Edge Neighborhood Contrastive Learning for Building Change Detection

Mingwei Zhang, Qiang Li, Yuan Yuan, *Senior Member, IEEE*, Qi Wang, *Senior Member, IEEE*

Abstract—**Building change detection aims to identify the change of buildings in the same geographic area. Recently, many methods based on deep learning (DL) have achieved encouraging performance. However, some challenges remain in effectively exploiting the temporal-spatial correlation and achieving good discrimination in the neighborhood of the edge. To relieve these issues, we develop a selective attention module (SAM) to model the relationship between the semantic and the state (i.e., unchanged or changed) of the pixel, which is integrated into an existing metric learning-based architecture. Moreover, inspired by recent advances in contrastive learning, we present a novel edge neighborhood contrastive learning method to force the network to learn discriminative and compact features, leading to improving the accuracy of building change detection. Experimental results demonstrate our method achieves competitive performance in terms of objective metrics and visual comparisons.**

Index Terms—**Building change detection, edge neighborhood, contrastive learning, selective attention.**

I. Introduction

BUILDING is an important element of human civilization, whose change reveals the expansion of the city, the development of society, and the use of land [1]–[3]. The dynamic monitoring of the building change is vitally helpful to urban planning, land source investigation, illegal construction surveying and the conservation of the ecology environment. Therefore, it is valuable and necessary to develop effective and intelligent building change detection technology.

Recently, owing to the breakthrough of computational power and the available large amount of data acquired by various space-borne and airborne sensors, remote sensing change detection (RSCD) techniques have made great progress [4], [5]. Wu *et al.* [6] propose a multi-scale graph convolutional network, which not only can be used for homogeneous images change detection, but also utilized for heterogeneous images. Besides, the convolutional neural network (CNN) is introduced in many advanced change detection methods due to its excellent feature extraction and representation abilities. Daudt *et al.* [7] present three fully convolutional network (FCN) based architectures for change detection. One is early fusion architecture, which feeds concatenated bitemporal images into the FCN directly. The other two are late fusion architectures. Instead, they adopt the Siamese FCN to extract the features from the bitemporal images, and then execute concatenation and difference operations respectively in feature-level to attain change information. However, these architectures show limited accuracy due to inadequate utilization of temporal-spatial correlation.

The temporal-spatial correlation is reflected mainly in two aspects: 1) the long-range dependency of different spatial locations in the bitemporal images, 2) the semantic consistency and the discrepancy of the same geographical location at different times. The former has been analyzed very well by the self-attention or the co-attention mechanism. For example, STANet [8] integrates a self-attention module in the process of feature extraction to model the spatial-temporal relationship. The Transformer founded on the multi-head self-attention mechanism is introduced as well. Bandara *et al.* [9] propose a pure Transformer-based Siamese architecture. Chen *et al.* [10] specially design a bitemporal Transformer. Yet, the potential of the latter is poorly explored, especially its benefits for building change detection. For the building change detection task, unchanged areas possess semantic consistency (i.e., building or non-building), which is undisturbed by low-level factors such as lighting differences and seasonal changes. Meanwhile, the semantics of land surface objects are usually mutually exclusive in changed areas. According to such useful characteristics, a selective attention module (SAM) is designed in this work. In particular, it is integrated into a metric learning-based architecture.

Depending on the semantic relation of the bitemporal pixels, SAM is possibly helpful to identify the change of the buildings, but it is weak to attain refined change detection results. In detail, because of the mutual interference between the background and the target features in the neighborhood of the edge, the discrimination of the features there usually is inadequate. It easily results in coarse predictions. Some methods introduce prior edge structure knowledge to mitigate this issue. EGRCNN [11] and EGDE-Net [12] learn to predict change and edge maps simultaneously. Besides, dense skip connections across multi-level features are proven to be effective, which preserve the structure information of the regions of change in deep features [13]. These means usually require a sophisticated network design, e.g., an extra network branch to predict the edge. Different from them, we explore a supervised contrastive learning strategy based on given labels to facilitate the learning of discriminative features. Contrastive learning has been employed in previous relevant studies. Saha *et al.*

This work was supported by the National Natural Science Foundation of China under Grant U21B2041, 61825603, National Key R&D Program of China 2020YFB2103902.

Mingwei Zhang is with the Unmanned System Research Institute and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China (e-mail: dlaizmw@gmail.com)

Qiang Li, Yuan Yuan, and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China (e-mail: liqmges@gmail.com, y.yuan1.ieee@gmail.com, crabwq@gmail.com) (*Corresponding author: Qi Wang.)*

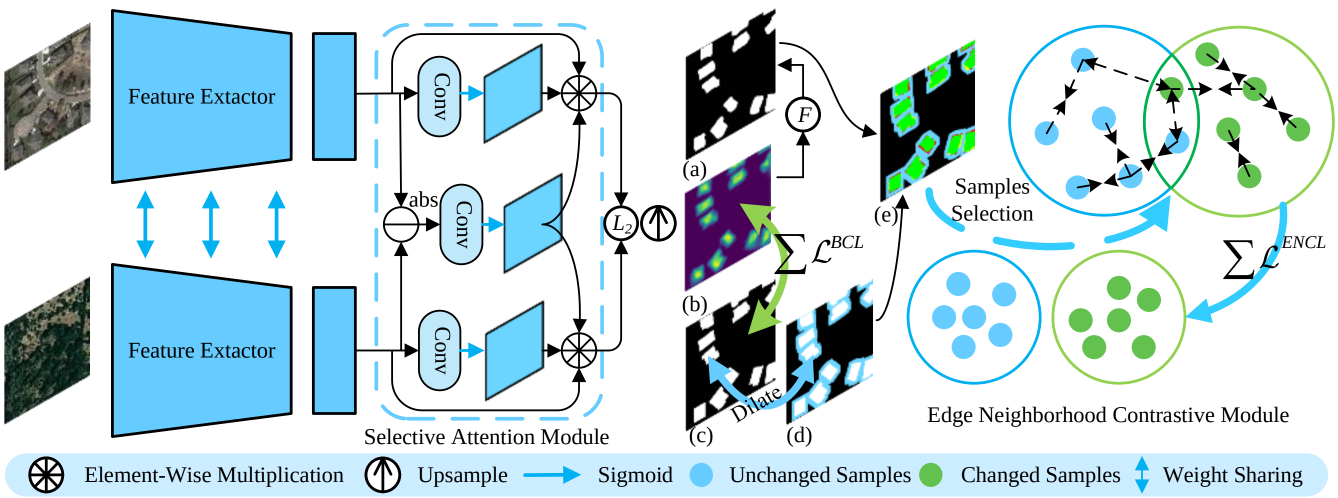


Fig. 1. The overall pipeline of our method for building change detection. (a) Change map, (b) Distance map, (c) Ground truth, (d) Edge neighborhood mask. The image (e) finely indicates the true positives (TPs), the false positives (FPs), the true negatives (TNs), and the false negatives (FNs) in the neighborhood of the edge, which are represented by the green, the yellow, the blue and the red respectively.

[14] propose an unsupervised heterogeneous image change detection method. It adopts the classical self-supervised contrastive fashion to constrain the learned representations, which effectively prevents the model from falling into a degenerate solution. Kang *et al.* [15] utilize the contrast between the pixels of the building and that of the background in latent space to improve the performance of the building extraction, which is highly inspired to our work. Differently, to help the acquisition of the refined change map, we focus on enhancing the discrimination of the features in the neighborhood of the edge by contrastive learning. Therefore, the pixels as contrastive samples are selected from the background near the edge and the target regions. In summary, this article makes the following contributions:

- We develop a selective attention module to adaptively improve the discrepancy and consistency of the bitemporal features. It essentially predicts the class confidence and the change confidence of pixels in the bitemporal images.
- We propose an edge neighborhood contrastive learning method for building change detection. It mainly learns a well discriminative feature space through the contrastive optimization between the background near the edge and the targets.
- Experiments demonstrate that our method achieves satisfactory performance on two widely used building change detection datasets.

II. Proposed Method

A. Overview

The overall pipeline of our approach is given in Fig. 1, which comprises three parts: a Siamese feature extractor, an attention module, and a contrastive learning module. First, the feature extractor is responsible to acquire representative features from the bitemporal inputs, which is a typical encoder-decoder architecture. The encoder is built based on the pre-trained ResNet-18 [16] on the ImageNet. The decoder is founded on a multi-scale features merging module presented in [17]. Next, the bitemporal features are modulated by the attention module to make them more refined than before. Then, the distance map (DM) shown in Fig .1(b) is generated by computing the Euclidean distance (i.e., L_2 norm) of the attentively refined bitemporal features for every pixel, where an upsample operation is included because of the lower size of the features than that of the bitemporal inputs. Finally, the states of pixels are decided by a decision function F, which can be formulated as follows:

$$\mathrm{CM}_i = \begin{cases} 1, & \text{if } d_i > T \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where CM_i denotes the value of the pixel i in the change map (CM) shown in Fig. 1(a), and d_i denotes that in the DM. 1 and 0 represent changed and unchanged states respectively. T is the decision threshold. As for the contrastive learning module, it is used to optimize the network during the training. More details on it are described shortly.

B. Selective Attention Module

According to Eq. 1, it can be intuitively seen that the features of the bitemporal images in unchanged areas are expected to be as consistent as possible. Instead, the more obvious the differences in changed areas, the better. However, there are some adverse factors existing constantly, such as lighting and color variations between the bitemporal images, resulting in weakly discriminative features. To solve this problem, SAM is developed to selectively suppress or enhance the feature discrepancy.

SAM is composed of three branches. One branch is applied to predict the change confidence of the land surface objects, which can be formulated as

$$\mathrm{A}_c = \sigma(g^{(k\times k)}(|\mathrm{F}_{pre} - \mathrm{F}_{post}|)), \tag{2}$$

where F_{pre} and F_{post} denote the features obtained by the feature extractor. σ represents the Sigmoid function. $g^{(k\times k)}$ is a convolutional operator and k is its kernel size. A_c indicates the change-confidence map, which roughly shows where changes

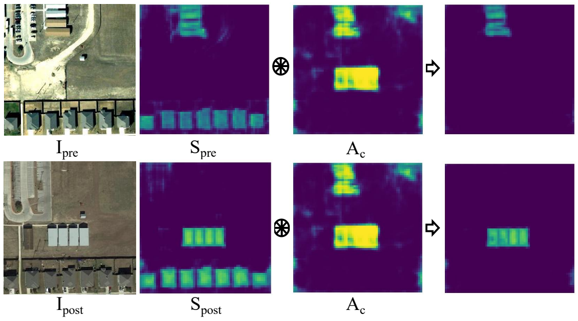


Fig. 2. The illustrations of the confidence maps attained by the SAM.

are most likely. The other two branches are utilized to compute the class confidence, which can be expressed as

$$\mathrm{S}_{pre} = \sigma(f^{(k\times k)}(\mathrm{F}_{pre})), \tag{3}$$

$$\mathrm{S}_{post} = \sigma(f^{(k\times k)}(\mathrm{F}_{post})), \tag{4}$$

where S_{pre} and S_{post} indicate the class-confidence maps of pre- and post-temporal images, and $f^{(k\times k)}$ owns the same structure as $g^{(k\times k)}$, with different weights. The class-confidence map is exploited to distinguish between buildings and non-buildings. Thereafter, F_{pre} and F_{post} are adjusted based on the change- and class-confidence maps. This process can be expressed as follows:

$$\mathrm{F}^*_{pre} = \mathrm{F}_{pre} \circledast (\mathrm{S}_{pre} \circledast \mathrm{A}_c), \tag{5}$$

$$\mathrm{F}^*_{post} = \mathrm{F}_{post} \circledast (\mathrm{S}_{post} \circledast \mathrm{A}_c), \tag{6}$$

where F^* implies the refined features. The operation $\circledast$ represents the element-wise multiplication in the spatial dimension along different channels. Inferred from an example given in Fig. 2, the differences of unchanged features and those of changed features are reduced and enhanced respectively.

C. Edge Neighborhood Contrastive Module

There are always a considerable number of false predictions in the neighborhood of the edge, which can be attributed to mutual mixing between the background and the target contexts during the feature extraction. Aiming at this problem, the current leading solution is to introduce the edge structure information, which effectively enhances the distinction of the features around the edge. Different from that solution, motivated by the latest advance in contrastive learning [18], we develop an edge neighborhood contrastive module (ENCM) in this work, which is made up of two steps: the generation of the edge neighborhood mask (ENM) and the computation of the edge neighborhood contrastive loss (ENCL).

As shown in Fig. 1(c) and (d), the ENM is first generated by dilating the ground truth (GT). It can be found that there are three kinds of colors in the ENM. Since the black region is the background far away from the edge, the discussion below is not relevant to it. The blue refers to the unchanged pixel and the white is the changed pixel. The ratio of the number of pixels of the two regions is 1. In addition, all changed pixels in the GT are reserved in the ENM. Then, a detailed predicted result in the two regions indicated by the ENM is obtained by exploiting the CM, which is shown in Fig .1(e). It is used to guide the selection of the samples used to calculate the ENCL. Note that selected samples can be from different images within the mini-batch. In detail, for the pixels of the two categories (i.e., changed and unchanged), we treat the pixels with incorrect predictions as hard samples and those with correct predictions as easy samples. Inspired by [18], half of the selected samples are the hard ones and half are the easy ones for every category. Accordingly, the ENCL can be computed as follows:

$$\mathcal{L}_j^{ENCL} = \max(\sum_{j^+\in\mathcal{P}^j}\frac{d_{jj^+}}{|\mathcal{P}^j|} - \sum_{j^-\in\mathcal{N}^j}\frac{d_{jj^-}}{|\mathcal{N}^j|} + \tau, 0), \tag{7}$$

$$d_{jj^+} = |d_j - d_{j^+}|, d_{jj^-} = |d_j - d_{j^-}|, \tag{8}$$

where $\mathcal{P}^j$ and $\mathcal{N}^j$ represent the collection of positive samples and that of negative samples corresponding to the sample j respectively. τ denotes the manually set margin. d_j equals the value of the pixel in the DM corresponding to the sample j. In particular, for the sample j with its ground-truth label c, other samples that also belong to the class c are the positive samples, i.e., j^+, while the samples belonging to the other classes $\bar{c}$ are the negatives, i.e., j^-. According to Eq. 7, it pulls the samples with the same class close and pushes those with different classes apart, which is beneficial to generate the representative embedding of the background near the edge and that of the targets, leading to reducing false predictions.

D. Overall Loss Function

Motivated by previous works [8], [17], the batch-balanced contrastive loss (BCL) is used for pixel-wise supervision. The overall training loss function in this work is

$$\mathcal{L}^{Overall} = \mathcal{L}^{BCL} + \frac{\alpha}{|\mathcal{M}|}\sum_{j\in\mathcal{M}}\mathcal{L}_j^{ENCL}, \tag{9}$$

where $\mathcal{M}$ is the contrastive sample set. α is the balanced factor. Here, the formula of $\mathcal{L}^{BCL}$ is

$$\begin{aligned}\mathcal{L}^{BCL} = \frac{1}{2}[&\frac{1}{|\mathcal{B}_0|}\sum_{i\in\mathcal{B}_0}(1-c_i)d_i^2 + \\ &\frac{1}{|\mathcal{B}_1|}\sum_{i\in\mathcal{B}_1}c_i\max(m-d_i,0)^2],\end{aligned} \tag{10}$$

where m is set to 2, and $c_i \in \{0,1\}$ indicates the ground-truth label of the pixel i. $\mathcal{B}_0$ and $\mathcal{B}_1$ represent the set of unchanged pixels and that of changed pixels within the mini-batch respectively.

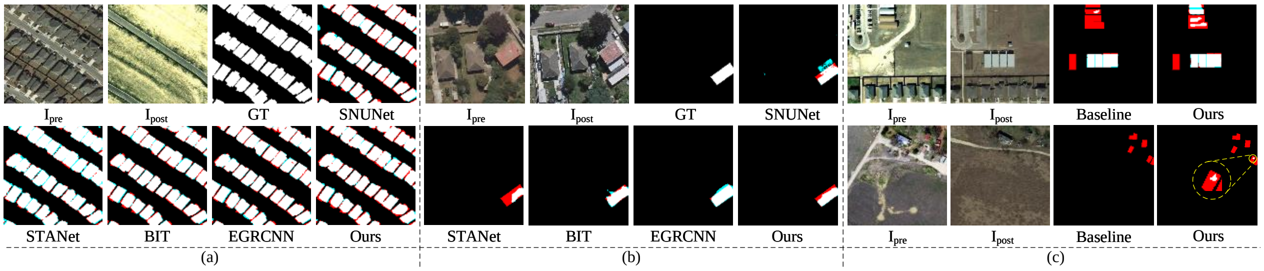
III. Experiment

A. Dataset and Evaluation Metrics

The proposed method is evaluated on two public building change detection datasets, LEVIR-CD [8] and WHU-CD [19]. The default split of the two datasets is retained. LEVIR-CD dataset contains 637 image pairs with the size of 1024×1024. Among them, 445 image pairs are used for training, 128 image pairs are for testing, and 64 image pairs are for validation. WHU-CD dataset contains two pairs of high-resolution aerial

TABLE I
EXPERIMENTAL RESULTS ON LEVIR-CD AND WHU-CD DATASETS (%). THE BOLD INDICATES THE BEST PERFORMANCE.

Method	Running Time(ms)	LEVIR-CD					WHU-CD				
		P	R	F1	IoU	OA	P	R	F1	IoU	OA
FC-EF [7]	**12.12**	86.09	83.05	84.77	73.57	98.47	82.84	78.02	80.36	67.16	98.62
FC-Siam-diff [7]	17.11	90.05	83.48	86.64	76.43	98.69	73.27	82.86	77.77	63.62	98.29
FC-Siam-conc [7]	16.85	90.46	83.84	87.02	77.03	98.73	65.41	86.32	74.42	59.27	97.85
SNUNet [13]	26.64	90.69	88.93	89.80	81.49	98.97	83.95	88.95	86.38	76.02	98.99
DSAMNet [17]	31.50	81.28	88.68	84.81	73.63	98.38	71.22	**92.28**	80.40	67.22	98.37
STANet [8]	31.15	85.01	**91.38**	88.07	78.69	98.74	88.59	85.18	86.86	76.76	99.07
ChangeFormer [9]	80.19	91.03	87.77	89.37	80.78	98.94	88.22	79.86	83.83	72.16	98.89
BIT [10]	34.74	**91.61**	88.74	90.15	82.07	99.01	78.33	89.21	83.42	71.55	98.72
EGRCNN [11]	67.18	88.58	91.13	89.84	81.55	98.85	90.92	89.41	90.16	82.08	99.29
Ours	13.90	91.27	89.72	**90.49**	**82.63**	**99.04**	**94.56**	86.77	**90.50**	**82.64**	**99.34**



(a) (b) (c)

Fig. 3. Visual comparisons on LEVIR-CD and WHU-CD datasets are shown in (a) and (b) respectively, where the FNs and the FPs are indicated by the red and the blue respectively. Some wrong results of our approach are given in (c).

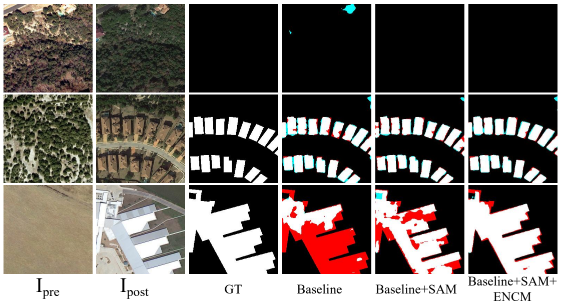


Fig. 4. The visual analysis of the effectiveness of proposed modules.

TABLE II
ABLATION STUDIES ABOUT THE MODULES (%).

Module		Evaluation Metrics				
ENCM	SAM	P	R	F1	IoU	OA
✗	✗	89.05	87.27	88.15	78.81	98.80
✓	✗	89.73	87.57	88.64	79.60	98.86
✗	✓	91.27	88.36	89.79	81.48	98.98
✓	#	88.79	89.74	89.26	80.61	98.90
✓	✓	91.27	89.72	90.49	82.63	99.04

TABLE III
SENSITIVITY EXPERIMENTS ON THE ENCM (%).

τ / ENM	0.5	1.0	1.5	2.0	4.0
✗	88.19	87.99	88.35	88.46	87.67
✓	88.48	88.64	88.48	88.54	88.37

images. One with the size of 21243 × 15354 is used to train and the other with the size of 11265 × 15354 is to test. We further crop the images in the two datasets into small patches with the size of 256 × 256 for training, testing or validation. Besides, to objectively assess different methods, the following five evaluation metrics are utilized: precision (P), recall (R), intersection over union (IoU) of the change category, F1-score (F1), and overall accuracy (OA).

B. Experimental Settings

For a fair comparison, all compared models are implemented on PyTorch following their default settings. In this work, the ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is employed to optimize our models. The initial learning rate is set to 10^{-4}, which is halved every 40 epochs until trained 200 epochs. τ and T are set to 1 and 2 respectively. α is set to 0.1. The size of the mini-batch is set to 8.

C. Comparison and Ablation Studies

First, nine advanced models are chosen to compare with our method. The DSAMNet and the STANet are metric-based models, and the rest are classification-based models. Especially, a model that uses edge maps to guide building change detection, namely EGRCNN, is chosen. Table I reports the quantitative results on the two datasets. The recall of our method is relatively low. Since there are some changes hard to perceive like the missed changed building instance shown in Fig. 2. Its appearance is similar to the land. Besides, compared to other metric-based methods, a more strict threshold is adopted to judge whether the building has been changed, which is of benefit to inhibit irrelevant changes but tends to reduce the recall. Nevertheless, the recall of our method is still higher than that of the DSAMNet on the LEVIR-CD dataset and that of the STANet on the WHU-CD dataset. Meanwhile, our method achieves a second high precision on the LEVIR-CD dataset and the highest on the WHU-CD dataset, which can be owing to the effect of the SAM. The principle of the SAM is shown in Fig. 2. For both datasets, it can be speculated that the SAM performs well in suppressing the discrepancy of the bitemporal features in the unchanged regions. In addition, our method yields a higher F1 of 90.49 % and that of 90.50 % than other methods. It achieves the highest overall accuracy with the second fast running speed. Fig. 3 shows that our method achieves comparable visual results. Compared to the STANet, on the LEVIR-CD dataset, dense changed buildings in the selected bitemporal images are separated. The visual results on the WHU-CD dataset demonstrate its effectiveness of change identification in the complex background.

Second, we investigate the effectiveness of the ENCM and the SAM. Table II presents the ablation studies on the LEVIR-CD dataset. Compared to the baseline model (i.e., w/o the two modules), the model equipped with the SAM and the ENCM improves the recall and the precision obviously. Notably, the performance enhancement of the SAM is higher than that of the ENCM, which is consistent with their motivations. The SAM aims to selectively reduce or enhance the discrepancy of the bitemporal features. According to the examples shown in Fig. 4, it effectively improves the ability of the model to identify changes and curb adverse factors, especially the unchanged buildings with different spectral characteristics. Here, we emphasize the importance of the change-confidence map by removing it in the SAM, i.e., only employing the class-confidence map. The results are given in the 4-th combination of Table. II. The ENCM aims to refine the change detection results by enhancing the distinction of the features around the edge. From Fig. 4, more precise results are acquired by introducing the ENCM, which is owing to the improvement of the representation ability.

Furthermore, the necessity of components of the ENCM is explored, while the SAM is not integrated into the experimental model. The results of F1 are summarized in Table III. It can be found that introducing contrastive learning can improve the performance of the model with a suitable hyperparameter. In particular, the ENCM with the ENM has better accuracy and is more robust to the variation of the margin τ. It further illustrates the performance of the model is sensitive to the representations around the edge. Enhancing them is helpful. In general, their effectiveness is demonstrated by quantitative ablation experiments and visualization discussions, but there are inevitably some deficiencies. To objectively assess our model, some wrongly segmented results are given in Fig. 3(c). It suffers from difficulties under some challenging scenarios, e.g., the change of small-scale or weakly discriminative buildings. In particular, the number of changed small-scale buildings is relatively few on the datasets, which impacts its generalization on them. A possible solution is to pretrain it on building extraction datasets, which directly transfers the building-relevant knowledge into it.

IV. Conclusion

In this work, a selective attention module is designed to model the temporal-spatial correlation in the bitemporal images from a new perspective, and an edge neighborhood contrastive learning method is proposed to improve the representation ability of the model. Our method shows competitive results on two widely used building change detection datasets. In the future, the application of the given method on multi-type semantic change detection will be explored.

References

[1] J. Tian, S. Cui, and P. Reinartz, “Building change detection based on satellite stereo imagery and digital surface models,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 406–417, 2013.

[2] Q. Li, Y. Yuan, X. Jia, and Q. Wang, “Dual-stage approach toward hyperspectral image super-resolution,” *IEEE Transactions on Image Processing*, vol. 31, pp. 7252–7263, 2022.

[3] H. Luo, C. Liu, C. Wu, and X. Guo, “Urban change detection based on dempster–shafer theory for multitemporal very high-resolution imagery,” *Remote Sensing*, vol. 10, no. 7, p. 980, 2018.

[4] Q. Wang, Z. Yuan, Q. Du, and X. Li, “GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 3–13, 2019.

[5] J. Wu, B. Li, Y. Qin, W. Ni, and H. Zhang, “An object-based graph model for unsupervised change detection in high resolution remote sensing images,” *International Journal of Remote Sensing*, vol. 42, no. 16, pp. 6209–6227, 2021.

[6] J. Wu, B. Li, Y. Qin, W. Ni, H. Zhang, R. Fu, and Y. Sun, “A multiscale graph convolutional network for change detection in homogeneous and heterogeneous remote sensing images,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 105, p. 102615, 2021.

[7] R. Daudt, B. Saux, and A. Boulch, “Fully convolutional siamese networks for change detection,” in *Proc. IEEE International Conference on Image Processing*, 2018, pp. 4063–4067.

[8] H. Chen and Z. Shi, “A spatial-temporal attention-based method and a new dataset for remote sensing image change detection,” *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[9] W. Bandara and V. Patel, “A transformer-based siamese network for change detection,” in *Proc. IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 207–210.

[10] H. Chen, Z. Qi, and Z. Shi, “Remote sensing image change detection with transformers,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[11] B. Bai, W. Fu, T. Lu, and S. Li, “Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[12] Z. Chen, Y. Zhou, B. Wang, X. Xu, N. He, S. Jin, and S. Jin, “EGDE-Net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 203–222, 2022.

[13] S. Fang, K. Li, J. Shao, and Z. Li, “SNUNet-CD: A densely connected siamese network for change detection of vhr images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[14] S. Saha, P. Ebel, and X. X. Zhu, “Self-supervised multisensor change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022.

[15] J. Kang, Z. Wang, R. Zhu, X. Sun, R. Fernandez-Beltran, and A. Plaza, “PiCoCo: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10 548–10 559, 2021.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, “A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[18] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, “Exploring cross-image pixel contrast for semantic segmentation,” in *Proc. IEEE International Conference on Computer Vision*, 2021, pp. 7303–7313.

[19] S. Ji, S. Wei, and M. Lu, “Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2018.