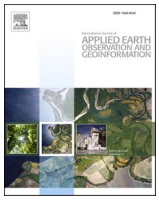
International Journal of Applied Earth Observation and Geoinformation 122 (2023) 103456

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



Double U-Net (W-Net): A change detection network with two heads for remote sensing imagery

Xue Wang [a,b], Xulan Yan [a,b], Kun Tan [a,b,*], Chen Pan [b,c], Jianwei Ding [d], Zhaoxian Liu [d], Xinfeng Dong [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] Shanghai Municipal Institute of Surveying and Mapping, Shanghai 200063, China
[d] The Second Surveying and Mapping Institute of Hebei, Shijiazhuang 050037, China
[e] China Aero Geophysical Survey and Remote Sensing Center for Natural Resources, Beijing 100083, China

ARTICLE INFO

Keywords:
Superpixel
Double head
Deep learning
Change detection

ABSTRACT

Recently, the deep learning algorithms have been increasingly utilized in remote sensing change detection. However, incomplete buildings and the blurred edges caused by the complex scenes in change detection applications make the detection results fail to describe the real land cover changes. Superpixels can be used to alleviate edge blurring, but the existing superpixel methods cannot be trained jointly with the models in change detection. In this work, we investigated an innovative double-head method using deep learning, called double U-Net (W-Net), which consists of a superpixel module and a change detection module. Due to the superpixel module, W-Net can handle building edges very well. In order to solve problem that multiple subtasks fail to achieve the optimal results, a two-branch multi-task coupling framework of change detection and superpixels is designed for W-Net, which enables the model to achieve a globally optimal detection performance. The advancement of the W-Net was demonstrated using three public datasets. The F1score on LEVIR-CD dataset was 0.9031 and kappa coefficient was 0.8969. The F1-score on WHU building dataset was 0.9172 and kappa coefficient was 0.9142. The F1-score on SYSU-CD dataset was 0.8167and and kappa coefficient was 0.7724. The experiments confirmed that the W-Net is capable to detect the edges of changed area better and outperforms the other advanced change detection methods.

1. Introduction

The dynamic monitoring by remote sensing technology is an extremely valuable technical tool in applications such as ecosystem monitoring with the background of global change (Ji et al., 2021), forestry resource management (Kim et al., 2014), damage assessment (Bovolo & Bruzzone, 2007), and agricultural surveying (Tan et al., 2021). However, the diverse background features and dense buildings bring challenges to building change detection.

Change detection is defined as the process of characterizing land surface changes and qualitatively analyzing them using remote sensing data derived from different times (Tan et al., 2019). The changed areas and types are obtained by feature extraction at the same location using images from different times. The traditional pixel-wised methods can be classified as clustering, transforming, image algebra and classification based (Shi et al., 2020). When object-oriented image analysis approaches were applied for processing the high-resolution data, the basic cells of the change areas had been gradually transformed from pixels to segmented objects. Compared with individual pixels, objects contain more complete contextual information. Pixel-based post-classification comparison methods and direct classification comparison methods have been implemented for object-oriented change detection, generally achieving a better accuracy than the pixel-level change detection methods. However, such methods heavily rely on the difference map, and the process of superpixel generation tends to lose detailed information, which leads to instability in the accuracy of the results.

Recently, the popularity of high-performance computing devices and the development of big data technology have driven the deep learning

* 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.2023.103456
Received 25 February 2023; Received in revised form 4 August 2023; Accepted 10 August 2023
Available online 14 August 2023

revolution, which has also been employed in the remote sensing image analysis, because of its remarkable ability of deep feature extraction (Niu et al., 2021; Wang et al., 2019). Convolutional neural network (CNN) models, which are a classical deep learning structure, can automatically learn the abstract spatio-spectral features from remote sensing images (Wang et al., 2022a). LeNet (LeCun et al., 1998) was one of the first CNNs released to recognize handwritten digits in images, achieving a comparable performance to the support vector machine method at the time and becoming the dominant method for supervised learning. Although LeNet can achieve good results on small datasets, it does not perform as well as the other machine learning methods on larger datasets.

However, the fully connected layer in CNN-based change detection methods has many parameters, is time-consuming, and has strict requirements on the size of the input and output. Therefore, fully convolutional networks (FCNs) employ the convolutional operation in the last layer to extend image-level classification to pixel-level classification (Long et al., 2015), which can be utilized in the remote sensing change detection. FCN based algorithms have also been applied for remote sensing segmentation. Various types of neural network models have emerged and a number of breakthroughs have been achieved. For example, Li et al. (Li et al., 2021) investigated the fully convolutional neural network by adding a multi-scale convolutional module, which has been demonstrated that the multi-scale features can promote the effectiveness of high-resolution image change detection. The U-Net architecture improves on FCNs by taking a completely different approach to feature fusion, concatenating features together in the channel dimension to produce more significant features than FCNs. Zheng et al. (2021) investigated a CNN by embedding a multi-scale cross block in U-Net that can merge the features in different scales and different levels to improve the information usage of the network. Peng et al. (2019) used multi-temporal imagery overlay as an input to UNet++ (Zhou et al., 2018) to fuse the change prediction maps using multi-scale characteristics. Zhou et al. (2019) improved the training structure of UNet++ to maintain the performance while also speeding up the inference. The well-known Siamese network architecture was originally used in change detection to compare the similarity between images from two periods (Zheng et al., 2022). Wang et al. (2020) devised a convolutional network using Siamese features to perform change detection by extracting the difference between features.

However, because of the complex scenes of the remote sensing imagery with high resolution, the current methods have the problems of incomplete buildings and blurred building edges when applied to building change detection, which leads to a limited model accuracy. Considering the above problems, scholars have adopted superpixel representation to correct the inaccurate edges. A superpixel is a small region consisting of a series of pixel points that are located next to each other and are of the same properties, such as brightness, texture and color. These regions keep valid characteristic for further processing with clear boundary information of the objects. In recent years, superpixel algorithms have gradually become widely utilized in change detection. For example, Sakurada and Okatani (Sakurada & Okatani, 2015) employed a CNN and superpixels for street image change detection; Zhang et al. (2021) also combined superpixels with a CNN to reduce the potential noise in the imagery.

Although these change detection methods have a good performance, the processes of change detection and superpixel generation are processed separately, and the final algorithm cannot obtain the globally optimal solution. To overcome these limitations, we designed the double U-Net (W-Net) architecture to achieve the global change discrimination and local superpixel coupling training objectives. The contributions are:

(1) We investigate a new coupled double-head model—W-Net—consisting of a change detection module and a superpixel module, which is capable to guarantee the completeness of the features and the edge information in the change detection results using the superpixel module.
(2) Integrating change detection and superpixel representation into the same network can obtain the globally optimal solution and can address the loss of detailed information and instability in the detection accuracy.
(3) The proposed W-Net method achieved a superior performance on two well-known open-source datasets when compared with the recent CNN-based detection methods.

This paper is organized as following. Methodology gives the previous works and W-Net. Experimental setup introduces the experimental setting and the change detection datasets utilized in the experiments. Experiments details the comparisons with other approaches. Discussion discusses the performances with different hyperparameters. Finally, we give the conclusions in Conclusion.

2. Methodology

In this part, we elaborate the details of the W-Net algorithm, to address the current problems of low accuracy, poor adaptability, and blurred edges in change detection. UNet++ is included as the basic framework, and the superpixel module is embedded to enrich the spatio-spectral characteristics and to correct the edges which can reduce the potential noise. In addition, a backbone and branch coupling structure are introduced to integrate the change detection and superpixel expression, to ensure that the different modules can converge to the global optimum.

2.1. Main module

As the representative end-to-end network for pixel-level prediction, FCN-based methods have become the basic framework for semantic segmentation.

In this work, UNet++ is included as the basic framework. UNet++ consists of intermediate features with different depths, whose decoders are refused by dense skip connections. UNet++ shares the same feature extractor during training, and all the modules have a common input and can train together, which improves the overall segmentation performance and efficiency. In addition, UNet++ does not suffer from the limitations of skip connections because its dense skip connections provide the multi-scale features during decoding.

The skip connection structure expression is as follows:

$$x^{i,j} = \begin{cases} \mathscr{H}\left(\mathscr{D}\left(x^{i-1,j}\right)\right), & j = 0 \\ \mathscr{H}\left(\left[\left[x^{i,k}\right]_{k=0}^{j-1}, \mathscr{U}\left(x^{i+1,j-1}\right)\right]\right), & j > 0 \end{cases} \tag{1}$$

where $\mathscr{H}$ represents the convolution operation, $\mathscr{D}$ is downsampling, and $\mathscr{U}$ is upsampling. $x^{i,j}$ denotes the output of node $[i.j]$, where i is the downsampling layer along the encoder index, and j is the convolutional layer along the skip connection index dense block.

However, UNet++ has a complex model structure, resulting in many model parameters. Therefore, deep supervision is introduced in UNet++ to calculate the loss function value for each sub-model. Each sub-model is named L1, L2, L3, and L4, respectively, according to their depth. The implementation of deep supervision is to add an output layer for binary change detection on each sub-model. Because each sub-model in UNet++ can predict the results, the loss can be propagated to L1, L2, L3, and L4. Deep supervision can prune the model during inference and maintain a high accuracy with a high inference efficiency. For better global optimization of the model, UNet++ defines a mixed segmentation loss function for each semantic scale, including pixel-level dice coefficient and cross-entropy loss function, which provides smooth gradients during the model training and alleviates the category imbalance issue. The outputs of the four scales of the proposed W-Net model

Fig. 1. Illustration of the superpixel.

are $\left[Q^1, Q^2, Q^3, Q^4\right] \in \mathbb{R}^{H\times W\times 2}$, and then the final output of the model Q^{all} is given by the average of the above branches as follows :

$$Q^{all} = \frac{1}{4} \bullet \sum_{i=1}^{4} Q^i \tag{2}$$

The true label of the data processed by the OneHot encoder is set as $G \in \mathbb{R}^{H\times W\times 2}$, and then the loss function is formulized as:

$$\mathscr{L}_{CE}(y, \hat{y}) = \begin{cases} -log(\hat{y}) & if\ y = 1 \\ -log(1-\hat{y}) & if\ y = 0 \end{cases} \tag{3}$$

$$\mathscr{L}\left(Q^{all}, G\right) = \frac{\mathscr{L}_{CE}\left(Q^{all}, G\right)}{2} + \frac{2 \bullet Q^{all} \bullet G}{Q^{all} + G} \tag{4}$$

where $\mathscr{L}_{CE}()$ represent the binary cross entropy.

2.2. Superpixels

Superpixels can effectively assist in tasks such as semantic segmentation (Gadde et al., 2016), classification (Mu et al., 2022), and change detection (Shuai et al., 2022). The simple linear iterative clustering (SLIC) algorithm (Achanta et al., 2012; Di et al., 2021) designs feature vectors containing position and color channel information and generates superpixels using k-means clustering. Scholars have investigated superpixel algorithms based on deep learning. The semantic sensor network (SSN) algorithm (Jampani et al., 2018) employed a neural network to generate the features in pixel-level, and then utilized a k-means algorithm to obtained superpixels. Although SSN implements an optimizable superpixel algorithm, the structure is complex. The superpixel-FCN (Spixel-FCN) (Yang et al., 2020) algorithm used a very concise structure to achieve highly accurate superpixel segmentation based on the direct generation of superpixel mappings by an FCN. A superpixel algorithm based entirely on deep learning algorithm provides an innovative idea to the superpixel module for change detection models. The conventional superpixel clustering operations have non-differentiable mathematical characteristics and cannot use back-propagation for deep learning. Spixel-FCN utilized an FCN with an upsampling-downsampling framework that enables fast generation of superpixels, which saves on the time required for clustering and has a simple structure for easy application.

Spixel-FCN is trained and optimized based on the two losses obtained from the prediction results, which are the superpixel internal reconstruction loss and the difference loss between labels. After dividing the image into a regular grid of size $H \times W$ and initializing the superpixel centers, only the nine superpixels around each pixel are considered. For example, the pixel in the green box in Fig. 1 only needs to be considered in association with the nine superpixels in the red box. A matrix $Q \in \mathbf{R}^{H\times W\times 9}$ is obtained by calculating the association of each pixel with the superpixels. After this, Spixel-FCN predicts Q by the neural network to obtain the superpixel results.

The features of the superpixel centers are obtained by weighted averaging according to the correlation between pixels and superpixels. The superpixel center s is denoted as $C_s = (u_s, I_s)$, where u_s represents the feature information and l_s represents the location information. The superpixel center is then calculated as:

$$\boldsymbol{u_s} = \frac{\sum_{\boldsymbol{x}:s\in N_x} x \bullet q_s(x)}{\sum_{\boldsymbol{x}:s\in N_x} q_s(x)}, \boldsymbol{l_s} = \frac{\sum_{\boldsymbol{x}:s\in N_x} x \bullet q_s(x)}{\sum_{\boldsymbol{x}:s\in N_x} q_s(x)} \tag{5}$$

where x denotes one pixel in the superpixel, f(x) represents the feature of x, $\mathscr{N}_x$ dentoes all the superpixels around the pixel, $q_s(x)$ represents the association of x with superpixel s.

After obtaining the superpixel centers, the association of pixels with superpixels is used to reconstruct the target features of each pixel on the basis of the superpixel features. The pixel reconstruction formula is:

$$\boldsymbol{f}'(\boldsymbol{p}) = \sum_{s\in N_p} \boldsymbol{u_s} \cdot q_s(\boldsymbol{p}), \boldsymbol{p}' = \sum_{s\in N_p} l_s \cdot q_s(\boldsymbol{p}) \tag{6}$$

where f′(p) is the reconstructed pixel feature, and p′ is the reconstructed pixel position information.

The general form of the loss function design is:

$$L(Q) = \sum_{p} dist(f(p), f'(p)) + \frac{m}{s}\|p - p'\|_2 \tag{7}$$

A form of loss function based on SLIC is used:

$$L_{SLIC}(Q) = \lambda_2 \sum_{p} \|f_{col}(p) - f'_{col}(p)\|_2 + \frac{m}{s}\|p - p'\|_2 \tag{8}$$

where the features are represented in CIELAB color space, and the distance is calculated using the L2 norm, λ_2 is the reconciliation hyperparameter.

2.3. Coupled double-head model

Most of the current methods for applying superpixels to change detection are based on using the superpixels to normalize the model predicted values. This is done by comparing the pixel and superpixel attributes based on the classification confidence, to determine the individual pixel attributes. However, the separation of superpixel generation from the inference of change detection can result in an optimization gap that limits the performance of the model, so it is desirable to implement end-to-end integration with an efficient framework. The implementation of multitasking with a single deep learning model has been studied in other fields. Wu et al. argued that the features generated by a single network model cannot achieve optimal results on multiple subtasks, so the same features should not be used to solve two subtasks simultaneously (Wu et al., 2020). Based on this idea, we designed a two-branch multi-task coupling framework for change detection and superpixels, which is a backbone and branch coupling structure integrating change detection and superpixel representation.

As shown in Fig. 2, the proposed W-Net is a coupled double-head model and includes two decoders sharing one encoder. Unlike the existing methods, W-Net divides the change detection and superpixels into different branches, which pass through their respective convolutional layers to produce two different features for the superpixel tasks

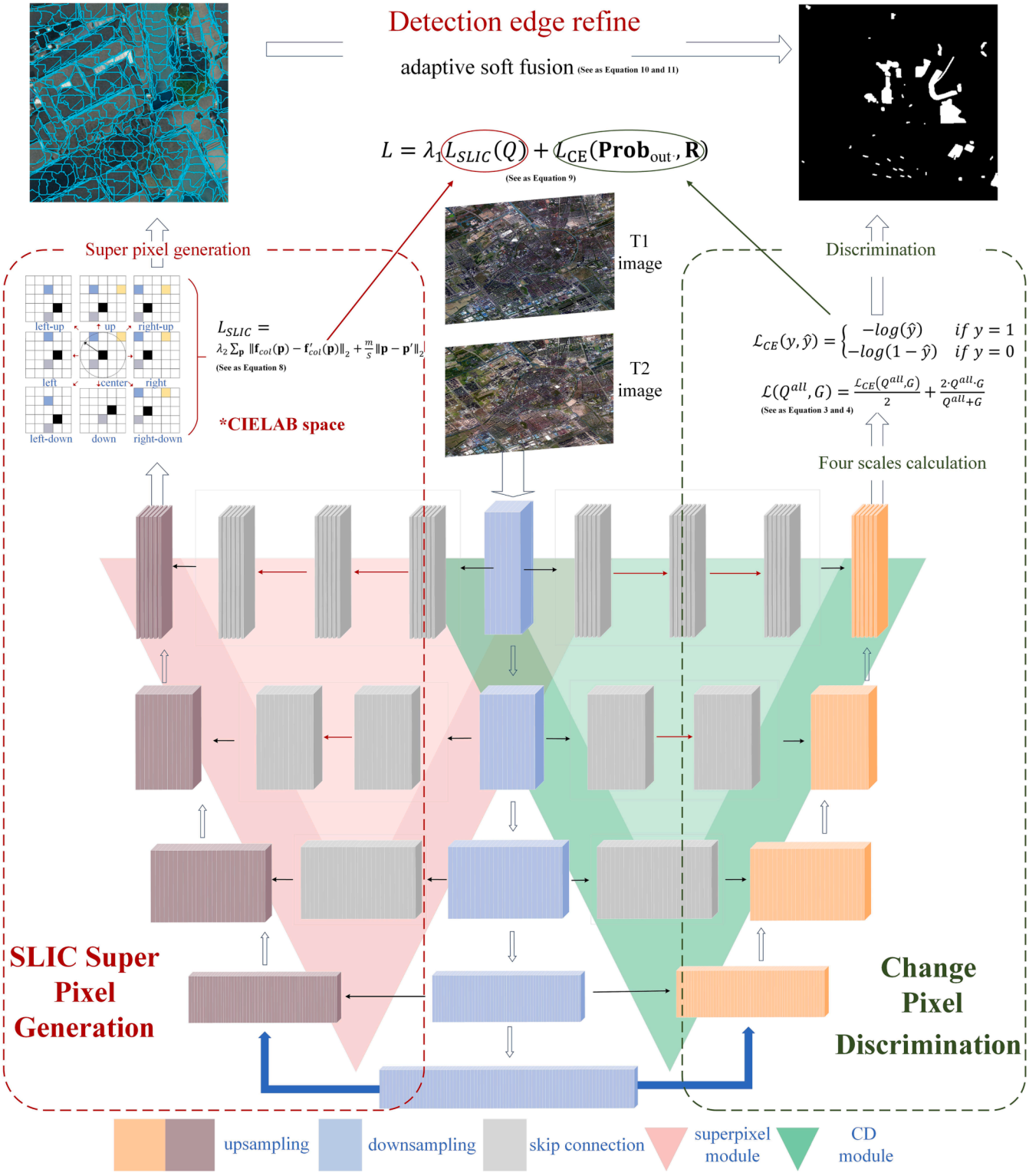


Fig. 2. The structure of the W-Net framework for remote sensing change detection.

Table 1
The performance of the different algorithms on L-CD dataset.

	FC	FC-S-C	FC-S-D	DeepLabv3	UNet++	HFANet	HDANet	Proposed
Precision	0.9016	**0.9394**	0.8911	0.9003	0.9144	0.8344	0.9226	0.9124
Recall	0.7255	0.7597	0.7756	0.8251	0.8524	0.8228	0.8761	**0.8921**
F1score	0.804	0.8401	0.8293	0.8611	0.8823	0.8286	0.8987	**0.9031**
Kappa	0.7947	0.8324	0.8209	0.8539	0.8762	0.8193	0.8934	**0.8969**

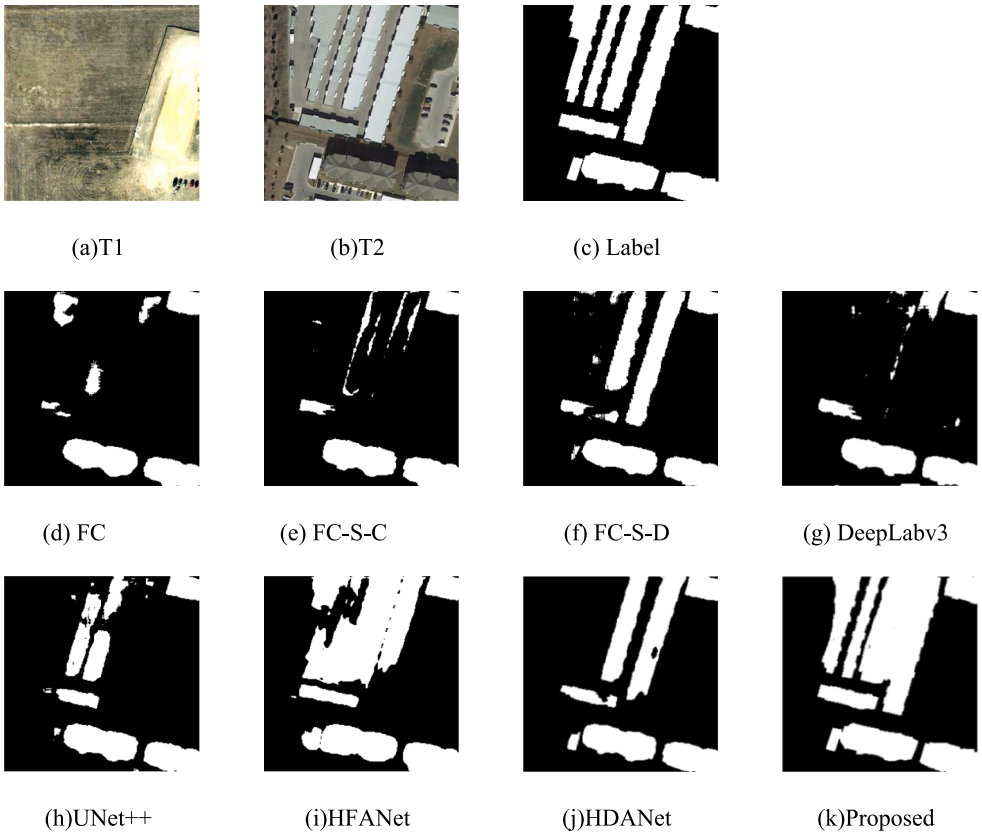


Fig. 3. The detection maps with A change scene on the L-CD dataset.

and change detection. Meanwhile, the two branches generate different types of loss function values. The differences between the two values are coordinated to jointly optimize the backbone network, so that the model achieves a globally optimal detection performance. Two kinds of losses have been integrated in W-Net, which comes from the sueprpixel generation part and the change discrimination part. The superpixel generation part employs two losses. The one is the distance with CIELAB features and the other is the spatial distance. Moreover, the change discrimination part employs the binary cross entropy loss. The balance parameter between these two parts has been optimized to set as 0.03. Only using CIELAB features or other unsupervised information cannot obtain the supervised semantic information. The superpixel generation part in this work is to improve the boundary integrity and reduce the pepper noise in the changed parcels, and the one-hot vector features is employed in the other head of the W-Net. Due to the lack of semantic information in the superpixel generation part, the generated superpixels are cracked, which can be modified using the sufficient semantic labels in the future research. All the SLIC loss and the supervised loss can be propagated to the encoder part of the W-Net.

To be more specific, two images of the same region at difference periods are acquired and normalized separately to obtain two images with size $H \times W \times 3$. These two images are input into the network and combined into an input with size $H \times W \times 6$ by channel-wise concatenation. The proposed network is a single-input, multi-output structure, and the input data are downsampled by the multi-segment convolutional structure of the network to generate multiple sets of depth feature maps with multi-scales $\{Hi \times Wi \times Ci\}$. In the upsampling part of the model, the feature map generates four sets of intermediate features by several sets of common skip connections and deconvolution layers. For the last set of upsampling, the proposed W-Net model sets up a set of parallel convolution modules, named "double heads", which are the change detection head and the superpixel head. The parallel modules perform the convolution calculation and softmax function on the input features and output the results. The change detection head outputs a classification vector with size H × W × 2, and the superpixel head outputs a superpixel prediction vector with size H × W × 9. Except for the parallel modules, where the two tasks are optimized separately, the other parts are optimized together. In the training step, the change detection head calculates cross entropy between output vector and label. Superpixel head calculates the reconstruction error within the superpixel prediction and the segmentation loss with the labels. The final loss is the sum of these items with different weights.

The final observation equation is expressed as:

$$L = L_{CE}(\boldsymbol{Prob}_{out}, \boldsymbol{R}) + \lambda_1 L_{SLIC}(Q) \tag{9}$$

where λ_1 is the reconciliation hyperparameter; $L_{\mathrm{CE}}(\mathbf{Prob}_{\mathrm{out}}, \mathbf{R})$ is the binary cross entropy between the result and the label. $L_{\mathrm{SLIC}}(Q)$ is the SLIC loss of the superpixel network.

In the test step, the model post-processes the classification vector to obtain the binary classification map and post-processes the superpixel prediction vector to obtain the pixel categorization data. Finally, the classification result map is adaptively soft fused with the pixel categorization data to achieve the detection map.

The adaptive soft fusion equation is given below. Assuming that the pixels in a superpixel are $\{[X_i, Y_i, Ppos_i, Pneg_i], i = 1,2,3,\cdots,N\}$, then the overall classification result for this superpixel is:

$$mode = \begin{cases} 0 & \sum_{i=1}^{N} Ppos_i > \sum_{i=1}^{N} Pneg_i \\ 1 & \sum_{i=1}^{N} Ppos_i \le \sum_{i=1}^{N} Pneg_i \end{cases} \tag{10}$$

where $[X_i, Y_i]$ is the superpixel position; $Ppos_i$ denotes the probability that the position labeling as the 'unchanged'; $Pneg_i$ denotes the probability that the position labeling as 'changed'; *mode* is determined by discriminating the confidence level of the two classes, which represents the class attribute of this superpixel. N is the number of the pixel in one superpixel.

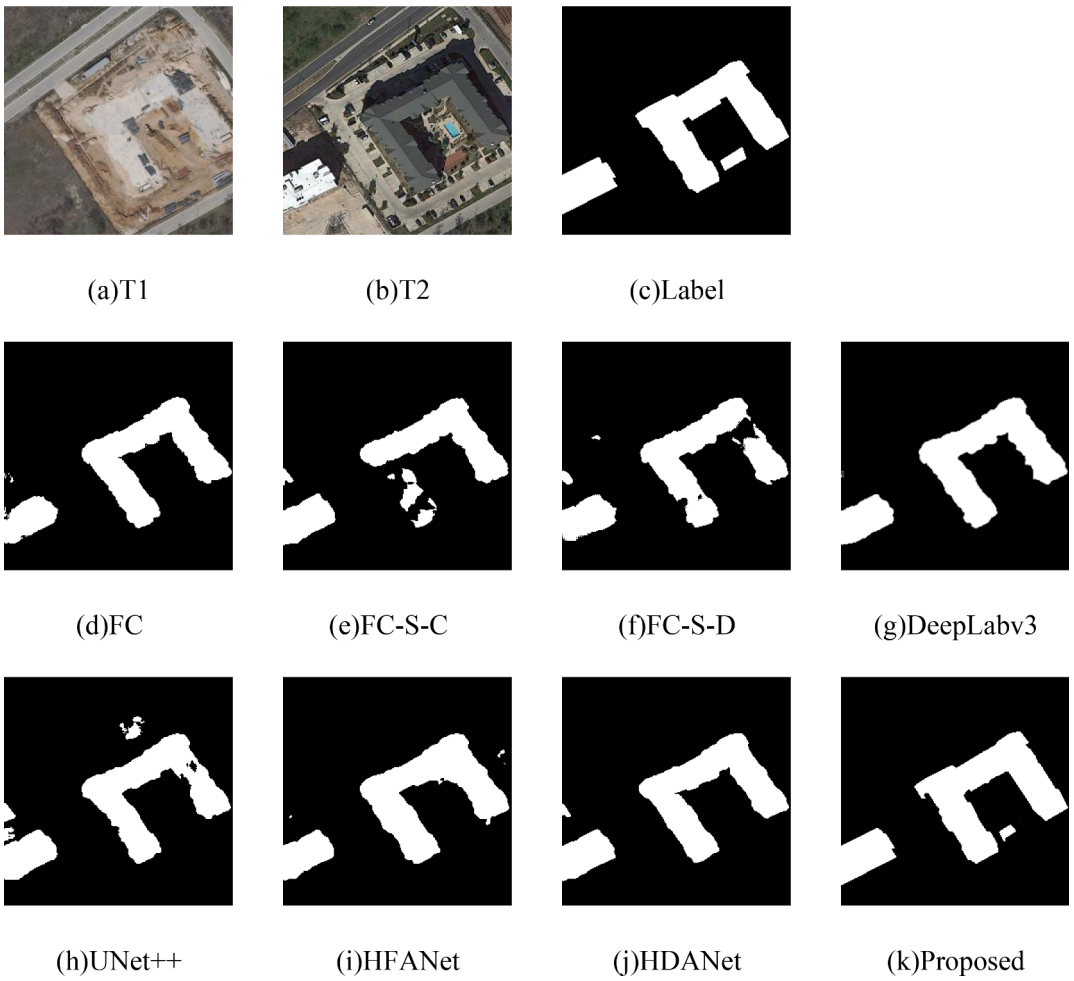


Fig. 4. The detection maps with B change scene on the L-CDdataset.

Table 2
The performance of the different algorithms on WB-CD dataset.

	FC	FC-S-C	FC-S-D	DeepLabv3	UNet++	HFANet	HDANet	Proposed
Precision	0.7623	0.8831	0.802	0.8256	0.8906	0.8344	0.8987	**0.9476**
Recall	0.7765	0.7261	0.7631	0.8197	0.7898	0.8228	0.8255	**0.8888**
F1score	0.7693	0.7969	0.7821	0.8226	0.8372	0.8286	0.8605	**0.9172**
Kappa	0.7603	0.7899	0.7739	0.8158	0.8313	0.8193	0.8554	**0.9142**

The category of each pixel is determined in turn, where the original classification result is assumed to be $CLSorig_i$ with confidence level P_i.

$$CLS_i = \begin{cases} CLSorig_i & P_i > \sum_{i=1}^{N} P_i * \frac{1}{N} \\ mode & otherwise \end{cases} \quad (11)$$

where P_i is the posterior probability for the original label, *mode* denodes the class attribute of this superpixel.

3. Experimental setup

To ensure the generality of the approach, we tested the proposed W-Net method using the WHU building and the LEVIR-CD dataset.

WHU building dataset (WB-CD dataset) is published by Wuhan University (Ji et al., 2018). The dataset was collected in Christchurch, New Zealand, with two periods of images acquired in 2012 and 2016. The original image has 15354 × 32507 pixels, which has been cropped into 256 × 256 non-overlapping images and divided in the ratio of 1:1:8 to get 743 images for the test set, 743 for the validation set and 5948 for the training set.

LEVIR-CD dataset (L-CD dataset) is a publicly available dataset targeting building change (Chen & Shi, 2020) between 2002 and 2018. The image in the original image set has 1024 × 1024 pixels and was sampled by the authors into a test dataset, a training dataset, and a validation dataset. We cropped the original image to 256 × 256 to obtain 2048 images to test the W-Net, 1024 images to validate the optimized model and 7120 images to train the W-Net.

SYSU-CD dataset is a batch of change detection datasets published by the Sun Yat-Sen University (Shi et al., 2022). The main change types of this dataset include changes in different natural objects. The dataset has been released with training set, validation set and test set by the authors. The number of images is 12000, 4000, and 4000 for the training, validation and test set respectively. The size of each image is 256 × 256.

3.1. Implementation details

The W-Net algorithm was conducted under the PyTorch 1.10

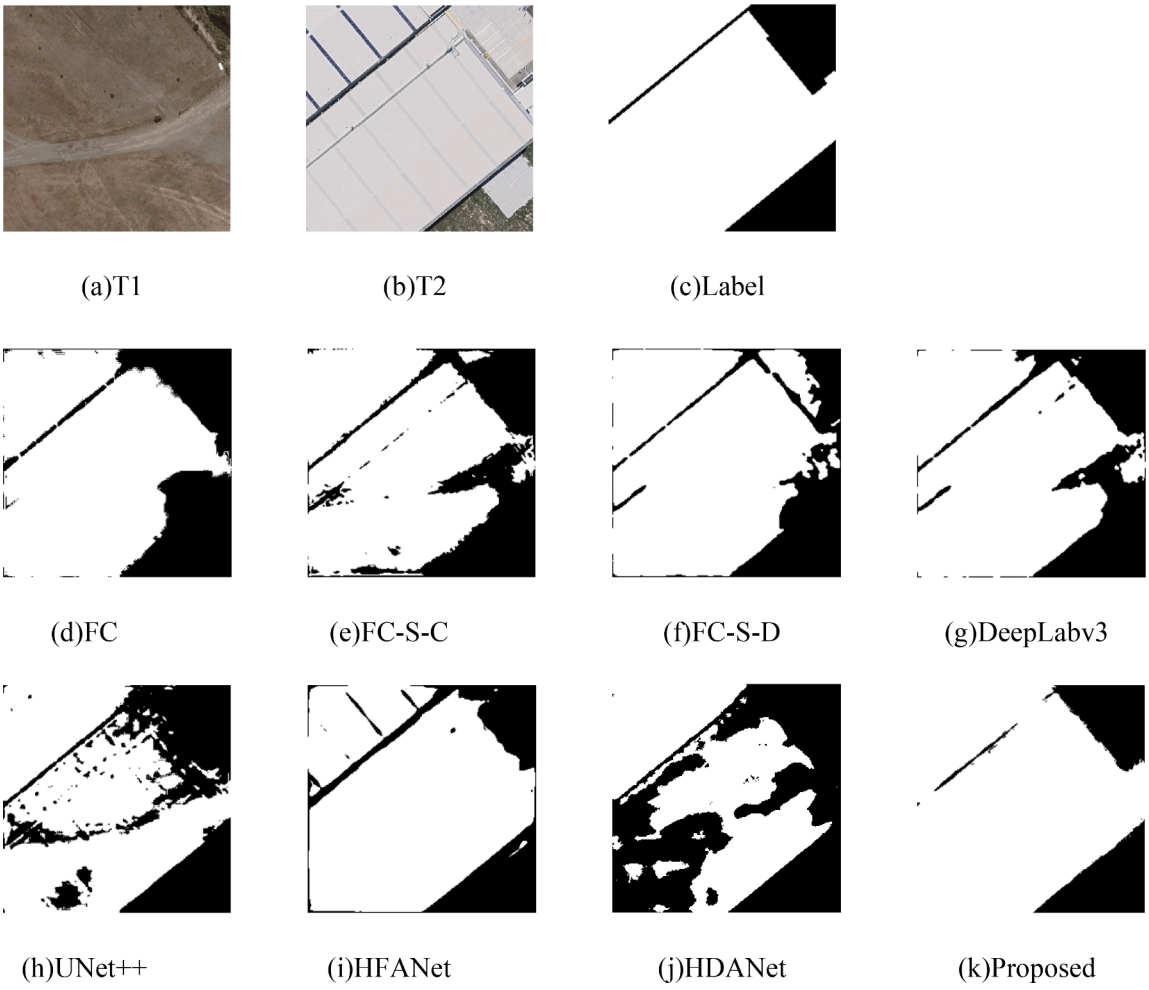


Fig. 5. The detection maps with A change scene on the WB-CD dataset.

framework. The optimizer is Adam optimizer with the initial learning rate of 0.005. The β1, andβ2 was set to 0.9 and 0.999, respectively. The weight decay was 0.01. The batch size was 8, and the training epochs were based on early stopping criterion. We observed in the experiments that the best results were achieved with λ1 and λ2 set to 0.03 and 0.001, respectively, in the loss function for the different settings. The details of the related experiments are provided in Section 5. The same training method was used on all the test datasets. The parameter settings used in the other comparison methods were essentially the same as the above. All the experiments in this study were conducted on an NVIDIA RTX 3090 GPU device and a Linux server.

3.2. Evaluation indicators

We selected the four indicators of F1-score, kappa coefficient, recall and precision to assess the accuracy, which are calculated by the following equations:

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (14)$$

$$OA = \frac{TP + TN}{TP + TN + FP + TN} \quad (15)$$

$$P = \frac{(TP + FP)(TP + FN) + (FN + TN)(FP + TN)}{(TP + TN + FP + TN)^2} \quad (16)$$

$$kappa = \frac{OA - P}{1 - P} \quad (17)$$

where FP denotes the amount of incorrectly recognized ‘changed’ label, TN is the amount of correctly recognized ‘unchanged’ label, FN is the number of incorrectly recognized ‘unchanged’ label, and TP represent the amount of correctly classified ‘changed’ label,

4. Experiments

To confirm the advantage of W-Net in terms of accuracy, seven other advanced algorhtims are compared the W-Net method, namely, FC, FC-S-D, FC-S-C, DeepLabv3, UNet++, HFANet, and HDANet.

FC (Daudt et al., 2018) uses U-Net to input the images of two periods into the network by fusing the features with the same size in each step, to recover the information that is lost in downsampling the feature maps. FC-S-C (Daudt et al., 2018), as an extended form, uses a Siamese-structured VGG network for feature extraction, processing the two images through two encoder branches with the same structure and the same parameters. The three features in each branch and the corresponding layer in the decoding are skip connected. The follows the encoders are integrated. FC-S-D (Daudt et al., 2018), which is also an extended form of FC_EF, utilizes the absolute difference of the subtracted features of the two decoders. DeepLabv3 (Chen et al., 2017) utilzes atrous spatial pyramid pooling structure with different expansion rates

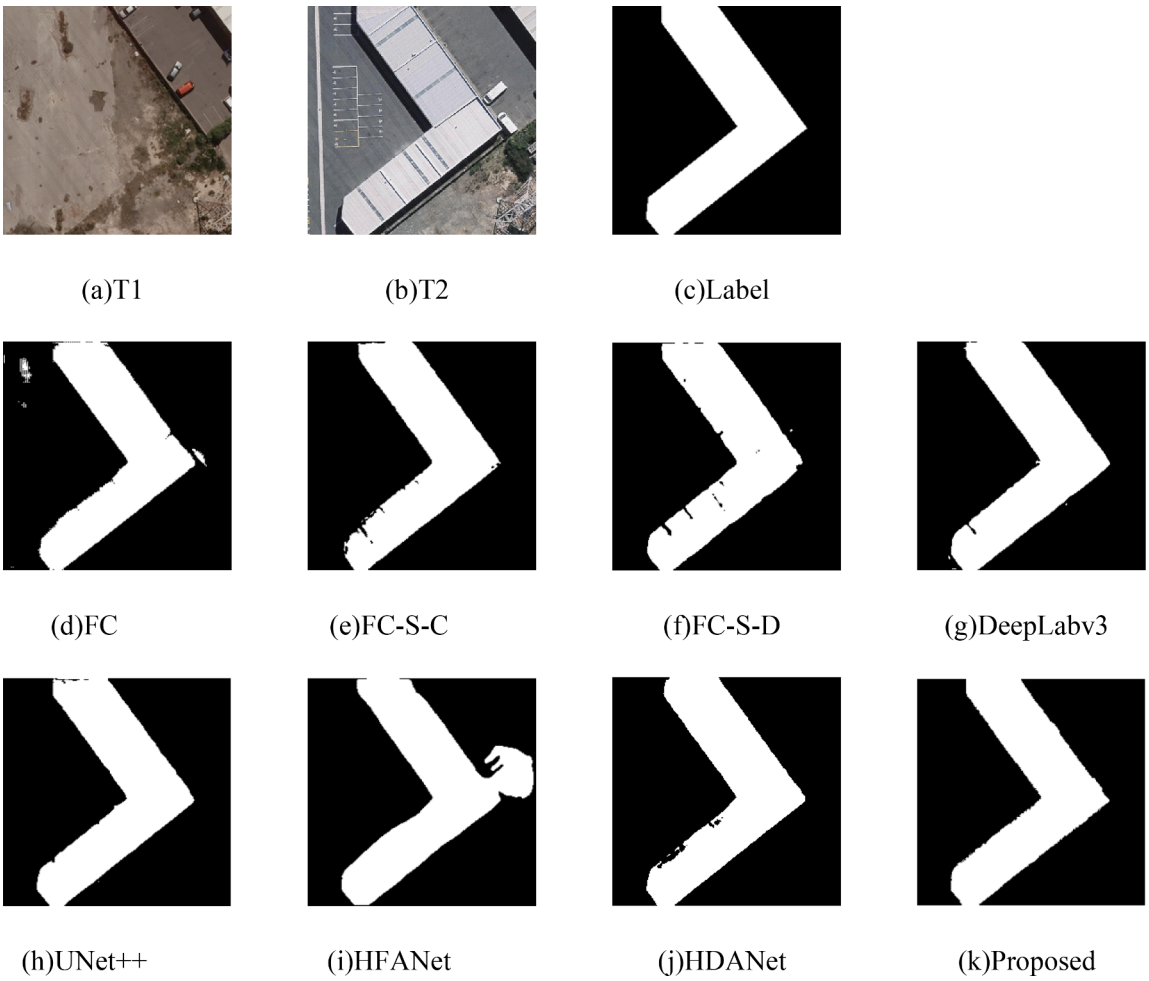


Fig. 6. The detection maps with B change scene on the WB-CD dataset.

Table 3
The performance of the different algorithms on SYSU-CD dataset.

	FC	FC-S-C	FC-S-D	DeepLabv3	UNet++	HFANet	HDANet	Proposed
Precision	0.8269	0.7792	**0.8492**	0.8099	0.8144	0.8013	0.7853	0.8302
Recall	0.6308	0.7208	0.643	0.7065	0.7466	0.7641	0.7988	**0.8037**
F1score	0.7156	0.7488	0.7319	0.7547	0.779	0.7823	0.792	**0.8167**
Kappa	0.6427	0.6752	0.6635	0.6856	0.7146	0.7617	0.7271	**0.7724**

of convolution in parallel, to learn the features of objects at different scales. We used ResNet-50 as the feature extraction network. UNet++ (Peng et al., 2019) uses two periods of image superposition as input and uses a more intensive jump connection approach, compared to U-Net. It also uses a deep supervision strategy to compute multiple outputs of different layers with simultaneous losses to improve the accuracy and training stability. HFANet (Zheng et al., 2022) uses a Siamese network as the backbone and applies a new attention module to capture the building information, enabling better detection of the edges of changed buildings. HDANet (Wang et al., 2022b) utilizes HRNet architecture to capture the features, connecting four different resolutions in parallel to achieve feature fusion between the different resolutions.

4.1. Experiments on the L-CD dataset

Table 1 reports the detection performance of each method on the L-CDdataset. W-Net algorithm outperforms the other methods. W-Net achieves the highest results in F1-score and kappa, with the value of 0.9031 and 0.8969, comparing with the other approahes. DeepLabv3, UNet++, and HDANet also achieve high detection accuracies, among which HDANet obtains the highest F1-score of 0.8987. The F1-score of W-Net is improved by 0.35% and the kappa coefficient is improved by 0.44% comparing with HDANet. The detection accuracy of all the algorithms in the FC, FC-S-D and FC-S-C is worse than that of the above algorithms. Although the precision of FC-S-C is the highest among all the methods, the recall is lower and the detection of changed objects is not comprehensive, resulting in a low F1-score of 0.8401. The F1-score of W-Net shows an improvement of 6.3% and the kappa coefficient is improved by 6.45% comparing with HDANet.

Fig. 3 and Fig. 4 show the visualization of the detection results under different scenes. From Fig. 3, all the methods, except for W-Net, show some omissions in this changing scene, among which FC-S-C, FC, and DeepLabv3 show the most serious omissions. W-Net algorithm has the highest target coverage and the best detection effect comparing with other approaches, and the generated change map is the most consistent to the real. For Fig. 4, FC, DeepLabv3, HDANet, and W-Net obtain the best detection results, FC-Siam-Diff and UNet++ have some voids in the changed regions, and HFANet shows false detection at the edges. Among

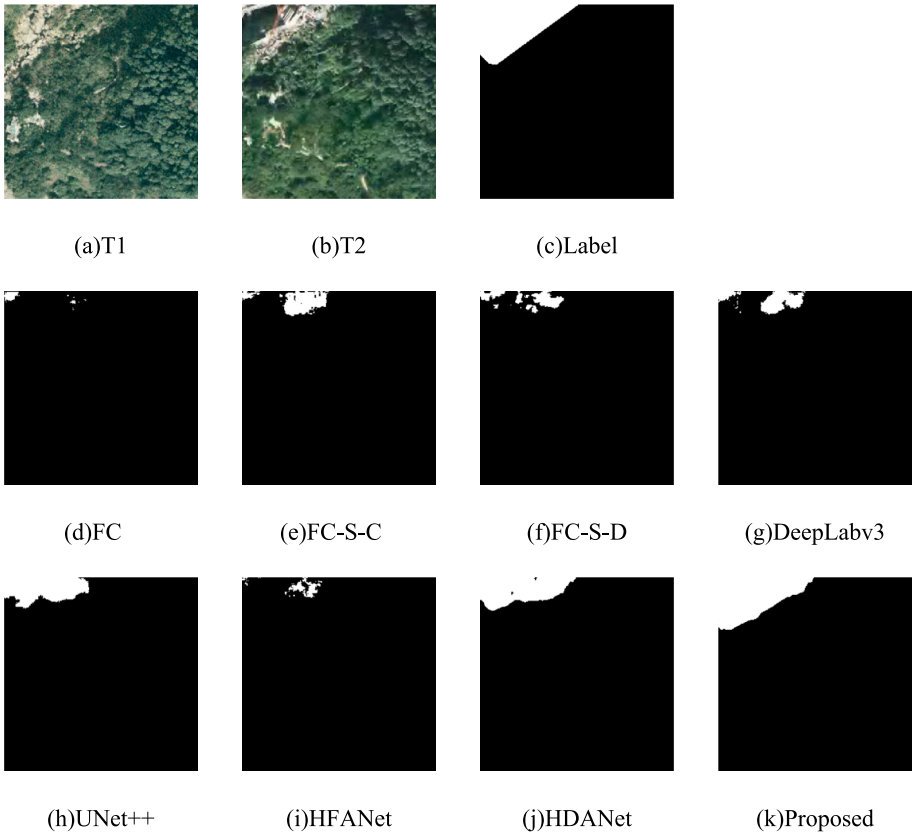


Fig. 7. The detection maps with A change scene on the SYSU-CD dataset.

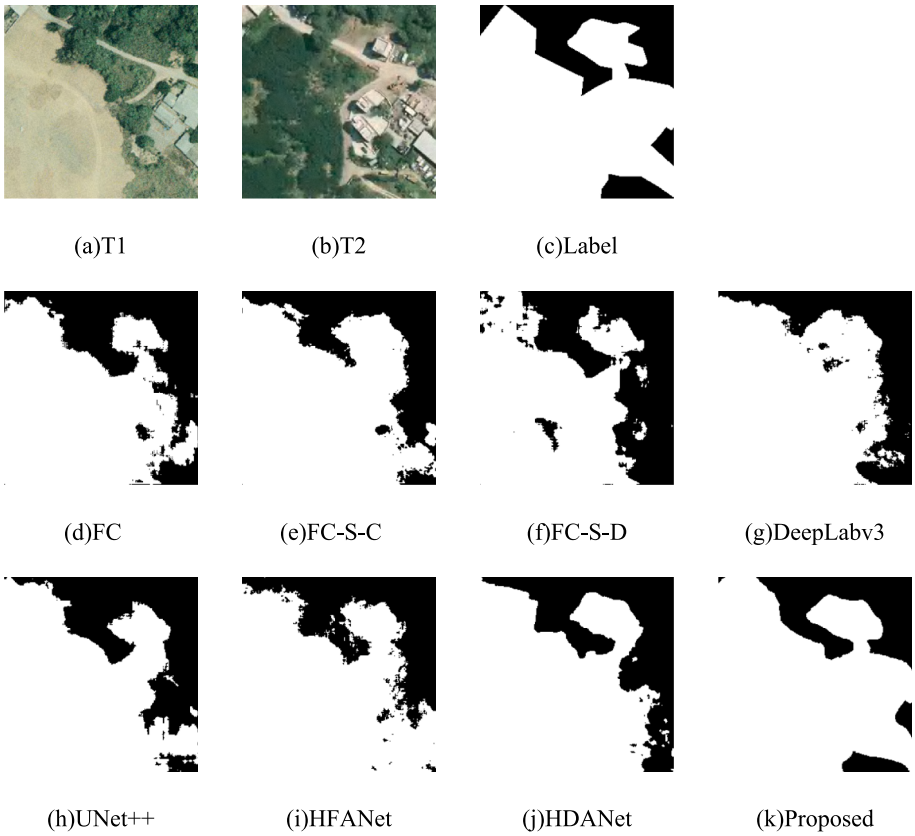


Fig. 8. The detection maps with B change scene on the SYSU-CD dataset.

Table 4
Results of the experiments with different superpixel methods.

Datasets	LEVIR-CD			WB-CD			SYSU-CD		
Methods	SLIC	SSN	Proposed	SLIC	SSN	Proposed	SLIC	SSN	Proposed
F1score	0.8863	0.8895	**0.9031**	0.8908	0.8941	**0.9172**	0.794	0.8042	**0.8167**
Kappa	0.885	0.8848	**0.8969**	0.8895	0.9002	**0.9142**	0.7441	0.767	**0.7724**

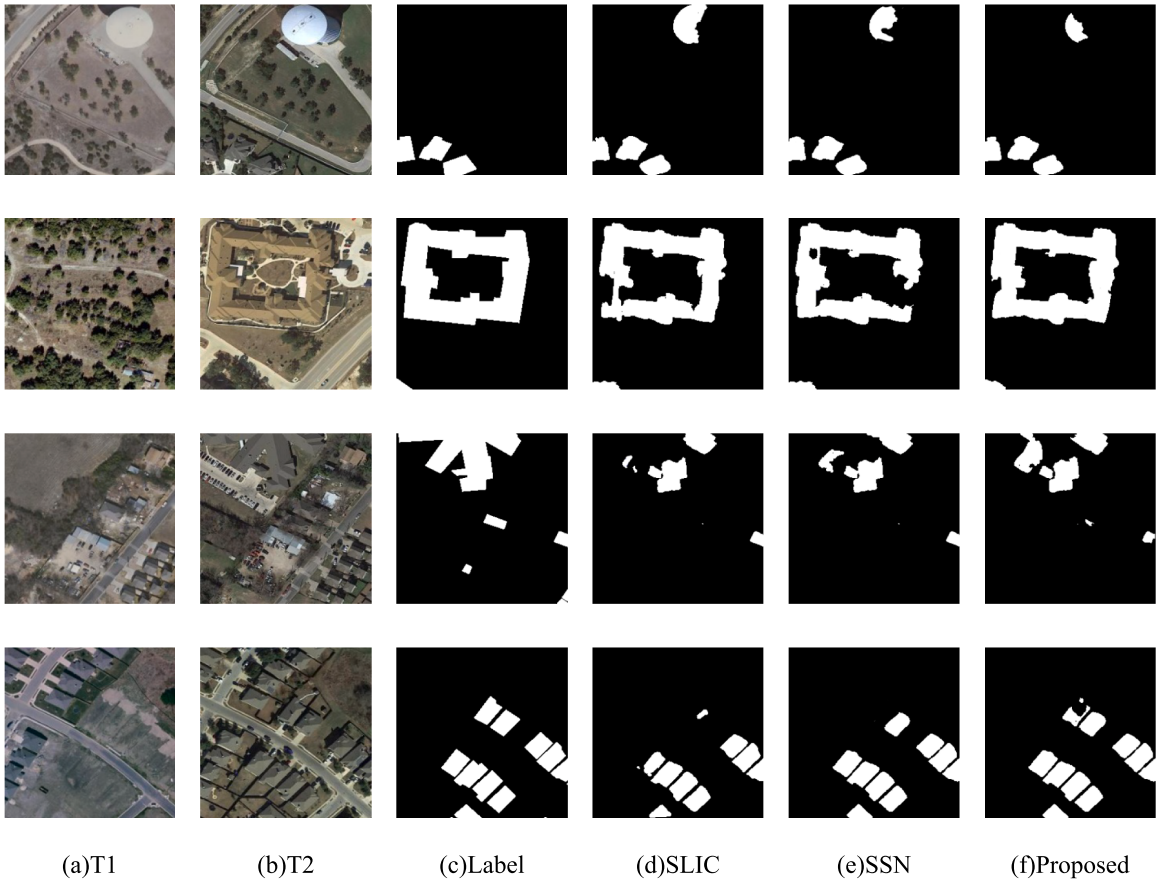


Fig. 9. The detection maps generated by the different superpixel algorithms on the L-CDdataset.

the different methods, W-Net effectively distinguishes the changed area, and the building edge contours of the detection results are the most complete.

4.2. Experiments on the WB-CD dataset

Table 2 reports the accuracy of all algorithms on the WB-CD dataset. W-Net outperforms the other method, with 0.9476 for the precision, 0.8888 for the recall, 0.9172 for the F1-score, and 0.9142 for the kappa coefficient, with each metric being the highest value among the different methods, quantitatively proving the superior of W-Net. The four counterpart methods of DeepLabv3, UNet++, HFANet, and HDANet all achieve a good accuracy, with HDANet performing the best. As with the previous dataset, the performance of the three methods in the FC series is poor, with FC obtaining the lowest accuracy. The F1-score of FC is lower than that of W-Net over 10%.

As shown in Figs. 5 and 6, two scenes are selected for visualization. From Fig. 5, the detection process for large objects results in the poor existence of the severe building damage, except for HFANet and W-Net. The HFANet segmentation results are more complete, but there are some missed detections affected by the texture of the target features. W-Net obtains the best results and demonstrates an optimal visual performance. In the second scene, all the methods perform better, overall, although HFANet shows a small area of false detection. W-Net again obtains the best change map, retains clearer boundaries, and further demonstrates its superior change detection capability.

4.3. Experiments on the SYSU-CD dataset

Different with the two datasets above, the main change types of this dataset include changes on different natural category objects, such as changes from forest land to building land, river banks expansion, the disappearance of ships in the water and the addition of buildings, etc.. The change of ground objects in this dataset is complicated, which brings the challenges to the change detection algorithms. Table 3 lists the accuracy of the detection results obtained for the SYSU-CD dataset. When tackling the forest changes or some changes in mountainous areas which have more irregular and unclear change boundary compared with the man-made building areas. W-Net achieved the highest values among the test results in terms of the kappa coefficient and F1-score, with an F1-score of 0.8167and a kappa coefficient of 0.7724. FC-S-D obtained the highest Precision, with a Precision 0.8492. FC obtained the lowest F1 score which was as with the previous two datasets.

Two natural resources change scenes have been selected for

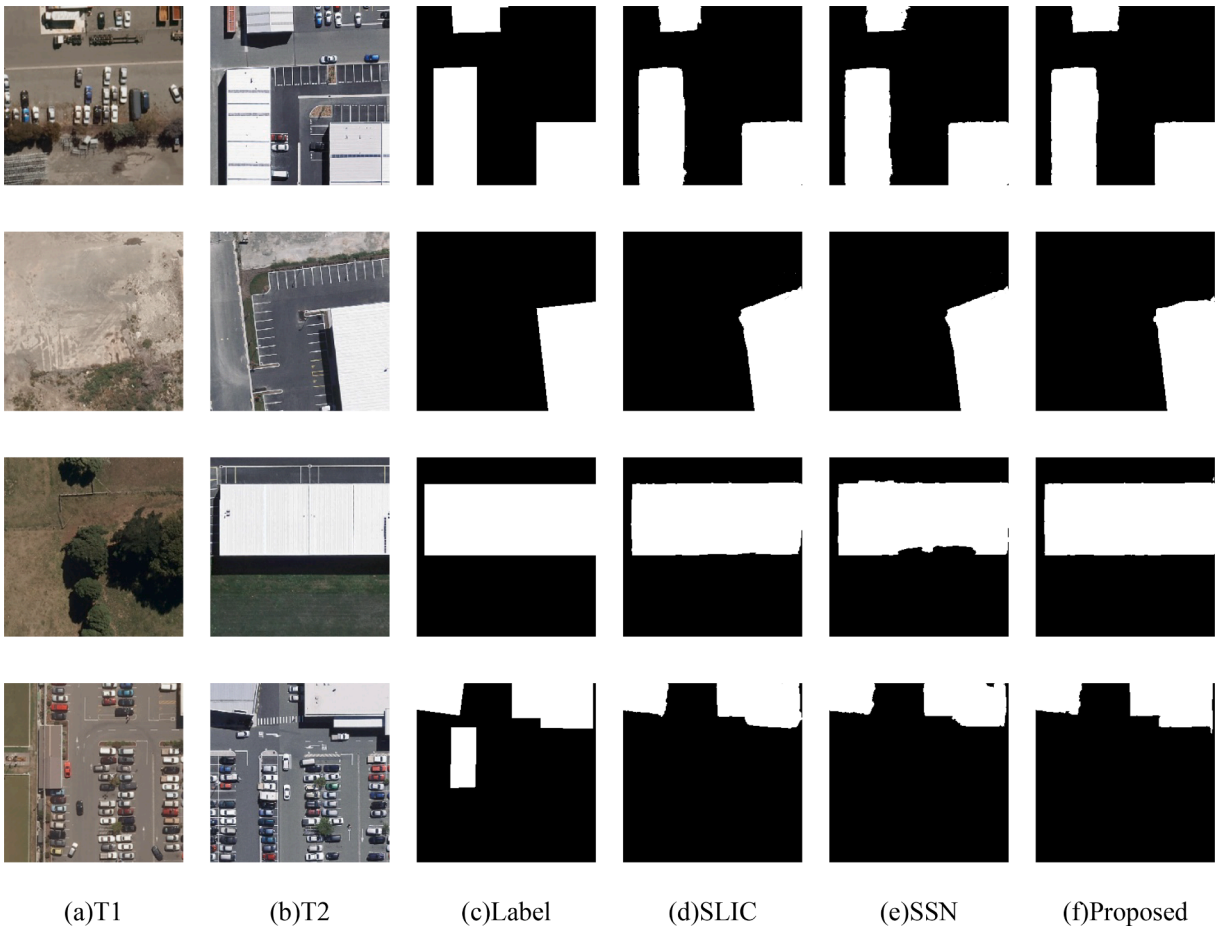


Fig. 10. The detection maps generated by the different superpixel algorithms on the WB-CD dataset.

visualization, shown in Figs. 7 and 8. From Fig. 7, the changed area in the exposed mountain has been detected well by W-Net, but the results obtained by other methods cannot detected the whole changed area. The HDANet segmentation results are more credible comparing with other contrast methods, but there is some pepper noise in the changed area comparing with the W-Net. The second scene shows that for the more complex change scenario, the five counterpart methods of FC, FC-S-C, FC-S-D, DeepLabv3 and HFANet have obtain the bad visual performance. Among the other methods, W-Net shows no obvious omission errors.

4.4. Experiments with other superpixel methods

The traditional superpixel approach normalizes the predicted values based on the plurality in the superpixel range, but this needs to be supported by better classification results. If the initial classification results are poor, superpixels may play a side role. Therefore, the method used in the proposed approach uses end-to-end superpixels and adaptive fusion to effectively alleviate the dependence of superpixels on classification quality. To ensure the validity of the proposed superpixel model, we selected four samples of difficult scenes on WB-CD dataset, L-CD dataset and SYSU-CD dataset for comparison experiments compared to other superpixel methods. We chose SLIC and SSN methods to be added to the backbone. The compare method SLIC and SSN methods are conventional superpixel methods. The detection results obtained by the backbone are refined by the superpixel results generated using SLIC and SSN. Table 4 lists the performance for each metric. Compared with SLIC, W-Net improves the performance of the F1-score by 1.68% and the kappa coefficient by 1.19% on the L-CDdataset comparing. W-Net improves the accuracy of the F1-score by 2.31% and the kappa coefficient by 1.40% on the WB-CD dataset comparing with SSN. W-Net improves the accuracy of the F1-score by 2.27% and the kappa coefficient by 3.13% on the SYSU-CD dataset comparing with SLIC. The F1-scores of the W-Net are consistently higher than the results of the other approaches on both datasets. It can also be illustrated in the Figs. 9–11 that the W-Net algorithm is effective in reducing the error rate, despite the poorer classification base, taking advantage of the double head and adaptive fusion in the training step to generate better detection maps.

5. Discussion

5.1. Hyperparameters

In this part, the effectiveness of the hyperparameters is disscussed. The loss function of W-Net introduces a new hyperparameter, the balance parameter λ_1, which controls the proportion of the loss values in the superpixel branch so that it is as balanced as possible with the classification branch. The superpixel branch is frozen when$\lambda_1 = 0$. The loss value of the superpixel branch will be up to hundreds when $\lambda_1 = 1$. The value of λ_1 has a significant influence on the detection results, which needs to be chosen reasonably, to ensure that both branches can maintain the same optimization progress in the training step. Since the ratio between the segmentation loss and cross entorpy loss of the superpixel module is fixed, we employed the same settings as Yang et al. [49] and fixed λ_2 to 0.001 while changing only the value of λ_1. For a fair comparison, extensive experiments on the value of λ_1 are conducted and the four metrics for W-Net on the two datasets at different values of λ_1 are evaluated.

It can be observed that, when the value of λ_1 is larger than 0.1, the superpixel branch has a severe inhibitory effect on the classification

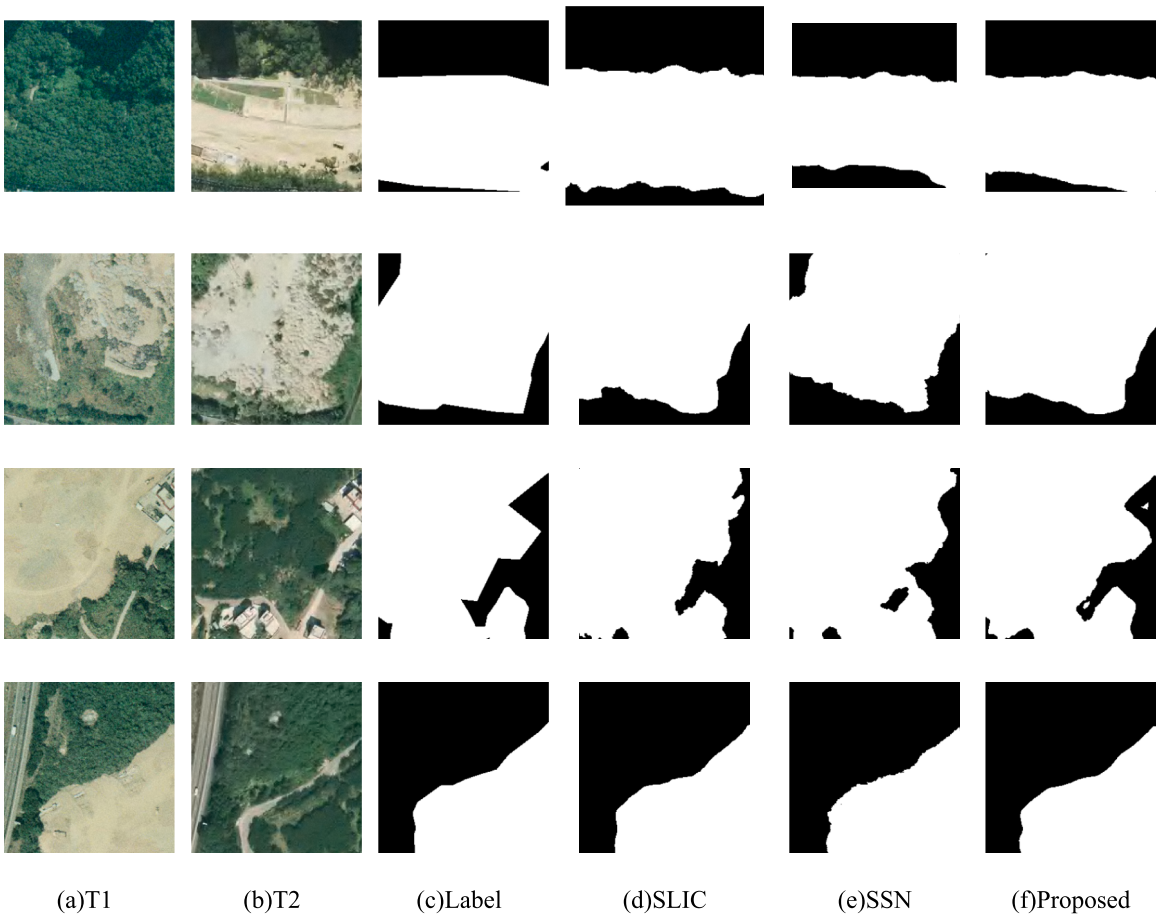


Fig. 11. The detection maps generated by the different superpixel algorithms on the SYSU-CD dataset.

Table 5
The results obtained by varying λ_1 for the W-Net loss function on the three datasets.

λ_1	L-CD		WB-CD		SYSU-CD	
	F1score	Kappa	F1score	Kappa	F1score	Kappa
0.001	0.8630	0.8468	0.8827	0.8736	0.7838	0.7433
0.01	**0.9041**	0.8938	0.9041	0.8938	0.8025	0.7668
0.03	0.9033	**0.8996**	**0.9132**	**0.9051**	**0.8167**	**0.7724**
0.1	0.8373	0.8141	0.8942	0.8827	0.7921	0.7644
1	0.7298	0.7129	0.7845	0.7579	0.797	0.7532

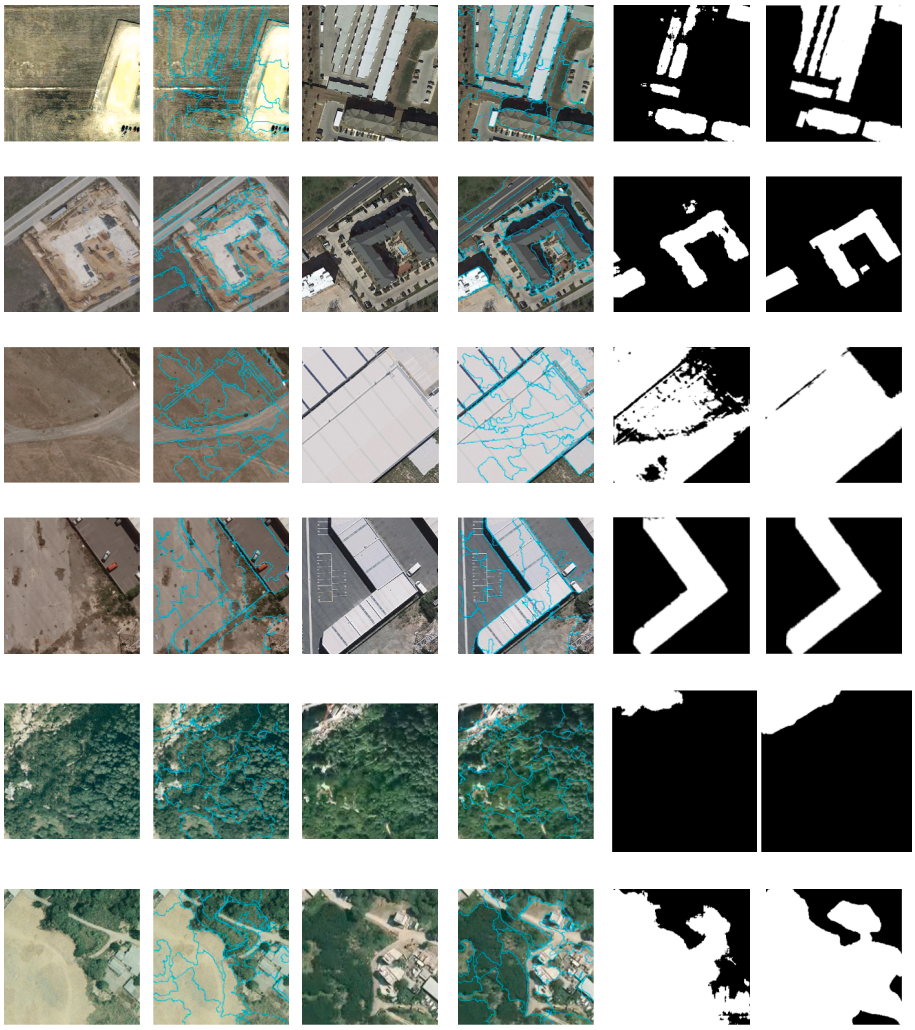
branch, and it is at the order of 0.01 that the balance between the two branches is maintained. In general, the value of λ_1 is selected to ensure that the loss function values of the two branches are in the same order of magnitude, and in fact the best λ_1 range is at the 0.01 level. In addition to the values, we tested values in the interval [0.01,0.09], with similar experimental results. The results for different values of λ_1 on three datasets are presented in Table 5, which can prove that the trained model yields the best performance when λ_1 is 0.03.

The proposed model has a single-input, multi-output structure, with the two branches having their own loss function computation modules. The two branches also have independent gradients in the double-head part, and the gradients are vectorially accumulated in the downsampling. Therefore, if there is an order of magnitude difference between the losses of the two branches, it causes the parameter update in the downsampling part of the model to be more biased toward the branch with the larger loss value, thus hindering the optimization process of the other branch. The hyperparameter λ_1 is what plays the role of adjusting the optimization step, so that the loss function values remain balanced at the early step of training and achieve the global optimum for the model.

5.2. Superpixel ablation study

We conducted ablation discussion on the two datasets to validate the validity of the W-Net. The traditional fusion method refers to the use of superpixels in the results and the comparison of pixel with superpixel attributes based on the classification confidence, to determine the individual pixel attributes. Since W-Net is derived from UNet++, we chose it as the baseline, which just utilized the multiscale features without the superpixel constraint and named "Without_SP". The generated superpixels have been overlaid on the images of two periods. Fig. 12 shows that the change detection results by W-Net are clearly better than that by the baseline which without superpixel generation part. Because the input data of the W-Net are the images of two periods, the generated superpixels represent bitemporal land cover characteristics. The superpixels align to many different landcover, and one building entity generally is commonly characterized by many superpixels. Because of the adaptive soft fusion, the superpixel which distributed in the changed area can constrain the change discrimination part to make the detected areas have regular shapes.

The superpixels were added to UNet++ in different ways to get the comparation models, and the performances are reported in Tables 6. After coupling the superpixels into the change detection algorithm, the model shows performance improvement in accuracy, and adding superpixels can reduce the noise and utilize more information in the scene.



(a)T1 (b)T1_SP (c)T2 (d)T2_SP (e)Without_SP (f) With_SP

Fig. 12. The illustration of the superpixel on the three datasets.

Table 6
The ablation study on the three dataset.

Methods	L-CD		WB-CD		SYSU-CD	
	F1score	Kappa	F1score	Kappa	F1score	Kappa
Without_SP	0.8823	0.8762	0.8372	0.8313	0.779	0.7146
With_SP	**0.9031**	**0.8969**	**0.9172**	**0.9142**	**0.8167**	**0.7724**

6. Conclusion

In this work, we have proposed a double-head model that couples a change detection branch and a superpixel branch. The proposed W-Net method effectively improves the building edge blurring through the auxiliary inference of superpixels and makes the model more holistic and fully optimized by coupling the training methods. The advancement of the W-Net was demonstrated using three public datasets. The F1score on LEVIR-CD dataset was 0.9031 and kappa coefficient was 0.8969. The F1-score on WHU building dataset was 0.9172 and kappa coefficient was 0.9142. The F1-score on SYSU-CD dataset was 0.8167and and kappa coefficient was 0.7724. The proposed W-Net method achieved a superior performance on three well-known open-source datasets when compared with the recent CNN-based and FCN-based change detection methods. In addition, we explored the role and association of the various modules of the W-Net. After coupling the superpixels into the change detection algorithm, the model shows performance improvement in accuracy, and adding superpixels can reduce the noise.

CRediT authorship contribution statement

Xue Wang: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Xulan Yan:** Data curation, Writing – original draft. **Kun Tan:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing. **Chen Pan:** Visualization, Investigation. **Jianwei Ding:** Resources, Supervision. **Zhaoxian Liu:** Investigation, Formal analysis. **Xinfeng Dong:** Software, Formal analysis.

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.

Data availability

Data will be made available on request.

Acknowledgment

This work is jointly supported by the Natural Science Foundation of China (grant nos. 42171335, 42001350), the Postdoctoral Science Foundation of China (grant nos. 2023T160218, 2021M691016), the Shanghai Municipal Science and Technology Major Project (grant no. 22511102800), the National Civil Aerospace Project of China (grant no. D040102) and Supported by the International Research Center of Big Data for Sustainable Development Goals (CBAS2022GSP07) .

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34 (11), 2274–2282.

Bovolo, F., Bruzzone, L., 2007. A split-based approach to unsupervised change detection in large-size multitemporal images: Application to tsunami-damage assessment. IEEE Trans. Geosci. Remote Sens. 45 (6), 1658–1670.

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. Retrieved from Remote Sens. (Basel) 12 (10), 1662. https://www.mdpi.com/2072-4292/12/10/1662.

Daudt, R. C., Le Saux, B., & Boulch, A. (2018). Fully convolutional siamese networks for change detection. Paper presented at the 2018 25th IEEE International Conference on Image Processing (ICIP).

Di, S., Liao, M., Zhao, Y., Li, Y., Zeng, Y., 2021. Image superpixel segmentation based on hierarchical multi-level LI-SLIC. Opt. Laser Technol. 135, 106703.

Gadde, R., Jampani, V., Kiefel, M., Kappler, D., & Gehler, P. V. (2016). Superpixel convolutional networks using bilateral inceptions. Paper presented at the European conference on computer vision.

Jampani, V., Sun, D., Liu, M.-Y., Yang, M.-H., & Kautz, J. (2018). Superpixel sampling networks. Paper presented at the Proceedings of the European Conference on Computer Vision (ECCV).

Ji, R., Tan, K., Wang, X., Pan, C., Xin, L., 2021. Spatiotemporal monitoring of a grassland ecosystem and its net primary production using Google Earth Engine: A case study of inner mongolia from 2000 to 2020. Remote Sens. (Basel) 13 (21), 4480.

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.

Kim, D.-H., Sexton, J.O., Noojipady, P., Huang, C., Anand, A., Channan, S., Townshend, J.R., 2014. Global, Landsat-based forest-cover change from 1990 to 2000. Remote Sens. Environ. 155, 178–193.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Li, X., He, M., Li, H., Shen, H., 2021. A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Mu, C., Dong, Z., Liu, Y., 2022. A Two-Branch Convolutional Neural Network Based on Multi-Spectral Entropy Rate Superpixel Segmentation for Hyperspectral Image Classification. Remote Sens. (Basel) 14 (7), 1569.

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 Sens. (Basel) 11 (11), 1382.

Sakurada, K., & Okatani, T. (2015). Change Detection from a Street Image Pair using CNN Features and Superpixel Segmentation. Paper presented at the BMVC.

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.

Shi, W., Zhang, M., Zhang, R., Chen, S., Zhan, Z., 2020. Change detection based on artificial intelligence: State-of-the-art and challenges. Remote Sens. (Basel) 12 (10), 1688.

Shuai, W., Jiang, F., Zheng, H., Li, J., 2022. MSGATN: A Superpixel-Based Multi-Scale Siamese Graph Attention Network for Change Detection in Remote Sensing Images. Appl. Sci. 12 (10), 5158.

Tan, K., Zhang, Y., Wang, X., Chen, Y., 2019. Object-based change detection using multiple classifiers and multi-scale uncertainty analysis. Remote Sens. (Basel) 11 (3), 359.

Tan, K., Ma, W., Chen, L., Wang, H., Du, Q., Du, P., Li, H., 2021. Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning. J. Hazard. Mater. 401, 123288.

Wang, X., Tan, K., Du, Q., Chen, Y., Du, P., 2019. Caps-TripleGAN: GAN-assisted CapsNet for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 57 (9), 7232–7245.

Wang, X., Du, J., Tan, K., Ding, J., Liu, Z., Pan, C., Han, B., 2022a. A high-resolution feature difference attention network for the application of building change detection. Int. J. Appl. Earth Obs. Geoinf. 112, 102950.

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. (Basel) 12 (2), 205.

Wang, X., Tan, K., Du, P., Pan, C., Ding, J., 2022b. A unified multiscale learning framework for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 60, 1–19.

Wu, Y., Chen, Y., Yuan, L., Liu, Z., Wang, L., Li, H., & Fu, Y. (2020). Rethinking classification and localization for object detection. Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Yang, F., Sun, Q., Jin, H., & Zhou, Z. (2020). Superpixel segmentation with fully convolutional networks. Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Zhang, H., Lin, M., Yang, G., Zhang, L., 2021. Escnet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images. IEEE Trans. Neural Networks Learn. Syst.

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

Zheng, Z., Wan, Y., Zhang, Y., Xiang, S., Peng, D., Zhang, B., 2021. CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 175, 247–267.

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 3–11.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging 39 (6), 1856–1867.