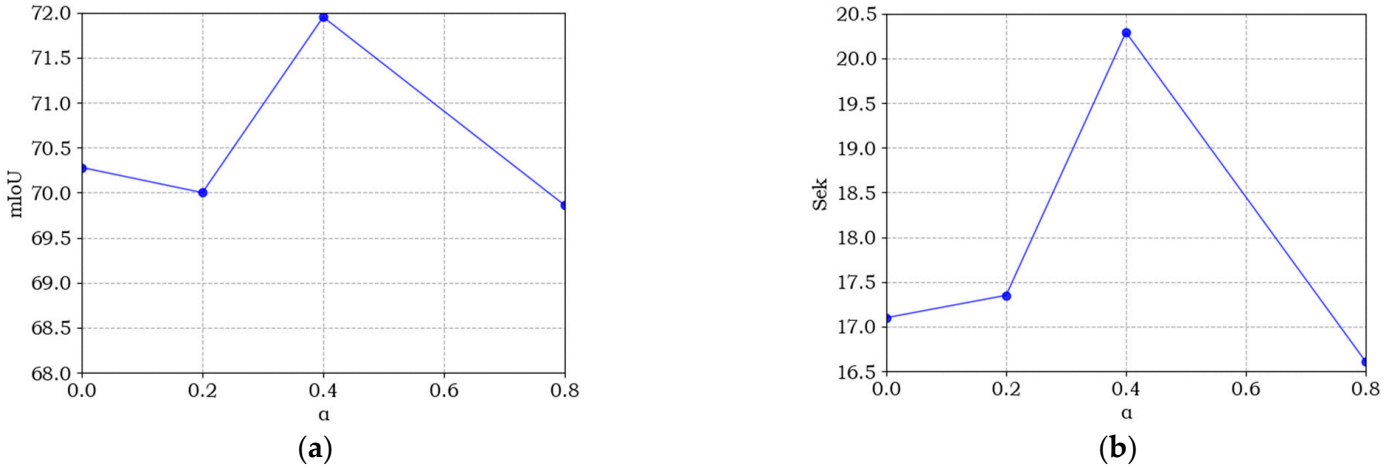

Figure 10. *Cont.*

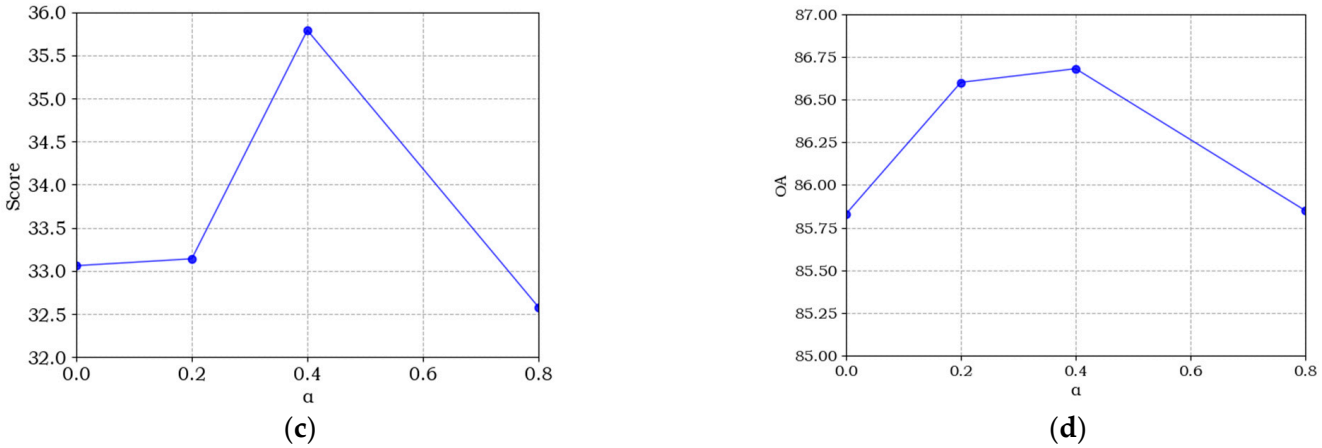

Figure 10. Sensitive analysis of α for the multi-task loss. (**a**–**d**) denote the accuracy curve of mIoU, Sek, Score, and OA, respectively.

6. Conclusions

SCD is a meaningful and challenging task in remote sensing field. It requires the full exploitation of the semantic and change features. In this paper, we propose SMNet, a novel CNN and transformer implementation model for remote sensing image SCD. It can output semantic and change information end-to-end. Based on an encoder–decoder architecture, we built the PRTB backbone, which uses the combination of CNN and transformer to extract image features with more global information. This improves the semantic expressability of our network. Then, we introduce the MCFM, which enhances the sensitivity of the model to change regions in complex backgrounds by fusing foreground, background, and global information. Finally, we use the multi-task loss to jointly guide network training, further improving the detection efficacy of SMNet. Comparative experimental results on the SECOND and Landsat-SCD datasets show that our new framework shows better accuracy. Ablation experiments then clearly demonstrate the necessity of each module for the detection results.

The SECOND and Landsat-SCD datasets are currently available datasets for remote sensing image SCD tasks with pixel-level annotations of large-scale images. The creation of such datasets is extremely time-consuming and laborious. Therefore, we plan to consider unsupervised or semi-supervised methods in the future to capture semantic change information through self-learning and achieve SCD tasks with comparable or even higher accuracy.

Author Contributions: Conceptualization, H.G. and J.L.; methodology, Y.N.; software, Y.N.; formal analysis, D.Y.; writing—original draft preparation, Y.N.; writing—review and editing, H.G., D.Y. and L.D.; funding acquisition, L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science Foundation of China, grant number 42201443.

Data Availability Statement: The datasets in our paper are public and available directly online.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Song, X.P.; Hansen, M.C.; Stehman, S.V.; Potapov, P.V.; Tyukavina, A.; Vermote, E.F.; Townshend, J.R. Global land change from 1982 to 2016. *Nature* **2018**, *560*, 639–643. [CrossRef]
2. Huang, X.; Schneider, A.; Friedl, M.A. Mapping sub-pixel urban expansion in China using Modis and DMSP/OLS nighttime lights. *Remote Sens. Environ.* **2016**, *175*, 92–108. [CrossRef]
3. Jin, S.; Yang, L.; Zhu, Z.; Homer, C. A land cover change detection and classification protocol for updating Alaska NLCD 2001 to 2011. *Remote Sens. Environ.* **2017**, *195*, 44–55. [CrossRef]

4. Huang, B.; Zhao, B.; Song, Y. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote-sensing imagery. *Remote Sens. Environ.* **2018**, *214*, 73–86. [CrossRef]
5. Zhang, C.; Sargent, I.; Pan, X.; Li, H.P.; Gardiner, A.; Hare, J.; Atkinson, P.M. Joint deep learning for land cover and land use classification. *Remote Sens. Environ.* **2019**, *221*, 173–187. [CrossRef]
6. Martins, V.S.; Kaleita, A.L.; Gelder, B.K.; da Silveira, H.L.; Abe, C.A. Exploring multiscale object-based convolutional neural network (multi-OCNN) for remote-sensing image classification at high spatial resolution. *ISPRS J. Photogramm.* **2020**, *168*, 56–73. [CrossRef]
7. Huang, W.; Zhao, Z.B.; Sun, L.; Ju, M. Dual-branch attention-assisted CNN for hyperspectral image classification. *Remote Sens.* **2022**, *14*, 6158. [CrossRef]
8. Wang, J.; Gong, Z.; Liu, X.; Guo, H.; Yu, D.; Ding, L. Object detection based on adaptive feature-aware method in optical remote sensing images. *Remote Sens.* **2022**, *14*, 3616. [CrossRef]
9. Dong, X.; Qin, Y.; Gao, Y.; Fu, R.; Liu, S.; Ye, Y. Attention-based multi-level feature fusion for object detection in remote sensing images. *Remote Sens.* **2022**, *14*, 3735. [CrossRef]
10. Dong, H.; Yu, B.; Wu, W.; He, C. Enhanced lightweight end-to-end semantic segmentation for high-resolution remote sensing images. *IEEE Access* **2022**, *10*, 70947–70954. [CrossRef]
11. Xiong, J.; Po, L.M.; Yu, W.Y.; Zhou, C.; Xian, P.; Ou, W. CSRNet: Cascaded selective resolution network for real-time semantic segmentation. *Expert Sys. Applic.* **2021**, *211*, 118537.
12. Daudt, R.C.; Le Saux, B.L.; Boulch, A. Fully convolutional Siamese networks for change detection. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067.
13. Zhou, Z.W.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J.M. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. [CrossRef]
14. Liu, R.; Jiang, D.; Zhang, L.; Zhang, Z. Deep Depthwise separable convolutional network for change detection in optical aerial images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1109–1118. [CrossRef]
15. Zhang, Y.; Fu, L.; Li, Y.; Zhang, Y. HDFNet: Hierarchical dynamic fusion network for change detection in optical aerial images. *Remote Sens.* **2021**, *13*, 1440. [CrossRef]
16. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
17. Ling, J.; Hu, L.; Cheng, L.; Chen, M.H.; Yang, X. IRA-MRSNet: A network model for change detection in high-resolution remote sensing images. *Remote Sens.* **2022**, *14*, 5598.
18. Chen, J.; Yuan, Z.Y.; Peng, J.; Chen, L.; Huang, H.Z.; Zhu, J.W.; Liu, Y.; Li, H. DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 1194–1206. [CrossRef]
19. Peng, X.; Zhong, R.; Li, Z.; Li, Q. Optical remote-sensing image change detection based on attention mechanism and image difference. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7296–7307. [CrossRef]
20. Guo, E.Q.; Fu, X.S.; Zhu, J.W.; Deng, M.; Liu, Y.; Zhu, Q.; Li, H.F. Learning to measure change: Fully convolutional Siamese metric networks for scene change detection. *arXiv* **2018**, arXiv:1810.09111.
21. Zhu, Q.; Guo, X.; Deng, W.; Shi, S.; Guan, Q.; Zhong, Y.; Zhang, L.; Li, D. Land-use/land-cover change detection based on a Siamese global learning framework for high spatial resolution remote-sensing imagery. *ISPRS J. Photogramm.* **2022**, *184*, 63–78. [CrossRef]
22. Gao, Y.; Zhou, M.; Metaxas, D.N. UTNet: A hybrid transformer architecture for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; pp. 61–71.
23. Li, G.Y.; Liu, Z.; Lin, W.S.; Ling, H.B. Multi-content complementation network for salient object detection in optical remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13.
24. Chen, P.; Zhang, B.; Hong, D.F.; Chen, Z.C.; Yang, X.; Li, B. FCCDN: Feature constraint network for VHR image change detection. *ISPRS J. Photogramm.* **2022**, *187*, 101–119. [CrossRef]
25. Lv, Z.Y.; Wang, F.J.; Cui, G.Q.; Benediktsson, J.A.; Lei, T.; Sun, W. Spatial-spectral attention network guided with change magnitude image for land cover change detection using remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
26. Lei, T.; Xue, D.; Ning, H.; Yang, S.; Lv, Z.; Nandi, A.K. Local and global feature learning with kernel scale-adaptive attention network for VHR remote sensing change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 7308–7322. [CrossRef]
27. Wei, H.; Chen, R.; Yu, C.; Yang, H.; An, S. BASNet: A boundary-aware Siamese network for accurate remote-sensing change detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
28. Liu, M.X.; Shi, Q.; Andrea, M.; He, D.; Liu, X.P.; Zhang, L.P. Super-resolution-based change detection network with stacked attention module for images with different resolutions. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18.
29. Peng, D.; Bruzzone, L.; Zhang, Y.J.; Guan, H.Y.; Ding, H.Y.; Huang, X. SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5891–5906. [CrossRef]

30. Tsutsui, S.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Semantic segmentation and change detection by multi-task U-net. In Proceedings of the IEEE International Conference on Image Processing, Anchorage, AK, USA, 19–22 September 2021; pp. 619–623. [CrossRef]
31. Peng, D.F.; Bruzzone, L.; Zhang, Y.J.; Guan, H.Y.; He, P.F. SCDNET: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102465. [CrossRef]
32. Liu, Y.I.; Pang, C.; Zhan, Z.; Zhang, X.; Yang, X. Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 811–815. [CrossRef]
33. Mou, L.C.; Bruzzone, L.; Zhu, X.X. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 924–935. [CrossRef]
34. Papadomanolaki, M.; Vakalopoulou, M.; Karantzalos, K. A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7651–7668.
35. Daudt, R.C.; Le, S.B.; Boulch, A.; Gousseau, Y. Multitask learning for large-scale semantic change detection. *Comput. Vis. Image Understand.* **2019**, *187*, 102783.
36. Yang, K.P.; Xia, G.S.; Liu, Z.C.; Du, B.; Yang, W.; Pelillo, M.; Zhang, L.P. Asymmetric Siamese networks for semantic change detection in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–18. [CrossRef]
37. Zheng, Z.; Zhong, Y.; Tian, S.; Ma, A.L.; Zhang, L. ChangeMask: Deep multitask encoder-transformer-decoder architecture for semantic change detection. *ISPRS J. Photogramm.* **2022**, *183*, 228–239. [CrossRef]
38. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Identity mappings in deep residual networks, Computer Vision—ECCV 2016. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 1995; Volume 9908. [CrossRef]
39. Yuan, P.; Zhao, Q.; Zhao, X.; Wang, X.; Long, X.; Zheng, Y. A transformer-based Siamese network and an open-optical dataset for semantic-change detection of remote sensing images. *Int. J. Digit. Earth* **2022**, *15*, 1506–1525.
40. Zhang, C.; Yue, P.; Tapete, D.; Jiang, L.; Shangguan, B.; Huang, L.; Liu, G. A deeply supervised image fusion network for change detection in high-resolution bitemporal remote-sensing images. *ISPRS J. Photogramm.* **2020**, *166*, 183–200. [CrossRef]
41. Ding, L.; Guo, H.T.; Liu, S.C.; Mou, L.C.; Zhang, J.; Lorenzo, B. Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14.
42. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14.