Sensors and Materials, Vol. 36, No. 10 (2024) 4409–4425
MYU Tokyo

S & M 3808

Change Detection for High-resolution Remote Sensing Images Based on a Siamese Structured UNet3+ Network

Chen Liang,[1,2*] Yi Zhang,[1,2] Zongxia Xu,[1,2] Yongxin Yu,[1,2] and Zhenwei Zhang[3]

[1]Beijing Institute of Surveying and Mapping, 60 Nanlishi Road, Beijing 100045, China
[2]Beijing Key Laboratory of Urban Spatial Information Engineering, 60 Nanlishi Road, Beijing 100045, China
[3]School of Remote Sensing and Geomatics Engineering,
Nanjing University of Information Science and Technology, No.219, Ningliu Road, Nanjing, China

(Received July 25, 2024; accepted October 10, 2024)

Keywords: change detection, deep learning, UNet3+, Siamese

The use of bi-temporal remote sensing images for detecting changes in land cover is an important means of obtaining surface change information, thus contributing to urban governance and ecological environment monitoring. In this article, we propose a deep learning model named Siam-UNet3+ for high-resolution remote sensing image change detection. This model integrates the full-scale skip connections and full-scale deep supervision of the network UNet3+, which can achieve the multi-scale feature fusion of remote sensing images, effectively avoiding the locality disadvantage of convolution operations. Different from UNet3+, Siam-UNet3+ has made major improvements, including the following: (1) incorporating a Siamese network in the encoder, which can process bi-temporal remote sensing images in parallel; (2) leveraging the residual module as the backbone, which can avoid gradient vanishing (or exploding) and model degradation problems; (3) adding a Triplet Attention module to the decoder, which can avoid information redundancy that may occur in full-scale skip connections and increase the ability to focus on changing patterns; and (4) designing a hybrid loss function consisting of focal loss and dice loss, which is more suitable for remote sensing image change detection tasks. In this study, we conducted change detection experiments using the publicly available LEVIR-CD dataset, as well as two local datasets in Beijing. Through comparative experiments with five other models and ablation experiments, the proposed model Siam-UNet3+ in this article demonstrated significant advantages and improvements in four evaluation metrics, namely, precision, recall, *F*1-score, and overall accuracy (OA), proving to have great potential in the application to high-resolution remote sensing image change detection tasks.

1. Introduction

The interaction between the natural environment and human activities has brought constant changes in land surface cover, especially in large cities where urbanization rapidly occurs. For example, Beijing, the capital of China, presents characteristics of diverse land coverage and

*Corresponding author: e-mail: chenliang90210@whu.edu.cn
https://doi.org/10.18494/SAM5250

ISSN 0914-4935

rapid spatiotemporal changes. Therefore, it is very necessary to carry out the dynamic monitoring of land cover change to meet the needs of urban management and ecological environment monitoring.

Remote sensing imagery has been widely used for monitoring and analyzing changes in land cover and land use patterns over time. The ability to detect and monitor these changes is crucial for various applications, including urban planning, environmental monitoring, land use, and disaster management.[1] Traditional change detection methods often rely on manual interpretation or simple pixel-based techniques, which are time-consuming and limited in their ability to handle complex and large-scale datasets. In recent years, deep learning algorithms have shown considerable potential in many computer vision tasks, including object detection, image classification, and semantic segmentation.[2] The application of deep learning techniques to remote sensing imagery has also attracted increasing attention, as they offer the ability to automatically learn and extract complex spatial and spectral features, facilitating more accurate and efficient change detections.[3] Many deep learning networks have been proposed throughout the years, and different networks have been used for change detection tasks, including autoencoders (AEs), deep belief networks (DBNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), and transformer networks.[4] Among these networks, CNNs are one of the most commonly used in change detection, especially since fully convolutional networks (FCNs) were proposed in 2015.[5] The emergence of FCNs is considered a milestone as it demonstrates how to achieve end-to-end training and learn feature information from input images of any size to achieve semantic segmentation. Despite the powerful functionality and flexibility of FCNs, they still have some drawbacks, such as not considering spatial context information and low computational efficiency for high-resolution image processing. Therefore, scholars have proposed various semantic segmentation models based on FCNs, such as SegNet, PSPNet, and UNet, which are subsequently used for change detection tasks. For example, Daudt *et al.*[6] designed the first end-to-end change detection model and proposed three effective architectures based on FCNs. Wiratama *et al.*[7] proposed a multi-spectral image change detection model based on feature-level UNet, which has a higher accuracy for detecting small changes. Deng *et al.*[8] developed a building change detection model based on UNet++ and EfficientNet-B1 to improve the accuracy and robustness in detecting land cover changes from high-resolution remote sensing images. Chen *et al.*[9] proposed a multi-input channel UNet model based on synthetic aperture radar (SAR) data for landslide detection. However, the convolution operation inherent in CNNs is weak in global information extraction. Consequently, researchers have come up with different approaches to tackle this drawback, such as using dilated convolution[10] and adding an attention mechanism.[11] Also, some scholars have attempted to integrate the CNN framework with the transformer backbone network for change detection, aiming to circumvent the locality limitation of CNNs, and have achieved promising results.[4,12]

The change detection task for remote sensing images can be essentially phrased as a semantic segmentation task. Nowadays, large models have emerged and demonstrated their effectiveness in semantic segmentation tasks, such as the Segment Anything Model (SAM).[13] However, large models face limitations such as struggles to segment small or peripheral objects and high

computational requirements. The latter is a significant obstacle that should be addressed before large models can be deployed and widely adopted in real-world applications. As a result, it is still of great value to explore relatively smaller models that can be deployed easily and efficiently and at the same time achieving satisfying segmentation precision.

From the research for change detection using deep learning networks, UNet is considered one of the standard CNN architectures for change detection tasks[14] and has a low computational cost, which makes it easy to train and deploy. Therefore, we explored the way to develop deep learning models based on the UNet network. The UNet network was initially proposed for the semantic segmentation of medical images,[15] and it is known for its classic U-shaped encoder-decoder structure. Subsequently, the UNet++ network was developed,[16] and later, Huang *et al.*[17] introduced the UNet3+ network, which incorporates full-scale skip connections and full-scale deep supervision to fully utilize the multi-scale features of images. Inspired by this, we propose a model specifically designed for remote sensing image change detection, which is a Siamese structured UNet3+ network named Siam-UNet3+. Our model aims to facilitate the extraction and integration of information from multi-scale remote sensing images. Moreover, the proposed model is trained and tested on three different datasets, including a publicly available change detection dataset and two local remote sensing image change detection datasets in Beijing. Through comparative experiments with five other models, it is demonstrated that the Siam-UNet3+ model designed in this article has prominent superiority in the evaluation metrics and has shown enormous potential in applications. The structure of the Siam-UNet3+ model and the experiments of the model will be elaborated in the following sections.

2. Methods

As stated earlier, change detection for remote sensing images can be phrased as a segmentation task. However, in remote sensing imagery, the presence of multi-scale features of ground objects poses much more challenges for image segmentation. Moreover, compared with general image segmentation tasks, change detection tasks involve two remote sensing images from different time phases, which further increases the difficulty of the task. In light of this, we propose a network named Siam-UNet3+, which integrates the advantages of Siamese and UNet networks. It is capable of simultaneously receiving two remote sensing images as input, utilizing full-scale skip connections to promote the integration of multi-scale features, and applying full-scale deep supervision to enhance the segmentation performance of the model. The structure of the network and the improvements we made will be detailed below.

2.1 Overview of model architecture

The proposed network Siam-UNet3+ in this article inherits the encoder-decoder architecture and hierarchical representation of the UNet series network. The encoder extracts features from the input image to obtain hierarchical feature representations, and then the decoder recovers the extracted features layer by layer to obtain the change patterns required for the change detection

task. The architecure of Siam-UNet3+ is shown in Fig. 1, which inherits the full-scale skip connections and full-scale deep supervision in UNet3+. The full-scale connections are used to aggregate feature maps from different levels in the encoder and decoder, achieving the multi-scale feature fusion of remote sensing images and effectively avoiding the localization disadvantage of convolution operations. The use of full-scale deep supervision to supervise the decoder output at each stage can effectively improve the training effect of the model. In addition, compared with UNet3+, Siam-UNet3+ has made many improvements and innovations, such as integrating the Siamese network structure in the encoder, using residual modules as the backbone, adding attention modules, and redesigning the loss function to be more suitable for remote sensing image change detection tasks. The following will provide a detailed explanation of the structure of the encoder and decoder.

The encoder of Siam-UNet3+ consists of five levels, each with a residual module for feature extraction. The maximum pooling operation is used for downsampling to extract hierarchical feature representation information at different levels. In addition, the encoder integrates the dual branch structure of the Siamese network, allowing the network to process the two input remote sensing images in parallel, making it more suitable for remote sensing image change detection tasks. Moreover, the features extracted from the two images are concatenated and fused.

The decoder of Siam-UNet3+ corresponds to the encoder and is divided into five levels, which perform information recovery step by step to achieve the goal of outputting changing patterns. The decoding stage of each level adopts full-scale skip connections for multi-scale feature fusion, and the output of each decoding stage is subjected to full-scale deep supervision. The design of full-scale skip connections inherits the advantages of UNet3+, where each level's decoder is connected to the same level's encoder, as well as shallow-level encoders and deep-level decoders. This mechanism aggregates feature maps not only from the same level, but also from the shallow and deep levels, enabling multi-scale feature fusion. In addition, the decoder at each stage produces an output, which is supervised using ground truth during training to achieve

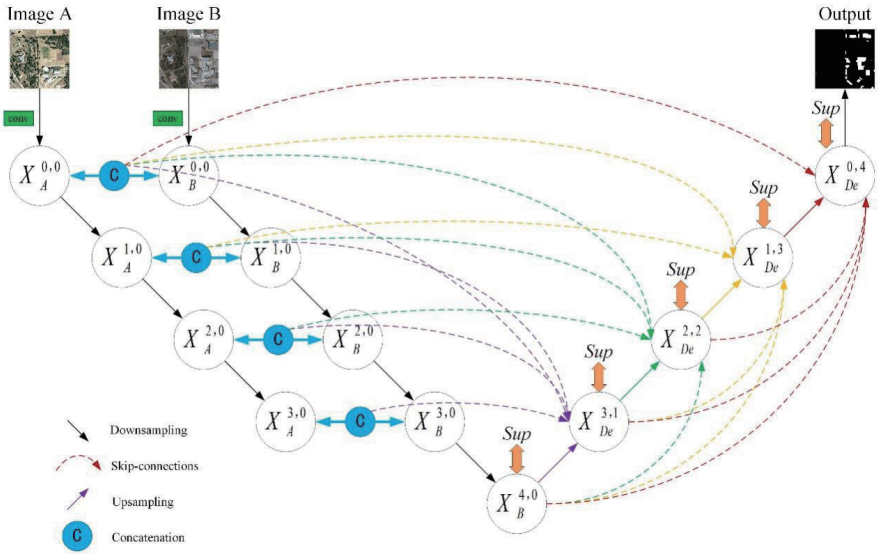

Fig. 1. (Color online) Architecture of Siam-UNet3+.

full-scale deep supervision. This enables shallow layers to receive more sufficient training, thus solving problems such as vanishing gradients and slow convergence, and improving model training effectiveness.

2.2 Backbone

The increase in network depth can lead to the issue of gradient vanishing or explosion. Furthermore, when deeper networks begin to converge, a degradation problem emerges: as the depth of the network increases, the accuracy reaches saturation and then degrades rapidly. To address this issue, He *et al.*[18] proposed the deep residual learning framework. In this paper, the proposed network Siam-UNet3+ incorporates the concept of residual learning, utilizing residual modules as the backbone for feature extraction, replacing the original single convolutional network layer in UNet3+ as the feature extraction module. As illustrated in Fig. 2, the residual module used in our model comprises two convolutional operations with batch normalization (BN) and ReLU activation in between, and an operation of addition to the initial value, which can effectively prevent gradient vanishing (or explosion) and model degradation.

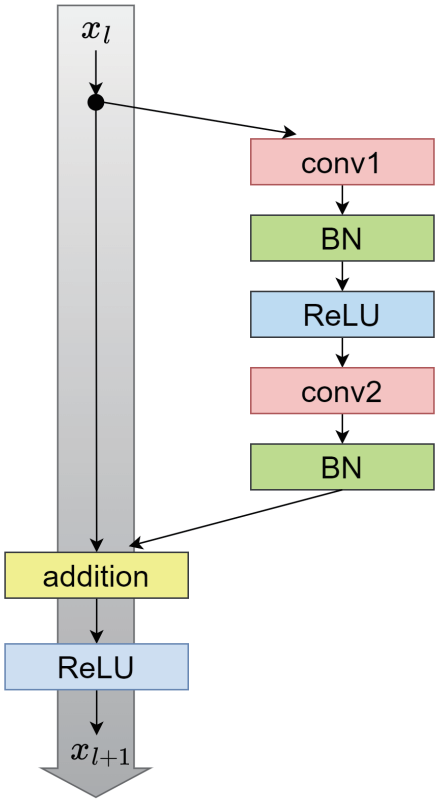

Fig. 2. (Color online) Structure of residual module.

2.3 Attention module

Since UNet3+ implements full-scale skip connections in the decoder stage, the concatenation of multiple feature maps may result in information redundancy, diluting the focus on target objects in the image. Therefore, we propose to add an attention module into Siam-UNet3+ to avoid information redundancy. The Triplet Attention module proposed by Misra *et al.*[19] aims to capture cross-dimensional interactive information without involving any dimensional reduction, providing significant network performance gains with negligible computational cost. In the Triplet Attention module, for a tensor with dimensions of $C \times H \times W$ (where C represents the number of channels, H the height, and W the width), three separate branches are utilized to capture the dependency relationships across the (C, H), (C, W), and (H, W) dimensions of the input tensor. As illustrated in Fig. 3, the outputs from all three branches are aggregated through a simple averaging approach.

In the designed network Siam-UNet3+, the Triplet Attention module is embedded after full-scale skip connections in each decoder stage to perform attention calculations. This eliminates potential information redundancy from multiple concatenations, focusing attention on target objects, and also helps overcome the locality limitation of convolutional operations.

2.4 Loss function

In remote sensing image change detection, the changed parts in the image often account for only a minority, so there is a significant disparity between the changed and unchanged parts, posing a formidable challenge for model training. Therefore, we propose a novel hybrid loss function tailored for remote sensing image change detection tasks, which is a combination of focal loss and dice loss, as shown in Eq. (1). In this equation, an adjustment factor λ is introduced to scale the focal loss and dice loss to the same order of magnitude.

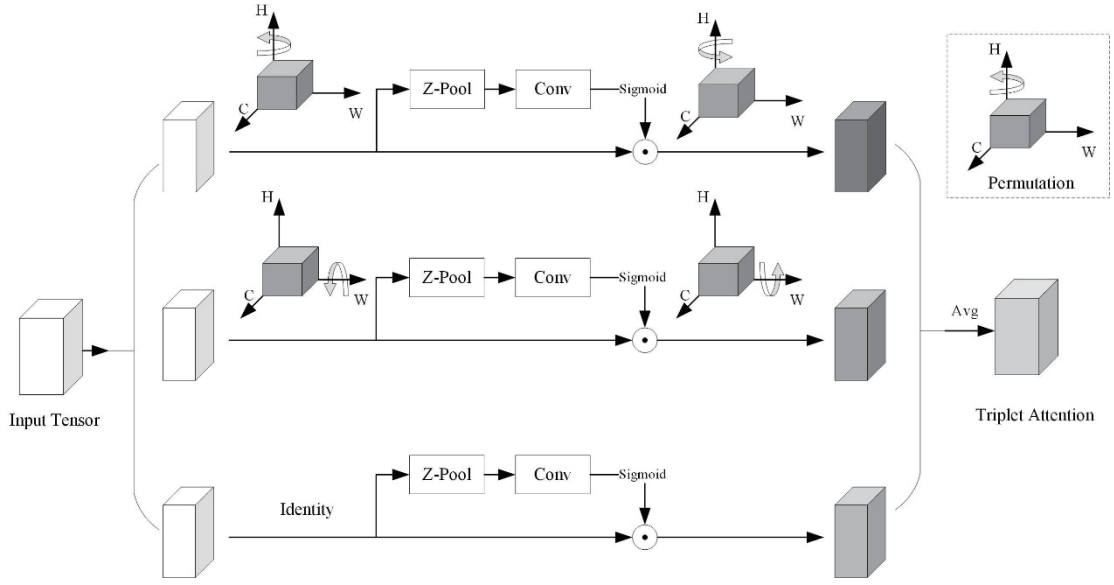

Fig. 3. Structure of Triplet Attention module.

$$L = \mathcal{L}_{dice} + \lambda \mathcal{L}_{focal} \tag{1}$$

Focal loss can address the issue of class imbalance in classification problems and focuses on samples that are difficult to classify, thereby contributing to the improvement of classification precision. Dice loss exhibits satisfying performance in scenarios with severe imbalance between positive and negative samples. In cases where there is a significant difference between the foreground and the background, dice loss can prioritize the exploration of foreground regions (target areas), making it suitable for remote sensing image change detection tasks. Furthermore, the proposed hybrid loss function can be extended to multi-class classification tasks, as shown in Eq. (2).

$$L = \frac{1}{M}\sum_{c=1}^{M} w_c \left(1 - \frac{2\sum_i^N p_{ic} g_{ic}}{\sum_i^N p_{ic} + \sum_i^N g_{ic}}\right) - \lambda \frac{1}{M}\sum_{c=1}^{M}\sum_{i=1}^{N} g_{ic}\alpha_c \left(1 - p_{ic}\right)^{\gamma} \log\left(p_{ic}\right) \tag{2}$$

Here, M represents the total number of classes, N represents the total number of pixels, p_{ic} denotes the probability predicted by the network that the i-th pixel belongs to the c-th class, g_{ic} indicates whether the i-th pixel truly belongs to the c-th class (0 for true and 1 for false), w_c and α_c respectively represent the weight values of the c-th class in dice loss and focal loss, and γ is referred to as the focusing parameter, which is an adjustable hyperparameter with the range of $\gamma \geq 0$.

3. Data, Experiments, and Discussion

3.1 Data

We conducted experiments to test the performance of deep learning models for remote sensing image change detection through three different datasets: the LEVIR-CD, Beijing Remote Sensing Image Change Detection (BJRS-CD), and Beijing Remote Sensing Image Multi-class Change Detection (BJRS-MCD) datasets. Among them, the LEVIR-CD dataset is an open benchmark dataset, whereas the other two are locally produced remote sensing image change detection datasets made by the Beijing Institute of Surveying and Mapping.

The LEVIR-CD dataset was proposed by Beihang University, and it is a novel large-scale public dataset for remote sensing building change detection and serves as a benchmark dataset for evaluating change detection algorithms.[20] Derived from Google Earth (GE) high-resolution (0.5 m/pixel) remote sensing imagery, this dataset comprises 637 image pairs, each with an image size of 1024 × 1024 pixels. The changed areas are annotated with binary labels (1 for change, 0 for no change), as illustrated in Fig. 4.

The BJRS-CD dataset was produced by the Beijing Institute of Surveying and Mapping and it is made from the Beijing-2 satellite images with a resolution of 0.8 m. This dataset comprehensively annotates all types of changes on Earth's surface, using a labeling system

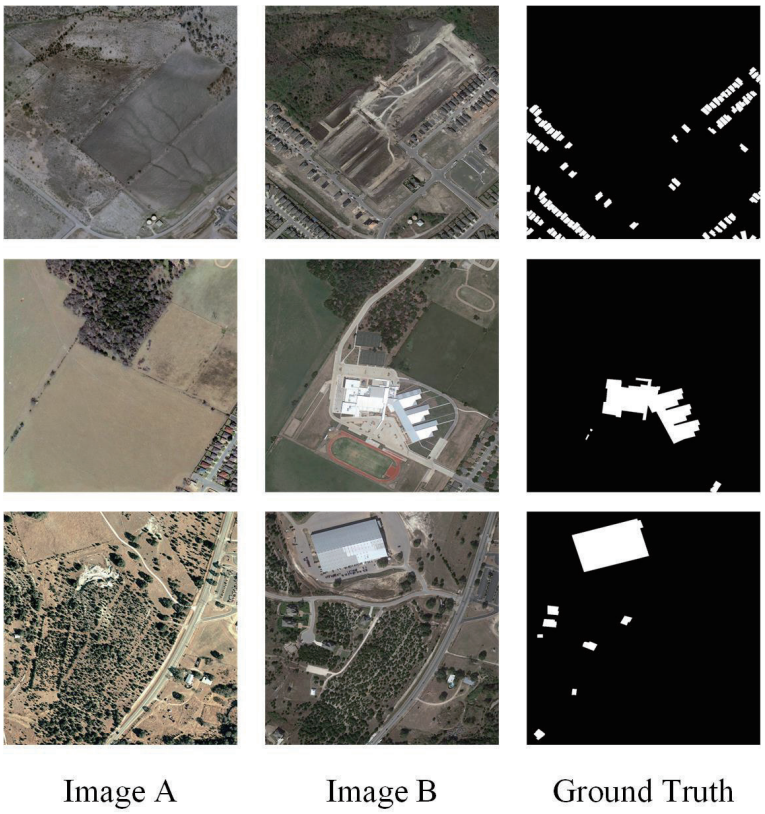

Fig. 4. (Color online) LEVIR-CD dataset.

where 0 signifies no change and 1 indicates a change. As depicted in Fig. 5, the size of each image is of 256 × 256 pixels. The training set comprises 1010 pairs of bi-temporal samples, whereas the validation and test sets each contain 50 pairs of samples.

The BJRS-MCD dataset was produced by the Beijing Institute of Surveying and Mapping, and it is also made from Beijing-2 satellite images with a resolution of 0.8 m. The BJRS-MCD dataset focuses on building changes, and each image has a size of 512 × 512 pixels. It consists of 5318 pairs of samples in the training set and 1772 pairs in both the validation and test sets. Distinguished from conventional change detection datasets, which serve for a binary classification task, the BJRS-MCD dataset serves for a multi-classification task, which further classifies the type of change. For two remote sensing images of the same area at different times, let t1 represent the image from the earlier time sequence and t2 the image from the later time sequence. The changes in buildings in the t2 image compared with those in the t1 image are categorized into two types, namely, increase and decrease, as shown in Fig. 6 with green indicating an increased number of buildings and red a decreased number of buildings in the label.

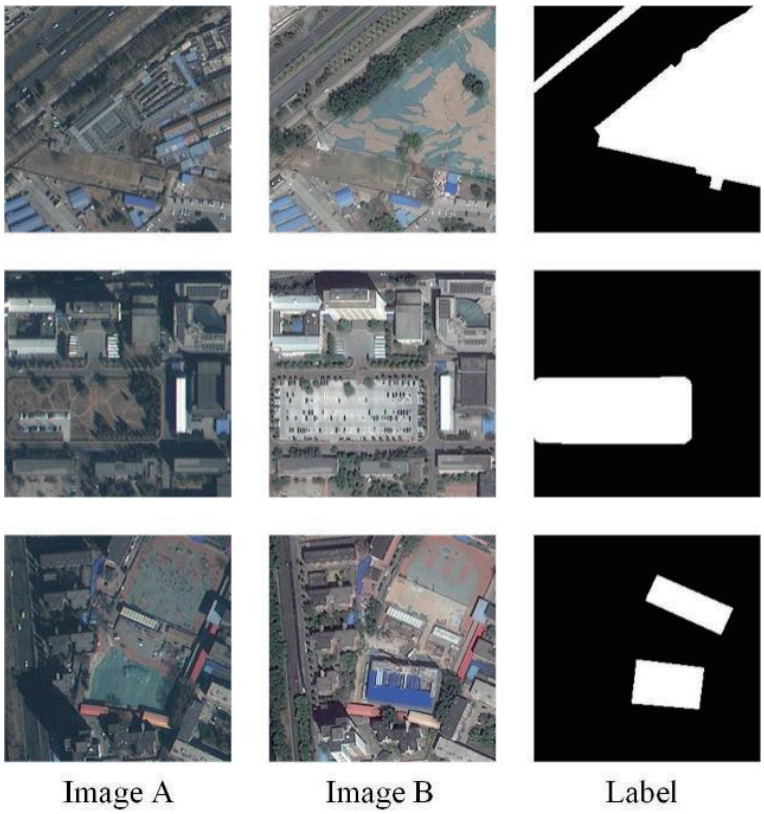

Fig. 5. (Color online) BJRS-CD dataset.

3.2 Model evaluation metrics

A typical change detection task is formulated as a binary classification problem. On the basis of commonly used evaluation metrics for binary classification accuracy in machine learning, we employed the following metrics to assess the performance of the proposed model:
(1) *Precision*: This is the ratio of the number of samples correctly classified as the positive class (change category) to the total number of samples classified as the positive class, which can also be referred to as the positive predictive value. The calculation method is shown in Eq. (3).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Here, *True Positive* (*TP*) represents the number of pixels that have actually changed and are correctly detected, *False Negative* (*FN*) represents the number of pixels in areas that have actually changed but are not detected, *False Positive* (*FP*) represents the number of pixels in areas that have not actually changed but are detected as changed, and *True Negative* (*TN*) represents the number of pixels in areas that have not actually changed and are correctly detected as unchanged.

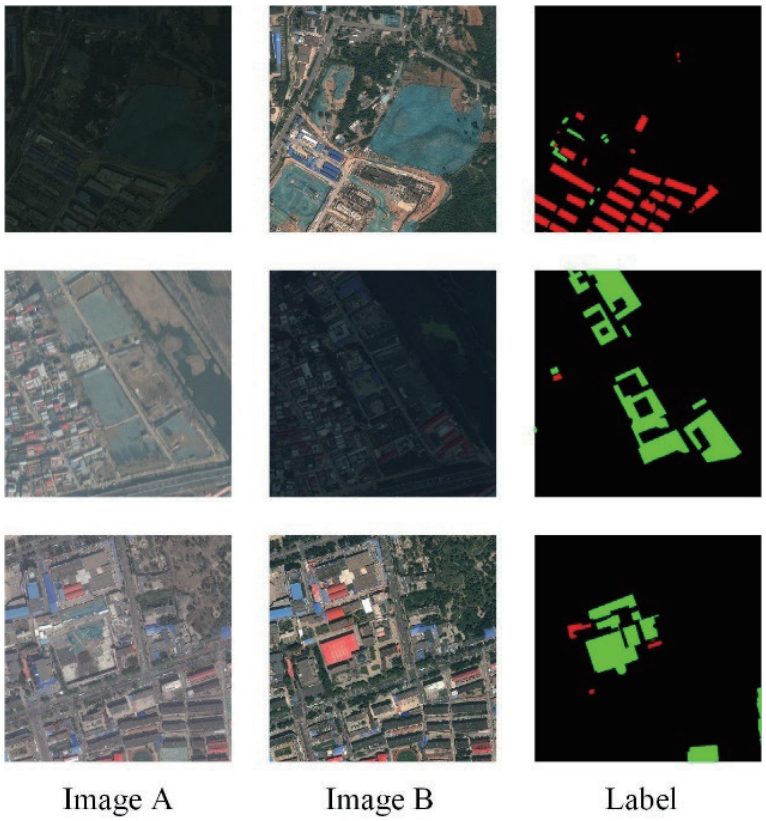

Fig. 6. (Color online) BJRS-MCD dataset.

(2) *Recall*: This is the ratio of the number of samples classified as positive (changed category) to the actual number of positive samples in the dataset. The calculation method is shown in Eq. (4).

$$Recall \quad \frac{TP}{TP \quad FN} \tag{4}$$

(3) *F*1-score: Since *precision* and *recall* are inversely proportional to each other, in order to find a balance between them, the metric of *F*1-score is introduced. *F*1-score considers both *precision* and *recall*, providing a better evaluation of the model's performance. The calculation method is shown in Eq. (5).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{5}$$

(4) *Overall Accuracy* (*OA*): This represents the ratio of the number of correctly classified pixels to the total number of pixels, and its calculation method is shown in Eq. (6).

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

The calculated values of the above four evaluation metrics range from 0 to 1. A higher value closer to 1 indicates a higher *prediction* performance of the model. These four metrics measure the precision level of the model in performing change detection tasks from different aspects. Among them, *F*1-score is considered the most important metric because it is a trade-off between *precision* and *recall*, comprehensively reflecting the prediction performance of the model.[21] The above four model evaluation metrics can also be extended to multi-classification scenarios. Among the four evaluation metrics, except for *OA*, the other three are designed for binary classification scenarios originally and can be extended to multi-class scenarios through macro-averaging. To achieve macro-averaging, the metric value is first calculated for each class individually and the arithmetic mean is then computed across all classes. Since, for each class, each pixel either belongs to that class or does not, each class can be viewed as a separate binary classification problem. Consequently, the evaluation metrics are calculated for each class separately, and the final values are obtained by averaging them. For instance, in a case with n classes (labeled from 0 to $n - 1$), *precision* can be computed as shown in Eq. (7), with *recall* and *F*1-score calculated similarly. In contrast, *OA* serves as a global indicator, involving the construction of a global confusion matrix based on the statistics of all pixels in the dataset. It is calculated by dividing the number of correctly classified pixels along the diagonal of the confusion matrix by the total number of pixels.

$$Precision = \frac{1}{n}\sum_{i=0}^{n-1}\frac{TP_i}{TP_i + FP_i} \tag{7}$$

3.3 Comparative experiments

The proposed Siam-UNet3+ model is trained and tested on the three datasets introduced above. To properly demonstrate the superiority of our model, we conducted extensive comparative experiments using five other deep learning models, namely, UNet, UNet++, FC-Siam-conc, FC-Siam-diff,[6] and SNUNet-CD.[11] These five comparative models were constructed using the source code provided by the original authors and trained and tested on each of the three datasets mentioned in Sect. 3.1, the same way as the Siam-UNet3+ model. The performance of each model was assessed according to the evaluation metrics stated in Sect. 3.2. All experiments were conducted under the PyTorch deep learning framework, utilizing an NVIDIA GeForce RTX 2080Ti GPU with 11 GB of memory. In the training process, the batch size was set to 10, the number of epochs was set to 300, the AdamW optimizer was used as the optimization algorithm, and the initial learning rate was configured to 1×10^{-3}. Figure 7 illustrates the change detection results predicted by the proposed Siam-UNet3+ model and the five comparative models on the LEVIR-CD, BJRS-CD, and BJRS-MCD datasets. As depicted in Fig. 7, the prediction results of the Siam-UNet3+ model are closest to the actual changes in the

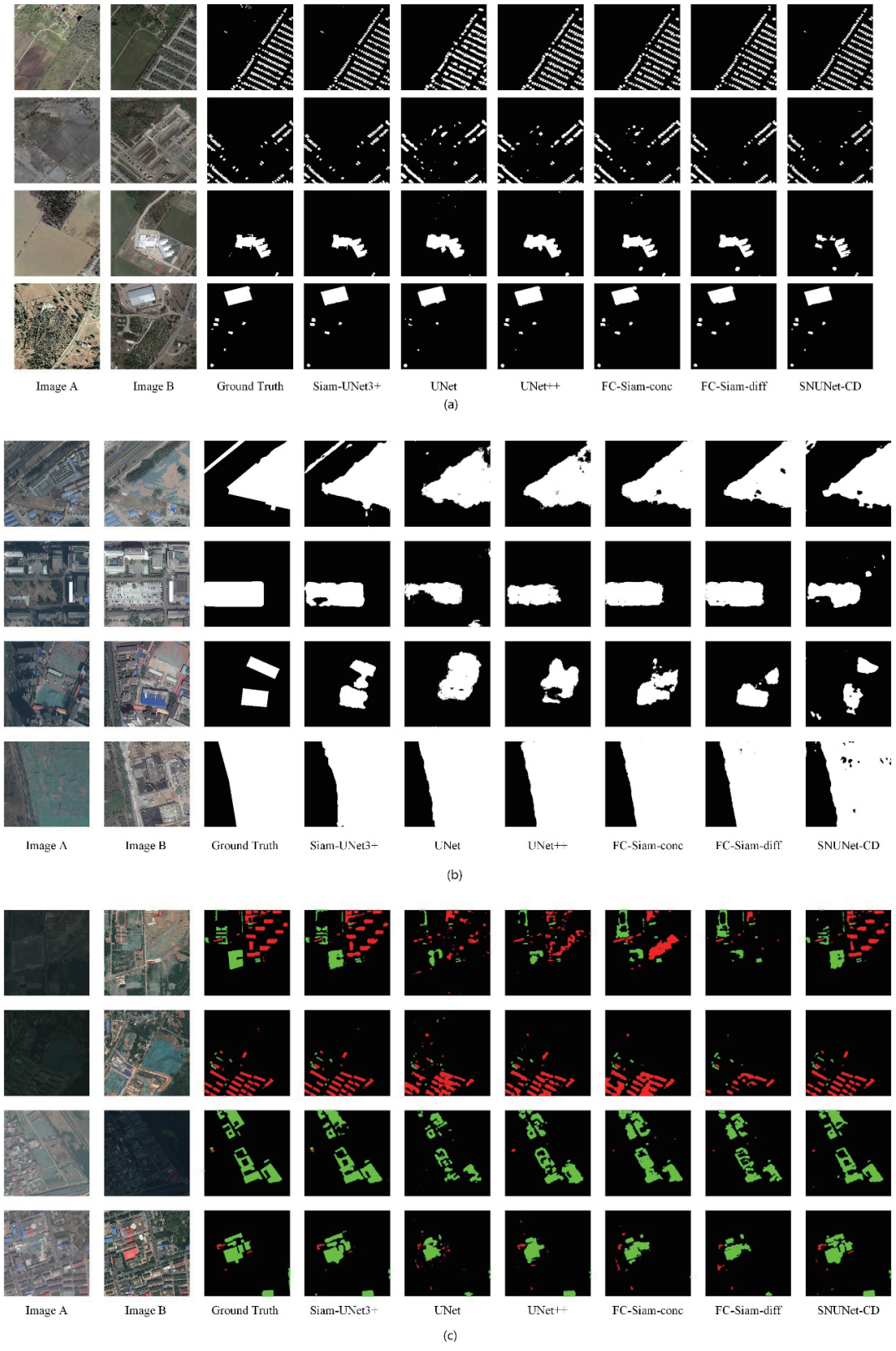

Fig. 7. (Color online) Results of model experiments on different datasets: (a) LEVIR-CD dataset, (b) BJRS-CD dataset, and (c) BJRS-MCD dataset.

imagery, whereas those of the other comparative models exhibit higher incidences of false changes and omissions.

Quantitatively, the evaluation metrics were calculated for the Siam-UNet3+ model and five comparative models on the three datasets, as shown in Tables 1–3. As evident from these tables, the Siam-UNet3+ model achieves the highest scores in terms of *precision*, *recall*, *F*1-score, and *OA* on both the LEVIR-CD and BJRS-MCD datasets. On the BJRS-CD dataset, the FC-Siam-diff model attains the highest *precision* score but ranks the lowest in recall, which means that the prediction results of the FC-Siam-diff model suffer the most omissions. In contrast, the Siam-UNet3+ model exhibits a *precision* that is second only to FC-Siam-diff, while its *recall*, *F*1-score, and *OA* surpass those of the other five comparative models by a large margin. Therefore, when considering all evaluation metrics comprehensively, compared with the other five models, the Siam-UNet3+ model proposed in this paper demonstrates the highest performance in the remote sensing change detection tasks.

3.4 Ablation study

The Siam-UNet3+ model presented in this article is an improvement based on the UNet3+ architecture, featuring the integration of a Siamese network structure within the encoder, the

Table 1
Results of evaluation metrics in comparative experiments on LEVIR-CD dataset.

Model	*Precision* (%)	*Recall* (%)	*F*1-score (%)	*OA* (%)
U-Net	74.11	74.36	74.02	97.14
UNet++	79.47	80.25	79.69	97.43
FC-Siam-conc	81.81	78.51	79.73	97.57
FC-Siam-diff	85.14	74.24	78.86	97.43
SNUNet-CD	87.74	85.00	86.28	98.14
Siam-UNet3+	90.89	86.28	88.47	98.29

Table 2
Results of evaluation metrics in comparative experiments on BJRS-CD dataset.

Model	*Precision* (%)	*Recall* (%)	*F*1-score (%)	*OA* (%)
U-Net	76.12	67.03	70.40	88.40
UNet++	79.62	60.21	67.27	88.60
FC-Siam-conc	77.03	67.75	71.38	88.60
FC-Siam-diff	86.20	55.99	67.21	88.80
SNUNet-CD	72.52	64.11	67.41	86.80
Siam-UNet3+	85.22	68.64	75.80	89.75

Table 3
Results of evaluation metrics in comparative experiments on BJRS-MCD dataset.

Model	*Precision* (%)	*Recall* (%)	*F*1-score (%)	*OA* (%)
U-Net	61.59	56.63	56.68	96.01
UNet++	67.98	63.03	63.85	96.60
FC-Siam-conc	65.05	61.90	61.85	96.44
FC-Siam-diff	69.73	52.78	56.58	96.28
SNUNet-CD	80.32	68.79	72.11	97.13
Siam-UNet3+	82.48	70.62	74.33	97.30

addition of an attention module in the decoder, and the utilization of a residual module as the backbone. To prove the effectiveness of these modifications, we carried out ablation studies on the LEVIR-CD dataset as an example.

For the ablation study on the Siamese network structure, we designed a variant of UNet3+, namely, UNet3+-EF, which eliminates the Siamese structure and instead adopts an early fusion approach.[6] In this approach, two remote sensing images are first stacked into a multi-channel image before being input into the model. During the ablation experiment, the UNet3+-EF model was trained using the same LEVIR-CD dataset as that for training the Siam-UNet3+ model, and its performance was evaluated on the test set. The calculated evaluation metrics are presented in Table 4.

Similarly, to conduct an ablation study on the attention module, the Triplet Attention module was removed to construct the Siam-UNet3+-AttentionAblation model. The results obtained from this ablation experiment are presented in Table 5. Additionally, to perform an ablation study on the residual module, the backbone of the original UNet3+ was restored instead of using the residual module, constructing the Siam-UNet3+-ResidualAblation model. The results of this ablation experiment are shown in Table 6.

As can be observed in Tables 4–6, the model improvements made in this study have led to significant enhancements in model performance. Therefore, the ablation experiments conducted above have proven the effectiveness of the improvements made in the Siam-UNet3+ model that we proposed.

4. Conclusions

The problem of change detection in remote sensing imagery can be reduced to a semantic segmentation task. Since CNNs are the most commonly used deep learning networks to handle semantic segmentation, within the CNN framework, we designed a model specifically tailored for remote sensing imagery change detection tasks named Siam-UNet3+. This model is inspired

Table 4
Ablation experiment results of Siamese network structure.

Model	*Precision* (%)	*Recall* (%)	*F*1-score (%)	*OA* (%)
UNet3+-EF	87.36	85.97	86.62	98.00
Siam-UNet3+	90.89	86.28	88.47	98.29

Table 5
Ablation experiment results of attention module.

Model	*Precision* (%)	*Recall* (%)	*F*1-score (%)	*OA* (%)
Siam-UNet3+- AttentionAblation	89.43	84.58	86.86	98.14
Siam-UNet3+	90.89	86.28	88.47	98.29

Table 6
Ablation experiment results of residual module.

Model	*Precision* (%)	*Recall* (%)	*F*1-score (%)	*OA* (%)
Siam-UNet3+-ResidualAblation	89.87	81.08	84.92	98.18
Siam-UNet3+	90.89	86.28	88.47	98.29

by UNet3+, and it inherits the full-scale skip connections and full-scale deep supervision of UNet3+, which can achieve the multi-scale feature fusion of remote sensing images and effectively avoid the localization disadvantage of convolution operations, thus enhancing the performance of change detection. However, different from the original UNet3+, multiple improvements have been made to the Siam-UNet3+ model. First, the Siamese networks are integrated into the encoder, and the dual-branch structure of Siamese networks can be used to process bi-temporal remote sensing images in parallel. Second, the residual module is used as the backbone, effectively avoiding model degradation problems. Third, a Triplet Attention module is added to the decoder, which has the characteristic of almost "zero parameters" and can mitigate information redundancy that may occur in full-scale skip connections and strengthen focus on the changed areas of the image. Last but not the least, a hybrid loss function consisting of focal loss and dice loss has been designed, which is more suitable for remote sensing image change detection tasks. We conducted a large number of model experiments, using five other models as comparative models. Through comparative experiments on three different change detection datasets and ablation experiments, the significant advantages and improvements of the Siam-UNet3+ model have been proven on the basis of four evaluation metrics: precision, recall, *F*1-score, and *OA*. In conclusion, the proposed Siam-UNet3+ model can serve as an effective tool for high-resolution remote sensing image change detection and has enormous potential in the application of geospatial monitoring, urban governance, and other related scenarios.

Acknowledgments

This work was supported by the project for the on-site verification ecological restoration of illegally demolished and vacated land in Beijing at the Beijing Institute of Surveying and Mapping.

References

1 K. Li, X. Cao, and D. Meng: IEEE Trans. Geosci. Remote Sens. **62** (2024) 1. https://doi.org/10.1109/TGRS.2024.3365825
2 S. Hao, Y. Zhou, and Y. Guo: Neurocomputing **406** (2020) 302. https://doi.org/10.1016/j.neucom.2019.11.118
3 B. Fang, L. Pan, and R. Kou: Remote Sens. **11** (2019) 1292. https://doi.org/10.3390/rs11111292
4 C. Liang, P. Chen, H. Liu, Xiaokun Zhu, Y. Geng, and Z. Zhang: Sens. Mater. **35** (2023) 183. https://doi.org/10.18494/SAM4180
5 J. Long, E. Shelhamer, and T. Darrell: Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition (2015) 3431. https://doi.org/10.1109/TPAMI.2016.2572683
6 R. C. Daudt, B. Le Saux, and A. Boulch: 2018 25th IEEE Int. Conf. Image Processing (ICIP) (2018) 4063. https://doi.org/10.1109/icip.2018.8451652
7 W. Wiratama, J. Lee, and D. Sim: IEEE Access **8** (2020) 12279. https://doi.org/10.1109/ACCESS.2020.2964798
8 L. Deng, Y. Wang, Q. Lan, and F. Chen: J. Appl. Remote Sens. **17** (2023) 034501. https://doi.org/10.1117/1.JRS.17.034501
9 H. Chen, Y. He, L. Zhang, W. Yang, Y. Liu, B. Gao, Q. Zhang, and J. Lu: IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **17** (2024) 1215. https://doi.org/10.1109/JSTARS.2023.3339294
10 M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun: IEEE Geosci. Remote Sens. Lett. **16** (2018) 266. https://doi.org/10.1109/lgrs.2018.2869608

11 S. Fang, K. Li, J. Shao, and Z. Li: IEEE Geosci. Remote Sens. Lett. **19** (2022) 1. https://doi.org/10.1109/LGRS.2021.3056416
12 Y. Tang, Z. Cao, N. Guo, and M. Jiang: Sci. Rep. **14** (2024) 4577. https://doi.org/10.1038/s41598-024-54096-8
13 A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, and W. Y. Lo: 2023 IEEE/CVF Int. Conf. Computer Vision (ICCV). https://doi.org/10.1109/ICCV51070.2023.00371
14 L. Khelifi and M. Mignotte: IEEE Access **8** (2020) 126385. https://doi.org/10.1109/access.2020.3008036
15 O. Ronneberger, P. Fischer, and T. Brox: Int. Conf. Medical Image Computing and Computer-assisted Intervention (2015) 234. https://doi.org/10.1007/978-3-319-24574-4_28
16 Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang: IEEE Trans. Med. Imaging. **39** (2019) 1856. https://doi.org/10.1109/TMI.2019.2959609
17 H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu: ICASSP 2020-2020 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP) (2020) 1055. https://doi.org/10.1109/icassp40776.2020.9053405
18 K. He, X. Zhang, S. Ren, and J. Sun: Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition (2016) 770. https://doi.org/10.1109/cvpr.2016.90
19 D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou: Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (2021) 3139. https://doi.org/10.48550/arXiv.2010.03045
20 H. Chen and Z. Shi: Remote Sens. **12** (2020) 1662. https://doi.org/10.3390/rs12101662
21 J. Dong, W. Zhao, and S. Wang: IEEE Geosci. Remote Sens. Lett. **19** (2021) 1. https://doi.org/10.1109/lgrs.2021.3121094

About the Authors

Chen Liang received his B.S. and Ph.D. degrees from Wuhan University, China, in 2012 and 2019, respectively. From 2019 to 2021, he was a lecturer at Pingdingshan University, China. From 2021 to 2023, he worked as a postdoctoral fellow at Beijing Normal University. Since 2023, he has been a senior engineer at the Beijing Institute of Surveying and Mapping. His research interests are in deep learning and geographic intelligent analysis. (chenliang90210@whu.edu.cn)

Zhang Yi was born in Xinxiang, Henan Province. She is a postgraduate student, and her research direction is 3S technology integration and application.
(1113611326@qq.com)

Zongxia Xu received her master's degree in geographic information engineering from Capital Normal University, Beijing, China, in 2019. Since 2019, she has been working at the Beijing Institute of Surveying and Mapping, and since 2021, she has been an engineer. Her research interest is remote sensing image interpretation. (xuzongxia123@163.com).

Yongxin Yu received his B.S. degree from Wuhan University in 2002 and M.S. degree from Peking University in 2013. He is now a professor-level senior engineer of the Beijing Institute of Surveying and Mapping. His research interests include urban territorial space monitoring, ecological environment monitoring, and GIS platform construction. (yyxgps@126.com)

Zhenwei Zhang received his B.S. and Ph.D. degrees from Wuhan University, China, in 2015 and 2021, respectively. Since 2022, he has been a lecturer at Nanjing University of Information Science and Technology. His research interests are in thermal remote sensing and geo-spatial modeling. (zhangzw@nuist.edu.cn)