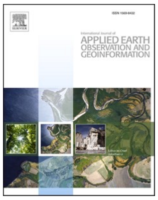
Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



A high-resolution feature difference attention network for the application of building change detection

Xue Wang [a,b], Junhan Du [a,b], Kun Tan [a,b,*], Jianwei Ding [c], Zhaoxian Liu [c], Chen Pan [d], Bo Han [e]

[a] *Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China*
[b] *Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), East China Normal University, Shanghai 200241, China*
[c] *The Second Surveying and Mapping Institute of Hebei, Shijiazhuang 050037, China*
[d] *Shanghai Municipal Institute of Surveying and Mapping, Shanghai 200063, China*
[e] *Institute of Remote Sensing Satellite, China Academy of Space Technology, Beijing 100094, China*

ARTICLE INFO

Keywords:
Difference attention module
Multi-scale feature
Change detection
Deep learning
Siamese network structure

ABSTRACT

Deep learning based change detection has brought a significant improvement in the accuracy and efficiency when compared with conventional machine learning methods. However, the issues of the lack of differential information and the diversity of the scale features of artificial objects are crucial barriers to the application of building change detection algorithms. A novel deep learning based approach named the high-resolution feature difference attention network (HDANet) is proposed in this work to solve these issues. HDANet can handle the change characteristics well, due to the Siamese network structure. To tackle the loss of the spatial features of buildings caused by the multiple successive down-sampling operations in the current change detection algorithms using fully convolutional networks (FCNs), a multi-resolution parallel structure is introduced in HDANet, and the image information with different resolutions is comprehensively employed, without any spatial information loss. Moreover, an innovative difference attention module is elaborated for the enhancement of the sensitivity to difference information, to keep the building change information. The experimental results obtained on building change detection datasets confirm that HDANet can improve the differential feature representation for change detection, and the performance of the building change detection is also superior to that of the other advanced change detection methods.

1. Introduction

Remote sensing technology is capable of obtaining information about the Earth's surface periodically and extracting the dynamic changes of the Earth's surface rapidly. As such, remote sensing is widely utilized in the fields of land resource surveying (Wang, Tan et al., 2022), urban building extraction (Deng, Shi et al., 2021), pollution detection (Niu, Tan et al., 2021), and military applications (Qin, Cai et al., 2021). How to identify building and land surface changes accurately has become one of the essential issues in remote sensing.

The difference features between images obtained by image preprocessing are utilized in the traditional change detection methods mainly utilize. The image preprocessing operations consist of radiation correction and image registration. Change vector analysis (CVA), which regards each band as a feature vector, is one of the traditional difference image feature generation approaches, and is used in many change detection methods (Bovolo and Bruzzone, 2006). Euclidean distance calculation of the feature vector can be utilized to indicate the change intensity of the corresponding pixel. The development of machine learning in many research fields has promoted the application of the related algorithms in change detection. Change detection can be treated as a classification task with binary categories, in which each pixel should be labeled as either "changed" or "unchanged". Other machine learning methods (Volpi, Tuia et al., 2013, Wessels, Van den Bergh et al., 2016) have also been investigated in change detection. Du and Liu (Du and Liu, 2012) noted that the utilization of rich features can boost the change detection performance. In the above-mentioned traditional approaches, the detection is typically divided into two processes—feature extraction and detection—and when the two processes cannot be combined as an entirety, the optimal solution will not be globally optimal.

* Corresponding author at: Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China.
E-mail address: tankuncu@gmail.com (K. Tan).

https://doi.org/10.1016/j.jag.2022.102950
Received 10 May 2022; Received in revised form 27 June 2022; Accepted 29 July 2022
Available online 6 August 2022

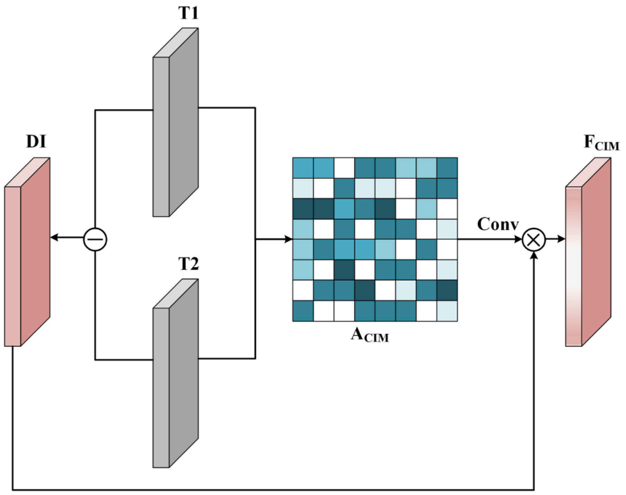


Fig. 1. Illustration of the difference attention module.

Deep learning has great feature mining and feature representation abilities. The deep learning application in change detection has become one of the hotspots in the remote sensing field (Hong, Gao et al., 2020, Hong, Gao et al., 2020). Convolutional neural networks (CNNs) (Albawi, Mohammed et al., 2017) can be utilized with remote sensing data through the convolution operation. When a patch-based method is implemented in change detection, the variable characteristics or the combination information of the bi-temporal images is input into the CNN to obtain the “changed” or “unchanged” label for the given patch. Wang et al., (Wang, Yuan et al., 2018) utilized the fusion feature of the difference image as the input of the CNN and found that the spatio-spectral features can achieve high-accuracy detection results. Some advanced CNN structure have been introduced in the remote sensing field, such as transformer (Hong, Han et al., 2021), Graph CNN(Gao, Hong et al., 2020) and Siamese network (Wang, Tan et al., 2020, Deng, Shi et al., 2021, Hong, Han et al., 2021, Zheng, Gong et al., 2022). The CNN-based Siamese network structure has been shown to be appropriate for change detection, due to its similarity comparison ability. Lin et al., (Lin, Li et al., 2019) proposed a feature map fusion approach for a Siamese CNN by multiplying the two branches of the Siamese feature matrix before the fully connected layer, and the results demonstrated that the feature fusion can enhance the change information. Mou et al., (Mou, Bruzzone et al., 2018) utilized the integration of a CNN-based Siamese network structure and a modified recurrent neural network to include bi-temporal image as the sequential features in the Siamese discrimination process, which outperformed the other mainstream deep learning methods. However, the flaw of the patch-based inference process is the many repeated steps, which cause the waste of computing resources and the low efficiency of the algorithms.

Remote sensing has entered the era of mass production, which significantly improves the situation of the image information lagging behind the spatio-temporal characteristics of ground objects (Hong, Gao et al., 2020). How to detect the changed areas rapidly and accurately for a large range of data is a hot issue. Unfortunately, it is time-consuming to infer pixel by pixel using a CNN with a fully connected layer. The fully convolutional networks (FCNs) can quickly infer the semantic segmentation and can label every pixel in an image concurrently, which makes them suitable for the change detection task. The encoder-decoder framework allows the FCN to be utilized in change detection in large scenes. Peng et al., (Peng, Zhang et al., 2019) utilized an improved FCN named UNet++, which employs a dense feature map shortcut link to promote the detection performance. Liu et al., (Liu, Jiang et al., 2020) replaced the traditional convolution operation with depthwise separable convolution and proposed a modified FCN, which improved the change detection accuracy and efficiency. To further boost the detection accuracy, Zhang et al., (Liu, Jiang et al., 2020) investigated a two-branch architecture in an FCN to learn the deep global features in two temporal images, which is followed by deeply supervised discriminating optimization. Ding et al., (Ding, Shao et al., 2021) proposed an FCN-based network with three-dimensional filters, which can effectively capture the spatio-spectral features of objects with different sizes. Some efforts have also been devoted to the construction of high-quality building change detection datasets (Ji, Wei et al., 2018, Chen and Shi, 2020, Shi, Liu et al., 2022), which have been utilized for the performance assessment of the innovative approaches. Moreover, the combination of Siamese and FCN have been investigated on change detection. Li et al., (Li, Yan et al., 2022) proposed a densely attentive refinement network to improve change detection, which employed the UNet encoder–decoder architecture with the Siamese network. Zheng et al., (Zheng, Wei et al., 2022) proposed a deep Siamese pairwise potential CRFs network which utilized the conditional random field method in the FCN and Siamese network to improve the change detection accuracy.

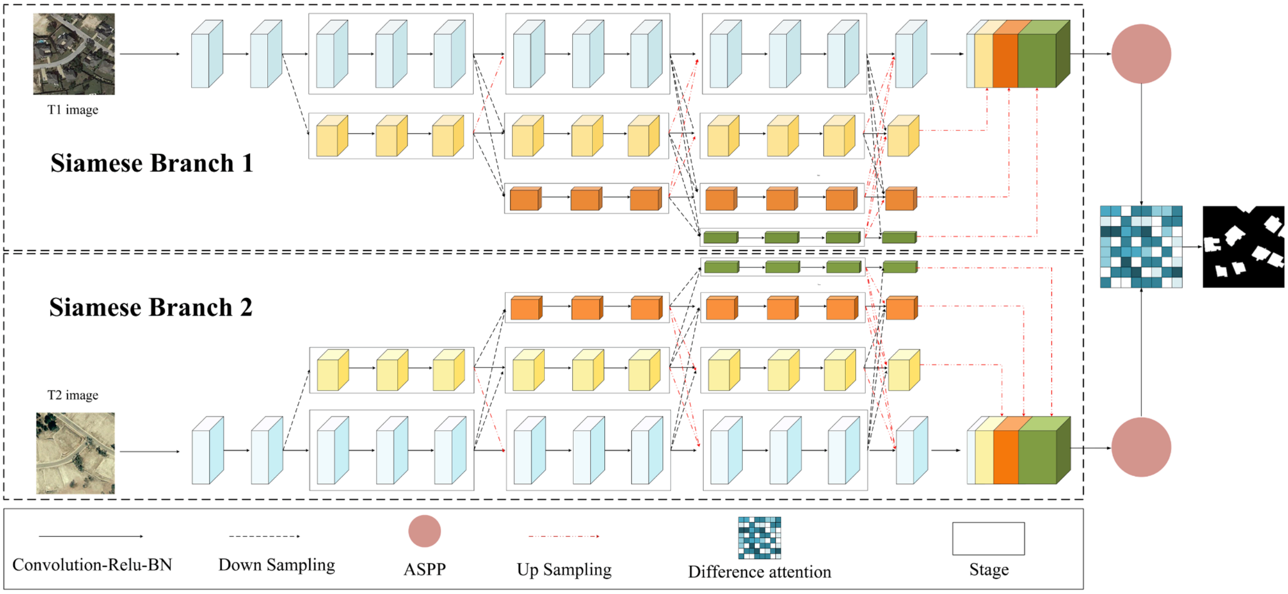


Fig. 2. The structure of the HDANet framework.

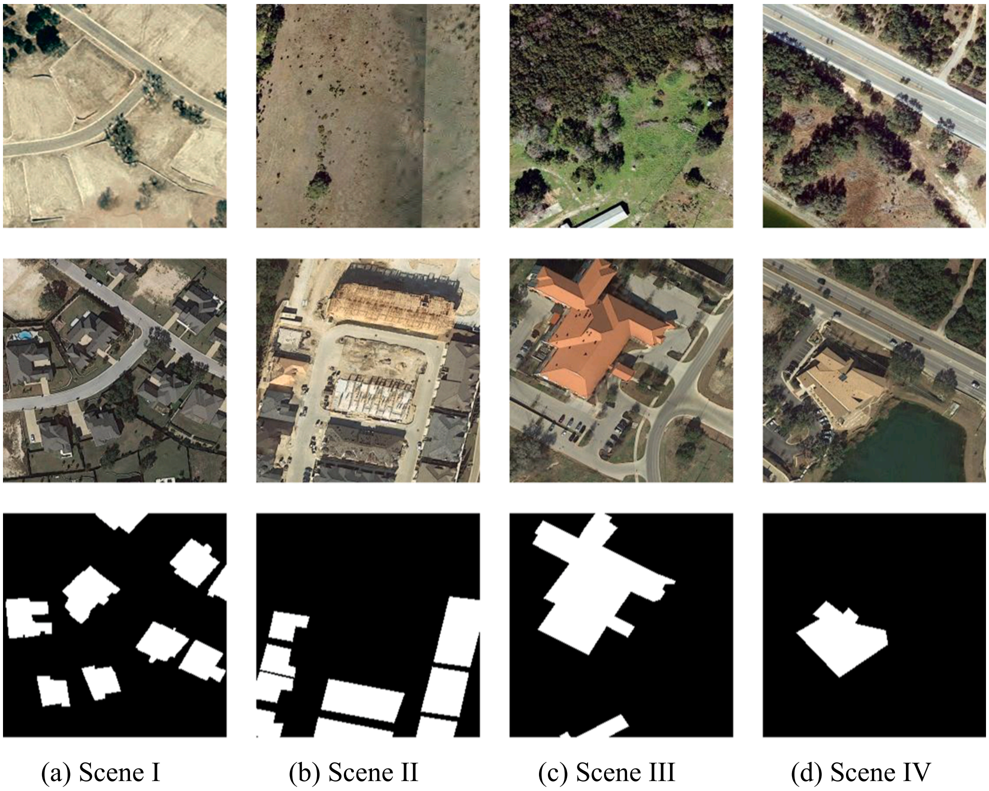


Fig. 3. Scenes from the LEVIR-CD dataset (the first row denotes T1 images, the second row denotes T2 images, and the last row denotes the label images).

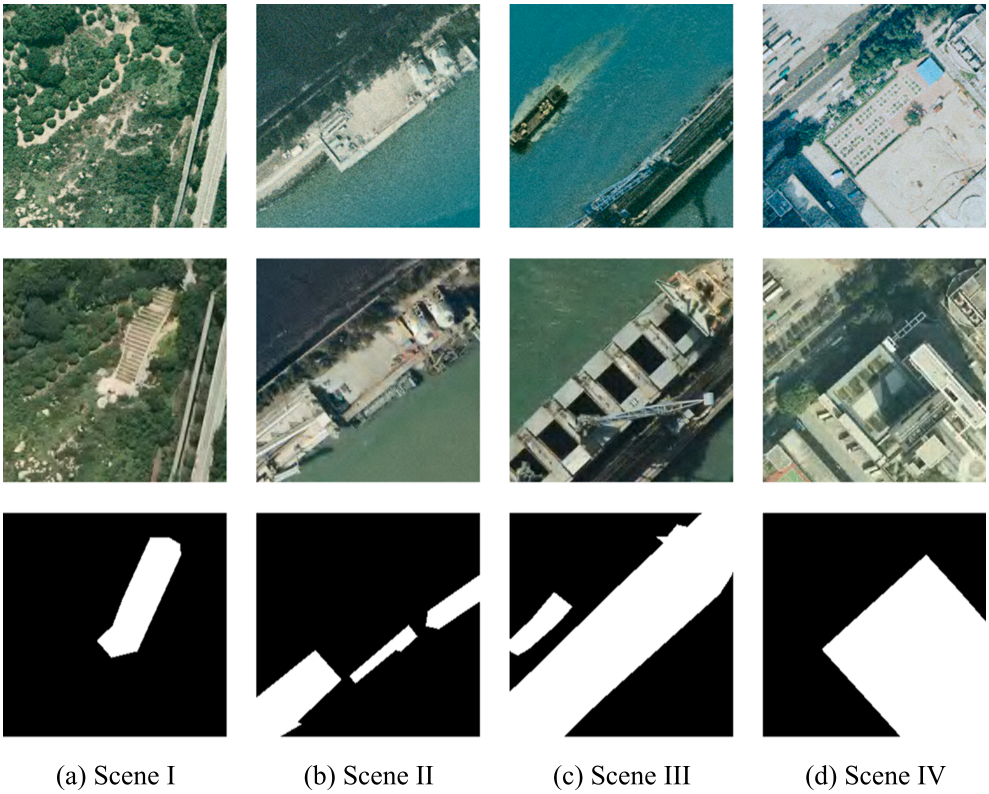


Fig. 4. Scenes from the SYSU-CD dataset (the first row denotes T1 images, the second row denotes T2 images, and the last row denotes the label images).

Zheng et al., (Zheng, Gong et al., 2022) investigated a high frequency attention-guided Siamese network to enhance high frequency information of changes and detect edges of changed area.

The above-mentioned deep learning methods can obtain high-precision detection results in most change detection tasks, and represent a significant improvement in accuracy and efficiency over the

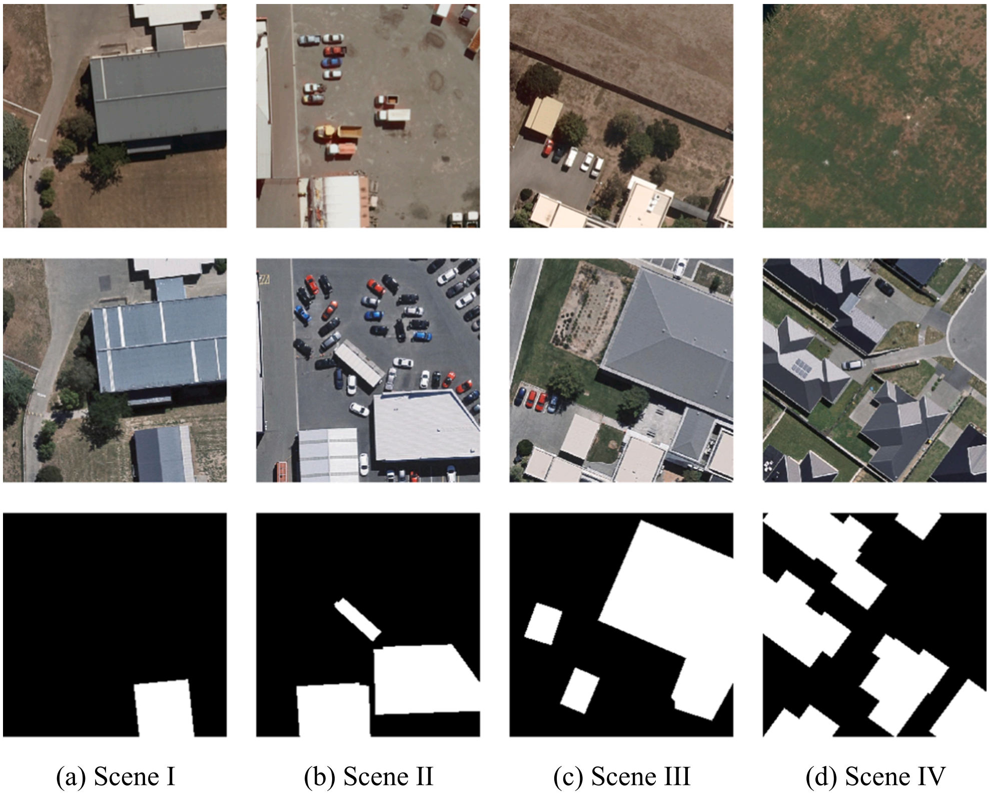


Fig. 5. Scenes from the WHU building dataset (the first row denotes the T1 images, the second row denotes the T2 images, and the last row denotes the label images).

Table 1
The accuracy of the different approaches for the LEVIR-CD dataset.

Method	Precision	Recall	F1-score	Kappa
FC	0.9016	0.7255	0.8040	0.7947
FC_S_C	**0.9394**	0.7597	0.8401	0.8324
FC_S_D	0.8911	0.7756	0.8293	0.8209
SegNet	0.9247	0.7094	0.8029	0.7938
DeepLabV3	0.9003	0.8251	0.8611	0.8539
DSIFN	0.9278	0.8211	0.8712	0.8647
STANet	0.9201	0.8333	0.8746	0.8682
UNet++	0.9144	0.8524	0.8823	0.8762
Siam_HRNet	0.9108	0.8479	0.8782	0.8719
HDANet	0.9226	**0.8761**	**0.8987**	**0.8934**

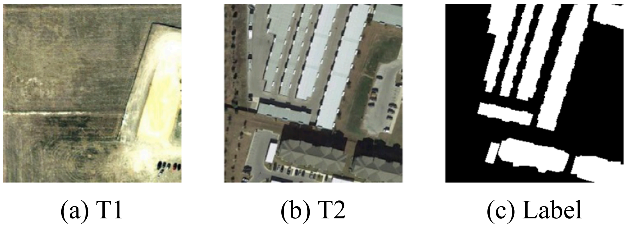


Fig. 6. The first change scene in the LEVIR-CD dataset.

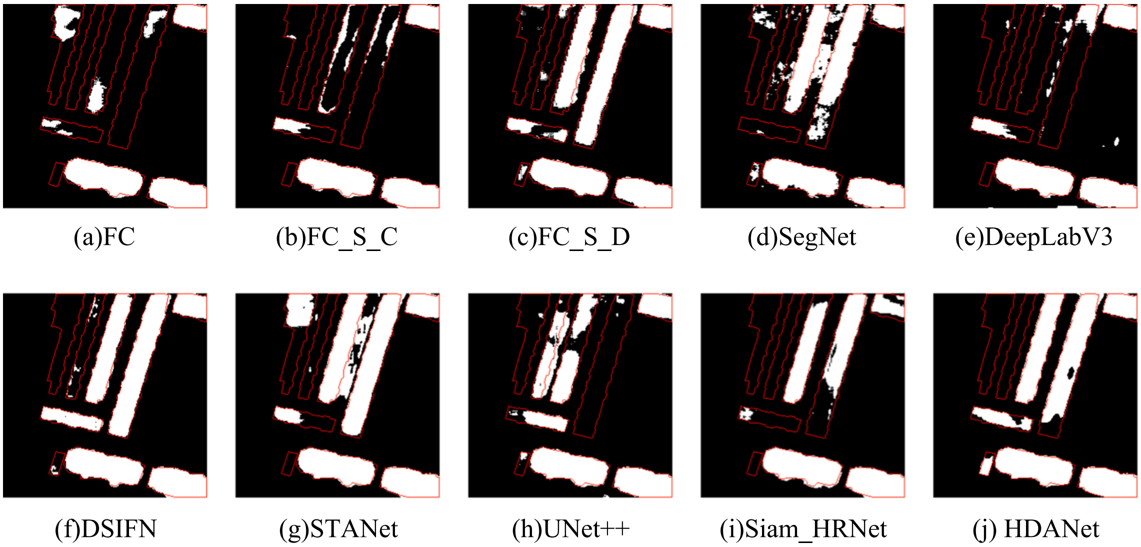
conventional machine learning methods. Building changes have regular shapes, close arrangements, and various scales, which represent a challenge to the deep learning based change detection methods because the differential information is non-significant and the performance of the feature extraction in different scales is poor. In this regard, an innovative FCN-based change detection algorithm named the high-resolution feature difference attention network (HDANet) is proposed in this work. To handle the loss of detailed context features caused by the successive down-sampling operations, a multi-resolution parallel structure is introduced in HDANet, in which the image information with different resolutions is comprehensively employed without any spatial information loss caused by multiple down-sampling. HDANet is carried out based on a Siamese network structure, which allows HDANet to represent the change characteristics well. Moreover, atrous spatial pyramid pooling (ASPP) is conducted by combining a group of convolution kernels in parallel with preset atrous rates, to represent the various features.

Although these approaches are able to handle change detection tasks with higher efficiency and accuracy, there still exist two main drawbacks that hinder the performance to be further improved, which can be described as follows:

(1) In the existing encoding–decoding change detection methods, fine-resolution features with successive down-sampling operation are lost when encoding high-level features and the lost spatial information cannot be effectively restored in the up-sampling process. As a result, the edge and interior area of the changes may be mistaken which leads to great loss in performance.
(2) The recently change detection methods are tended to lose the change information such as the diversity of the scale features and the difference characteristics which will cause the low change detection accuracy.

To deal with the aforementioned problems, a novel deep learning based approach named the high-resolution feature difference attention network (HDANet) is proposed for change detection using bi-temporal remote sensing images. The main contributions of this work are given as follows.

(1) We investigate a novel HDANet for the change detection in an encoder-decoder manner. The network employs a Siamese


(a)FC (b)FC_S_C (c)FC_S_D (d)SegNet (e)DeepLabV3

(f)DSIFN (g)STANet (h)UNet++ (i)Siam_HRNet (j) HDANet

Fig. 7. The results obtained by the different methods with the first change scene in the LEVIR-CD dataset.

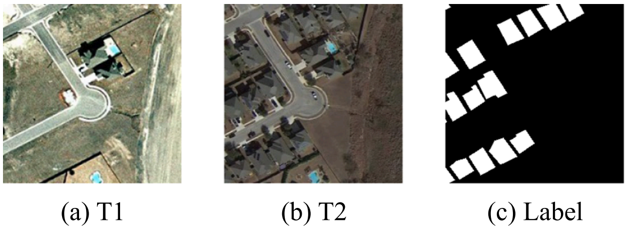

(a) T1 (b) T2 (c) Label

Fig. 8. The second change scene in the LEVIR-CD dataset.

feature discriminant structure with shared weights to extract the difference features with two period remote sensing image.

(2) We integrate the feature representation with different resolutions and scales in parallel, and the detailed change information is well kept for the final detection.

(3) We devise a differential attention module (DAM) based on change intensity, which can extract the differential features of two different temporal images effectively.

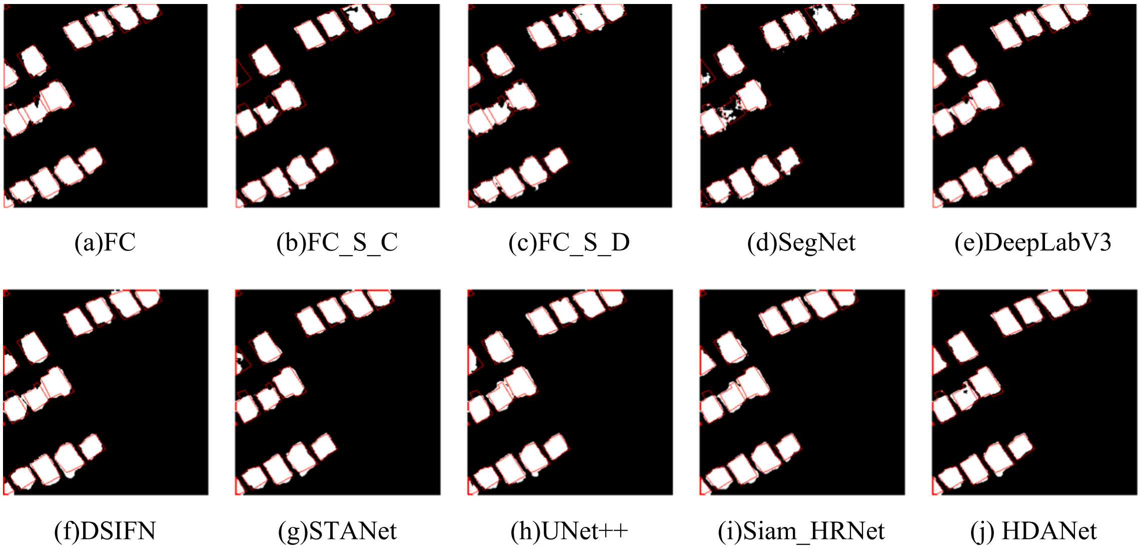
(4) From comprehensive comparisons among the recent CNN-based change detection approaches, our proposed method is able to achieve state-of-the-art performance.

2. Previous works

2.1. Conventional CNNs and FCNs

In the conventional CNNs the inference model utilizes the patches of the whole image as the input. The change results can be obtained by repetitive iterations though all the pixels, one after another. However, there will be a lot of overlapping areas between the patches of adjacent pixels, so that the efficiency of the convolution operation per pixel is relatively low. The convolutional layer can only extract the features of local areas, resulting in a poor detection performance. The output of a single layer in a CNN is followed by the fully connected layer and after the transformation of the fully connected layer, the softmax is employed.

$$lgt = z^{fc} \tag{1}$$


(a)FC (b)FC_S_C (c)FC_S_D (d)SegNet (e)DeepLabV3

(f)DSIFN (g)STANet (h)UNet++ (i)Siam_HRNet (j) HDANet

Fig. 9. The results obtained by the different methods with the second change scene in the LEVIR-CD dataset.

Table 2
The accuracy of the different approaches for the SYSU-CD dataset.

Method	Precision	Recall	F1-score	Kappa
FC_EF	0.8269	0.6308	0.7156	0.6427
FC_Siam_conc	0.7792	0.7208	0.7488	0.6752
FC_Siam_diff	**0.8492**	0.6430	0.7319	0.6635
SegNet	0.8235	0.6629	0.7345	0.6638
DeepLab v3	0.8099	0.7065	0.7547	0.6856
DSIFN	0.7932	0.7285	0.7595	0.6894
STANet	0.8038	0.7475	0.7746	0.7084
UNet++	0.8144	0.7466	0.7790	0.7146
Siam_HRNet	0.8095	0.7391	0.7727	0.7066
HDANet	0.7853	**0.7988**	**0.7920**	**0.7271**
Method	**Precision**	**Recall**	**F1-score**	**Kappa**
FC	0.8269	0.6308	0.7156	0.6427
FC_S_C	0.7792	0.7208	0.7488	0.6752
FC_S_D	**0.8492**	0.6430	0.7319	0.6635
SegNet	0.8235	0.6629	0.7345	0.6638
DeepLabV3	0.8099	0.7065	0.7547	0.6856
DSIFN	0.7932	0.7285	0.7595	0.6894
STANet	0.8038	0.7475	0.7746	0.7084
UNet++	0.8144	0.7466	0.7790	0.7146
Siam_HRNet	0.8095	0.7391	0.7727	0.7066
HDANet	0.7853	**0.7988**	**0.7920**	**0.7271**

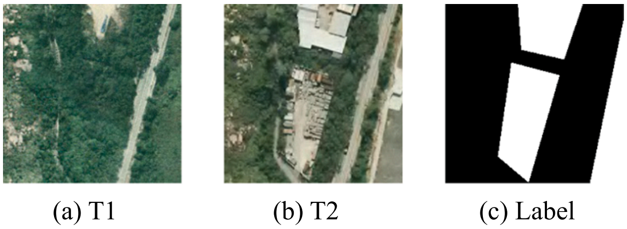

(a) T1 (b) T2 (c) Label

Fig. 10. The first change scene in the SYSU-CD dataset.

$$C_i = e^{lgt_i} / \sum_j e^{lgt_j} \tag{2}$$

where fc represents the transformation of the fully connected layer; and $lgt = (0, 1)$ indicates the logit, which describes the confidence toward "changed" and "unchanged".C_i denotes the output from the softmax layer.

Compared with the traditional CNN, the FCN abandons the following fully connected layer and replaces it with a convolutional layer, which forms an encoder-decoder framework. The input of the FCN does not need to be of a fixed image size. After the processing by the softmax function, the value of each channel represents the probability of the label.

2.2. Siamese neural networks

A Siamese neural network is a special network structure used to compare the similarity of input samples, where the characteristics of the different input data are obtained through a set of weight-sharing networks (Wu, Wang et al., 2018).

The main characteristic of a Siamese network is that it is composed of two sub-networks whose weights and structures are identical. According to the requirement of the task, a Siamese network can be made up of any basic neural network model, such as a CNN or an FCN. In a Siamese network, the backbone network contains two sub-networks, which can be regarded as two branches to process two sets of different data. The two network branches encode the input data to generate high-level features of the original data, where the aim is to represent the feature similarity of the bi-temporal data. The input data are generally the images of two periods in change detection. In the conventional change detection methods, the input is the fusion feature of bi-temporal images, which is called early fusion. The disadvantage of early fusion is that it is hard to keep the independence of the detailed features because of the mutual interference between each band of the two periods. Meanwhile, in a Siamese network, the images of the two periods are utilized as independent inputs, so that the network can retain the independence of the features through the shallow and deep layers from the two periods, without causing the two images to affect each other.

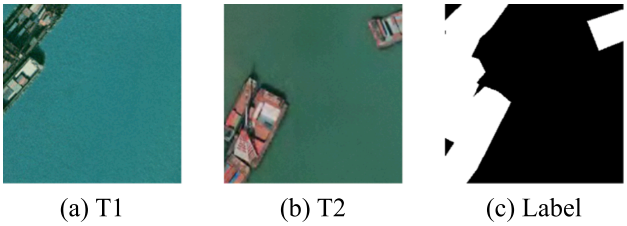

(a) T1 (b) T2 (c) Label

Fig. 12. The second change scene in the SYSU-CD dataset.

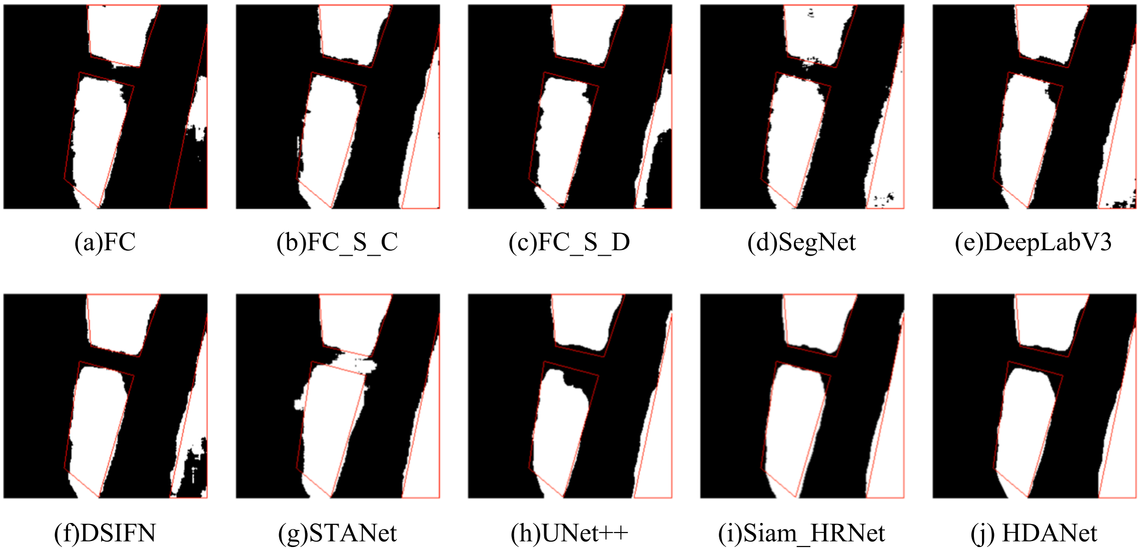

(a)FC (b)FC_S_C (c)FC_S_D (d)SegNet (e)DeepLabV3
(f)DSIFN (g)STANet (h)UNet++ (i)Siam_HRNet (j) HDANet

Fig. 11. The results obtained by the different methods with the first change scene in the SYSU-CD dataset.

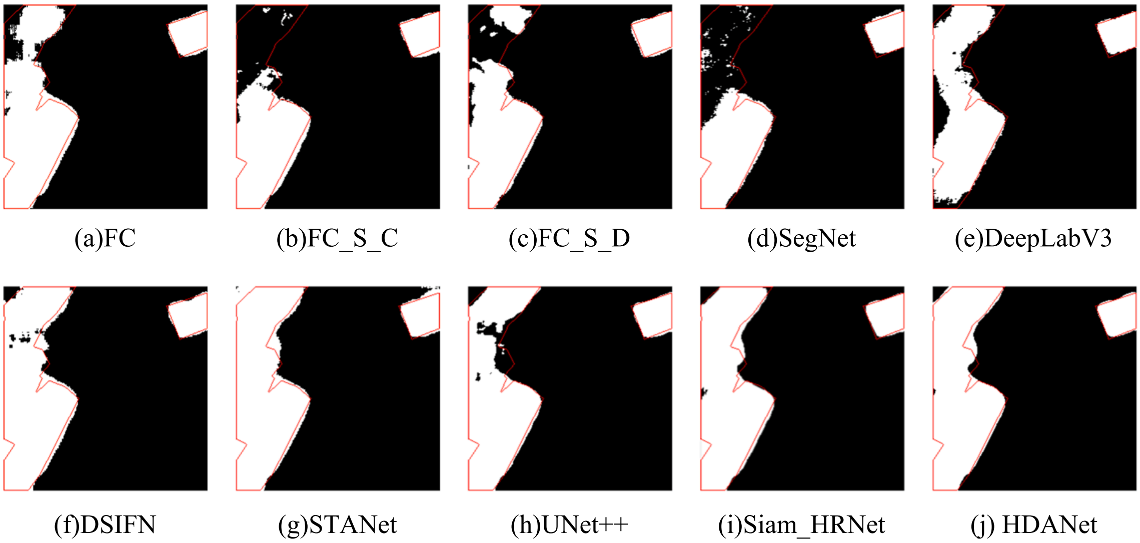


Fig. 13. The results obtained by the different methods with the second change scene in the SYSU-CD dataset.

Table 3
The accuracy of the different approaches for the WHU building dataset.

Method	Precision	Recall	F1-score	Kappa
FC_EF	0.7623	0.7765	0.7693	0.7603
FC_Siam_conc	0.8831	0.7261	0.7969	0.7899
FC_Siam_diff	0.8020	0.7631	0.7821	0.7739
SegNet	0.7813	0.6878	0.7316	0.7219
DeepLab v3	0.8256	0.8197	0.8226	0.8158
DSIFN	0.8686	0.8093	0.8379	0.8319
STANet	0.8601	**0.8340**	0.8468	0.8410
UNet++	0.8906	0.7898	0.8372	0.8313
Siam_HRNet	0.8806	0.8098	0.8437	0.8380
HDANet	**0.8987**	0.8255	**0.8605**	**0.8554**
Method	**Precision**	**Recall**	**F1-score**	**Kappa**
FC	0.7623	0.7765	0.7693	0.7603
FC_S_C	0.8831	0.7261	0.7969	0.7899
FC_S_D	0.8020	0.7631	0.7821	0.7739
SegNet	0.7813	0.6878	0.7316	0.7219
DeepLabV3	0.8256	0.8197	0.8226	0.8158
DSIFN	0.8686	0.8093	0.8379	0.8319
STANet	0.8601	**0.8340**	0.8468	0.8410
UNet++	0.8906	0.7898	0.8372	0.8313
Siam_HRNet	0.8806	0.8098	0.8437	0.8380
HDANet	**0.8987**	0.8255	**0.8605**	**0.8554**

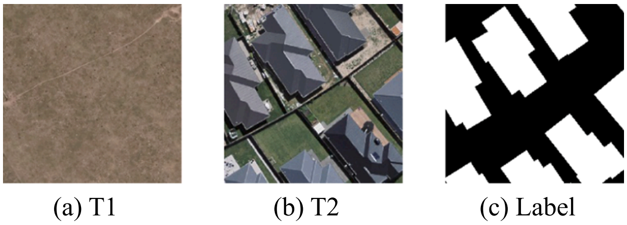


Fig. 14. The first change scene in the WHU building dataset.

2.3. Accuracy evaluation

In change detection, the result is to label each pixel into a changed or unchanged category, which is a pixel-level binary classification task. Therefore, the common accuracy evaluation indicators used in classification can be used to assess the detection performance, i.e., kappa coefficient, F1-score, recall, and precision. These indicators are calculated as follows:

$$Precision = \frac{PC}{PC + PM} \tag{3}$$

$$Recall = \frac{PC}{PC + NM} \tag{4}$$

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

$$OA = \frac{PC + NC}{PC + NC + PM + NC} \tag{6}$$

$$P = \frac{(PC + PM)(PC + NM) + (NM + NC)(PF + NC)}{(PC + NC + PM + NC)^2} \tag{7}$$

$$kappa = \frac{OA - P}{1 - P} \tag{8}$$

where PC represents the number of correctly classified “changed” labels. PM denotes the count of misclassified “changed” classes. NC is the count of classified “unchanged” labels correctly, which represents the pixels that have not changed and are classified as unchanged. NM denotes the number of misclassified “unchanged” classes.

3. Proposed method

To improve the detection performance and allow the model to handle high-precision change detection with various complex artificial objects, a new method based on high-resolution feature representation and a difference attention module is proposed. The HRNet architecture (Sun, Xiao et al., 2019) with a Siamese network structure is introduced as the basic skeleton of the network, where the difference features for the bi-temporal images are enhanced through the difference attention module, based on the intensity of the changes.

3.1. Difference attention module

The aim of an attention module is to filter out the important information and ignore the unimportant background information from the redundant features. It is often utilized to enhance the feature information. A spatial attention is utilized to highlight the information in different positions of the feature map, and the purpose of a channel attention module is to enhance the characteristics in the different channels, giving different weights to each channel (Hu, Shen et al.,

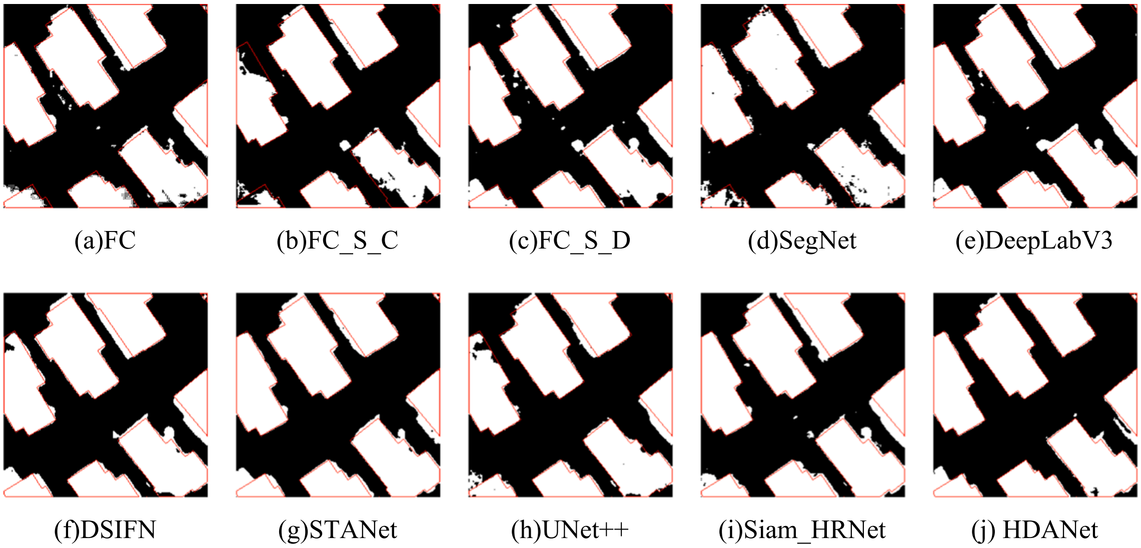


Fig. 15. The results obtained by the different methods with the first change scene in the WHU building dataset.

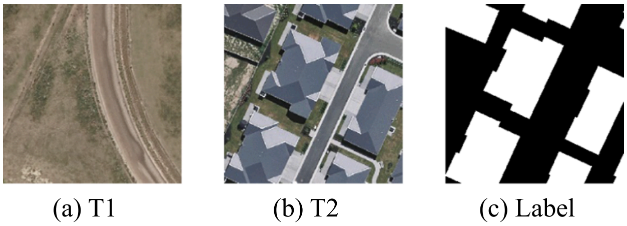


Fig. 16. The second change scene in the WHU building dataset.

2018). The two modules can be represented by:

$$A_c = \sigma(MLP(Maxpool(F)) + MLP(Avgpool(F))) \tag{9}$$

$$A_s = \sigma(Conv_{7\times7}(Concat(Maxpool(F), Avgpool(F)))) \tag{10}$$

where A_c denotes the channel and A_s is the spatial attention module. F represents the input features,$Concat()$ represents the concatenation operation, and σ is the activation function.

In a change detection algorithm using a Siamese network structure, a difference feature map is also constructed when extracting the image features of the two periods, so as to enhance the ability to discriminate the difference information. The characteristics of the difference map can be indicated by the magnitude of the change intensity of the bi-temporal images. Regions with a higher intensity have larger values in the difference map and should be given larger weights as these regions are more likely to change. However, the spatial attention module only integrates the feature map on the channel dimension, and cannot reflect the difference strength well. In this regard, a difference attention module (DAM) is proposed, which is illustrated in Fig. 1. The Euclidean distance between the feature map from the bi-temporal data in a pixel-wise manner is calculated to represent the change intensity map (CIM). The channel number of the change intensity map is 1. Each pixel in the change intensity map denotes the change intensity of the corresponding location in the bi-temporal images. Convolution with a 3 × 3 kernel is implemented on the change intensity map and is followed with a sigmoid function, which can generate the difference attention weights. After this, the channel attention is integrated with the difference attention weight to obtain the output attention, which can emphasize the intensity of the difference of each pixel in the spatial dimension and assign greater weights to the positions with larger differences and

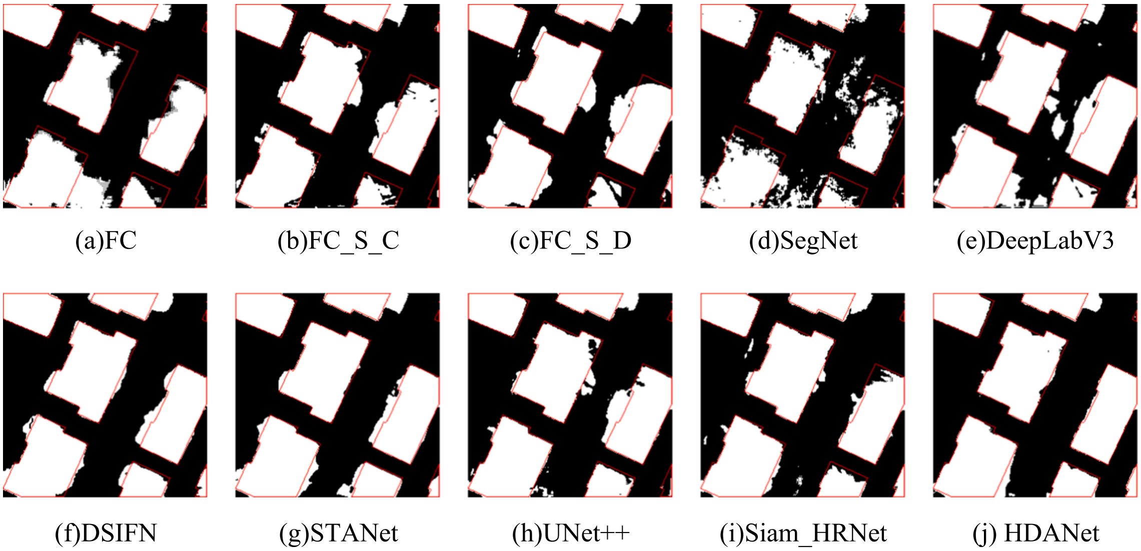


Fig. 17. The results obtained by the different methods with the second change scene in the WHU building dataset.

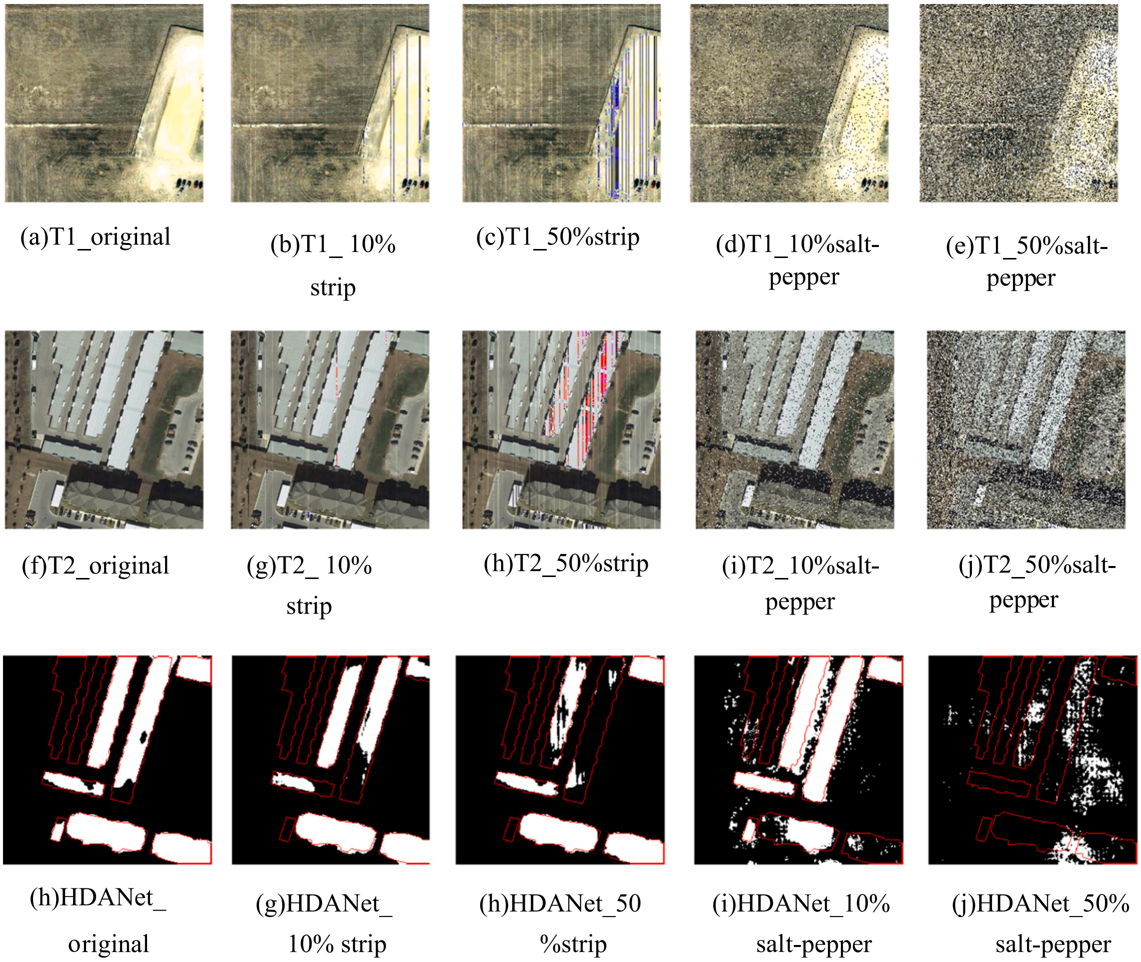

(a)T1_original
(b)T1_ 10% strip
(c)T1_50%strip
(d)T1_10%salt-pepper
(e)T1_50%salt-pepper
(f)T2_original
(g)T2_ 10% strip
(h)T2_50%strip
(i)T2_10%salt-pepper
(j)T2_50%salt-pepper
(h)HDANet_ original
(g)HDANet_ 10% strip
(h)HDANet_50 %strip
(i)HDANet_10% salt-pepper
(j)HDANet_50% salt-pepper

Fig. 18. The results obtained on the first change scene in the LEVIR-CD dataset with the different data noise.

smaller weights to the positions with smaller differences. The proposed DAM can help to keep the building change information and the calculation is:

$$CIM = \sqrt{\sum_{c=1}^{n} (T_{1c} - T_{2c})^2} \tag{11}$$

$$A_{CIM} = \sigma(Conv_{3\times3}(CIM)) \tag{12}$$

where T_{1c} and T_{1c} represent the features in the c th channel for the two periods, respectively. n is the channels count. A_{CIM} represents the difference attention weight map, and σ denotes the sigmoid function, which can make the range of A_{CIM} be 0–1. The final attention map is then obtained by weighting the influence from all the channels:

$$DI = |T_1 - T_2| \tag{13}$$

$$F_{CIM} = A_{CIM} \otimes DI \tag{14}$$

where DI is the difference on each channel after the Siamese network, and the size of DI is kept the same as T_1 or T_2.

3.2. Multi-scale feature learning

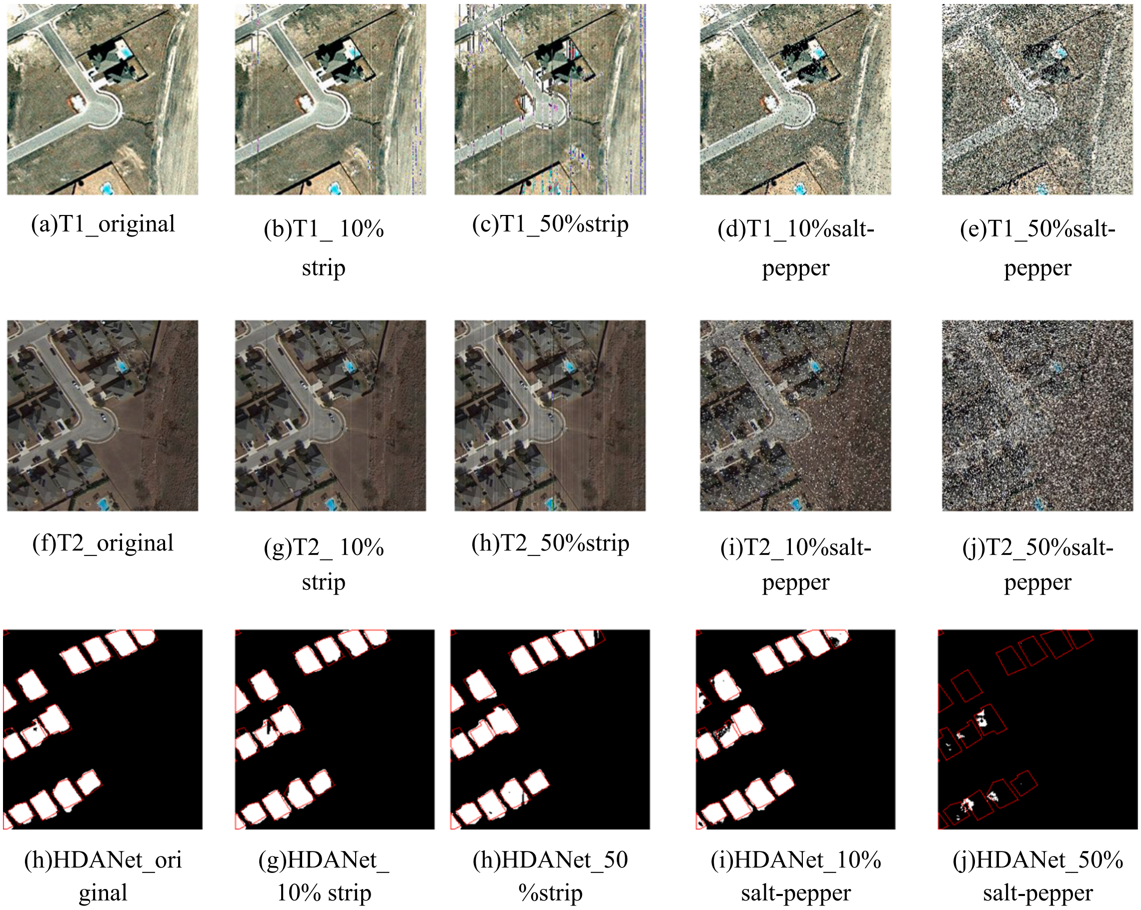
Building changes have regular shapes, close arrangements, and various scales, and the features should be represented under different scales to support the change discrimination. In a vanilla CNN, the kernel size is 3 × 3, and a larger kernel can increase the receptive field. The GoogLeNet Inception module (Szegedy, Liu et al., 2015) uses convolution kernels with different sizes. Larger convolution kernels (7 × 7 or larger) improve the parameter amount of the model, and can be replaced with a cascade of smaller-size convolution kernels. Dilated convolution has the bigger receptive field without additional parameters. With the same receptive field, the 3 × 3 dilated convolution requires significantly less computation than the 5 × 5 convolution with 25 parameters and the 7 × 7 convolution with 49 parameters.

The ASPP module of the DeepLab series of networks (Chen, Papandreou et al., 2017) is included for the multi-scale feature extraction. The convolution process includes three convolutional layers with kernel size of 3 × 3 and dilation rates of 1, 6, and 12, respectively, and one 1 × 1 convolutional layer. In order to generate multi-scale features after the Siamese network, the features of different scales are concatenated together.

3.3. The proposed HDANet framework

HDANet consists of a fully convolutional part, a multi-scale difference feature learning part, and a detection part. The majority of the existing FCN-based approaches utilize convolution and pooling for the feature extraction and down-sampling, and then employ interpolation or transposed convolution to carry out the up-sampling, to restore the high resolution. Different-resolution features are connected in cascade, and the feature maps go through the feature extraction without losing the spatial information, which can be achieved by including skip connections.

HDANet utilizes the high-resolution HRNet architecture to learn the feature maps, which connects four different resolutions in parallel.



(a)T1_original (b)T1_ 10% strip (c)T1_50%strip (d)T1_10%salt-pepper (e)T1_50%salt-pepper

(f)T2_original (g)T2_ 10% strip (h)T2_50%strip (i)T2_10%salt-pepper (j)T2_50%salt-pepper

(h)HDANet_original (g)HDANet_ 10% strip (h)HDANet_50 %strip (i)HDANet_10% salt-pepper (j)HDANet_50% salt-pepper

Fig. 19. The results obtained on the second change scene in the LEVIR-CD dataset with the different data noise.

Differing from the traditional FCN, the feature map resolution is maintained on the same branch in HRNet and, in this way, some feature branches always keep the high resolution, which is beneficial for the detailed information learning. There are many stages to learn the different-level features. At the end of a stage, the feature is down-sampled and the resolution is reduced to half of the original with doubled channels. The feature fusion is then implemented among the different resolutions, in which down-sampling is performed when the high resolution is fused to a low resolution and up-sampling is performed when the low resolution is fused to a high resolution. Fig. 2 illustrates the HDANet framework.

In HDANet, HRNet is included as the basic skeleton of the Siamese network. Both branches of the Siamese network structure have four branches of different resolutions, and the low-resolution branch is restored to the maximum resolution by up-sampling in the final stage, which can be regarded as a fusion feature map combined with multiple resolutions. ASPP is appended after the Siamese high-resolution feature extraction for the multi-scale feature learning. The Siamese features are fed into the ASPP to learn the difference information. After this, the difference map is constructed using the multi-scale feature maps, and is followed with the DAM to reinforce the feature representation. Finally, the softmax operation is carried out to transform the values of the last feature map to the probability of belonging to the changed or unchanged category, so as to judge whether the corresponding pixel has changed. The parallel different-resolution feature extraction and the multi-scale difference feature map can emphasize the feature representation under different scales of buildings, and the DAM can keep the subtle changes, which is beneficial to the building change detection. The floating-point-operations (FLOPs) of HDANet is 284.33G when the input size of bi-temporal images is 256 × 256 × 3 and the total parameter size is 6.77 MB.

4. Experiments

4.1. Dataset description

4.1.1. The LEVIR-CD dataset

The LEVIR-CD dataset (Chen and Shi 2020) was published by Beihang University. The original imagery of this dataset is made up of Google Earth images of Texas, USA, taken between 2002 and 2018, with spatial resolution of 0.5 m. The dataset has been released with training, validation, and test sets, for which the image number is 445, 64, and 128, respectively. The size of each image in the original dataset is 1024 × 1024. Considering the limitation of GPU memory size, each image was divided into 16 equal parts with size of 256 × 256, obtaining a total of 7120, 1024, and 2048 images in the training, validation and test sets, respectively. Fig. 3 shows some scenes from the LEVIR-CD dataset.

4.1.2. The SYSU-CD dataset

The SYSU-CD dataset (Shi, Liu et al., 2022) was published by Sun Yat-Sen University. The main change types of this dataset include changes in different urban category objects, such as changes from forest land to building land, river bank expansion, the disappearance of ships in the water, and the addition of buildings. The changes of the ground objects in this dataset are complex. The image size is 256 × 256. The number of images is 12,000, 4,000, and 4,000 for the training, validation, and test sets, respectively. Fig. 4 shows some scenes from the SYSU-CD dataset.

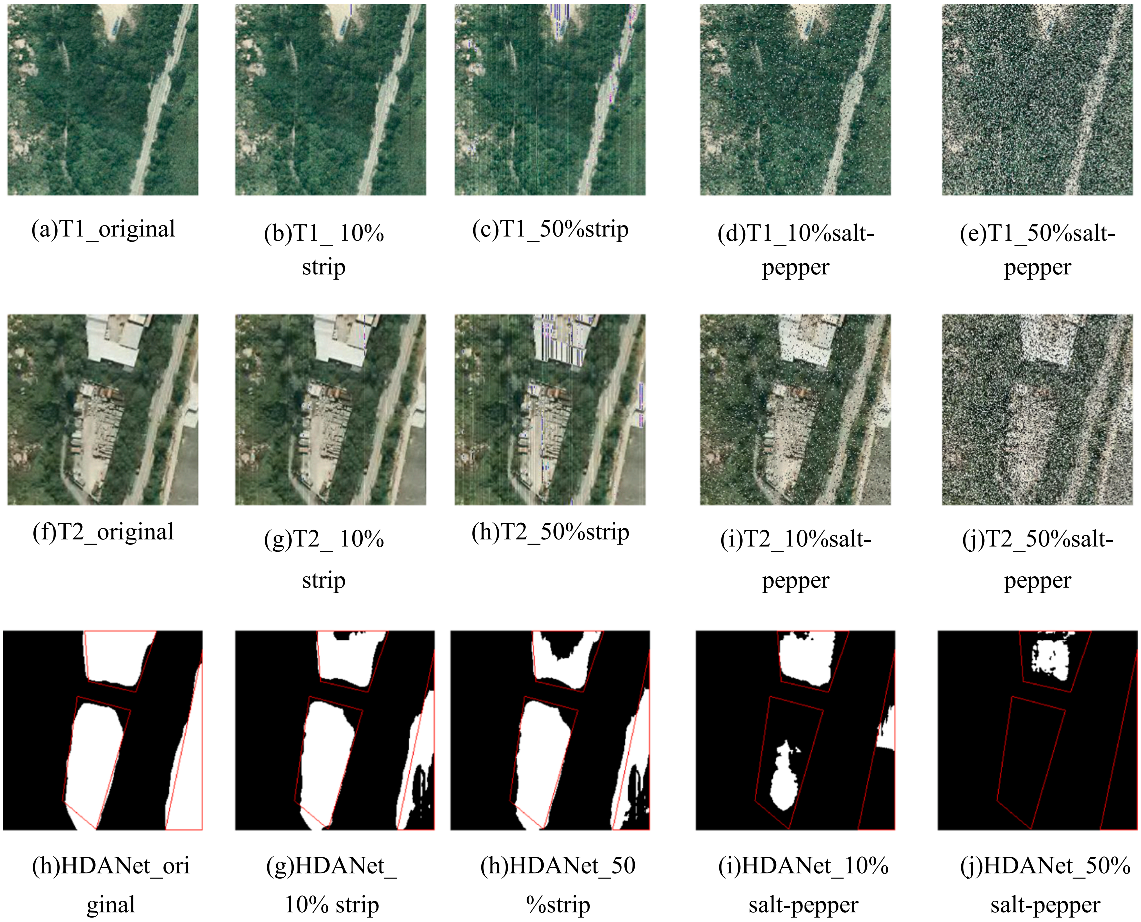


Fig. 20. The results obtained on the first change scene in the SYSU-CD dataset with the different data noise.

4.1.3. *The WHU building dataset*

The WHU building dataset (Ji, Wei et al., 2018) was published by Wuhan University. The original image size of this dataset is 15354 × 32507, and the location is Christchurch, New Zealand. The image pairs were obtained in 2012 and 2016. An earthquake in 2011 caused widespread damage to homes in the area, so the main types of changes are the construction of post-disaster housing and the conversion of bare soil to buildings. We cropped the entire image without overlapping, with a window size of 256 × 256, and obtained a total of 7434 images of 256 × 256 in size. The original dataset was divided by the ratio of 8:1:1 to form the training, validation, and test sets, for which the image number is 5948, 743, and 743 respectively. Fig. 5 shows some scenes from the WHU building dataset.

4.2. *Experimental settings*

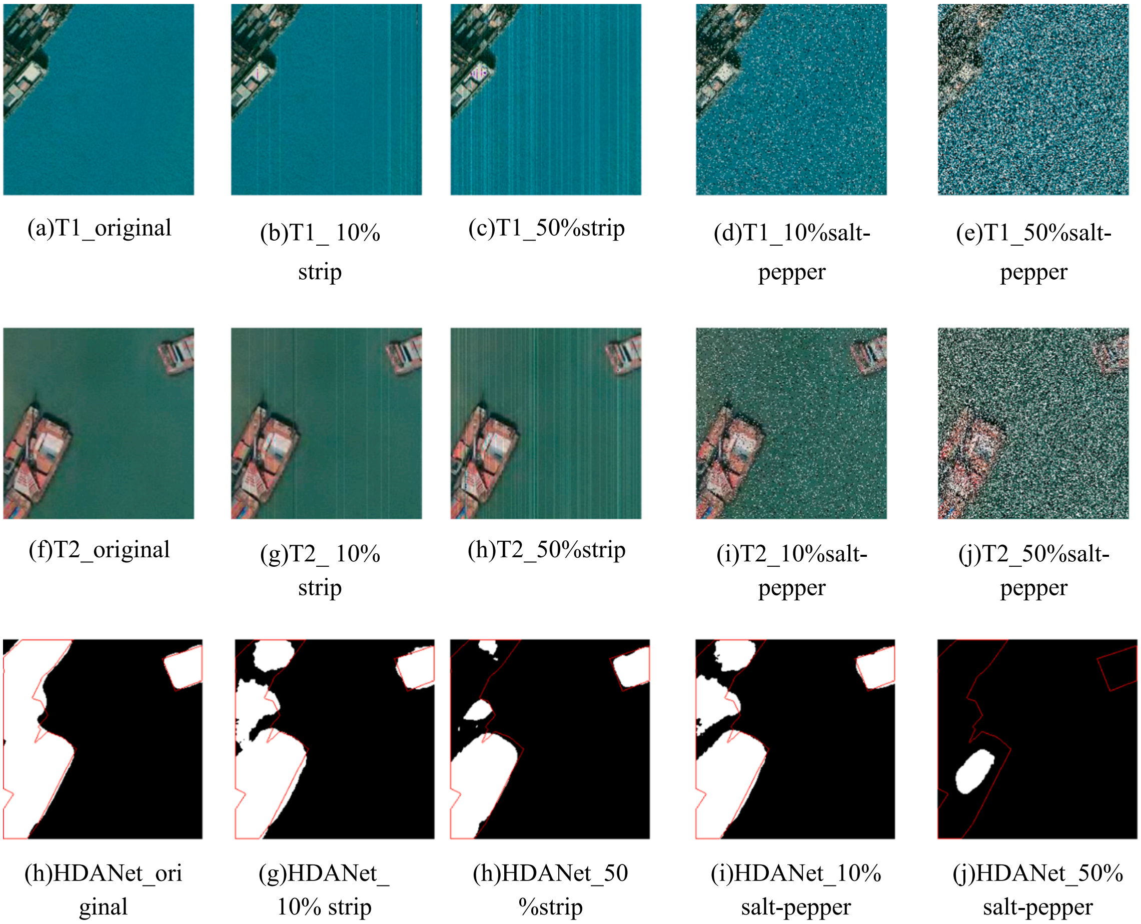
Six current mainstream deep learning based methods were utilized as the counter parts: fully convolutional early fusion (FC) (Daudt, Le Saux et al., 2018), fully convolutional Siamese-concatenation (FC_S_C) (Daudt, Le Saux et al., 2018), fully convolutional Siamese-difference (FC_S_D) (Daudt, Le Saux et al., 2018), deeply supervised image fusion network (DSIFN) (Liu, Jiang et al., 2020), spatial–temporal attention neural network (STANet) (Chen and Shi, 2020), UNet++ (Peng, Zhang et al., 2019), and the two classic semantic segmentation models of SegNet (Badrinarayanan, Kendall et al., 2017) and DeepLab v3 (Chen, Papandreou et al., 2017). HRNet with the Siamese network structure (Siam_HRNet) was also used as a comparison method. Experiments were implemented with the LEVIR-CD, SYSU-CD, and WHU building datasets. All the models in the experiments were built using the PyTorch framework. The hardware device used in the experiments was an NVIDIA GeForce RTX 3090 GPU. In the parameter setting of the network model training, considering the limitation of the model size and GPU memory, the batch training size was set to 4. The maximum epochs for the model training was 100. In order to prevent overfitting during training, an early stopping method was used to end the training process.

4.3. *Experiments with the LEVIR-CD dataset*

Table 1 lists the results obtained on the LEVIR-CD dataset. The F1-score of HDANet is 0.8987 and the kappa coefficient is 0.8934, which are the highest scores among all the methods. In the results of the comparison algorithms, the F1-scores of the four methods of UNet++, STANet, DSIFN, and DeepLab v3 are all higher than 0.85. Among these methods, UNet++ obtains the highest accuracy among the comparison methods, with an F1-score of 0.8823 and a kappa coefficient of 0.8762, which are 0.0164 lower than the F1-score of HDANet and 0.0172 lower than the kappa coefficient.

Siam_HRNet, as a method that uses the HRNet architecture with the Siamese network structure, achieves an F1-score of higher than 0.87 and the accuracy is only lower than that of UNet++ among all the comparison algorithms. The detection accuracy of the three methods based on the FC method is lower than that of the other methods. Among these methods, the FC method obtains the lowest accuracy, with a kappa of 0.7947 and an F1-score of 0.8040, which are 0.0947 lower than the F1-score of HDANet and 0.0987 lower than the kappa coefficient. The Siamese network structure based FC_S_C and FC_S_D methods show their superiority over the methods without the Siamese network structure in change feature extraction.

As shown in Fig. 6 and Fig. 8, images of two scenes were chosen from the LEVIR-CD test set for visualization of the detection results. The



(a)T1_original (b)T1_ 10% strip (c)T1_50%strip (d)T1_10%salt-pepper (e)T1_50%salt-pepper

(f)T2_original (g)T2_ 10% strip (h)T2_50%strip (i)T2_10%salt-pepper (j)T2_50%salt-pepper

(h)HDANet_original (g)HDANet_ 10% strip (h)HDANet_50 %strip (i)HDANet_10% salt-pepper (j)HDANet_50% salt-pepper

Fig. 21. The results obtained on the second change scene in the SYSU-CD dataset with the different data noise.

change type of the first scene is the change from bare ground to buildings. As shown in Fig. 7, the results of the FC, FC_S_C, and DeepLab methods show serious missed detections, and some false detections occur between the buildings for SegNet and STANet. DSIFN and HDANet show the best detection effect in this scene, and the boundaries of the buildings are kept completely. Small buildings in this scene are well detected in the result of HDANet, which performs better than Siam_HRNet. The second scene is the change of bare soil and vegetation to residential land. As shown in Fig. 9, the building changes are missed in the results of the SegNet method, and the FC, FC_S_D, DeepLab, DSIFN, UNet++, and Siam_HRNet methods cannot handle the detection in adjacent buildings, where the results do not show sharp boundaries between adjacent buildings. HDANet shows no obvious missed detection in the change boundaries, and the detailed outlines of the building changes are more complete.

4.4. Experiments with the SYSU-CD dataset

Table 2 reports the accuracy of the detection results obtained for the SYSU-CD dataset. Due to the complexity of the various change scenarios, the accuracies of the detection results of each method in the SYSU-CD dataset are lower than for the LEVIR-CD dataset. The HDANet method again achieves the highest values among the test results in terms of the kappa coefficient and F1-score, with an F1-score of 0.7920 and a kappa coefficient of 0.7271. The five comparison algorithms of UNet++, Siam_HRNet, STANet, DSIFN, and DeepLab all achieve F1-scores of over 0.75. UNet++ performs the best among the other algorithms, with an F1-score of 0.7790 and a kappa coefficient of 0.7146. Compared with HDANet, the F1-score is lower by 0.0130 and the kappa coefficient is lower by 0.0125. The detection accuracy of the three FC-based approaches and SegNet is relatively low. The performance of the FC method is the worst, with an F1-score of 0.7156 and a kappa coefficient of 0.6427, which are 0.0764 and 0.0844 lower than those of HDANet, respectively. Compared with the FC method, the better detection accuracy of FC_S_C and FC_S_D demonstrates that the Siamese network structure is a valid way to handle building change detection.

The images of two scenes in the test set were utilized to visualize the detailed results, as shown in Fig. 10 and Fig. 12. The change type of the first scene is the change from forest land to building land. From the detection results of each method shown in Fig. 11, it can be seen that all the methods show a certain degree of omission, among which the FC and FC_S_D methods are the most obvious, and the results include many “holes” caused by missed detection. The results of HDANet, Siam_HRNet, and UNet++ show better integrity for the land objects, among which the results of Siam_HRNet and HDANet are relatively close, but the boundaries of the buildings in the result of HDANet are smoother. The second scene 2 features the reconstruction of a shipping wharf, which includes the mutual transformation between water body and building. Fig. 13 gives the detection results of all the methods, where it can be seen that, in this change scenario, all the methods show a certain degree of missed detection, among which the SegNet and FC_S_C methods show the most serious omission in the area where the water body changes to building. The overall performance of the STANet, Siam_HRNet, and HDANet methods is better than that of the other methods. HDANet shows no obvious omission errors.

4.5. Experiments with the WHU building dataset

Table 3 reports the results for the WHU building dataset. The performance of the HDANet method is the best in terms of the F1-score and

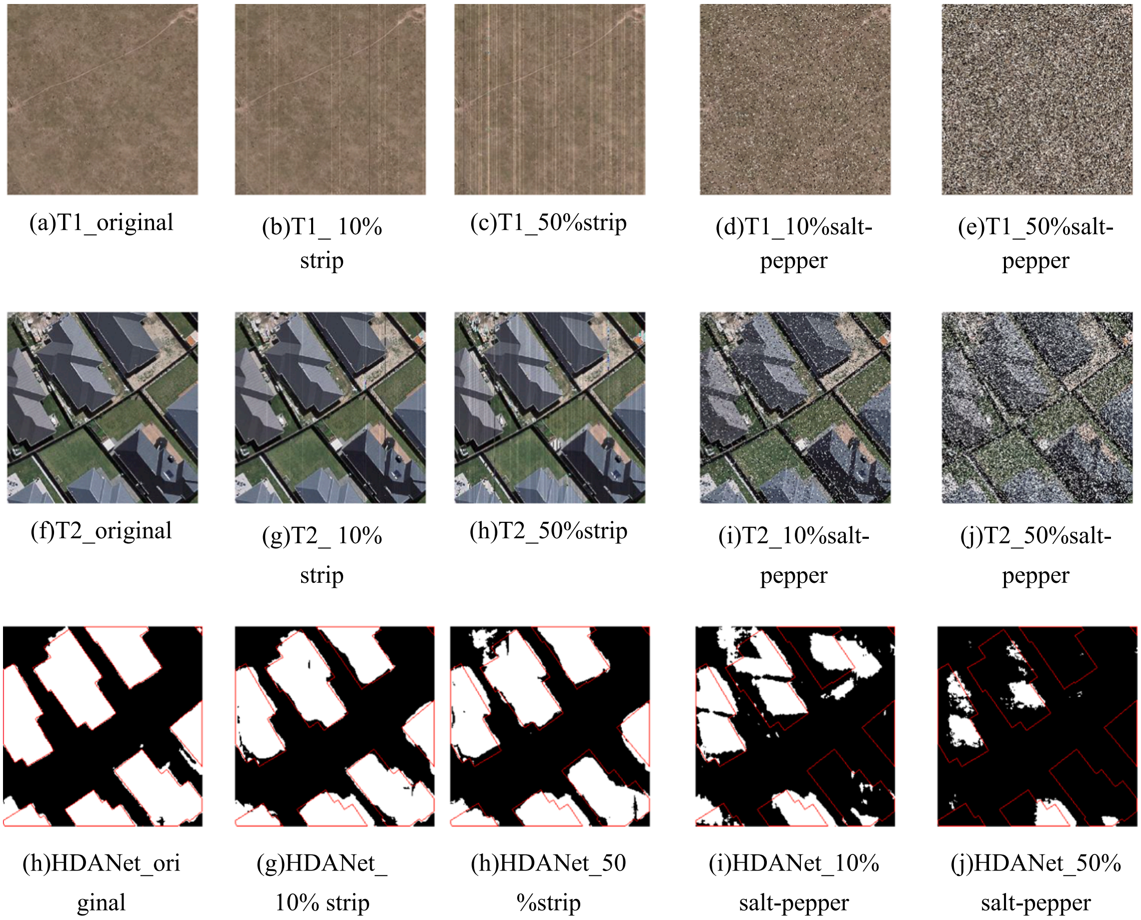


Fig. 22. The results obtained on the second change scene in the WHU building dataset with the different data noise.

kappa coefficient, with a kappa coefficient of 0.8554 and an F1-score of 0.8605. In the results of the other methods, STANet obtains the best performance, with an F1-score of 0.8468 and a kappa coefficient of 0.8410, which are 0.0137 and 0.0144 lower than HDANet, respectively. SegNet obtains the lowest accuracy among the comparison algorithm, with a kappa coefficient of 0.7219 and an F1-score of 0.7316. The accuracy of the FC_S_C and FC_S_D methods is still higher than that of the FC method with a non-Siamese network structure, as with the first two datasets.

As shown in Fig. 14 and Fig. 16, two images of different scenes were chosen to explore the performance of all approaches. The first scene is the change from bare soil to residential area. As shown in Fig. 15, HDANet shows a superior ability to preserve the edges of the buildings, and there is no “horn” caused by obvious false detection. However, for the other comparison methods, there are many omission errors in the interior of the buildings. The second scene also features the change from bare soil to residential area, which is shown in Fig. 17. The buildings are clearly scattered in the result of the SegNet method, and the FC method cannot keep an unbroken shape for the buildings. The result of HDANet shows the best effect in maintaining the changed boundaries of the buildings in this scene, and there is no obvious missed detection or broken buildings.

The result of HDANet shows the best effect in maintaining the changed boundaries of the buildings in this scene, and there is no obvious missed detection or broken buildings.

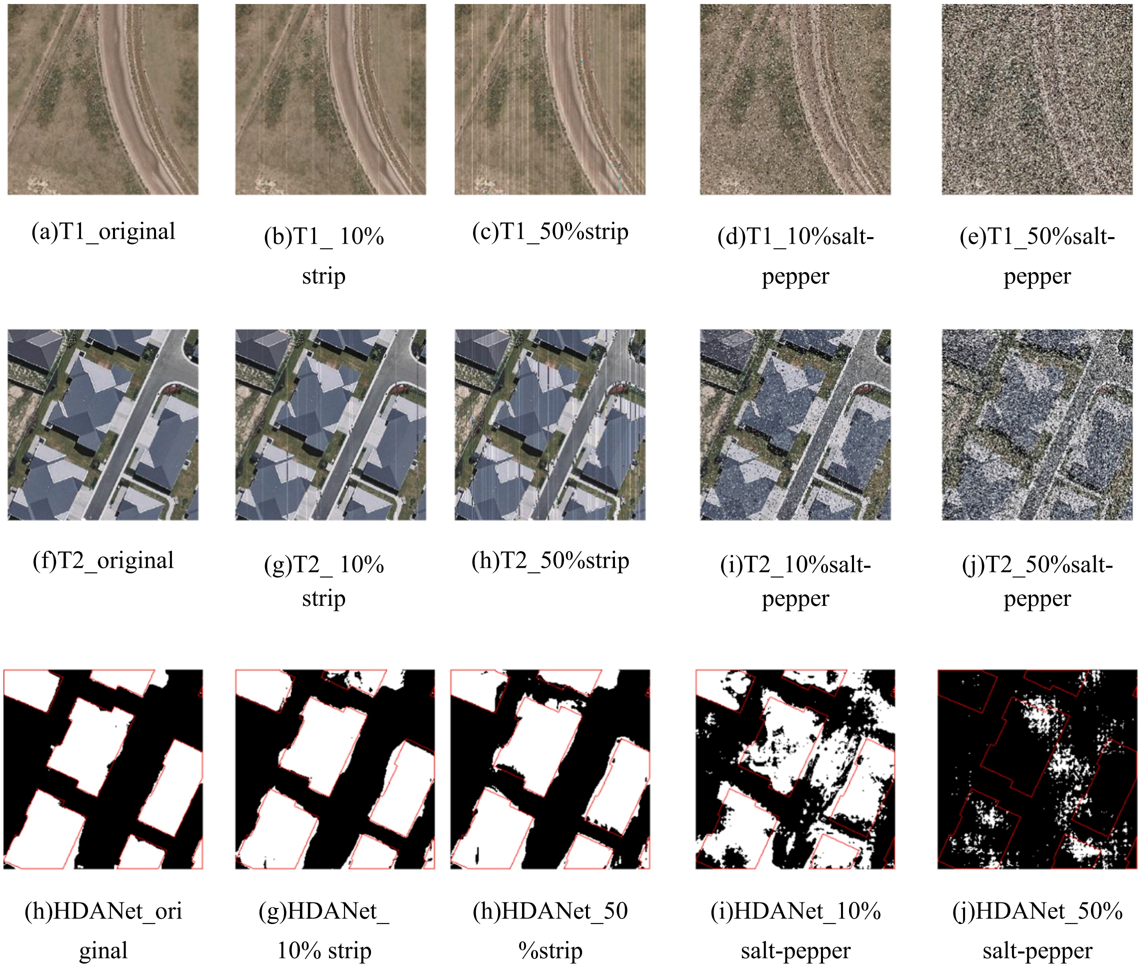
4.6. Experiments with the degradation dataset

The noise of remote sensing image will impact the performance of the remote sensing image processing (Hong, Yokoya et al., 2018). In this section, the noise immunity of the proposed HDANet has been analyzed. Firstly, salt-and-pepper noise and stripe noise, which are two kinds of common noise in remote sensing image, are utilized as the simulated noise to added to the original dataset. Salt-and-pepper noise is a kind of spot-like noise and the value of the affected pixels is the maximum or minimum value in the value range of the image. Stripe noise can be one pixel wide or multiple pixels wide. The gray value of the bright stripe is higher than that of the surrounding normal pixels, while gray value of the dark stripe is lower than that of the surrounding normal pixels. In this work, the noise ratio is set to 10 % and 50 %.

Figs. 18-23 show the original images, the noise added images and the corresponding change detection results on three datasets. Table 4 lists the detection accuracy for the different datasets. Salt-and-pepper noise and stripe noise caused a loss of the detection accuracy. HDANet on 10 % of salt-and-pepper noise and stripe noise datasets obtained the higher accuracy compared with the 50 % of salt-and-pepper noise and stripe noise datasets. All the F1 of HADNet on these degradation datasets are lower than 0.88. From the detection map, the degraded impacts from salt-and-pepper noise are worse than that from stripe noise. Most change areas can not be well detected when the alt-and-pepper noise ratio is 50 %. We have added the corresponding discussion on the manuscript.

5. Conclusion

In this work, a new method named HDANet has been proposed to handle building change detection. The parallel multi-resolution structure of HRNet is utilized as the basic skeleton of the network. The feature maps of different resolutions are then fused to represent various characteristics. The HDANet framework is carried out using a Siamese network structure, which allows HDANet to represent the change


(a)T1_original (b)T1_ 10% strip (c)T1_50%strip (d)T1_10%salt-pepper (e)T1_50%salt-pepper

(f)T2_original (g)T2_ 10% strip (h)T2_50%strip (i)T2_10%salt-pepper (j)T2_50%salt-pepper

(h)HDANet_original (g)HDANet_ 10% strip (h)HDANet_50 %strip (i)HDANet_10% salt-pepper (j)HDANet_50% salt-pepper

Fig. 23. The results obtained on the second change scene in the WHU building dataset with the different data noise.

Table 4
The accuracy of the HDANet with the different data noise.

Dataset		Precision	Recall	F1-score	Kappa
LEVIR-CD dataset	HDANet_SPN10%	0.8810	0.8679	0.8744	0.8635
	HDANet_ SPN 50 %	0.8120	0.8007	0.8063	0.7902
	HDANet_SN10%	0.8714	0.8795	0.8753	0.8670
	HDANet_SN50%	0.7834	0.7215	0.7512	0.7320
SYSU-CD dataset	HDANet_SPN10%	0.7517	0.7342	0.7428	0.7219
	HDANet_ SPN 50 %	0.6829	0.6315	0.6562	0.6488
	HDANet_SN10%	0.7494	0.7537	0.7515	0.7371
	HDANet_SN50%	0.6352	0.6150	0.6249	0.6052
WHU building dataset	HDANet_SPN10%	0.8742	0.8681	0.8711	0.8533
	HDANet_ SPN 50 %	0.8251	0.8339	0.8295	0.8140
	HDANet_SN10%	0.8506	0.8240	0.8371	0.8219
	HDANet_SN50%	0.7995	0.7460	0.7718	0.7538

characteristics well. Moreover, a differential attention module based on change intensity is proposed to enhance the differential representation, which enhances the differential features of two different temporal images. For the case of there being land objects in the same image with different scales, the parallel ASPP module with preset dilation rates is used for the multi-scale features extraction. The experimental results obtained in this study showed that the HDANet can achieve a high building change detection accuracy, compared with the current mainstream methods, with public building change detection datasets.

In the aspect of future studies, two key points can be further pursued. Firstly, more advanced networks in computer vision field can be investigated to learn the fine-grant features. Moreover, the proposed method is based on the supervised learning, which needs a large amount of the annotated samples. How to combine the deep learning methods with semi-supervised and unsupervised methods to solve the change detection task is the other future research direction.

CRediT authorship contribution statement

Xue Wang: Methodology, Writing - original draft, Writing - review & editing. **Junhan Du:** Methodology, Software. **Kun Tan:** Conceptualization, Methodology, Writing – review & editing. **Jianwei Ding:** Data curation, Writing - review & editing. **Zhaoxian Liu:** Data curation, Formal analysis. **Chen Pan:** Data curation. **Bo Han:** Investigation, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is jointly supported by the Natural Science Foundation of China (grant nos. 41871337, 42171335,42001350), by the Postdoctoral Science Foundation of China (grant no. 2021M691016), and by Key

Laboratory of Spatial-temporal Big Data Analysis and Application of Natural Resources in Megacities, MNR (grant no. KFKT-2022-07).

References

Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. Understanding of a convolutional neural network. 2017 International Conference on Engineering and Technology (ICET), Ieee.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Bovolo, F., Bruzzone, L., 2006. A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. IEEE Trans. Geosci. Remote Sens. 45 (1), 218–236.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

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.

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.

Deng, W., Shi, Q., Li, J., 2021. Attention-Gate-Based Encoder–Decoder Network for Automatical Building Extraction. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 2611–2620.

Ding, Q., Shao, Z., Huang, X., Altan, O., 2021. DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 105, 102591.

Du, P., Liu, S., 2012. Change detection from multi-temporal remote sensing images by integrating multiple features. National Remote Sensing Bulletin 16 (4), 663–677.

Gao, L., Hong, D., Yao, J., Zhang, B., Gamba, P., Chanussot, J. J. I. T. o. G., Sensing, R., 2020. Spectral superresolution of multispectral imagery with joint sparse and low-rank learning, 59(3): 2269-2280.

Hong, D., Yokoya, N., Chanussot, J., Zhu, X. X. J. I. T. o. I. P., 2018. An augmented linear mixing model to address spectral variability for hyperspectral unmixing. 28(4): 1923-1938.

Hong, D., Gao, L., Yao, J., Zhang, B., Plaza, A., Chanussot, J. J. I. T. o. G., Sensing, R., 2020. “Graph convolutional networks for hyperspectral image classification.” 59(7): 5966-5978.

Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B., 2020b. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE Trans. Geosci. Remote Sens. 59 (5), 4340–4354.

Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., Chanussot, J., 2021. SpectralFormer: Rethinking hyperspectral image classification with transformers. IEEE Trans. Geosci. Remote Sens. 60, 1–15.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 57 (1), 574–586.

Li, Z., Yan, C., Sun, Y., Xin, Q., 2022. A Densely Attentive Refinement Network for Change Detection Based on Very-High-Resolution Bitemporal Remote Sensing Images. IEEE Trans. Geosci. Remote Sens. 60, 1–18.

Lin, Y., Li, S., Fang, L., Ghamisi, P., 2019. Multispectral change detection with bilinear convolutional neural networks. IEEE Geosci. Remote Sens. Lett. 17 (10), 1757–1761.

Liu, R., Jiang, D., Zhang, L., Zhang, Z., 2020. Deep depthwise separable convolutional network for change detection in optical aerial images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 1109–1118.

Mou, L., Bruzzone, L., Zhu, X.X., 2018. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. IEEE Trans. Geosci. Remote Sens. 57 (2), 924–935.

Niu, C., Tan, K., Jia, X., Wang, X., 2021. Deep learning based regression for optically inactive inland water quality parameter estimation using airborne hyperspectral imagery. Environ. Pollut. 286, 117534.

Peng, D., Zhang, Y., Guan, H., 2019. End-to-end change detection for high resolution satellite images using improved UNet++. Remote Sensing 11 (11), 1382.

Qin, P., Cai, Y., Liu, J., Fan, P., Sun, M., 2021. Multilayer Feature Extraction Network for Military Ship Detection From High-Resolution Optical Remote Sensing Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 11058–11069.

Shi, Q., Liu, M., Li, S., Liu, X., Wang, F., Zhang, L., 2022. A Deeply Supervised Attention Metric-Based Network and an Open Aerial Image Dataset for Remote Sensing Change Detection. IEEE Trans. Geosci. Remote Sens. 60, 1–16.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Volpi, M., Tuia, D., Bovolo, F., Kanevski, M., Bruzzone, L., 2013. Supervised change detection in VHR images using contextual information and support vector machines. Int. J. Appl. Earth Obs. Geoinf. 20, 77–85.

Wang, M., Tan, K., Jia, X., Wang, X., Chen, Y., 2020. A deep siamese network with hybrid convolutional feature extraction module for change detection based on multi-sensor remote sensing images. Remote Sens. 12 (2), 205.

Wang, X., Tan, K., Du, P., Pan, C., Ding, J., 2022. A Unified Multiscale Learning Framework for Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 60, 1–19.

Wang, Q., Yuan, Z., Du, Q., Li, X., 2018. GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection. IEEE Trans. Geosci. Remote Sens. 57 (1), 3–13.

Wessels, K.J., Van den Bergh, F., Roy, D.P., Salmon, B.P., Steenkamp, K.C., MacAlister, B., Swanepoel, D., Jewitt, D., 2016. Rapid land cover map updates using change detection and robust random forest classifiers. Remote Sens. 8 (11), 888.

Wu, L., Wang, Y., Gao, J., Li, X., 2018. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. IEEE Trans. Multimedia 21 (6), 1412–1424.

Zheng, H., Gong, M., Liu, T., Jiang, F., Zhan, T., Lu, D., Zhang, M., 2022b. HFA-Net: High frequency attention siamese network for building change detection in VHR remote sensing images. Pattern Recogn. 129, 108717.

Zheng, D., Wei, Z., Wu, Z., Liu, J., 2022a. Learning Pairwise Potential CRFs in Deep Siamese Network for Change Detection. Remote Sens. 14 (4), 841.