BUILDINGS CHANGE DETECTION USING HIGH-RESOLUTION REMOTE SENSING IMAGES WITH SELF-ATTENTION KNOWLEDGE DISTILLATION AND MULTISCALE CHANGE-AWARE MODULE

Qinglie Yuan[a*], Ningjun Wang[a]

[a.] Department of Geographic Information and Remote Sensing Research Centre, School of Civil and Architecture Engineering, Panzhihua University, Panzhihua 617000, China- (*yuanqinglie, wangningjun) @pzhu.edu.cn.

KEY WORDS: building change detection, convolutional neural network, deep learning, self-attention, knowledge distillation.

ABSTRACT

Building change detection from remote sensing images is vital for many applications such as urban planning and dynamic monitoring, smart city construction, and geographical information census. In recent years, with the improvement of artificial intelligence and computer vision techniques, deep learning algorithms, especially convolutional neural networks (CNN), provide automatic detection and extraction methods. Unlike traditional approaches relying on shallow manual features, CNN can generate the deep semantic features by fusing spatial and spectral information, which is conducive to identifying building change regions. However, most deep CNN models directly fuse different level features and recover spatial details, which probably introduce redundant background information and noise from shallow layers. Considering the building's multiscale, convolution operation with fixed receptive fields cannot obtain a strong global feature response to changed regions. To address the above problems, we develop a CNN framework for automatic building region change detection using dual-temporal high-resolution remote sensing images. To refine the shallow features, self-attention knowledge distillation strategies are introduced to FCN. Furthermore, we propose the multiscale feature change-aware module to integrate the globally changed information in different decoders. Finally, the model aggregates the multi-scale regional change information and outputs the prediction map. The results of comparative visual analysis and quantitative evaluation in two public datasets demonstrate that the proposed network model can improve the accuracy and efficiency of the automatic building change detection (85.69 IoU, 97.56 OA on the WHU dataset, and 83.72 IoU, 97.64 OA on the LEVIR-CD dataset).

1. INTRODUCTION

Buildings as an artificial feature closely related to human beings have gradually become an active geographical entity with the continuous advancement of industrialization and the acceleration of urbanization. Building and urban landcover change are the comparatively obvious characteristics in the process of urbanization. Therefore, the urbanization information collection of building change has been an urgent need for government management, economic construction, and sociological research. It is significant for the acquisition of buildings change information using the accurate, rapid, and automatic method. Additionally, spatial information of buildings provides abundant prior knowledge in some fields such as urban planning, disaster emergency deployment, military detection, basic geographic information update, and other applications (Yuan et al., 2021).

With the gradual expansion of application requirements, humans put forward higher requirements for the timeliness and automatic level for the building information extraction methods. However, traditionally, manual census and sampling inspection from specific institutions have problems, such as time-consuming, laborious, and low efficiency, which are not suitable for the current urbanization information collection (Yuan et al., 2021). Although manual field survey is still the dominant operation approach in actual production, this is not competent for the needs of accurate and rapid extraction of building information in large areas and complex scenes. It is particularly urgent to complete new intelligent building identification and change detection methods.

In a few decades, the advanced technologies and platforms have made extensive application and deep progress, such as the LiDAR system, high spatial resolution earth observation satellite, UAV system, artificial intelligence (AI), which have broken the shackles and provided abundant 2D or 3D interpret information from the geographical objects (Paoletti et al.,2019). In particular, compared with traditional medium and low-resolution remote sensing images, high-resolution remote sensing images can capture abundant scene detail information, including color and texture. Also, the topology structure of ground objects can be represented correctly. Therefore, automatic buildings change detection using high-resolution remote sensing images has become the frontier in current research.

A large number of change detection methods have been proposed using multi-temporal remote sensing images. In the early days, representative change detection methods mainly focused on the spectral difference or ratio and regression analysis methods. To further explore the spectral information of images, some methods based on image transformation have been developed, such as spectral angle change analysis and principal component analysis (PCA). These direct comparison methods in pixel-wise have the advantages of simple principle and high computational efficiency, which are often used in low

and medium-resolution satellite remote sensing images. However, since the semantic correlations within surrounding pixels are not considered, they are sensitive to noise and are not suitable for high-resolution images. In addition, complex scene information and environment bring obstacles to detecting changing built-up areas due to some factors such as shadow, brightness change, and variable geometry-shape in multi-temporal images.

With the development of machine learning, many algorithms such as Support vector machine (SVM) (Suthaharan, 2016), random forest (RF) (Belgiu & Drăgu, 2016), and artificial neural networks (ANN) have greatly improved the accuracy of remote sensing image classification, stimulating the emergence of change detection methods based on classification processing, and gradually constructed the prototype framework for the remote sensing data change detection. Distinguishing with the temporal sequence of classification processing, the change detection methods based on machine learning can be divided into two categories, including comparison methods before or after classification. On the one hand, the comparison methods before classification first concatenate the multi-temporal images as inputs, and the classifiers learn and predict the changing targets. On the other hand, the comparison methods after classification identify the categories from the dual-temporal images independently and then compares the classification results to obtain the final change regions. In terms of the latter, the multiple classification errors and the inevitable registration errors can cumulatively propagate between multi-temporal images. Moreover, independence between images in different phases during classification weakens the robustness. Although the former can avoid these problems, the application in practical scenes is limited due to the manual design features and classifiers with nonlinear feature mapping.

In recent years, deep learning algorithms as the progress of AI techniques, such as convolutional neural network (CNN), recurrent neural network (RNN), and generative adversarial network (GAN), perform stronger adaptive feature extraction ability than conventional methods (Paoletti et al., 2019). The deep learning framework is suitable for the analysis and representation of applying massive data samples, which have been applied to remote sensing data processing. Therefore, buildings detection based on deep learning has been widely applied, especially CNN is used with remote sensing images for classification and object detection. For instance, Rodrigo et al. proposed three FCN structures to detect change regions for multi-spectral images. Zhang et al. combined semantic segmentation and object detection methods to identify changing objects. Chen et al. introduced the attention mechanism and modeled the spatiotemporal relationship to enhance the feature correlation in the changed region at different spatial locations and times.

Although the above methods can effectively improve results for the building change detection, there are still some challenges that need to be solved. Firstly, CNN can generate hierarchical features, but where shallow layers contain redundant information with fewer semantic features. If the shallow features are fused directly, the model will introduce interference information to reduce the optimization efficiency. Secondly, Multi-scale context can enhance the ability to change perception for the encoder of the model. Unfortunately, the existing models focus on the correlation of pixels but ignore the feature relationship between different regions with insufficient multi-scale regional context. Third, buildings present scale variation with different sizes and shape while convolutional operation cannot capture multi-scale information with the fixed size of kernels.

To solve the above issues, we develop a CNN framework combing with the novel network structures for the automatic building region change detection from high-resolution remote sensing images. In particular, the dual-temporal remote sensing images are transmitted into the network encoders and extract features complying with the weight sharing principle. To enhance multi-scale feature fusion, the Siamese feature pyramid network is constructed (Yann et al., 2005). Meanwhile, we introduce self-attention knowledge distillation training mode by constructing a spatial attention map. This strategy can assist the network to enhance the representation ability for the dual-temporal building regions and refine the shallow features through interactive learning within the model. Furthermore, we propose the multiscale feature change-aware module to enhance the processing ability of the global changed information in network decoders. This module performs the difference operation and feature transformation on the Siamese feature pyramid from two temporal remote sensing images. Also, the multi-scale spatial pooling operators are introduced to generate the similarity map and further strengthen the feature response in change regions.

2. THE PROPOSED NETWORK MODEL

2.1 The overall architecture of the model

The overall architecture of the model is shown in Figure 1. The proposed network is mainly composed of three parts: weight-shared backbone network, multi-scale change-aware module (MCA), feature pyramid fusion network, and self-attention semantic distillation (SASD) strategy.

In the dual temporal datasets, the remote sensing images as inputs with high spatial resolution can be obtained and transmitted to the proposed model. The weight-shared backbone network is used to extract the parallel features in different scales from the model encoders. These encoded feature maps construct a Siamese feature pyramid network structure with various spatial resolutions, which corresponds to image inputs in different phases. Furthermore, the MCA module-guided feature fusion unit captures the fine-grained difference features and aids the network to focus on global feature correlation the pixel-wise and enhance the features response between different scale regions. Meanwhile, the spatial attention map from the Siamese feature pyramid network can be generated. From shallow to deep semantic information, a knowledge distillation strategy can refine features and improve the efficiency of model training. Therefore, the network model establish the semantic distillation loss equation in the training stage by calculating the multi-scale spatial attention map. Then, the different features in multiple scales are fed into the change detection decoders and finally predict the change maps.

2.2 Backbone network and feature encoders

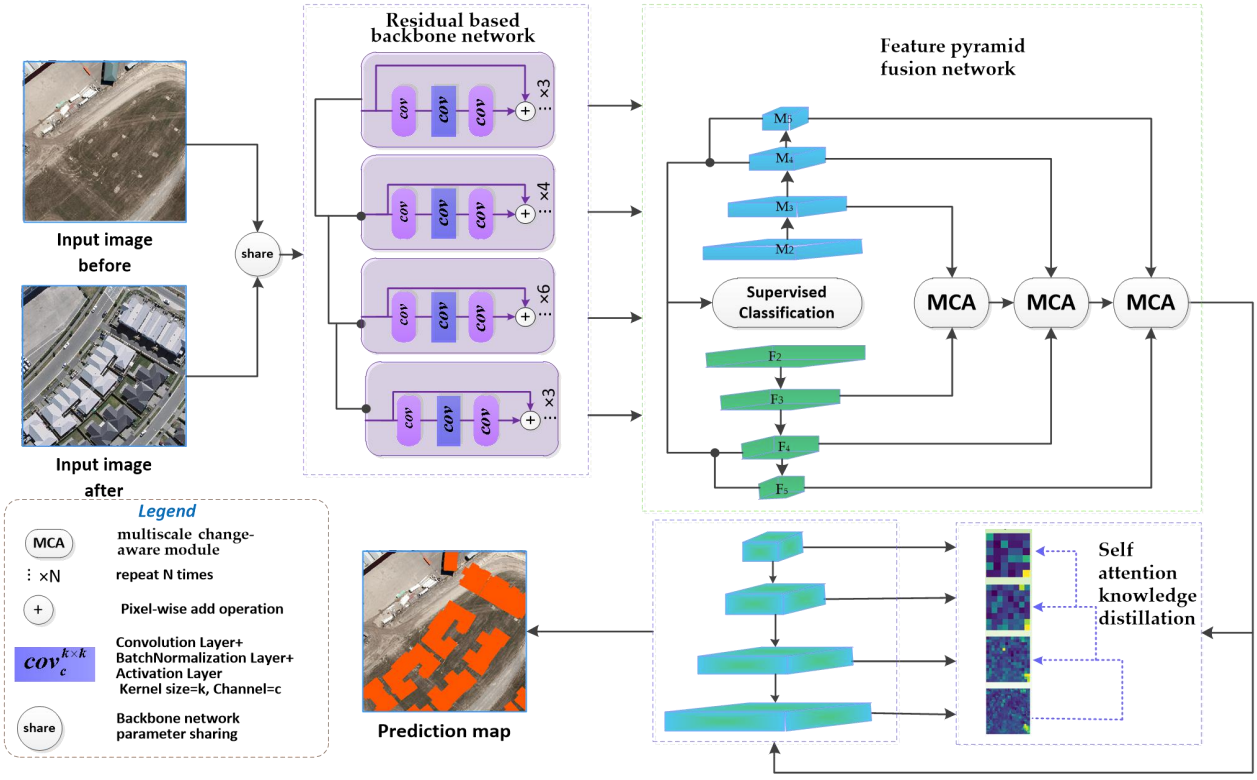


Figure 1. The proposed network framework.

The residual network is the special convolutional neural network proposed by He et al. (2016). It won the champion of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2015. The residual network is easy to optimize and can improve the accuracy by increasing a considerable depth. Its internal residual block uses skip-connection, which alleviates the problem of gradient disappearance caused by increasing layers in the deep neural network. Therefore, for the weight-shared backbone network, the modified ResNet50 as the backbone network is used to extract the multi-level features, which contains four residual convolution blocks (resBlock1~resBlock4 as presented in Figure 1).

2.3 Multiscale change-aware module

The convolution operation has a fixed receptive field and a convolution kernel with a certain size, which limits the ability to obtain global context information. Although CNN can obtain large receptive fields and comprehensive deep features as the increase of network layers, it is struggling to capture the response of similar features over a long distance. For the building change areas, the global context information can produce synergy benefit, where the irrelevant background features should be suppressed and the target features can be enhanced. In addition, due to the complexity of spectral changes of ground objects, many ground objects have intra-class variability and inter-class similarity, which brings some interference and interpretation difficulties to change detection. In particular, high-resolution remote sensing images are sensitive to this heterogeneity. Probably, some misdetection occurs when buildings are in a complex remote sensing background. Therefore, we construct a multi-scale change-aware module (MCA) to address these problems and enhance feature representation ability.

For acquiring global context information, a non-local network, proposed by Wang et al., is a classical structure that can be regarded as a spatial self-attention mechanism. Non-local neural network (Wang et al.,2018) is developed from the non-local means method. Ordinary filter with convolution kernel of $K\times K$ (K is the size of convolution kernel) that slides and calculates on the whole image, processing local information. The non-local means operation combines a relatively large search range and is weighted. Similarly, the non-local in CNN operation can increase the receptive field to the global scope, rather than being limited to a local field. Non-local operations directly capture long-distance dependencies by calculating the interaction between any two locations, instead of being adjacent points. It is equivalent to constructing a convolution kernel as large as the size of the feature map to maintain more information. A non-local network can be used as a component and combined with other network structures. This process can be modeled as follows:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i + x_j) g(x_j) , \quad (1)$$

where, x is the feature map and i represents the output location, such as the index of space, time, or space-time. Its response should enumerate J and then calculate it. The similarity between i and j is calculated by $f(x_i,x_j)$; The $g(x_j)$ calculates the representation of the feature map at position j. Finally, y_i can be obtained after standardization of response factor $C(x)$.

In the MCA module, the non-local operation is introduced to generate global context and enhance feature response in the change areas. Concretely, a pair of the encoder features with F_i and M_i (where i is the level index of residual blocks) from the Siamese feature pyramid network are converted into embedding space by conv1×1. In equation(2), $G(\cdot)$ is the feature conversion function using conv1×1 and the number C of feature channels is decreased to C/r where r is the ratio. f_i and m_i are the new

features from the result of $G(\cdot)$, respectively. Meanwhile, a difference feature X_i can be obtained by subtraction operation between f_i and m_i. Although non-local operation can capture long-distance feature response by equation (1), the burden of computation is large, which is not suitable for the shadow layer features with high spatial resolution. Additionally, the long-distance context in changed areas not only are related to any spatial position but correlates with multi-scale regions. Unfortunately, the non-local network only establishes pixel-wise spatial correlations and ignores the multiple regions. Therefore, the MCA module applies the spatial pooling pyramid (SPP) structure with multiple scale pooling layers to aggregate

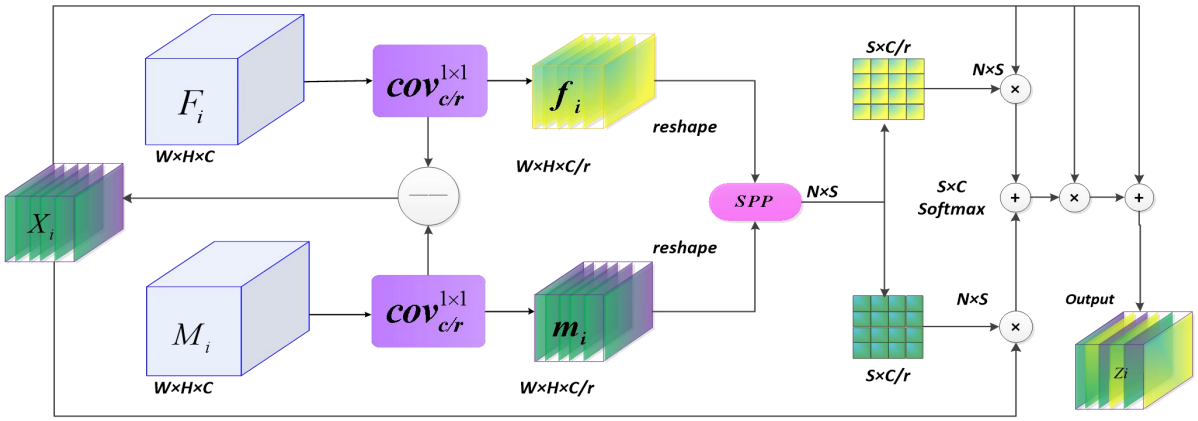


Figure 2. Multiscale change-aware module

context from the different regions (He et al., 2015). In addition, *SPP* can reduce computational complexity with max-pooling or average pooling operation. Based on SPP operation, multiple features with $F_i{}^p$ and $M_i{}^p$ are concatenated into vectors with the scale of S (S depends on pooling ratio), respectively.

$$f_i=G(F_i),\ m_i=G(M_i), \tag{2}$$

$$X_i=|\ f_i - m_i\ |, \tag{3}$$

$$F_i{}^p = SPP(f_i),\quad M_i{}^p = SPP(m_i), \tag{4}$$

$$Y_i = soft\,max(\,X_i \otimes (\,F_i^p + M_i{}^p\,))\,, \tag{5}$$

To obtain a similarity map between the difference feature and transformation feature, X_i is reshaped into two dimensions and multiplied by $F_i{}^p$ and $M_i{}^p$, respectively, as shown in equation (5) where, $\otimes$denotes matrix multiplication. Then, feature Y_i can be obtained by fusing two feature maps and normalized by the *Softmax* function. In this perspective, the differential X_i represents the feature response of the change regions and establishes a global correlation with the dual-temporal feature by the similarity matrix. Furthermore, Y_i integrates the contextual dependence from arbitrary spatial locations to multi-scale global regions. In the model structure, the MCA module is embedded into the residual connection to prevent gradient degradation. Finally, the enhanced feature Z_i can be generated by equation (6).

$$Z_i = Y_i \otimes X_i + X_i \tag{6}$$

2.4 Knowledge distillation strategy based on self-attention

Features from encoders are converted into multiple-scale layers by MCA, which constructs a spatial pyramid structure. Generally, the shadow layers have more details than the deep layers, in turn, the high-level features contain an abundant semantic feature that is easy to be classified in the decoders. To decode these features correctly, many methods apply a hierarchical fusion paradigm to aggregate information flow via up-bottom or bottom-up. This process can integrate semantic information and recover fine-grained structures for segmentation. However, if the network fuses these features directly, the shadow layers probably bring some noise information in the process of model training, which can cause instability and reduce efficiency. To alleviate this issue, we use a knowledge distillation strategy in the model training process.

Knowledge distillation is a common method of model transformation and feature refinement (Gou et al., 2021). For the former, it refers to the "knowledge distillation" of the features learned by the complex and strong learning ability network in the teacher-student network framework where useful information from the teacher network is transmitted to the student network with small parameters and weak learning ability. Furthermore, we can get a robust and powerful network, thus which is a conceptual model compression scheme. On the other hand, for the latter, distillation can make the student network learn "more soft knowledge" in the teacher network, which contains information between categories that are not available in the traditional one-hot label. Due to the nature of softening labels, distillation can also be considered a regularization strategy. The high-level semantic information from teachers' networks can guide students in selecting features effectively. In conclusion, knowledge distillation can learn not only the feature representation ability of large models but also the inter-category information.

Based on different distillation objects, the existing knowledge distillation methods mainly include the Teacher-student interaction network, Attention-based information transfer network, and the Self-attention knowledge distillation network. The first two networks require building a strong enough deep neural network as the teacher's role with a cumbersome parameter. Knowledge can be transferred from a teacher network to a small student network with a small number of parameters. However, unlike natural scene image data sets, remote sensing data samples are limited, especially for some special problems and building objects, it is not enough to construct a teacher network and cater to specific problems. Therefore, in the network structure, we introduce the knowledge distillation paradigm based on self-attention to refine the

shallow features.

Specifically, for lane detection, Hou et al. proposed self-attention knowledge distillation (SAD) that learns context relations in itself at different stages and effectively achieves improvement without any additional supervision from a cumbersome teacher network. Inspired by the SAD model, the hierarchical attention maps are constructed from the feature pyramid, as illustrated in Figure 1. In the equation (7), $\Phi^{p}_{sum}(Z_i)$ denotes the feature mapping function that transmits feature Z_i from equation (6) into 2D spatial attention maps along the channel dimension, where C_m presents *m-th* channel and *j* is the slice from Z_i; γ is set to 2 based on experiments; *Ai* is *i-th* layer attention map.

$$A_i = \Phi^{\gamma}_{sum}(Z_i^j) = \sum_{j=0}^{C_m} \left| Z_i^j \right|^{\gamma} \tag{7}$$

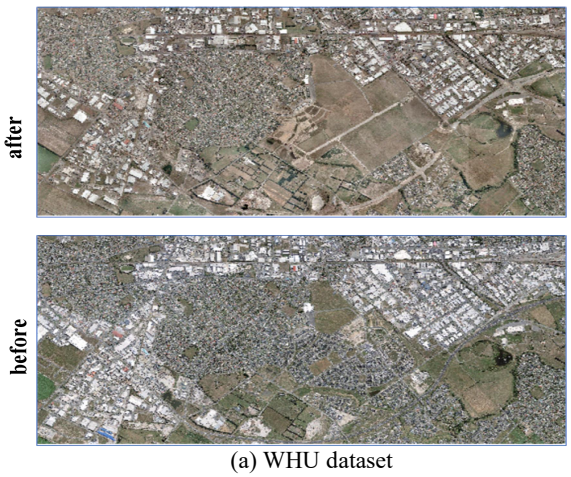
The spatial resolution of adjacent attention maps keeps consistent by bilinear interpolation for the computation of distillation loss. Finally, the distillation loss equation can be defined between adjacent attention maps(A_i, A_{i+1}) as follows:

$$Loss^{dis}(A_i, A_{i+1}) = \sum_{i=1}^{n} L_2(\varphi(A_i), \varphi(A_{i+1})) \tag{8}$$

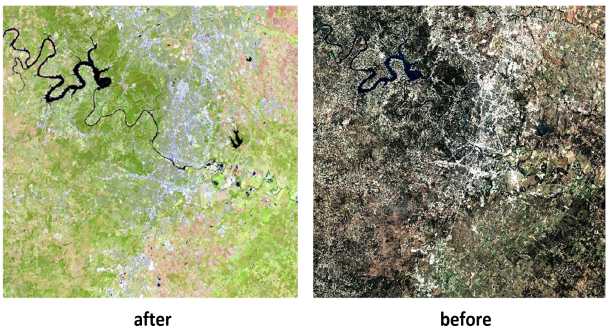
where L_2 denotes L2-norm loss function and $\varphi(\cdot)$ present distillation objects; n is set to 4. In the current path of the feature pyramid, the flow direction of knowledge distillation is from the shallow layer to the deep layer as follows: resBlock2 *mimic* resBlock3, resBlock3 *mimic* resBlock4, resBlock4 *mimic* resBlock5.

3. EXPERIMENTAL RESULTS AND DISCUSSION

To verify the effectiveness of the proposed method, the open-source building change detection datasets including the WHU dataset (Ji et al., 2018) and the LEVIR-CD dataset (Chen et al., 2020) are selected for the experimental analysis. The two data sets contain complex image scene coverage, which not only has high spatial resolution and high-quality semantic annotation but also covers buildings with complex styles, various types, and diverse structures. They are applied in many methods, providing accurate and objective ground truth for the comparison between different methods. Some samples of the experimental datasets and corresponding extent are shown in figure 3.



(a) WHU dataset



(b) LEVIR-CD dataset

Figure 3. Experimental data preview of spatial distribution in the dual-temporal images.

3.1 Data Description and experimental configuration

The WHU dataset contains the aerial images that covered 20.5km^2 of buildings and the change areas that occurred in April 2012 and 2016 years, including 12796 and 16077 buildings respectively. 30 GCPs are selected manually on the ground surface. In the experiment, the sub-dataset is extracted and the number of training, verification, and test data sets are set as 6:3:1, respectively. The whole image is divided into 1040 patches with 512 × 512 pixels and downsampled to a spatial resolution of 0.02m.

LEVIR-CD dataset contains 31,333 individual change buildings in the google satellite dual-temporal images with a period of 5 to 14 years (from 2002 to 2018). In the experiment, the number of training, verification, and test data sets are set as 6:3:1, respectively. we extract 525 images containing changed regions as the final sub-datasets, which are divided into 1050 patches with 512 × 512 pixels and a spatial resolution of 0.05m.

The Keras framework based on TensorFlow is our experimental platform, which uses the Adam optimization algorithm with an initial 10-3 learning rate, 0.1decay, and 15 batch size. The backbone model ResNet50 is initialized using the weights trained by ImageNet and trained with 200 epochs. Binary classification loss and distillation loss are used as model training total loss.

3.2 Building detection results for the proposed network

To verify the effectiveness of the proposed module, two groups of ablation experiments were completed. Meanwhile, the overall accuracy (OA) and F1 score are applied as the accuracy evaluation indicators. In the first group of experiments, the MCA module is regarded as the ablation object and SAD structures are removed. Similarly, in the second group of experiments, the SAD module is regarded as the ablation object and MCA structures are removed.

Table 1 displays experimental results in the two public datasets. Overall, in the WHU datasets, the proposed method (with RESNET+MCA+SAD) obtained overall detection accuracy of 97.56% with the F1 score of 95.83% and IoU of 85.69% and outperformed RESNET with an overall accuracy of 96.54%, IoU 82.57%, F1 score of 91.03%. Compared with the backbone network, although OA does not have a significant increase, the application of the MCA model can improve the accuracy of IOU by 0.59 and F1 by 3.18. In the LEVIR-CD datasets, the proposed method obtained overall detection accuracy of 97.64% with the F1 score of 87.67% and IoU of 83.72% and outperformed RESNET. Compared with the backbone network,

the application of the SAD strategy can improve the accuracy of IOU by 1.4 and F1 by 3.63. In conclusion, the combination of MCA and SAD can significantly improve the accuracy of model detection.

Method	Datasets					
	WHU			LEVIR-CD		
	IoU (%)	OA (%)	F1 (%)	IoU (%)	OA (%)	F1 (%)
RESNET	82.57	96.54	91.03	82.32	97.55	83.97
RESNET +MCA	83.16	96.52	94.21	82.44	97.50	84.16
RESNET +SAD	84.47	97.41	95.74	83.57	97.57	85.22
RESNET +MCA+SAD	**85.69**	**97.56**	**95.83**	**83.72**	**97.64**	**87.67**

Table 1. Comparison accuracy on the different datasets in the ablation result of modules. The bold values denote the best result.

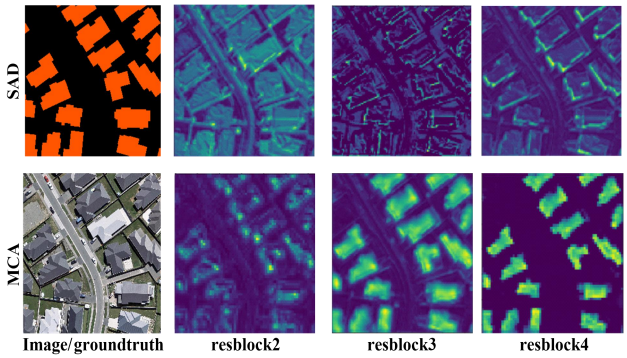


Figure 4. Heatmap in different module applications.

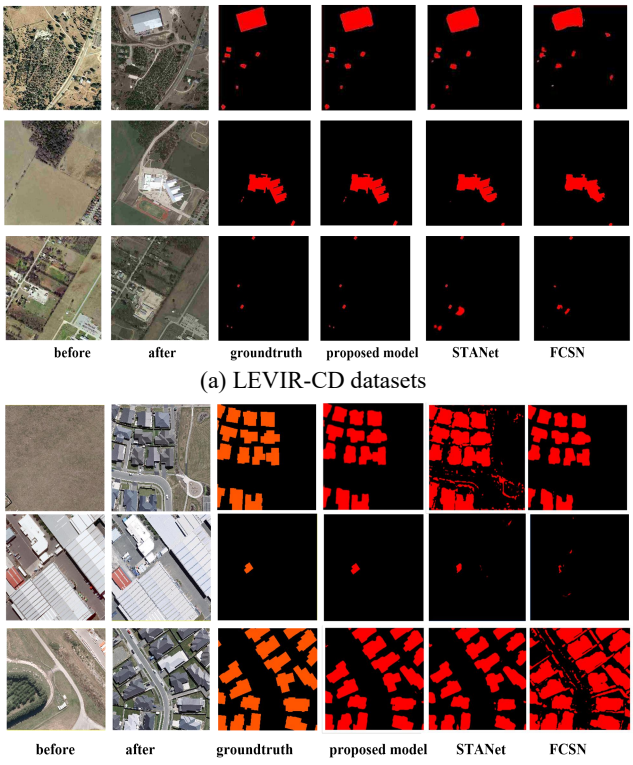
The heat maps in the spatial domain display the response before and after feature transformation via MCA or SAD, as shown in figure 4. We calculate the average fused feature from residual block2 to block4 in the channel dimension. Compared with heatmaps in different stages, most of the background-related information is suppressed after knowledge distillation. In addition, the large-scale building change area has a more significant holistic response than the previous local attention.

Furthermore, the proposed model is compared and analyzed with recent methods. Daudt et al. (2018) proposed fully convolutional Siamese networks (FCSN) for change Detection adopting three network structures including fully convolutional fusion, Siamese concatenation, and fully convolutional Siamese difference. In the experiment, the last network structure is applied for comparison. Chen et al. developed a spatial-temporal attention neural network (STANet) for change detection using remote sensing images. This module introduces the attention mechanism and exploits spatial-temporal correlation to enhance the ability of feature representation.

Table 2 shows the accuracy of the proposed methods and other detectors for the evaluation results on the different datasets. It can be observed that FCSN has the lowest IoU in LEVIR-CD, which indicates large false changed areas are detected. Although STANet is slighter low than ours in F1 and IoU, the proposed methods achieve better performance in OA. Visually, FCSN has a weak ability to detect small buildings as shown in figure 5. STANet obtained a better result in multiscale buildings than FCSN, but some false detection and missed detection occur in some buildings. Overall, the proposed method outperformed others in different scale regions, but the details of buildings have false detection in parts of unchanged areas.

Method	Datasets					
	WHU			LEVIR-CD		
	IoU (%)	OA (%)	F1 (%)	IoU (%)	OA (%)	F1 (%)
FCSN	85.47	96.85	95.76	72.32	93.55	83.97
STANet	85.29	96.21	94.21	82.44	94.50	87.30
Proposed method	**85.69**	**97.56**	**95.83**	**83.72**	**97.64**	**87.67**

Table 2. Comparison Accuracy using different methods. The bold values denote the best result.



(a) LEVIR-CD datasets

(b) WHU datasets

Figure 5. buildings change detection results compared to other methods in different datasets

4. CONCLUSION

This paper develops a fully convolution neural network framework for building change detection using high spatial resolution dual-temporal remote sensing images. To enhance the ability of feature extraction, two novel modules are proposed. The multi-scale change aware module can aggregate the global context from multiple spatial feature regions. The knowledge distillation strategy refines shallow semantic information and improves the detection ability of small buildings. The backbone network adopts the Siamese feature pyramid combined with the two models, which can improve the accuracy of the results. In future research, we will further explore the feature relationship of multi-temporal images, and extend the knowledge distillation

method to the multi-scale decoders to enhance the perception of changing regions.

REFERENCES

Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114, 24-31.

Chen, H., Shi, Z., 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10), 1662.

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1, 539-546.

Daudt, R. C., Le Saux, B., Boulch, A., 2018. Fully convolutional siamese networks for change detection. *2018 25th IEEE International Conference on Image Processing* (*ICIP*). IEEE, 4063-4067.

Gou, J., Yu, B., Maybank, S. J., & Tao, D., 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6), 1789-1819.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. *European conference on computer vision*, Springer, Cham, 630-645.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9), 1904-1916.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1), 574-586.

Paoletti, M. E., Haut, J. M., Plaza, J., Plaza, A., 2019. Deep learning classifiers for hyperspectral imaging: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158, 279-317.

Suthaharan, S., 2016: Support vector machine. *Machine learning models and algorithms for big data classification*. Springer, Boston, MA, 207-235.

Wang, X., Girshick, R., Gupta, A., & He, K., 2018. Non-local neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*,7794-7803.

Yuan, Q., Mohd Shafri, H. Z., Alias, A. H., Hashim, S. J. B., 2021. Multiscale semantic feature optimization and fusion network for building extraction using high-resolution aerial images and LiDAR data. *Remote Sensing*, *13*(13), 2473.

Yuan, Q., Ang, Y., & Shafri, H. Z. M., 2021. Hyperspectral Image Classification Using Residual 2D and 3D Convolutional Neural Network Joint Attention Model. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 44, 187-193.

Zhang, L., Hu, X., Zhang, M., Shu, Z., & Zhou, H., 2021. Object-level change detection with a dual correlation attention-guided detector. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177, 147-160.