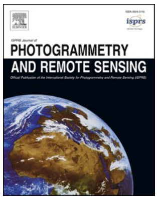
Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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



CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery

Zhi Zheng [a], Yi Wan [a,*], Yongjun Zhang [a,*], Sizhe Xiang [a], Daifeng Peng [b,c], Bin Zhang [a]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[b] School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
[c] Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy

ARTICLE INFO

Keywords:
Change detection
Optical remote sensing image
Deep convolutional neural networks
Cross-Layer Block (CLB)
Cross-Layer Network (CLNet)
UNet

ABSTRACT

Change detection plays a crucial role in observing earth surface transition and has been widely investigated using deep learning methods. However, the current deep learning methods for pixel-wise change detection still suffer from limited accuracy, mainly due to their insufficient feature extraction and context aggregation. To address this limitation, we propose a novel Cross Layer convolutional neural Network (CLNet) in this paper, where the UNet structure is used as the backbone and newly designed Cross Layer Blocks (CLBs) are embedded to incorporate the multi-scale features and multi-level context information. The designed CLB starts with one input and then split into two parallel but asymmetric branches, which are leveraged to extract the multi-scale features by using different strides; and the feature maps, which come from the opposite branches but have the same size, are concatenated to incorporate multi-level context information. The designed CLBs aggregate the multi-scale features and multi-level context information so that the proposed CLNet can reuse extracted feature information and capture accurate pixel-wise change in complex scenes. Quantitative and qualitative experiments were conducted on a public very-high-resolution satellite image dataset (VHR-Dataset), a newly released building change detection dataset (LEVIR-CD Dataset) and an aerial building change detection dataset (WHU Building Dataset). The CLNet reached an F1-score of 0.921 and an overall accuracy of 98.1% with the VHR-Dataset, an F1-score of 0.900 and an overall accuracy of 98.9% with the LEVIR-CD Dataset, and an F1-score of 0.963 and an overall accuracy of 99.7% with the WHU Building Dataset. The experimental results with all the selected datasets showed that the proposed CLNet outperformed several state-of-the-art (SOTA) methods and achieved competitive accuracy and efficiency trade-offs. The code of CLNet will be released soon at: https://skyearth.org/publication/project/CLNet.

1. Introduction

Land cover change detection is a crucial problem in earth observation, land-use monitoring, urban expansion, resource management, etc. (Akcay et al.,2010; P. Zhang et al., 2016; Chen et al.,2013; Hulley et al.,2014; Stramondo et al., 2006; Yang et al., 2012; Xian and Homer, 2010; Liang and Weng, 2010). Change detection via multitemporal images has been widely investigated over the past decades. Before the recent explosive development of deep learning methods, researchers mainly solved the change detection problem by manually designing the complicated feature extractors, which required a great deal of expert domain knowledge, and the accuracy was hard to be improved. Deep learning methods, leveraging stacked neurons for empirical features encoding, have greatly decreased the demand of expert domain knowledge. Characteristics like deep feature representation and nonlinear modeling ability make deep learning methods more suitable for complex image understanding and therefore are attracting the attention of the remote sensing image change detection community (L. Zhang et al., 2016; Zhu et al., 2017).

Pixel-wise optical remote sensing imagery change detection (ORSICD) is one of the most important branches of change detection in the remote sensing community. Essentially, pixel-wise ORSICD is to predict the pixels into changed/unchanged labels and then obtain the binary change maps. Therefore, this task can be regarded as dense pixel

[*] Corresponding authors.
E-mail addresses: zhengzhi@whu.edu.cn (Z. Zheng), yi.wan@whu.edu.cn (Y. Wan), zhangyj@whu.edu.cn (Y. Zhang), xiangsizhe@whu.edu.cn (S. Xiang), daifeng@nuist.edu.cn (D. Peng), bin.zhang@whu.edu.cn (B. Zhang).

https://doi.org/10.1016/j.isprsjprs.2021.03.005
Received 3 August 2020; Received in revised form 26 February 2021; Accepted 1 March 2021
Available online 30 March 2021

classification in the deep learning field, and thus the successful experience of semantic segmentation can be embraced and transferred to deal with the ORSICD (Alcantarilla et al., 2018; Wang et al.,2018; Peng and Guan, 2019; R.C. Daudt et al., 2018a, 2018b). In recent years, many methods (Alcantarilla et al., 2018; Caye Daudt et al., 2018; R.C. Daudt et al., 2018a, 2018b) were proposed based on the fully convolutional network (FCN) (Long et al., 2015) and have been proved to be effective for ORSICD.

It should be noted that there is some difference between the fundamental problem of bi/multitemporal ORSICD and the single-image dense pixel classification (i.e., semantic segmentation). Therefore, though the networks that proposed for single-image dense pixel classification can be transferred or modified to deal with ORSICD, some special characteristics should also be taken into consideration. First, the concatenated input image/fused features are not internally consistent due to the existence of change, that is, the same location of different feature channels might represent different semantic content. Second, how to define and detect accurate change in complex remote sensing scenes is also a crucial problem. Real changes under the negative influences (i.e., scale/season differences) should be distinguished with the determined training labels (i.e., whether regarding the changes of cars' appearance/disappearance as change). These special characteristics make the bi/multitemporal change detection much more complicated than the single-image dense pixel prediction task and also put forward higher requirements of comprehensive feature representation for accurate change detection.

Currently, acquiring context information and incorporating multi-scale features of change areas in bi-/multitemporal images is proved to be effective to predict fine changes and improve change detection accuracy. Therefore, some works try to explore the combination of FCN and the excellent feature extraction blocks proposed for vision tasks for high accuracy ORSICD. Lei et al. (2019) proposed to use spatial pyramid pooling (SPP) (Lin et al., 2017) for better landslide inventory mapping. Zhang et al. (2019) applied atrous spatial pyramid pooling (ASPP) (L.C. Chen et al., 2017) for change detection, which combined the advantages of dilated convolution (Yu and Koltun, 2015) and feature pyramid network (Lin et al., 2017). Chen et al. (2019) proposed a multi-scale feature convolution unit and designed two novel deep Siamese convolutional networks for unsupervised and supervised multitemporal change detection. Most of the strategies used in ORSICD for accuracy improvement can be concluded into two categories: one is using different-size convolutions (i.e., dilated convolution) or multi-scale blocks (i.e., SPP) for receptive field increment; the other is aggregating multi-level context information incorporation, i.e., UNet++ (Zhou et al., 2018). The multi-scale strategy can expand the receptive field or information incorporation while it only extracts features at the same feature level and neglects the relationships between different levels. On the other hand, although UNet++ can achieve multi-level context incorporation, the up-sampling operation it used greatly increased the memory occupation and thus limits the model's learning ability. Thus, both strategies are still unable to incorporate image information sufficiently and cannot generate change maps with enough accuracy.

To ease the contradiction between the feature representation requirements and insufficient feature extraction problem in ORSICD, this paper designs a novel Cross Layer Block (CLB) for more sufficient feature extraction and representation, and proposes a novel network called Cross Layer Convolutional Neural Network (CLNet). This approach contributes to the remote sensing community in the following three major aspects:

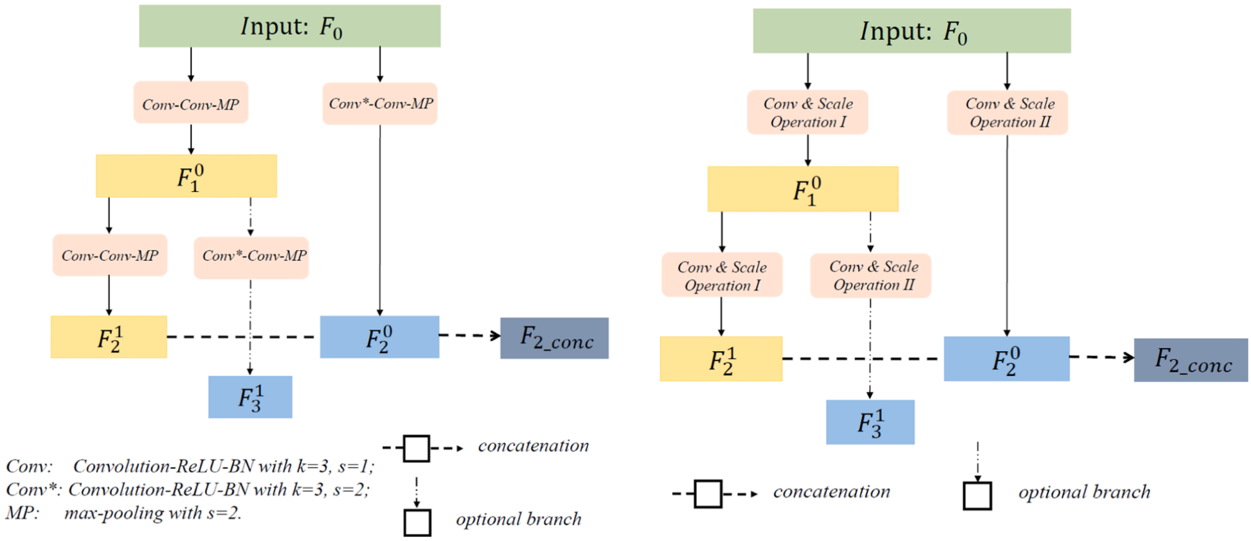
1. A novel Cross Layer Block (CLB) is designed to integrate multi-scale features and multi-level semantic context information. The CLB starts with the concatenated images/fused feature maps and then splits into two parallel but asymmetric branches, and thus achieves the comprehensive and sufficient feature representation.
2. A novel end-to-end Cross Layer convolutional neural Network (CLNet) that is modified from UNet (Ronneberger et al., 2015) is proposed for accurate pixel-wise ORSICD, which contains two CLBs and a dimension compression operation at the encoder part. The stacked CLBs gradually aggregate the image information and the dimension compression operation eliminate the side influence introduced by abundant high-level features. Both of the strategies increase the feature representation ability and thus improved the accuracy for ORSICD.
3. According to experimental results on two types of ORSICD tasks (all-objects change detection and building change detection), CLNet attains new state-of-the-art (SOTA) accuracy and achieves competitive accuracy and efficiency trade-offs compared to several SOTA methods, which demonstrated its effectiveness and robustness for ORSICD.

The remainder of this paper is organized as follows. Section 2 is an overview of the change detection literature. Section 3 provides details about the designed CLB and illustrates the architecture of CLNet. Section 4 presents our comprehensive investigation of the superiority of CLNet along with a comparison of it to several SOTA FCN-based methods. Finally, Section 5 concludes our findings and future works.

2. Related works

Traditional pixel-wise change detection methods can be categorized generally as either pixel-based methods or object-based methods, depending on their specific procedures and optimizing targets. The pixel-based change detection algorithms (Bruzzone and Prieto, 2000; Benedek and Szirányi, 2009; Bovolo and Bruzzone, 2006; Zanetti et al., 2015; Ghosh et al., 2009) prefer to generate a difference map by comparing the corresponding pixels in the given bitemporal images, and then the change map is generated by threshold segmentation or other decision strategies. The object-based methods (Hussain et al., 2013; Desclée et al., 2006; R.C. Daudt et al., 2018a, 2018b; Leichtle et al., 2017; Yu et al., 2016; Xiao et al.,2017; Peng and Zhang, 2017), on the other hand, segment pixels into disjoint homogeneous objects under predetermined conditions, and then obtain the change detection results according to the segmented objects. Different from the traditional methods, the learning-based methods directly predict pixel-level classification maps as change detection results under an end-to-end framework, which blurs the boundary between the pixel-based approaches and the object-based approaches (Zhang et al., 2019).

In recent years, tremendous efforts have been made to exploit the deep learning methods for remote sensing imagery change detection. Based on convolutional neural network (CNN), Gong et al. (2015) trained a network for synthetic aperture radar (SAR) image, which could suppress speckle noise effectively, and generated a difference image with good performance. Wang et al. (2018) proposed a general end-to-end 2D CNN framework to handle the high dimension problem and explore abundant information in hyperspectral images. In their work, hybrid unmixing spectral information and abundant information were separately extracted and then combined to form mixed-affinity matrices. Since the image spectral affinity and the abundance affinity were distributed in the top-left and bottom-right corners of the mixed-affinity matrices, they utilized two different convolution kernels to explore the spectral features and abundant features. At last, fully connected layers were employed to fuse the features and predict the change maps. Liu et al. (2019) developed two approaches based on FCNs for multitemporal change detection. In their methods, the task of change detection was performed as a post-classification comparison, which allows multiple transitions of multitemporal data. Yang et al. (2020) presented an asymmetric siamese network to locate and identify semantic changes through feature pairs obtained from modules of widely different structures. Peng et al. (2019) proposed an unsupervised approach for ORSICD based on saliency analysis and deep feature representation. Due to the



(a) Interior structure of a CLB

(b) General version of a CLB

Fig. 1. Structure of the Designed CLB.

superiority in sequential data processing, recurrent neural network (RNN) has been considered for multitemporal change detection tasks. For example, Lyu et al. (2016) proposed an end-to-end RNN for multispectral/hyperspectral change detection tasks, where a long short-term memory (LSTM)-based RNN was employed. Generative adversarial network (GAN) (Goodfellow, 2016) is also applied for change detection, for example, Niu et al. (2018) proposed using conditional generative adversarial network (cGAN) for change detection between optical and SAR images. In addition, the superiority of combining different networks has been exploited as well. Wiratama et al. (2018) proposed a dual-dense convolutional network (DCN), which jointed two deep convolution networks and improved the change detection accuracy by optimizing a contrastive loss. Mou et al. (2018) proposed a joint CNN and RNN framework (ReCNN) to learn joint spectral-spatial-temporal features for multitemporal change detection. Moreover, considering the absence of training data, some researchers investigated the problem of change detection labels acquisition. Gevaert et al. (2020) combined rule-based (expert knowledge) methods to obtain the training label for deep learning methods, which avoids the cost of manual labelling and can still obtain reasonably accurate results. Gong et al. (2019) proposed a generative discriminatory classified network (GDCN) for multispectral image change detection, in which the labeled data, unlabeled data, and new fake data generated by GAN were all used as training samples.

Due to the high relevance between pixel-wise change detection and dense pixel classification, fully convolutional networks (FCN) (Long et al., 2015) based methods have become one of the most developed branches of pixel-wise change detection. To deal with pixel-wise change detection in street view scenes, a change detection network (CDNet) was investigated with stacking contraction and expansion blocks (Alcantarilla et al., 2018). R.C. Daudt et al. (2018a, 2018b) proposed three FCN networks for change detection in satellite images, namely, fully convolutional-early fusion network (FC-EF), fully convolutional Siamese-concatenation network (FC-Siam-conc), and fully convolutional Siamese-difference network (FC-Siam-diff). The FC-EF stacked the image pairs and then fed the six-channel images into the network, in which the joint features were learned. The siamese network took two parallel encoding streams with a weight-sharing process for better weight reuse, which reduced the number of parameters. Therefore, using the Siamese network as a part of the feature extraction, the FC-EF was extended to the FC-Siam-conc and the FC-Siam-diff. The difference between FC-Siam-conc and FC-Siam-diff is in the network input. FC-Siam-conc took the element-add results as the network input while FC-Siam-diff took the difference of the image pairs. In addition, Peng et al. (2019) employed an effective deep supervision strategy on the UNet++ structure (Zhou et al., 2018) for remote sensing change detection, which was proven better than the several SOTA FCN-based methods by capturing the minimal changes.

The current methods of ORSICD mainly utilize either the multi-scale features or the multi-level context information to improve the change detection accuracy. The proposed CLNet aggregates the multi-level context information with the multi-scale features through two parallel but asymmetric branches. Experimental results demonstrate the effectiveness of CLNet on pixel-wise ORSICD.

3. Methodology

This section firstly describes the detailed settings of the designed CLB, and its general version that might be modified for other tasks is extended. Next, the whole network architecture is illustrated, which is modified from the typical UNet backbone, contains two CLBs and a dimension compression operation in the encoder part. At last, the loss function is introduced.

3.1. Cross-Layer Block (CLB)

3.1.1. Motivation and interior structure of CLBs

In deep works, multi-scale feature representation can enlarge receptive fields for better parsing the scenes, while multi-level feature representation can aggregate semantic context information and spatial details (Zhao et al., 2017). Therefore, it is a natural idea to incorporate them for improving the performance of dense pixel prediction tasks. From this point, the designed CLB tried to use two parallel but asymmetric branches to simultaneously extract the multi-scale and multi-level features and aggregate them for comprehensive feature representation.

Fig. 1(a) shows the interior structure of the CLBs. Suppose F_0 is the concatenated images or fused feature maps of a network, then the multi-scale feature maps F_1^0 and F_2^0 are extracted through the opposite branches. Both branches of F_0 are performed with two conv units (equipped with the sequence 3×3Conv-ReLU-BN) and a max-pooling with stride 2. The only difference is the stride of the first conv unit of the two branches, of which the stride is 1 for the left-branch and 2 for the opposite. After that, F_1^0 repeats the operation of F_0 to extract higher-level multi-scale features F_2^1 and F_3^1. According to the above operations, the

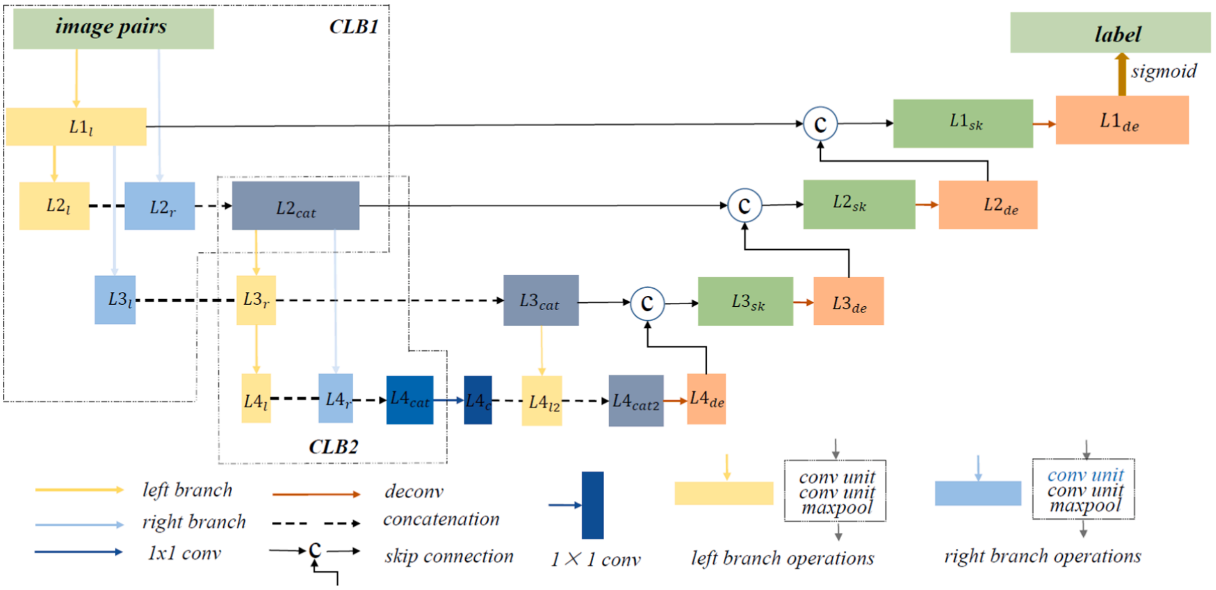


Fig. 2. Architecture overview of the proposed CLNet. (The detailed model parameters are provided in Appendix A.)

size of F_1^0, F_2^0, F_2^1 and F_3^1 are 1/2, 1/4, 1/4 and 1/8 of F_0, respectively. F_2^0 and F_2^1 are with the same size and thus can be concatenated as F_{2_conc}. F_2^1 obtains higher-level context information while F_2^0 maintains larger receptive fields by bigger stride, thus F_{2_conc} achieves aggregation of the multi-scale features and multi-level context information. Besides, the highest-level feature map F_3^1 is an optional branch, which indicates that it can whether be concatenated with other feature maps or be removed if it is unnecessary. Since the concatenation operation is always with feature maps from different layers (F_2^0 and F_2^1in Fig. 1(a)), we named the designed block as Cross-Layer Block (CLB).

3.1.2. *General version of CLB*

Except using the CLB described in Fig. 1(a) for ORSICD in this paper, we further summarize the general version of CLB which may be helpful for other dense pixel prediction tasks. The general version of CLB requires the following constraints to be satisfied:

$$\begin{cases} F_2^1 : F_0 \overset{\Delta}{\rightarrow} F_1^0 \overset{\Delta}{\rightarrow} F_2^1 \\ F_2^0 : F_0 \overset{\nabla}{\rightarrow} F_2^0 \\ F_{2\_conc} : F_2^1 \oplus F_2^0, if\ \ Size\left(F_2^0\right) = Size\left(F_2^1\right) \\ F_3^1 : F_1^0 \overset{\nabla}{\rightarrow} F_3^1, optional \end{cases} \tag{1}$$

where Δ represents *Conv&ScaleOperationsI*; ∇ represents *Conv&ScaleOperationsII*; $\oplus$ represents concatenation.

In deep networks, the size of the output feature map (denoted as *I*) can be calculated with the size of the input feature map (denoted as *i*) by:

$$I = \left\lfloor \frac{i + 2 \times p - k}{s} \right\rfloor + 1 \tag{2}$$

where *k* is the size of the convolution kernel; *s* is the adopted stride; and *p* is the size of the padding operation.

According to Eq. (1), F_2^0 and F_2^1 must have same size to achieve feature aggregation. Suppose the size of F_0 as I_0, the scale factor of Δ and ∇ as $s_1$and s_2, then $s_1$and $s_2$must satisfy the following constraints with proper *p* and *k*:

$$\begin{cases} s_2 = \alpha s_1 (\alpha > 1); if\ \ s_1 = 1 \\ s_2 = {s_1}^2; if\ \ s_1 > 1 \end{cases} \tag{3}$$

Therefore, the outputs of CLB meet either of the following situations and thus, $F_2^0$and F_2^1 can be concatenated as F_{2_conc}:

(1) When $s_1 > 1$, the size of F_1^0, F_2^0, F_2^1 and F_3^1 are $\frac{1}{s_1}I_0$, $\frac{1}{s_1^2}I_0$, $\frac{1}{s_1^2}I_0$ and $\frac{1}{s_1^3}I_0$, respectively.
(2) When $s_1 = 1$, the size of F_1^0, F_2^0, F_2^1 and F_3^1 are $\frac{1}{\alpha}I_0$, $\frac{1}{\alpha^2}$,$\frac{1}{\alpha^2}$ and $\frac{1}{\alpha^3}$, respectively.

Notice that other convolution units and pooling operations, such as ResNet Block (He et al., 2016) and average pooing can also be implemented in CLB, but in our experiments we observe no significant accuracy improvement for ORSICD.

3.1.3. *Learning ability of CLB*

Overall, the designed *CLB* has following abilities:

i) The two asymmetric branches of F_0 enable multi-scale feature extraction, as well as the two asymmetric branches of F_1^0. The feature mapsF_2^1 and F_2^0 enable higher-level and lower-level context information representation, respectively.
ii) The concatenation operation of F_2^0 and F_2^1 further aggregates multi-level context information, which enables comprehensive feature representation. Moreover, the optional branch F_3^1 provides potential possibility of multi-level feature aggregation.
iii) The designed CLB directly concatenates the extracted feature maps that come from different branches but have the same size, which is different from most multi-scale blocks which up-sample features for aggregation. This operation concatenates feature maps at a smaller size and reduces the memory occupation.

By combining these learning abilities, the designed *CLB* can leverage both multi-scale features and multi-level context information. Experiments in **Ablation Studies** demonstrated that such architecture can exploit more image information and boost the accuracy of ORSICD.

3.2. *Cross-Layer network (CLNet) architecture*

The CLNet is proposed under the structure of UNet (Ronneberger et al.,2015), which contains two CLBs and a dimension compression operation in the encoder part. Fig. 2 displays the specific architecture of CLNet.

3.2.1. *Encoder part*

Stack of CLBs: As indicated in R.C. Daudt et al. (2018a, 2018b) and

Peng et al. (2019), the input image pairs are concatenated to a six-channel image as network input. Then a CLB with branch F_3^1 (named *CLB1* in Fig. 2) is implemented to extract multi-scale features and incorporate multi-level context information. In *CLB1*, $L2_l$ and $L2_r$ are concatenated to $L2_{cat}$, which aggregates image features for the first time. With feature maps $L2_{cat}$ as input, a CLB without branch F_3^1 (named *CLB2* in Fig. 2) is performed to further exploit image information, where the feature maps $L4_l$ and $L4_r$ are concatenated as $L4_{cat}$ to aggregate image features for the second time. In addition, the feature maps $L3_l$ and $L3_r$ that namely extracted by *CLB1* and *CLB2* are also concatenated. These operations achieve comprehensive information aggregation at different feature stages, which greatly increases the feature representation ability of CLNet and is demonstrated to be effective for ORSICD by the experiments.

Due to the GPU memory limitation, the training samples were usually cropped to small-size patches (i.e., 256 × 256 pixels). With using two CLBs described as Section 3.1.1, the deepest layers of CLNet already down-sample the feature maps' size to 1/16 of the input patches. In order to avoid excessive information loss, no extra operations are implemented to extract deeper features and thus, the branch F_3^1 of *CLB2* is removed. As a substitution, $L4_{l2}$ is added to the end of the encoder part.

Dimension compression: The network is performed with 24 channels to extract the preliminary features and the channels of higher-level feature maps are set as twice of the feature maps where they come from. As a result, the channels of $L1_l$, $L2_r$, $L2_l$ and $L3_l$ are 24, 24, 48 and 48. The channels of $L2_{cat}$ are 72. The channels of $L3_r$,$L4_l$ and $L4_r$ are 144, 288 and 144. The channels of $L3_{cat}$ and $L4_{cat}$ are 192 and 432. To enhance the representation ability, a convolution unit with kernel size 1 is added to the end of the encoder to compress $L4_{cat}$ to 144 channels, which can ease the side influence of abundant information introduced by overmuch high-level features.

3.2.2. Decoder part

The decoder part is similar to other UNet-based networks, which leverages conv-transpose operations (equipped with the sequence 3 × 3Deconv-ReLU-BN) to generate 2 × up-sampled feature maps. After that, the feature maps of the encoder and decoder parts at the same layer are concatenated by skip connections. Finally, a classifier, consisting of a 3 × 3 convolution and a sigmoid function, is used to generate the final change map.

3.3. Loss function

Binary cross-entropy loss is widely used in the binary classification tasks. As the pixel-wise change detection is treated as a binary classification task in this paper, we selected weighted binary cross-entropy loss as a part of our loss function, which is as follows:

$$E_{bce} = \frac{1}{N} \times \left[\alpha \sum_{y_n=1} y_n \times \log(p_n) + (1-\alpha) \sum_{y_n=0} (1-y_n) \times \log(1-p_n) \right] \quad (4)$$

where N is the number of pixels in an image patch; α is used to balance the changed and unchanged areas in the given dataset; y_n is the state of the n-th pixel with $y_{n=1}$ representing the changed and $y_{n=0}$ representing the unchanged.p_n is the possibility of change.

Meanwhile, the severe proportion imbalance of the changed/unchanged area in the remote sensing images need to be addressed. Therefore, the dice coefficient loss E_{dc} is selected as the other part of the final loss to eliminate the negative effect of the data imbalance to some extent and to improve the classification performance. The dice coefficient loss is defined as:

$$E_{dc} = 1 - \frac{2Y\widehat{Y}}{Y + \widehat{Y}} \quad (5)$$

where Y is the ground-truth change map and $\widehat{Y}$ indicates the predicted change map.

The overall loss function is the combination of E_{bce} and E_{dc}:

$$E = E_{bce} + \lambda E_{dc} \quad (6)$$

where λ is used to balance E_{bce} and E_{dc}.

4. Experiments

This section presents the comprehensive experiments on three public datasets to evaluate CLNet's performance of dealing with two types pixel-wise ORSICD tasks. Experiments on a public VHR remote sensing dataset (named VHR dataset (Lebedev et al., 2018)) was conducted to observe its performance on all-objects change detection. Experiments on LEVIR-CD dataset (Chen and Shi, 2020) and WHU Building dataset (Ji et al., 2019) were conducted to observe its performance of detecting small-and-dense building change and large-and-sparse building change, respectively. As a result, all experiments demonstrate the superior accuracy performance and competitive efficiency of the proposed CLNet, as well as the robustness.

4.1. Data description, implementation details, comparison methods and evaluation indicators

4.1.1. Data description

The VHR dataset collected eleven VHR remote sensing image pairs from Google Earth, which contains the all-object change in each image pairs. The seasonal radiometric differences and varied resolutions make the VHR dataset a challenging dataset for all-object change detection. In this dataset, the objects change caused by seasonal radiometric differences was not considered as change (e.g., trees in different seasons), while the appearance/disappearance of cars was regarded as change.

Both the LEVIR-CD dataset and the WHU Building dataset are served for building change detection, where the former focuses on small-and-dense building change while the latter focuses on large-and-sparse building change. The LEVIR-CD dataset contains 637 image pairs with 0.5 m resolution whose acquisition dates varied from 2002 to 2018. These images were collected from Google Earth in Austin, Lakeway, Bee Cave, and other cities of Texas, US. The WHU Building dataset covers an area reconstructed after a 6.3-magnitude earthquake. The collected image pairs were taken in 2012 and 2016 and whose resolution is 1.6 m. In this dataset, the appearance/disappearance of cars is neglected. Detailed data description and some data samples can be found in Appendix B.

4.1.2. Implementation details

The network was implemented using Keras with Tensorflow as the backend. The network parameters were initialized with the initializer proposed in He et al. (2016). Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) was selected as the optimizer, and sigmoid was used as the activation function of the last fully connected layer for binary classification. In all experiments, α in the weighted binary cross-entropy loss function (E_{bce}) was set as 0.5 by default; and λ in the overall loss function (E) was set as 0.5. Notice that other values can also be used for α and λ, but we observe no significant improvement with different settings in our experiments. The end-to-end training was implemented under Ubuntu16.04 on a workstation with two Inter Xeon(R) E5-2620 v4 cores at 2.10 GHz and 32 GB RAM memory and a single NVIDIA Titan X (Pascal) GPU with 12 GB GPU memory.

4.1.3. Comparison methods

Several SOTA FCN-based methods were selected for comparison, including UNet (Ronneberger et al.,2015), DeepLabv3 (Liang-Chieh Chen et al., 2017), CDNet (Alcantarilla et al.,2018), FC-EF (R.C. Daudt et al., 2018a, 2018b), FC-Siam-conc (R.C. Daudt et al., 2018a, 2018b),

Table 1
The quantitative comparison on the **augmented** VHR dataset. (The best performance is emphasized in bold.)

Methods	Precision (%)	Recall (%)	F1 Score	OA (%)
UNet (Ronneberger et al., 2015)	*84.6*	*70.0*	*0.765*	*94.7*
DeepLabv3 (Liang-Chieh Chen et al., 2017)	*86.7*	*76.0*	*0.810*	*95.6*
CDNet (Alcantarilla et al., 2018)	87.4	73.3	0.792	95.4
FC-EF (R.C. Daudt et al., 2018a, 2018b)	79.3	62.1	0.697	93.3
FC-Siam-conc (R.C. Daudt et al., 2018a, 2018b)	89.2	87.3	0.882	97.1
FC-Siam-diff (R.C. Daudt et al., 2018a, 2018b)	89.3	84.8	0.870	96.9
FCN-PP (Lei et al., 2019)	87.9	69.2	0.775	95.0
UNet + ASPP	86.9	80.3	0.833	96.0
Peng et al. (Peng et al., 2019)	87.6	85.9	0.868	96.7
CLNet	**94.7**	**89.7**	**0.921**	**98.1**

FC-Siam-diff (R.C. Daudt et al., 2018a, 2018b), FCN-PP (Lei et al.,2019), UNet (Ronneberger et al.,2015) + ASPP (L.C. Chen et al., 2017), and Peng et al.(2019). Among these methods, UNet and DeepLabv3 are typical semantic segmentation network, and CDNet is a typical and basic FCN-based network. FC-EF, FC-Siam-conc and FC-Siam-diff are FCN-based Siamese networks with different network input. FCN-PP and UNet + ASPP utilize spatial pyramid pooling for multi-level feature extraction. The method of Peng et al. can be regarded as a kind of multi-scale and multi-level method, where the UNet++ network incorporates the multi-level context information and the multiple side out fused deep supervision (MSOF) (Xie and Tu, 2015) strategy aggregates the multi-scale outputs.

For fairy comparison, all the methods were reproduced in the same experiment environment and were trained from the scratch. The hyperparameters of the selected methods were strictly followed the descriptions in their original literature, except the setting of the batch size. The batch size of the compared methods was also set to 20 if there was enough GPU memory. Otherwise, it was set as large as possible under the GPU memory limitation.

4.1.4. *Evaluation indicators*

In order to evaluate the performance of CLNet, four indicators were selected for accuracy evaluation (namely, *overall accuracy, precision, recall,* and *F1 score*). The range of all four indicators is [0, 1], where higher values represent better model performance.

4.2. *All-objects change detection*

Experiments were conducted on the VHR dataset to evaluate CLNet's performance for all-objects change detection. In the experiments, the model was trained from the scratch for 15 epochs with an initial learning rate of 0.0001, and the batch size was set as 20. After 10 epochs, the learning rate was decreased by 10%.

Data augmentation was applied to the raw VHR dataset to avoid the potential overfitting problem caused by the lack of data. Specifically, the raw training and validation samples were augmented by shifting, scaling, and rotating with an angle of 90°, 180°, or 270°, and flipping horizontally or vertically. Thus, we had seven times the number of raw samples in the augmented VHR dataset. (The experiments conducted on the raw VHR Dataset are displayed in Appendix C. Even though CLNet outperformed the compared methods, its detection accuracy still lied at a low level.).

4.2.1. *Quantitative evaluation*

Table 1 lists the change detection results for the augmented VHR dataset with the augmented VHR dataset, CLNet outperformed all the

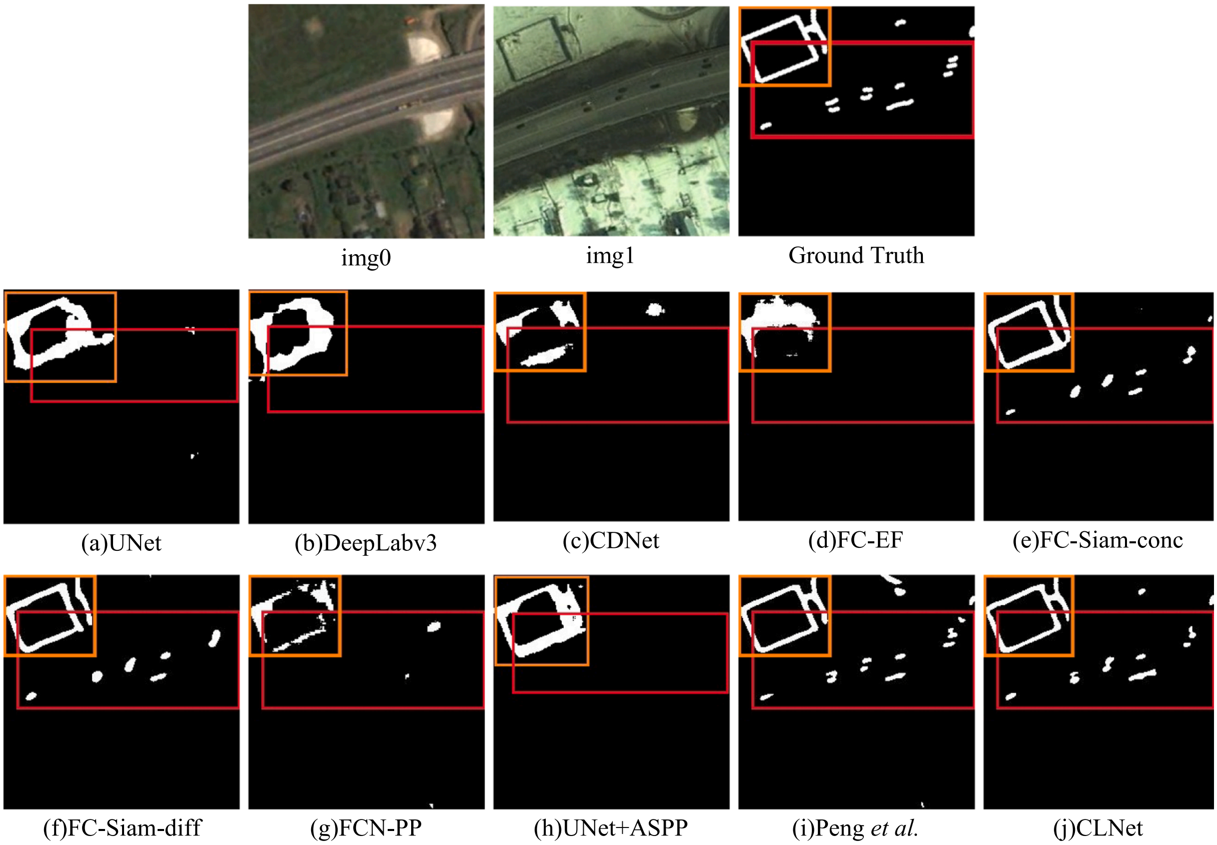

img0
img1
Ground Truth
(a)UNet
(b)DeepLabv3
(c)CDNet
(d)FC-EF
(e)FC-Siam-conc
(f)FC-Siam-diff
(g)FCN-PP
(h)UNet+ASPP
(i)Peng *et al.*
(j)CLNet

Fig. 3. Visual Comparison of generated change maps of small objects.

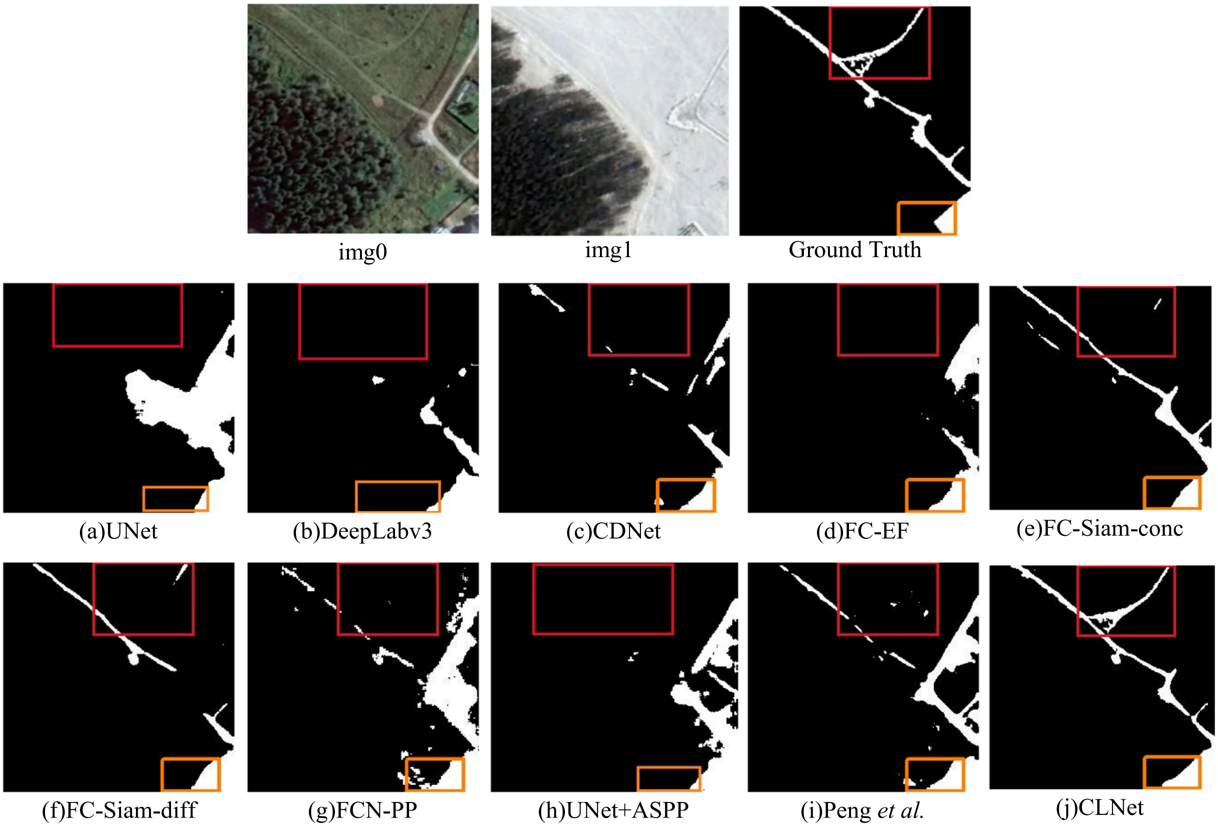


Fig.4. Visual Comparison of generated change maps of a thin road.

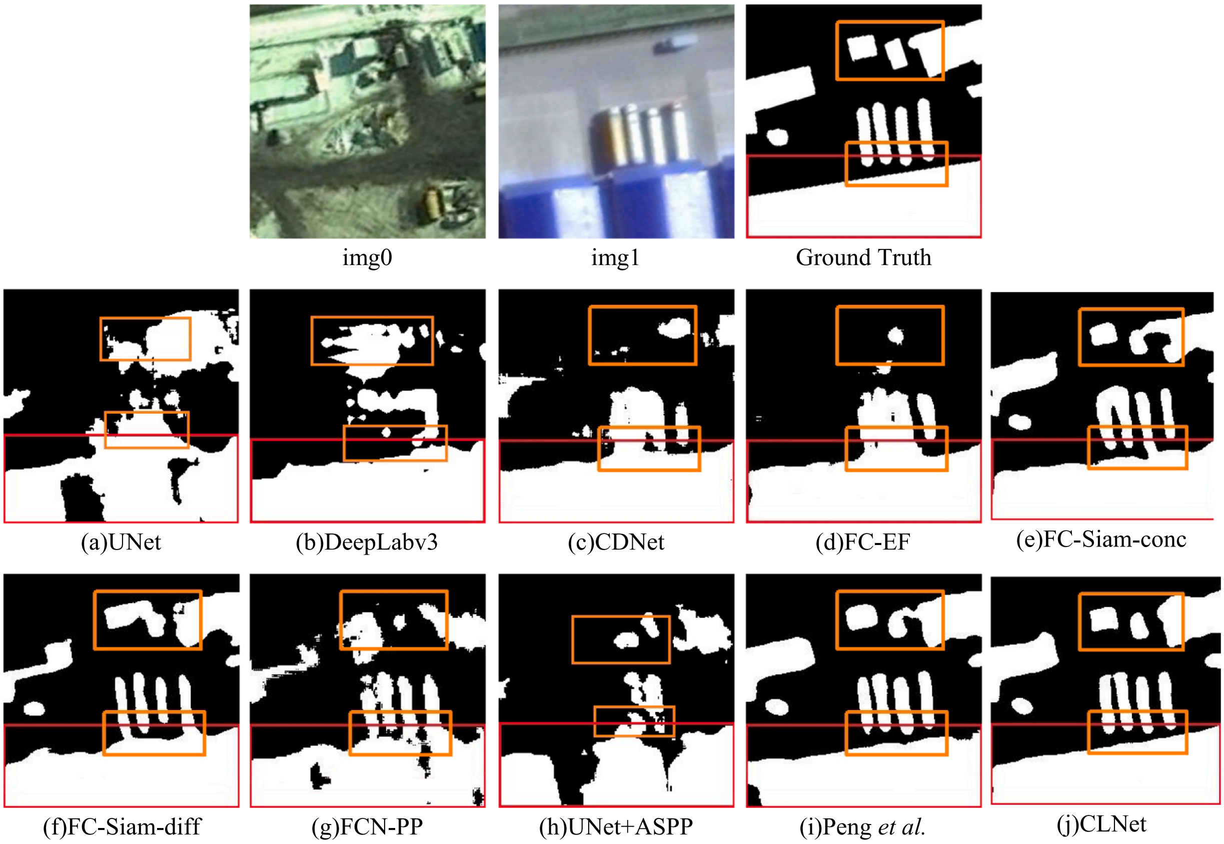


Fig. 5. Visual Comparison of generated change maps of large objects.

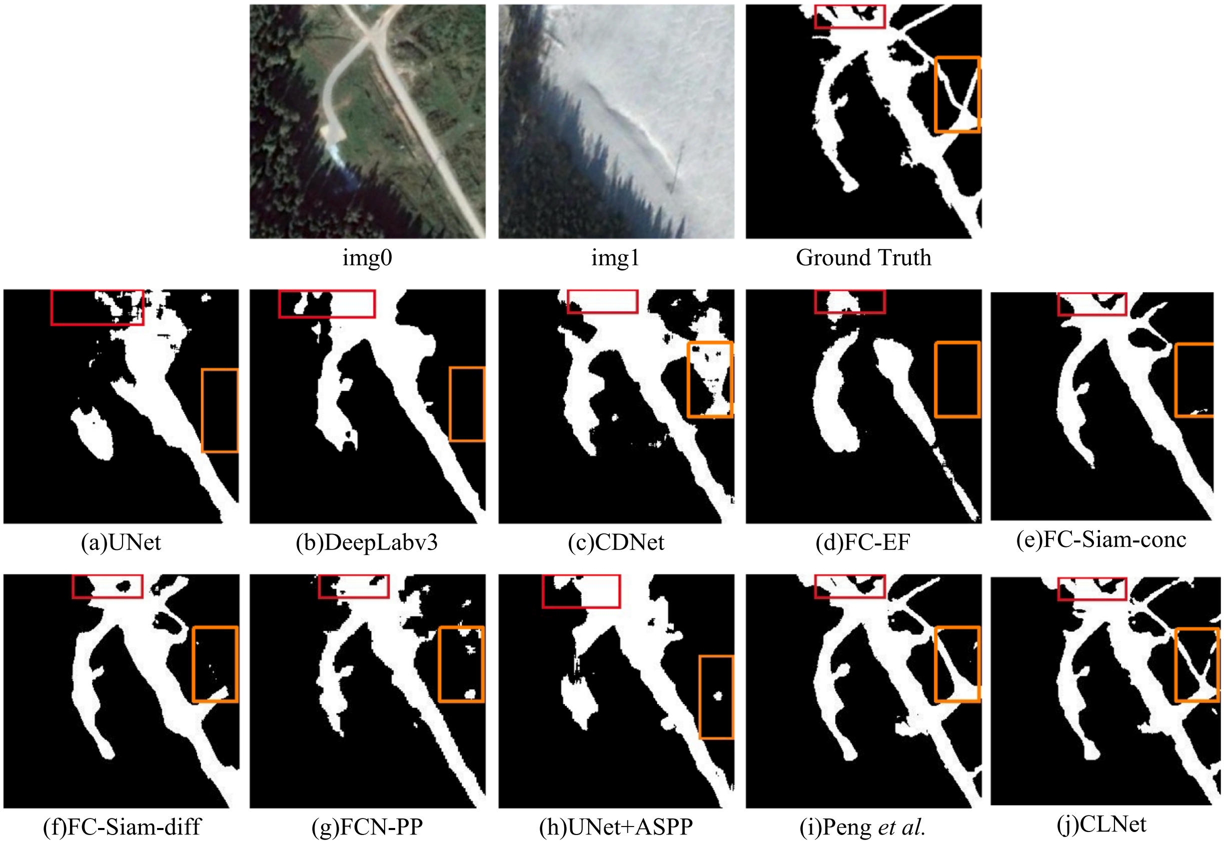


Fig. 6. Visual Comparison of generated change maps of complex scenes.

compared methods on all the indicators. CLNet increased the accuracy of the *F1 Score* by 0.039 and the *OA* by 1%, in comparison with FC-Siam-conc; by 0.051 of the *F1 Score* and 1.2% of the *OA* in comparison with FC-Siam-diff; and even more compared to the other methods. The above quantitative results demonstrated the superior performance of the proposed CLNet for all-objects change detection.

4.2.2. *Qualitative evaluation*

The obtained change maps of the four typical scenes (the changes of small objects, thin objects, large objects, and complex scenes) were selected to qualitatively evaluate the performance of CLNet and are displayed from Figs. 3–6.

Fig. 3 shows the change detection results of small objects and a thin closed wall. The proposed CLNet and Peng et al. (2019) detected almost all the small object change (see the red box in each change map), while the others missed some of them. In addition, the change maps obtained by CLNet and Peng et al. (2019) in the orange boxes are almost identical to the ground truth change map.

Fig. 4 shows the change detection results of thin roads. The proposed CLNet detected almost all the detailed changes in the red box, and the change map it generated was the most similar to the ground truth, except for the small piece of false alarms in the orange box. All the other methods obtained poor change detection results for thin roads.

Fig. 5 shows the change detection results of large objects (e.g., buildings). The proposed CLNet and Peng et al. (2019) obtained better change maps compared to those obtained by the other methods. CLNet reserved a sharper boundary and did not produce an inexplicable hole inside the changed area (compared to the others as shown in the red boxes in Fig. 5). In addition, CLNet effectively distinguished the unchanged area between the changed areas (see the orange boxes in Fig. 5).

Fig. 6 shows the change maps of a complex scene, where the proposed CLNet once again exhibited the optimal visual performance, especially in the red and orange boxes.

Figs. 3–6 verified the robustness of the proposed CLNet as well. The seasonal radiometric differences, all-objects change and different resolutions make VHR dataset a complex dataset. Even so, the proposed CLNet performed well and always generated better change maps compared to the other methods. However, none of the other methods obtained good performance in all the selected situations. For example, Peng et al. (2019) did not detect the complete and consistent roads in the thin-road scene (see Fig. 4(i)).

4.3. *Building change detection*

To evaluate the performance of CLNet on building change detection, experiments were conducted on LEVIR-CD dataset and WHU Building dataset. The former dataset was to verify the performance of detecting small-and-dense building change, and the latter one aimed at detecting large-and-sparse building change. Sections 4.3.1 and 4.3.2 displayed part of experimental results on the two datasets and illustrated the performance of CLNet. (More experimental results were displayed in Appendix D.)

4.3.1. *Experiments on LEVIR-CD dataset*

On this dataset, the model was trained from the scratch for 20 epochs with an initial learning rate of 0.001, and the batch size was set as 12. After 10 epochs, the learning rate was decreased by 10% each 5 epochs. It needs to be mentioned that we implemented all the compared methods with settings described in their original literatures, while the accuracy performance was always in a relatively low level (see Appendix D). For a fair comparison, the loss functions of all the compared methods were modified to the one used in CLNet to eliminate the performance difference caused by the loss function. The quantitative assessment of

Table 2
Quantitative performance comparison on the LEVIR-CD dataset. (The best performance is emphasized in bold.)

Methods	Precision (%)	Recall (%)	F1 Score	OA (%)
modified_UNet	84.6	85.2	0.849	98.4
modified_DeepLabv3	84.4	86.0	0.851	98.5
modified_CDNet	82.9	88.9	0.858	98.3
modified_FC-EF	81.0	88.4	0.845	98.2
modified_FC-Siam-conc	89.1	84.4	0.867	98.5
modified_FC-Siam-diff	84.7	89.7	0.871	98.5
modified_FCN-PP	89.1	84.0	0.866	98.4
modified_UNet + ASPP	85.3	85.4	0.853	98.5
modified_Peng et al.	86.7	86.9	0.868	98.5
CLNet	**89.8**	**90.3**	**0.900**	**98.9**

the modified methods is listed in Table 2 and several selected samples for qualitative comparison are displayed in Figs. 7 and 8.

Quantitative evaluation: Compared to the results of the modified methods, the proposed CLNet still achieved the best performance with the highest *precision* (89.8%), *recall* (90.3%), *F1 Score* (0.900) and *OA* (98.9%) on the LEVIR-CD dataset, which verified its superior ability of detecting small change and further reflected its robustness.

Qualitative evaluation: Two groups of change maps were selected from the test set of the LEVIR-CD dataset for qualitatively evaluating the performance of CLNet. As shown from Figs. 7 and 8, CLNet achieves better visual results compared to the other methods with less false detection and less misdetection (see the orange boxes). In addition, the changed areas detected by CLNet are with sharp boundaries, while the other methods (i.e., FC-Siam-conc, FC-Siam-diff, Peng et al.) tend to misclassify more unchanged pixels as changed pixels and thus obtained inaccurate building boundaries (see the red boxes).

4.3.2. *Experiments on WHU building dataset*

As for the WHU Building dataset, the buildings in this dataset are much larger than the ones in LEVIR-CD dataset and the specific structure of each changed building is also enlarged because of the higher image resolution. On this dataset, the model was trained from the scratch for 40 epochs with an initial learning rate of 0.0001 and the batch size was set as 20. The learning rate was decreased by 10% each 5 epochs. The quantitative assessment on this dataset is listed in Table 3 and several selected samples for qualitative comparison are displayed in Figs. 9 and 10.

Quantitative evaluation: Table 3 lists the change detection results for the WHU Building dataset. It can be seen that the proposed CLNet still achieved the best performance with the highest *precision* (96.9%), *recall* (95.7%), *F1 Score* (0.963) and *OA* (99.7%). It needs to be noted that the *recall* of CLNet is much higher than the compared methods (1.6% compared to the result of Peng et al. (2019), which is the second-best result), which indicated that CLNet detected more changed areas when little change occurred.

Qualitative evaluation: Two groups of the obtained change maps were selected from the test set of the WHU Building dataset to qualitatively evaluate the performance of CLNet. As shown from Figs. 9 and 10, CLNet achieves better change detection results compared to the other methods by generating the most similar change maps to the ground truth. Fig. 9 showed the CLNet's ability of obtaining sharper building boundaries (Fig. 9) and finding real change (i.e., neglecting the unconcerned cars' appearance/ disappearance in the scene). Fig. 10 showed that CLNet could greatly preserve the actual shape of changed objects and generated more accurate change maps.

4.4. *Accuracy/efficiency trade-offs*

We first evaluated the model efficiency on the VHR dataset according

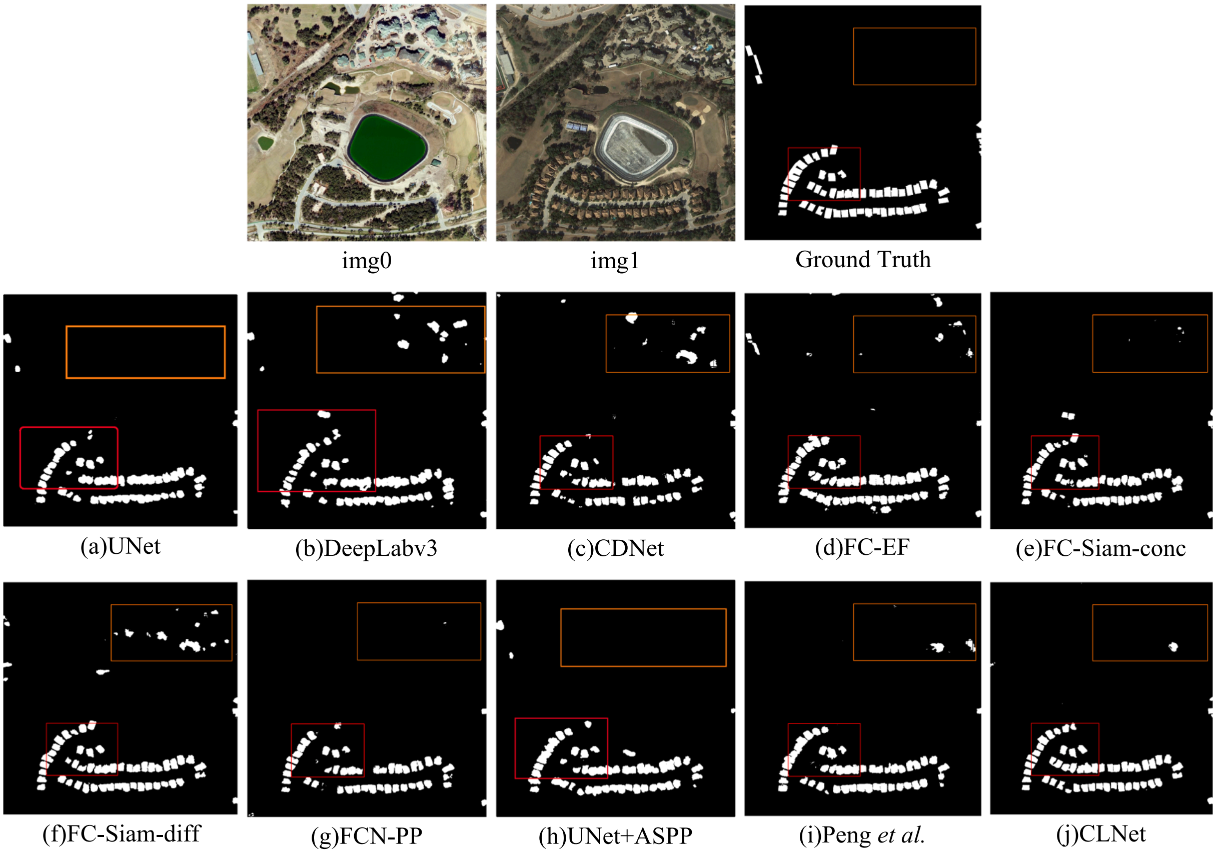


Fig. 7. Visual Comparison of generated change maps in the LEVIR-CD dataset (image index: test_19).

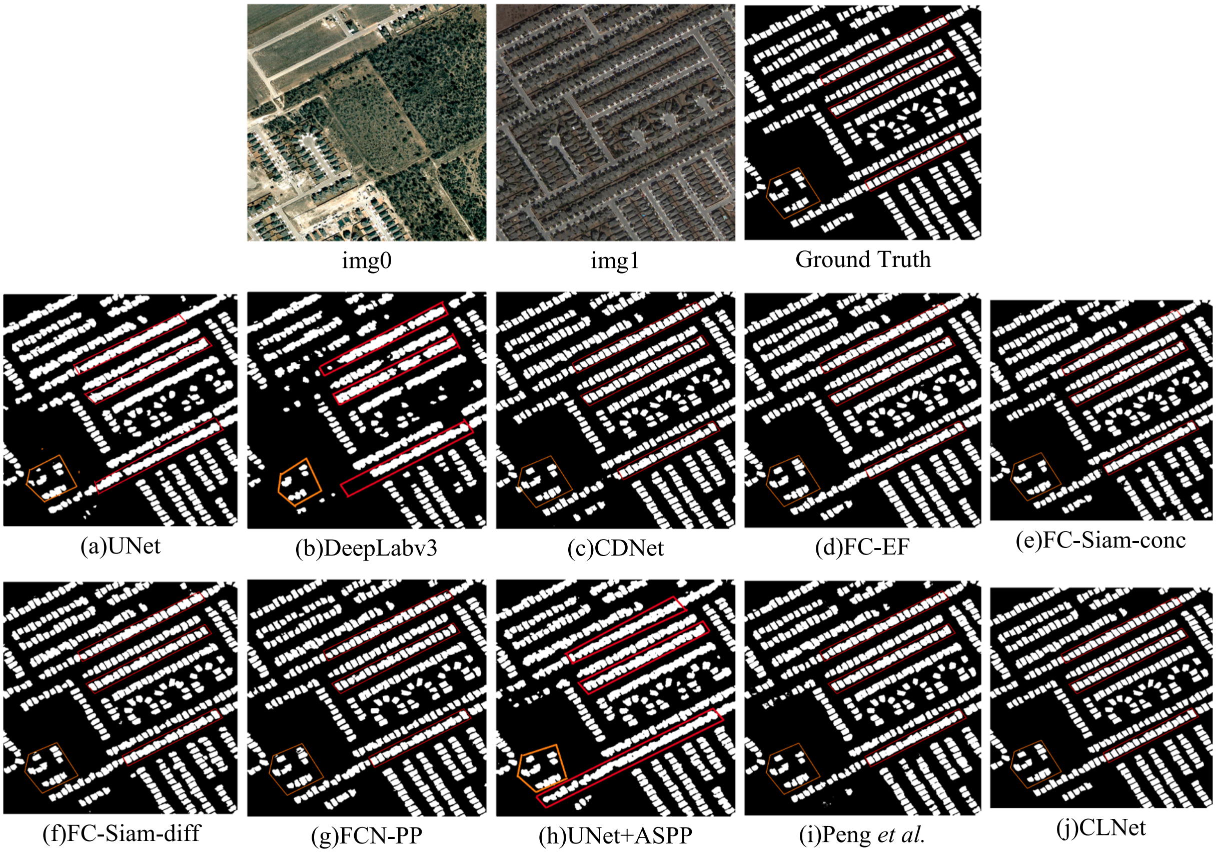


Fig. 8. Visual Comparison of generated change maps in the LEVIR-CD dataset (image index: test_45).

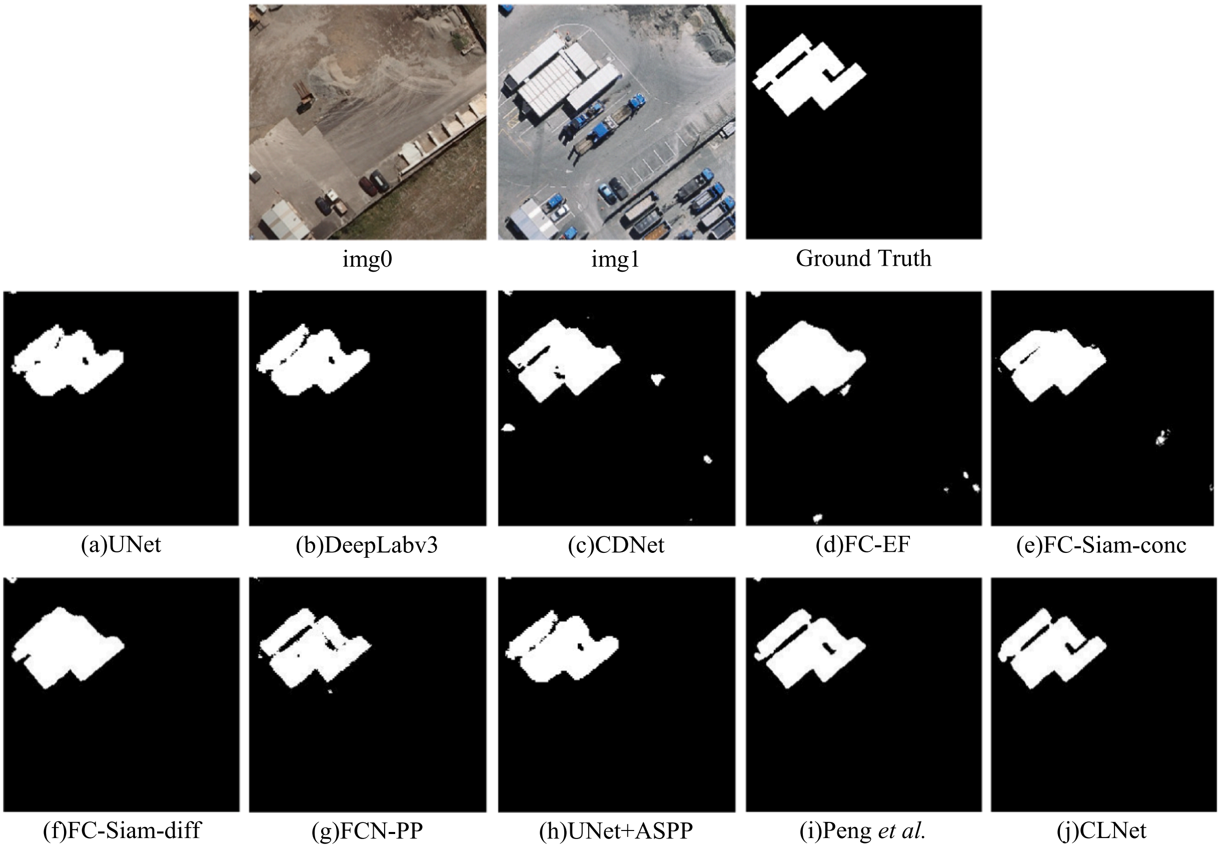
Table 3
Quantitative performance comparison on the WHU building dataset. (The best performance is emphasized in bold.)

Methods	Precision (%)	Recal l(%)	F1 Score	OA (%)
modified_UNet	84.1	84.4	0.842	98.9
modified_DeepLabv3	85.1	85.7	0.854	99.0
CDNet (Alcantarilla et al., 2018)	82.8	85.8	0.843	98.9
FC-EF (R.C. Daudt et al., 2018a, 2018b)	77.6	88.3	0.826	98.7
FC-Siam-conc (R.C. Daudt et al., 2018a, 2018b)	79.0	90.9	0.845	98.9
FC-Siam-diff (R.C. Daudt et al., 2018a, 2018b)	78.2	87.6	0.826	98.9
FCN-PP (T. Lei et al.,2019)	93.9	88.6	0.909	99.4
UNet + ASPP	94.3	84.0	0.889	99.3
Peng et al. (2019)	96.0	94.1	0.951	99.6
CLNet	**96.9**	**95.7**	**0.963**	**99.7**

to the time complexity and space complexity. In this paper, time complexity was represented by the average time cost for each epoch of the training procedure as well as the entire prediction time cost; and the space complexity was represented by the number of parameters. Table 4 lists the time consumption and model parameters for all methods. Table 4 lists the time consumption and model parameters for all methods. According to the model parameters, the compared methods were divided into two groups: 1) lightweight models (UNet, DeepLabv3, CDNet, FC-EF, FC-Siam-conc,FC-Siam-diff and UNet + ASPP) and 2) heavyweight models (FCN-PP, Peng et al. (2019) and the proposed CLNet).

For ease of illustration, the indicator ’time/parameters’ was calculated in Table 4 to represent the model efficiency, where lower values represent better trade-offs between time complexity and space complexity. In addition, *F1 Score* and *OA* were selected to reflect the model accuracy, since they have a higher tolerance to data imbalance and can better illustrate the comprehensive performance of a model. As shown in Table 4, the lightweight models had lower time complexity, but they barely achieved enough detection accuracy, which indicated that their feature representation abilities were insufficient in complex remote sensing scenes. Among the lightweight models, FC-Siam-conc achieved the optimal accuracy. Compared with FC-Siam-conc, Peng et al. (2019) achieved similar accuracy but had a high memory cost (5.5 ×) and time cost (4.2 ×). FCN-PP reached the optimal trade-offs, but its performance was not as good as FC-Siam-conc. The proposed CLNet reached an acceptable trade-offs, in that the number of parameters was 4.9 × of FC-Siam-conc while its time consumption was only 1.2 × of FC-Siam-conc. In addition, the proposed CLNet only cost about 450 s to generate the change maps for the entire test set (about 0.15 s on average for each 256 × 256 change map), which was acceptable for most change detection tasks. In conclusion, CLNet’s efficiency was competitive compared with the several SOTA methods.

Fig. 11 displays the trade-offs between accuracy and efficiency intuitively. The indicator ’time/parameters’ in Table 4 was selected as the x-axis, and the *F1 Score* and *OA* were selected as the y-axis (see Fig. 11(a) and (b)). A lower time/parameters ratio and a higher accuracy

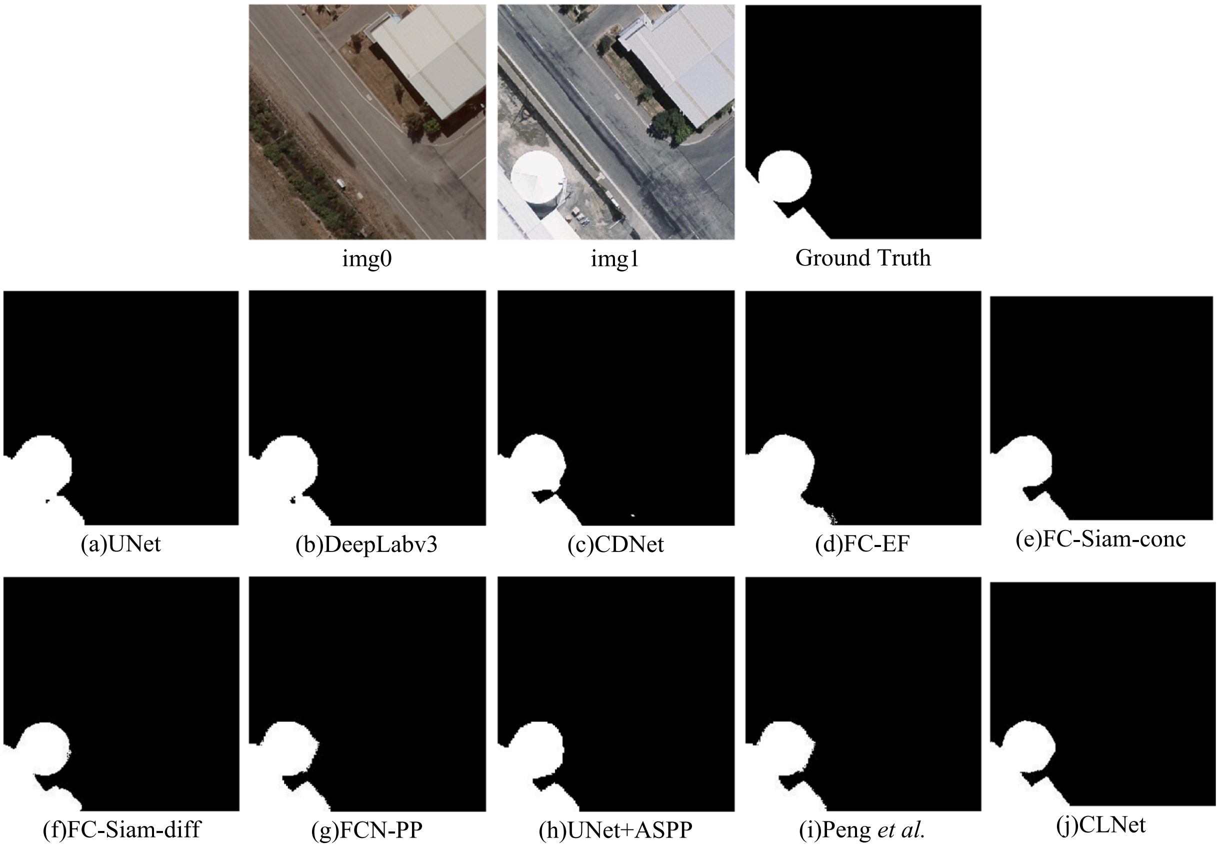


img0 img1 Ground Truth

(a)UNet (b)DeepLabv3 (c)CDNet (d)FC-EF (e)FC-Siam-conc

(f)FC-Siam-diff (g)FCN-PP (h)UNet+ASPP (i)Peng *et al.* (j)CLNet

Fig. 9. Visual Comparison of generated change maps in the WHU Building dataset.



img0 img1 Ground Truth

(a)UNet (b)DeepLabv3 (c)CDNet (d)FC-EF (e)FC-Siam-conc

(f)FC-Siam-diff (g)FCN-PP (h)UNet+ASPP (i)Peng *et al.* (j)CLNet

Fig. 10. Visual Comparison of generated change maps in the WHU Building dataset.

Table 4

Performance and speed trade-offs. (The best performance is emphasized in bold.)

	Train					Test
Methods	F1 Score	OA (%)	T/E	Para.	T/P ($\times 10^2$ s/M)	Test time (3000 images)
UNet (Ronneberger et al.,2015)	*0.765*	*94.7*	~*540 s*	~*0.63* M	*8.57*	~*190 s*
DeepLabv3 (Liang-Chieh Chen et al., 2017)	*0.810*	*95.6*	~*1315 s*	~*2.14* M	*6.14*	~*300 s*
CDNet (Alcantarilla et al.,2018)	0.792	95.4	~2250 s	**~1.24 M**	18.06	~1040 s
FC-EF (R.C. Daudt et al., 2018a, 2018b)	0.697	93.3	**~1100 s**	~1.44 M	7.64	~250 s
FC-Siam-conc (R.C. Daudt et al., 2018a, 2018b)	0.882	97.1	~1320 s	~1.63 M	8.10	~290 s
FC-Siam-diff (R.C. Daudt et al., 2018a, 2018b)	0.870	96.9	~1270 s	~1.50 M	8.47	~290 s
FCN-PP (Lei et al.,2019)	0.775	95.0	~1450 s	~9.95 M	**1.46**	**~150 s**
UNet + ASPP	0.833	96.0	~650 s	~1.74 M	3.74	~210 s
Peng et al. (2019)	0.868	96.7	~5570 s	~9.00 M	6.19	~1530 s
CLNet	**0.921**	**98.1**	~1550 s	~8.00 M	1.94	~450 s

T/E* represents time/epoch; ***Para.* represents parameter of models; ***T/E*** represents time/parameters.

in Fig. 11 indicate better accuracy/efficiency trade-offs. Therefore, the lowest time/parameters ratio and the highest accuracy (the left-top corner in Fig. 11(a) and (b)) represent the optimal accuracy/ efficiency trade-offs while the highest time/parameters ratio and lowest accuracy (the right-bottom corner in Fig. 11(a) and (b)) represent the worst accuracy/efficiency trade-offs, respectively. As shown in Fig. 11, the proposed CLNet was the closest one to the left-top corner both in Fig. 11(a) and (b), while FCN-PP was the closest one to the left-bottom corner; the CDNet was the closest one to the right-bottom corner, and the other methods were near the central axis. Fig. 11 shows that the proposed CLNet had competitive accuracy/efficiency trade-offs compared to the SOTA methods.

4.5. Ablation study on VHR dataset

To verify the effectiveness of the proposed CLNet, we conducted experiments on VHR dataset with several settings including the usage of *CLB1, CLB2* and dimension compression operation (DC). Since the CLNet is derived from the UNet (Ronneberger et al., 2015), we selected UNet as the baseline method. Different network settings were added to the UNet to obtain the model variants, and the quantitative results were listed in Table 5. As shown in Table 5, the proposed CLB module shows a significant performance improvement for the change detection accuracy, especially on the indicator *recall* (i.e., see the comparison between *UNet* and *UNet + CLB1*). The results demonstrated that the designed CLB enhanced the scene understanding ability and could capture more change information in the scene. After combining the *CLB1* and the *CLB2*, the *precision* improved a little (i.e., 1.5% to *UNet + CLB2*), while the *recall* dropped (i.e., 2.1% to *UNet + CLB2*). This might be caused by too much high-level features, as CLB module introduced more high-level features compared to the single branch network. The proposed CLNet demonstrated the assumption. After the dimension compression operation was added to the *UNet + CLB1 + CLB2,* which reduced the channels of high-level features, the model's accuracy improved a lot and the best performance compared to the other settings were achieved.

4.6. Discussion

4.6.1. Difference between CLB and several multi-scale blocks

As shown in Fig. 12, Inception Block (Szegedy et al., 2015) enables multi-scale feature extraction with different-size convolution kernels. ASPP (L.C. Chen et al., 2017) consists of parallel dilated convolutions with different rates to obtain multi-scale features. Different from Inception Block and ASPP, PPM (Zhao et al., 2017) implements pyramid pooling to replace convolutions for multi-scale feature extractions. All these blocks enlarge the receptive fields with different strategies and then concatenate the feature maps at the size of input feature maps.

The proposed CLB simplifies the multi-scale feature extraction by using parallel branches with different strides, and it further incorporates multi-level context information with the two asymmetric branches. Moreover, CLB allows feature concatenation at a smaller size rather than up-samples features for concatenation, which naturally reduces memory consumption and improves efficiency. As a result, more image information is exploited and thus the ability of comprehensive feature representation is boosted. By stacking CLBs in networks, features extracted from different-scale and different-level can be gradually aggregated.

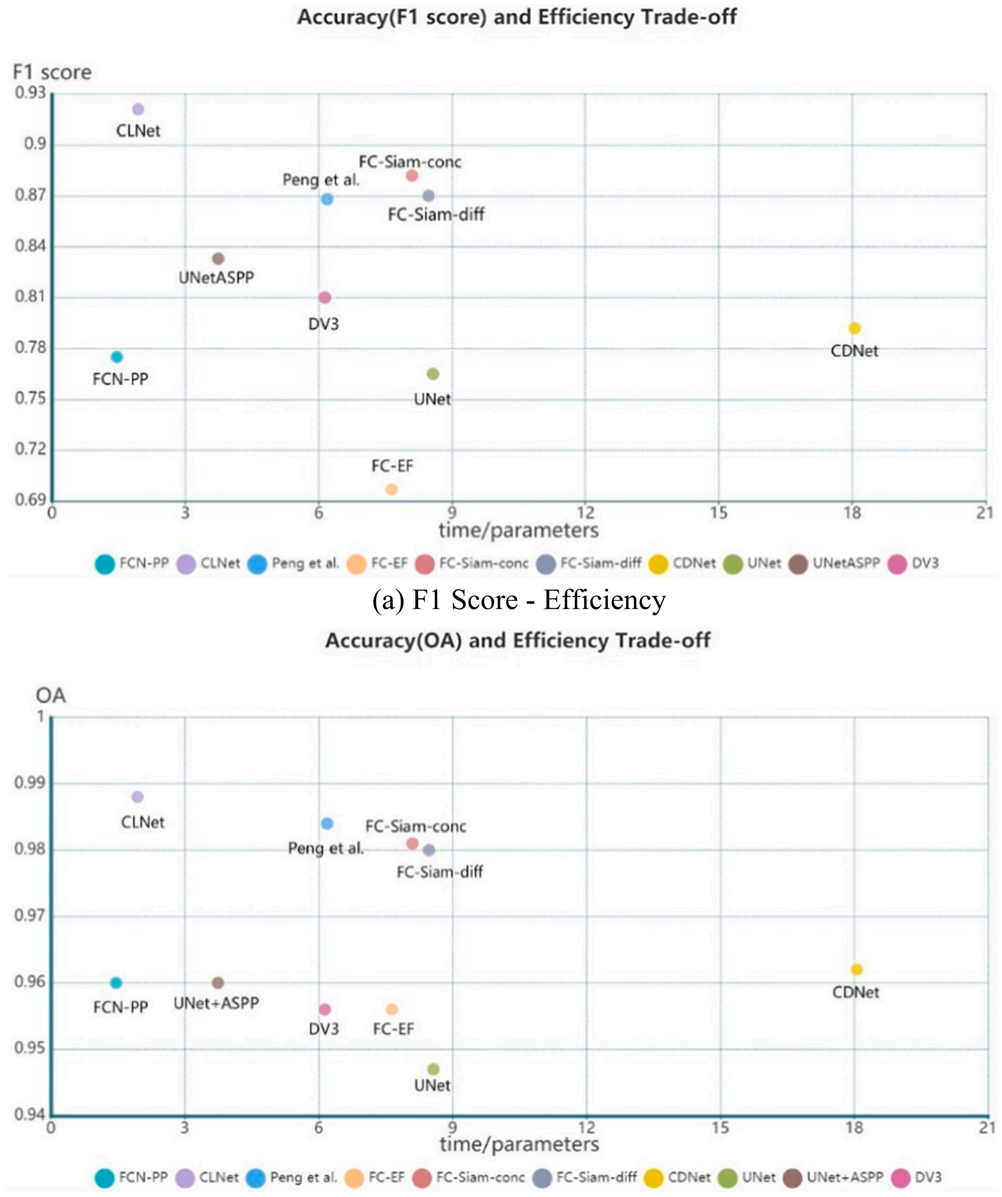
Experiments demonstrated the effectiveness of our work. Among the compared methods, UNet + ASPP is an implementation to test the performance of ASPP, and FCN-PP is performed with FCN and PPM. The experimental results show that the proposed CLNet always achieves higher accuracy and is more robust when it comes to deal with different change detection tasks.

4.6.2. Performance discussion

The ablation studies demonstrated that the designed CLBs are effective for ORSICD and can improve the detection accuracy. The quantitative results show that the proposed CLNet obtained higher accuracy compared to the several SOTA FCN-based methods, which demonstrated that the strategy of aggregating the multi-scale features and multi-level context information could exploit more image information and boost feature representation ability.

According to the qualitative experiments, the change maps of UNet and DeepLabv3 were not as good as other methods' in most selected scenes, which indicated that just transferring semantic segmentation networks into the change detection field might unable to achieve expected results. Peng et al. (2019) obtained the second most satisfactory results in most situations on the VHR dataset (see Fig. 4(i) and Fig. 5(i)) and all the situation on the WHU Building dataset, which indicated that the (MSOF) (Xie and Tu, 2015) strategy was also useful for accuracy improvement, while its performance was not as good as the other methods on the LEVIR-CD dataset. As for the proposed CLNet, all three datasets showed its ability of capturing accurate change. More precisely, experiments on the VHR dataset and LEVIR-CD dataset demonstrate that the proposed CLNet is sensitive to the change of small targets and can detect fine change, i.e., small and thin targets. Experiments on VHR dataset and WHU Building dataset indicate that it can also find the real changes under negative influences, such as the scale and season varying in VHR dataset and the appearance/disappearance of cars in WHU Building dataset.

The calculated indicator 'time/parameters' on the VHR dataset of the proposed CLNet was 1.94, which was very close to the indicator of FCN-PP (1.46) and much smaller than the other compared methods (8.57, 6.14, 18.06, 7.64, 8.1, 8.47, 3.74 and 6.19 for UNet, DeepLabv3, CDNet, FC-EF, FC-Siam-conc, FC-Siam-diff, UNet + ASPP and Peng et al. (2019), respectively). These results indicate that the proposed cross-layer strategy of aggregating multi-scale features and multi-level context



(a) F1 Score - Efficiency

(a) OA - Efficiency

Fig. 11. Trade-offs between Accuracy and Efficiency (the left-top corner indicates the best trade-offs).

Table 5
Evaluation of CLNet with different settings. We computed the quantitative accuracy of the model variants on the VHR dataset.

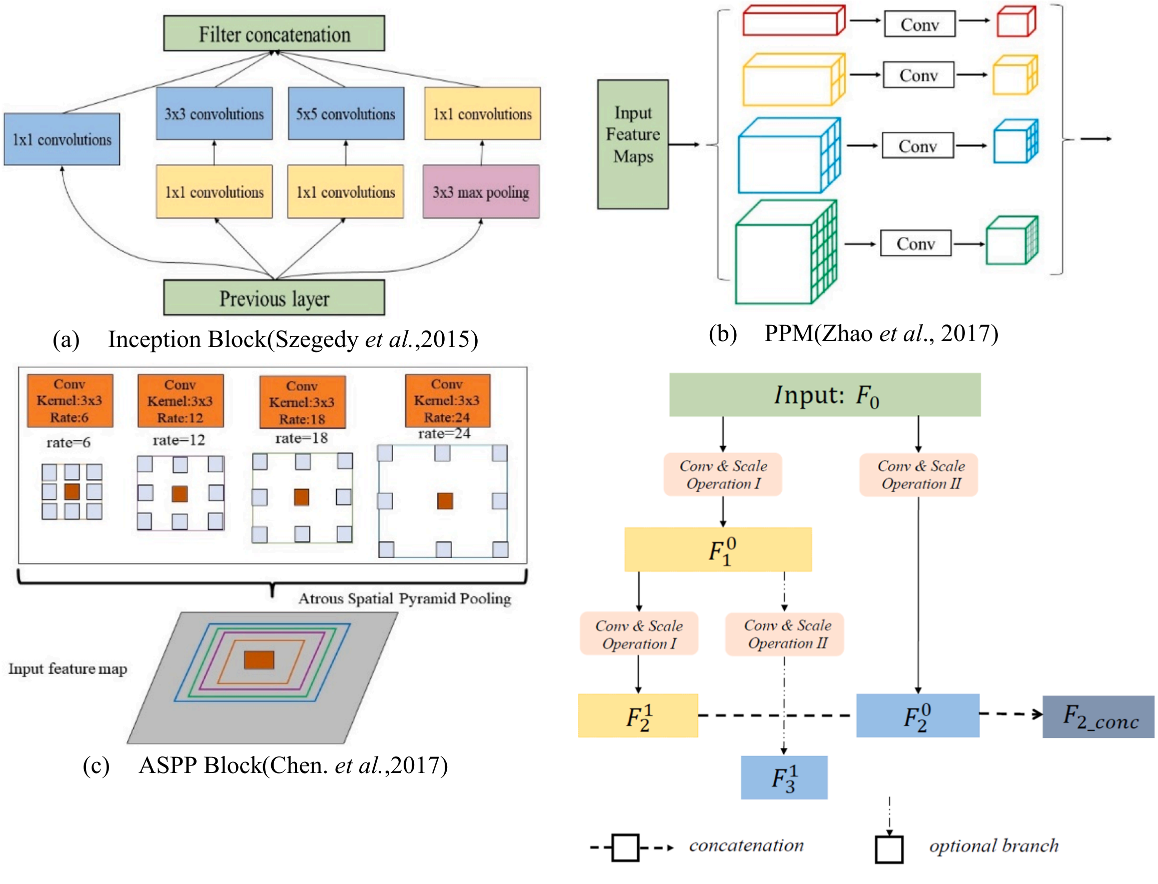
	Network Setting			Experimental Results on VHR Dataset			
	CLB1	CLB2	DC	Precision (%)	Recall (%)	F1-Score	OA (%)
UNet	×	×	×	84.6	70.0	0.765	94.7
UNet + CLB1	√	×	×	88.5	86.4	0.875	96.9
UNet + CLB2	×	√	×	89.8	87.0	0.884	97.2
UNet + CLB1 + CLB2	√	√	×	91.3	84.9	0.880	97.1
Proposed CLNet	√	√	√	**94.7**	**89.7**	**0.921**	**98.1**

(DC indicates dimension compression).

information could greatly reuse the extracted feature information and reduce the time cost.

In all the selected scenes on the three datasets, the proposed CLNet achieved better accuracy and visualization performance than the compared FCN-based methods, demonstrating that the proposed CLNet is a better choice since it could obtain better change detection results more efficiently.

Although good performance was achieved on the experimental dataset, the proposed CLNet was also found to have some limitations. The concatenation operations at each layer resulted in rapid increases in the feature channels and a large number of model parameters, which increased the GPU memory cost and hindered the model's efficiency.



(a) Inception Block(Szegedy *et al.*,2015)

(b) PPM(Zhao *et al.*, 2017)

(c) ASPP Block(Chen. *et al.*,2017)

(d) Our CLB

Fig. 12. A comparison about Inception Block, PPM, ASPP and our CLB.

5. Conclusion

In this paper, a new end-to-end convolution neural network called CLNet was proposed for bitemporal ORSICD, in which two novel CLBs were embedded. The designed CLB can effectively aggregate the multi-scale features and multi-level context information and is capable of reusing information with minimal extra memory requirements.

The experiments on a public VHR dataset and two building change detection datasets show that the proposed CLNet obtained higher accuracy, generated better change maps, and achieved competitive accuracy/efficiency trade-offs compared to several SOTA methods, which demonstrated the prominent performance and robustness of the proposed CLNet. Moreover, since the input image pairs are integrated together as the network input, it holds great natural potential for extension to the change detection tasks of multitemporal image sequences.

The designed CLB is a general module and thus can be extended with some basic/advanced blocks. Besides, the proposed CLNet in this paper still have some limitations (i.e., its large number of parameters, which increases the GPU memory requirements). Future works will focus on transferring the designed CLB into other remote sensing tasks and designing new lightweight ORSICD architectures.

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.

Acknowledgements

The authors are graceful to the authors of the three change detection datasets for providing the public remote sensing change detection data. This work was supported in part by the National Natural Science Foundation of China under Grant No. 42030102, 42001406, and 41801386, the Fund for Innovative Research Groups of the Hubei Natural Science Foundation under Grant No. 2020CFA003, and China Postdoctoral Science Foundation under Grant No. 2020M672416. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

Appendix A. The detailed network architecture

See Table 6.

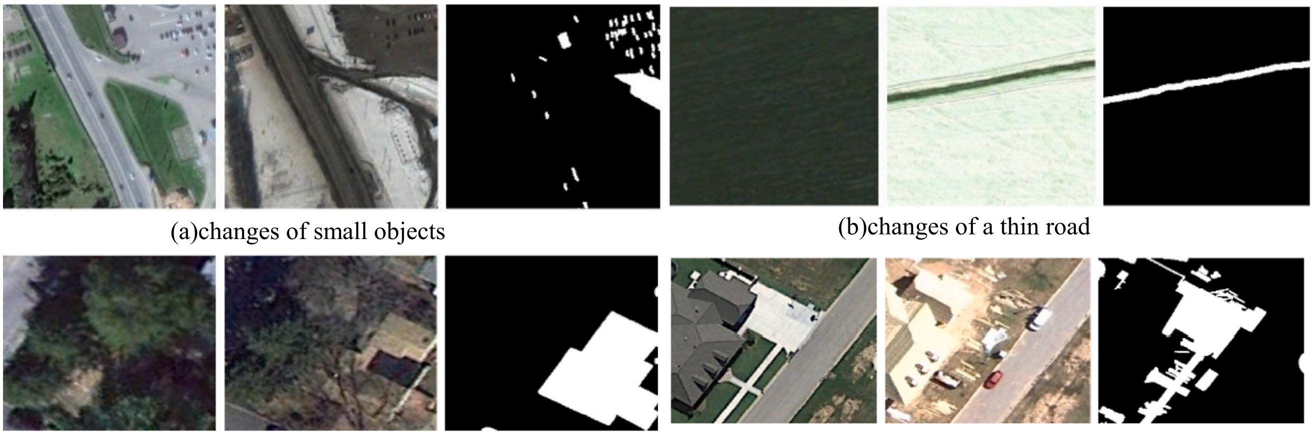
Table 6
Parameters of the Proposed CLNet Architecture. Construction of basic blocks in Fig. 2 are designated in brackets. The usage of ReLU and batch normalization follows Peng et al. (2019). The layer names correspond to the description in Fig. 2 in main manuscript. H and W denote the height and width of the input image.

Blocks	Input	Layer Settings	Output Dimension
/	*Input images*	*Concat*	H × W × 6
$L1_l$	*input*	$\begin{bmatrix}3\times3,24,s=1\\3\times3,24,s=1\\Maxpool,2\times2\end{bmatrix}$	$\frac{1}{2}H\times\frac{1}{2}W\times24$
$L2_l$	$L1_l$	$\begin{bmatrix}3\times3,48,s=1\\3\times3,48,s=1\\Maxpool,2\times2\end{bmatrix}$	$\frac{1}{4}H\times\frac{1}{4}W\times48$
$L2_r$	*input*	$\begin{bmatrix}3\times3,24,s=2\\3\times3,24,s=1\\Maxpool,2\times2\end{bmatrix}$	$\frac{1}{4}H\times\frac{1}{4}W\times24$
$L2_{cat}$	$L2_l,L2_r$	***concat***	$\frac{1}{4}H\times\frac{1}{4}W\times72$
$L3_l$	$L1_l$	$\begin{bmatrix}3\times3,48,s=2\\3\times3,48,s=1\\Maxpool,2\times2\end{bmatrix}$	$\frac{1}{8}H\times\frac{1}{8}W\times48$
$L3_r$	$L2_{cat}$	$\begin{bmatrix}3\times3,144,s=1\\3\times3,144,s=1\\Maxpool,2\times2\end{bmatrix}$	$\frac{1}{8}H\times\frac{1}{8}W\times144$
$L3_{cat}$	$L3_l,L3_r$	***concat***	$\frac{1}{8}H\times\frac{1}{8}W\times192$
$L4_l$	$L3_r$	$\begin{bmatrix}3\times3,288,s=1\\3\times3,288,s=1\\Maxpool,2\times2\end{bmatrix}$	$\frac{1}{16}H\times\frac{1}{16}W\times288$
$L4_r$	$L2_{cat}$	$\begin{bmatrix}3\times3,144,s=2\\3\times3,144,s=1\\Maxpool,2\times2\end{bmatrix}$	$\frac{1}{16}H\times\frac{1}{16}W\times144$
$L4_{cat}$	$L4_l,L4_r$	***concat***	$\frac{1}{16}H\times\frac{1}{16}W\times432$
$L4_c$	$L4_{cat}$	$\boldsymbol{conv,}1\times1\boldsymbol{,\ 144}$	$\frac{1}{16}H\times\frac{1}{16}W\times144$
$L4_{l2}$	$L3_{cat}$	$\begin{bmatrix}3\times3,384,s=1\\3\times3,384,s=1\\Maxpool,2\times2\end{bmatrix}$	$\frac{1}{16}H\times\frac{1}{16}W\times384$
$L4_{cat2}$	$L4_{conv},L4_{l2}$	***concat***	$\frac{1}{16}H\times\frac{1}{16}W\times528$
$L4_{de}$	$L4_{cat2}$	$\begin{bmatrix}3\times3,384,s=1\\3\times3,384,s=1\end{bmatrix}$	$\frac{1}{16}H\times\frac{1}{16}W\times384$
$L3_{sk}$	$L4_{de}$	$\boldsymbol{deconv,}3\times3\boldsymbol{,\ 192}$	$\frac{1}{8}H\times\frac{1}{8}W\times384$
	$L4_{de},L3_{cat}$	***skipconnection***	
$L3_{de}$	$L3_{sk}$	$\begin{bmatrix}3\times3,144,s=1\\3\times3,144,s=1\end{bmatrix}$	$\frac{1}{8}H\times\frac{1}{8}W\times144$
$L2_{sk}$	$L3_{de}$	$\boldsymbol{deconv,}3\times3\boldsymbol{,\ 72}$	$\frac{1}{4}H\times\frac{1}{4}W\times144$
	$L3_{de},L2_{cat}$	***skipconnection***	
$L2_{de}$	$L2_{sk}$	$\begin{bmatrix}3\times3,48,s=1\\3\times3,48,s=1\end{bmatrix}$	$\frac{1}{4}H\times\frac{1}{4}W\times48$
$L1_{sk}$	$L2_{de}$	$\boldsymbol{deconv,}3\times3\boldsymbol{,\ 24}$	$\frac{1}{2}H\times\frac{1}{2}W\times48$
	$L2_{de},L1_l$	***skipconnection***	
$L1_{de}$	$L1_{sk}$	$\begin{bmatrix}3\times3,24,s=1\\3\times3,24,s=1\end{bmatrix}$	$\frac{1}{2}H\times\frac{1}{2}W\times24$
		$\boldsymbol{deconv,}3\times3\boldsymbol{,\ 24}$	$H\times W\times24$
		$\boldsymbol{conv,}3\times3\boldsymbol{,\ 1}$	$H\times W\times1$
FC	$L1_{de}$	***sigmoid***	$H\times W$

Appendix B. Description of datasets

All-Object Change Detection: The VHR-Dataset is a real VHR remote sensing image change detection dataset from Lebedev et al. (2018), which contains all-objects change. The dataset contained 11 images pairs collected from Google Earth. The dataset publisher used seven images (4725 × 2700 pixels) to create manual ground truth and the other four images (1900 × 1000 pixels) to add additional objects manually. The dataset was randomly cropped to 16,000 256 × 256 patches by Lebedev et al. (2018), and each patch had at least one changed object to meet the input requirements of the deep learning methods. Among the patches, 10,000 were used as the training set, 3000 were used as thevalidation set and the other 3000 were used as the test set. The spatial resolutions of the dataset varied from 3 cm to 100 cm. In this dataset, the seasonal radiometric differences of the same objects (like trees or bare land) were not considered as change, while the appearance/disappearance of cars was regarded as change. Fig. 13illustrates several of the change-types of the VHR dataset, including the changes of small objects, thin objects, large objects, and complex scenes.

Building Change Detection: The LEVIR-CD dataset, which focuses on the small-and-dense buildings change, is a large and challenging building change detection dataset that contains 637 pairs of very-high-resolution (0.5 m/pixel) image patch with a size of 1024 × 1024 pixels (Chen and Shi, 2020). These bitemporal images were collected from Google Earth in Austin, Lakeway, Bee Cave, and other cities of Texas, US. The acquisition dates vary from 2002 to 2018. In this dataset, there are extra spectral differences caused by seasonal changes and illumination changes, which made it more challenging to distinguish real changes. The dataset was randomly split into three parts, where 70% samples for training (445 image pairs), 10% for validation (64 image pairs), and 20% for testing (128 image pairs) by Chen and Shi (2020). In addition, each sample in both the training and validation parts was split into 16



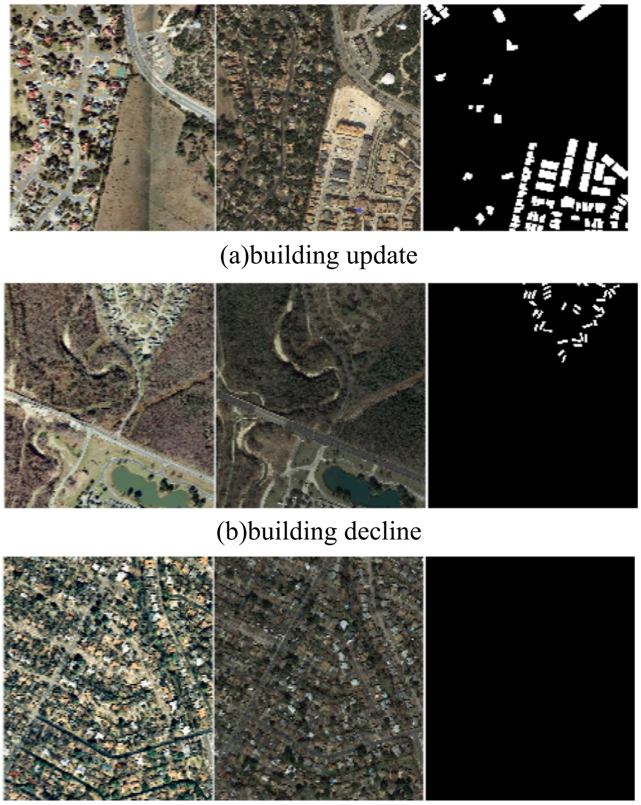
(a)changes of small objects

(b)changes of a thin road

(c)changes of a building

(d)changes of a complex scene

Fig. 13. Several samples of the change-types in the VHR dataset. Subfigures (a), (b), (c) and (d) display the bitemporal images and ground truth change maps of small objects (i.e., car), a thin road, large objects (i.e., building) and a complex scene, respectively.

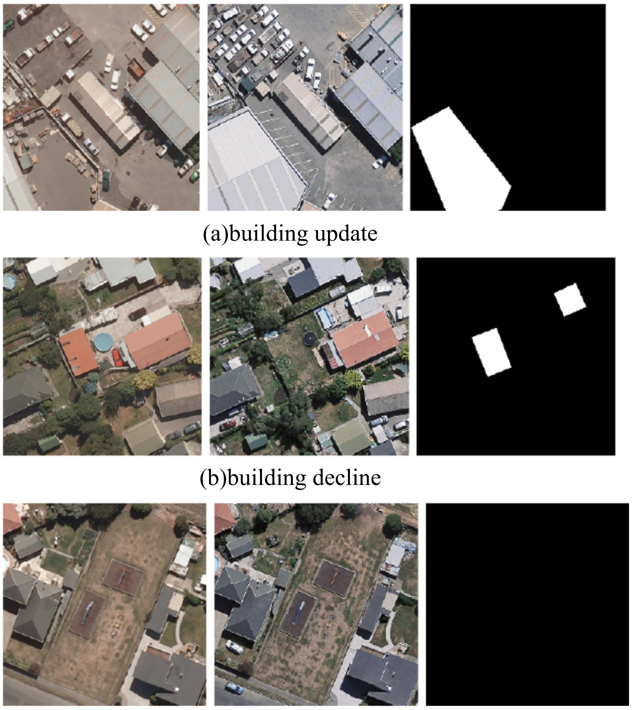


(a)building update

(b)building decline

(c)no change scene

Fig. 14. Several samples of the change-types in the LEVIR-CD dataset. Subfigures (a), (b) and (c) display the bitemporal images and ground truth change maps of building update, building decline and no change, respectively.



(a)building update

(b)building decline

(c)no change scene

Fig. 15. Several samples of the change-types in the WHU Building dataset. Subfigures (a), (b) and (c) display the bitemporal images and ground truth change maps of building update, building decline and no change, respectively.

small patches with a size of 256 × 256 pixels. WHU Building dataset, which is a hybrid dataset of aerial images and satellite images, is mainly designed for building extraction, but part of it can be extended for building change detection (Ji et al., 2019). The images used for change detection cover an area where a 6.3-magnitude earthquake occurred in February 2011 and then reconstructed in the following years. The images with 1.6 m-resolution were acquired in April 2012 (with 12,796 buildings in 20.5 km^2) and 2016 (with 16,077 buildings in the same area), respectively. This dataset focuses on the large-and-sparse buildings change, and the appearance/disappearance of cars is neglected. The dataset was randomly cropped into 2260 small patches with a size of 256 × 256 pixels, of which 1622 were used as the training set, 169 were used as the validation set and the rest 169 were used as the test set.

Figs. 14 and 15 illustrate the change of building growth, building decline and no change areas in LEVIR-CD dataset and WHU Building dataset, respectively.

Appendix C. Experiments on raw VHR dataset

Table 7 lists the change detection results for the raw VHR dataset in our experiments. The proposed CLNet outperformed the compared methods on all the indexes for the dataset. The CLNet results were acceptable but still unsatisfactory. Although the *precision* and *OA* reached 91.4% and 96.5%, the *recall* and *F1 Score* were only 79.5% and 0.851. The low *recall* indicated that a large amount of missed detection occurred (see Fig. 16).

Table 7
The quantitative comparison on the raw VHR dataset. (The best performance is emphasized in bold.)

Methods	Precision (%)	Recall (%)	F1 Score	OA (%)
UNet (O. Ronneberger *et al.*)	*79.3*	*63.4*	*0.693*	*93.2*
DeepLabv3 (Chen, Liang-Chieh, et al.)	*69.0*	*48.8*	*0.573*	*91.0*
CDNet (P.F. Alcantarilla et al.)	79.6	70.0	0.745	94.0
FC-EF (R.C. Daudt et al.)	72.6	36.6	0.486	90.5
FC-Siam-conc (R.C. Daudt et al.)	84.7	60.0	0.700	93.7
FC-Siam-diff (R.C. Daudt et al.)	84.7	35.5	0.500	91.2
FCN-PP (T. Lei et al.)	79.2	72.1	0.754	94.2
UNet + ASPP	*77.9*	*66.7*	*0.718*	*92.9*
Peng et al. (D. Peng et al.)	87.9	76.4	0.818	95.8
CLNet	**91.4**	**79.5**	**0.851**	**96.5**

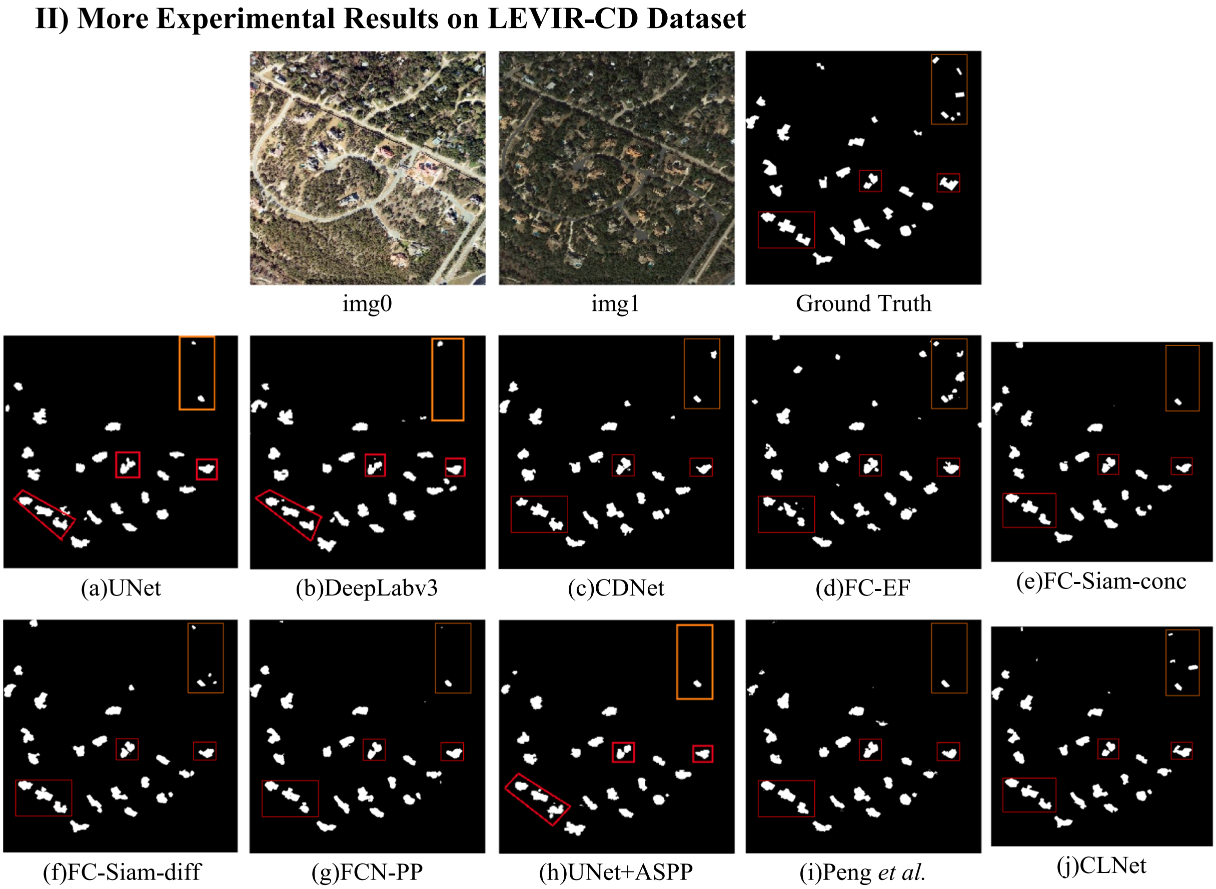


Fig. 16. Visual Comparison of generated change maps in the LEVIR-CD dataset (image index: test_25).

Appendix D. More experimental results of building change detection

I) Experiments of Comparison methods on LEVIR-CD dataset with the settings in their original literatures.

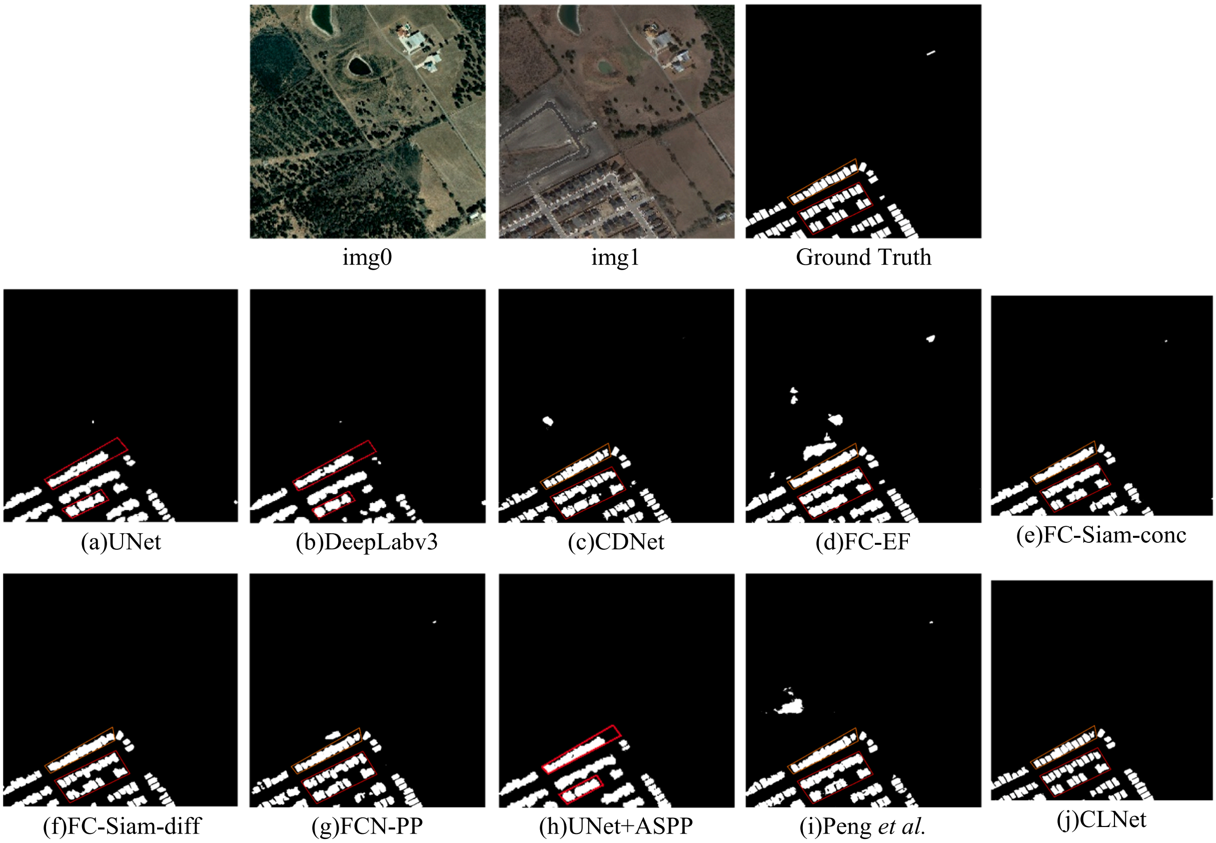
Table 8 shows the experiment results of the comparing methods under the settings described in the original literature. It can be found that their accuracy lied at a relatively low level, especially in the *precision* and *F1 Score*. For a fair comparison, the loss functions of all the compared methods were modified to the one used in CLNet to eliminate the performance difference caused by the loss function. After modifying the loss function, the accuracy of all the compared methods increased to a large margin (i.e., the *precision* and *F1 Score* of UNet were increased by 20.2% and 0.094).

II) More experimental results on LEVIR-CD dataset

See Fig. 17.

Table 8
Quantitative performance comparison on the LEVIR-CD dataset with the settings in their original literatures.

Methods	Precision (%)	Recall (%)	F1 Score	OA (%)
UNet (Ronneberger et al., 2015)	*64.4*	*89.8*	*0.755*	*97.1*
DeepLabv3 (Liang-Chieh Chen et al., 2017)	77.3	86.0	0.821	98.1
CDNet (Alcantarilla et al., 2018)	74.6	89.1	0.812	97.1
FC-EF (R.C. Daudt et al., 2018a, 2018b)	76.5	85.6	0.808	97.9
FC-Siam-conc (R.C. Daudt et al., 2018a, 2018b)	75.1	88.4	0.712	97.9
FC-Siam-diff (R.C. Daudt et al., 2018a, 2018b)	73.5	84.9	0.782	97.6
FCN-PP (Lei et al., 2019)	84.4	79.3	0.818	98.2
UNet + ASPP	78.6	84.2	0.809	98.0
Peng et al. (2019)	79.8	82.1	0.810	98.0


img0 img1 Ground Truth

(a)UNet (b)DeepLabv3 (c)CDNet (d)FC-EF (e)FC-Siam-conc

(f)FC-Siam-diff (g)FCN-PP (h)UNet+ASPP (i)Peng *et al.* (j)CLNet

Fig. 17. Visual Comparison of generated change maps in the LEVIR-CD dataset (image index: test_39).

III) More experimental results on WHU building dataset

See Figs. 18 and 19.

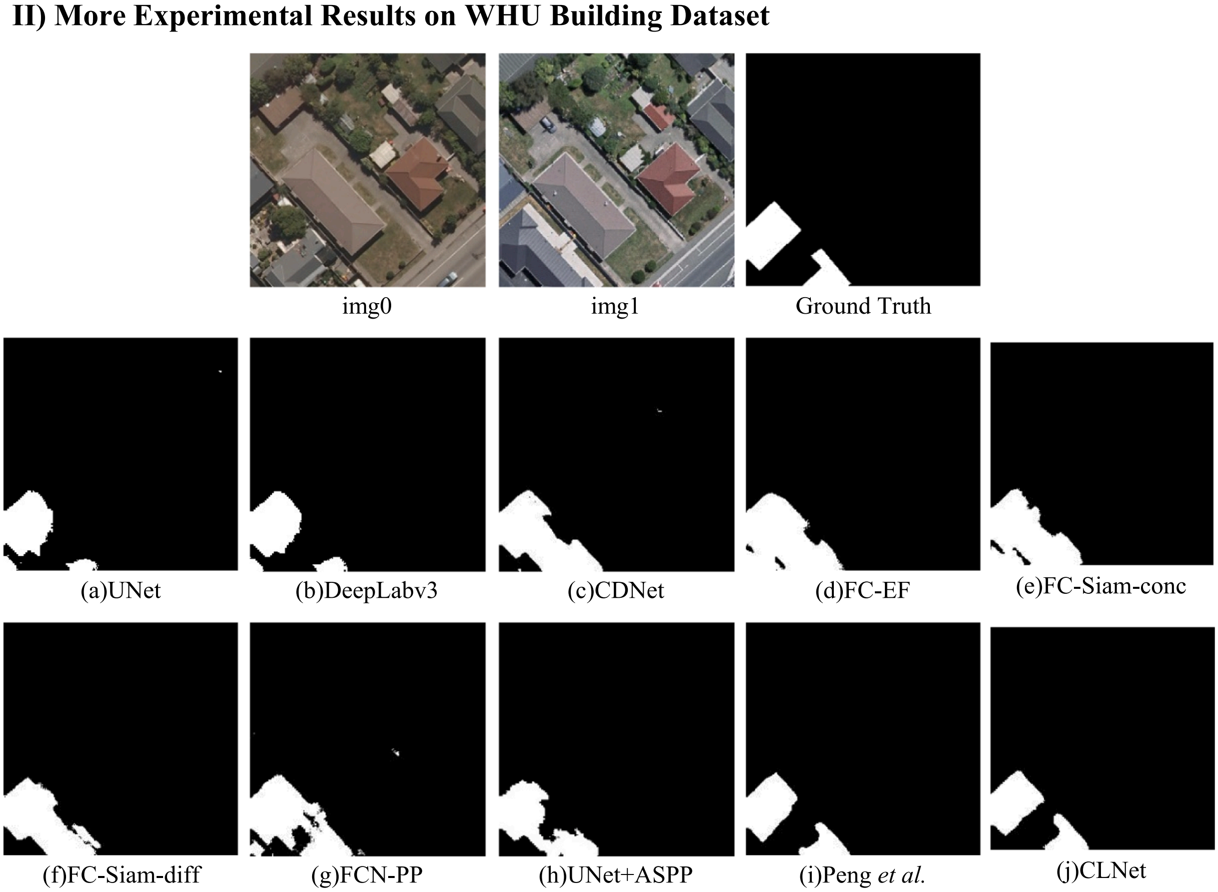


Fig. 18. Visual Comparison of generated change maps in the WHU Building dataset.

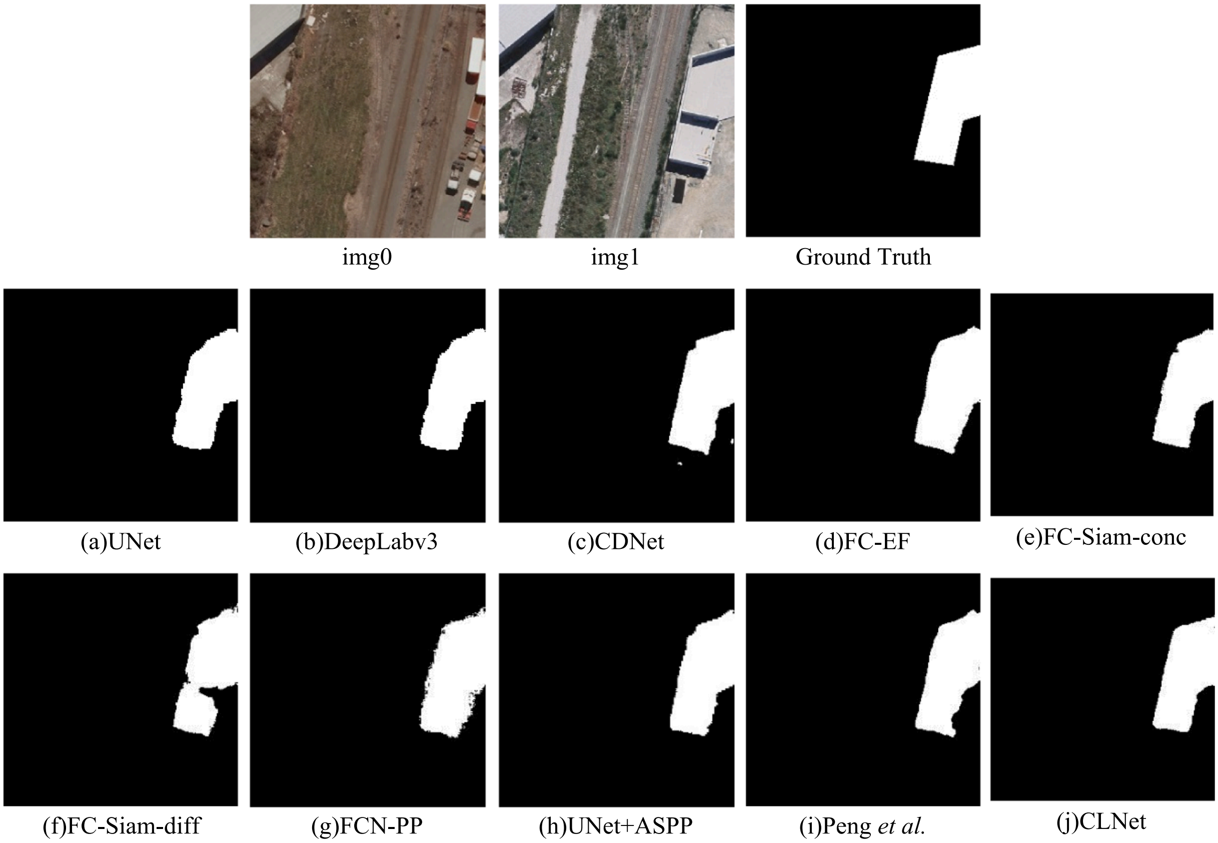


Fig. 19. Visual Comparison of generated change maps in the WHU Building dataset.

References

Akcay, Huseyin Gokhan, Aksoy, S., 2010. Building detection using directional spatial constraints. In: 2010 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1932–1935.

Alcantarilla, P.F., Stent, S., Ros, G., Arroyo, R., Gherardi, R., 2018. Street-view change detection with deconvolutional networks. Autonomous Robots 42 (7), 1301–1322.

Benedek, C., Szirányi, T., 2009. Change detection in optical aerial images by a multilayer conditional mixed markov model. IEEE Trans. Geosci. Remote Sensing 47(10), 3416–3430.

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 Sensing 45 (1), 218–236.

Bruzzone, L., Prieto, D.F., 2000. Automatic analysis of the difference image for unsupervised change detection. IEEE Trans. Geosci. Remote Sensing 38 (3), 1171–1182.

Chen, Liang-Chieh, et al., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint. arXiv:1706.05587.

Chen, H., Wu, C., Du, B., Zhang, L., 2019. Deep siamese multi-scale convolutional network for change detection in multi-temporal VHR images. In: 2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp). IEEE, pp. 1–4.

Chen, J., Lu, M., Chen, X., Chen, J., Chen, L., 2013. A spectral gradient difference based approach for land cover change detection. ISPRS J. Photogram. Remote Sensing 85, 1–12.

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. Intelligence 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, Caye, Rodrigo, Bertrand Le Saux, Boulch, Alexandre, 2018. Fully convolutional Siamese networks for change detection. arXiv preprint. arXiv:1810.08462v1.

Daudt, R.C., Saux, B.L., Boulch, A., Gousseau, Y., 2018a. High resolution semantic change detection. CoRR, vol. abs/1810.08452.

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

Desclée, B., Bogaert, P., Defourny, P., 2006. Forest change detection by statistical object-based method. Remote Sensing Environ. 102(1–2), 1–11.

Gevaert, C.M., Persello, C., Sliuzas, R., Vosselman, G., 2020. Monitoring household upgrading in unplanned settlements with unmanned aerial vehicles. Int. J. Appl. Earth Observ. Geoinform. 90, 102117.

Ghosh, S., Patra, S., Ghosh, A., 2009. An unsupervised context-sensitive change detection technique based on modified self-organizing feature map neural network. Int. J. Approximate Reason. 50 (1), 37–50.

Gong, M., Zhao, J., Liu, J., Miao, Q., Jiao, L., 2015. Change detection in synthetic aperture radar images based on deep neural networks. IEEE Trans. Neural Networks Learn. Syst. 27 (1), 125–138.

Gong, M., Yang, Y., Zhan, T., Niu, X., Li, S., 2019. A generative discriminatory classified network for change detection in multispectral imagery. IEEE J. Selected Topics Appl. Earth Observ. Remote Sensing 12 (1), 321–333.

Goodfellow, Ian, 2016. NIPS 2016 tutorial: Generative adversarial networks. arXiv preprint. arXiv:1701. 00160.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hulley, G., Veraverbeke, S., Hook, S., 2014. Thermal-based techniques for land cover change detection using a new dynamic modis multispectral emissivity product (mod21). Remote Sensing Environ. 140, 755–765.

Hussain, M., Chen, D., Cheng, A., Wei, H., Stanley, D., 2013. Change detection from remotely sensed images: from pixel-based to object-based approaches. ISPRS J. Photogram. Remote Sensing 80, 91–106.

Ji, S., Wei, S., Lu, M., 2019. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sensing.

Lebedev, M., Vizilter, Y.V., Vygolov, O., Knyaz, V., Rubis, A.Y., 2018. Change detection in remote sensing images using conditional adversarial networks. Int. Arch. Photogram., Remote Sensing & Spatial Inform. Sci. 42(2).

Lei, T., Zhang, Y., Lv, Z., Li, S., Liu, S., Nandi, A.K., 2019. Landslide inventory mapping from bitemporal images using deep convolutional neural networks. IEEE Geosci. Remote Sensing Lett. 16 (6), 982–986.

Leichtle, T., Geiß, C., Wurm, M., Lakes, T., Taubenbӧck, H., 2017. Unsupervised change detection in VHR remote sensing imagery–an object-based clustering approach in a dynamic urban environment. Int. J. Appl. Earth Observ. Geoinform. 54, 15–27.

Liang, B., Weng, Q., 2010. Assessing urban environmental quality change of indianapolis, united states, by the remote sensing and gis integration. IEEE J. Selected Topics Appl. Earth Observ. Remote Sensing 4 (1), 43–55.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125.

Liu, R., Kuffer, M., Persello, C., 2019. The temporal dynamics of slums employing a CNN-based change detection approach. Remote Sensing 11 (23), 2844.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Lyu, H., Lu, H., Mou, L., 2016. Learning a transferable change rule from a recurrent neural network for land cover change detection. Remote Sensing 8 (6), 506.

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 Sensing 57 (2), 924–935.

Niu, X., Gong, M., Zhan, T., Yang, Y., 2018. A conditional adversarial network for change detection in heterogeneous images. IEEE Geosci. Remote Sensing Lett. 16 (1), 45–49.

Peng, D., Guan, H., 2019. Unsupervised change detection method based on saliency analysis and convolutional neural network. J. Appl. Remote Sensing 13 (2), 024512.

Peng, D., Zhang, Y., 2017. Object-based change detection from satellite imagery by segmentation optimization and multi-features fusion. Int. J. Remote Sensing 38 (13), 3886–3905.

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.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Stramondo, S., Bignami, C., Chini, M., Pierdicca, N., Tertulliani, A., 2006. Satellite radar and optical remote sensing for earthquake damage detection: results from different case studies. Int. J. Remote Sensing 27 (20), 4433–4447.

Szegedy, Christian, et al., 2015. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

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 Sensing 57 (1), 3–13.

Wiratama, W., Lee, J., Park, S.E., Sim, D., 2018. Dual-dense convolution network for change detection of high-resolution panchromatic imagery. Appl. Sci. 8 (10), 1785.

Xian, G., Homer, C., 2010. Updating the 2001 national land cover database impervious surface products to 2006 using landsat imagery change detection methods. Remote Sensing Environ. 114 (8), 1676–1686.

Xiao, P., Yuan, M., Zhang, X., Feng, X., Guo, Y., 2017. Cosegmentation for object-based building change detection from high-resolution remotely sensed images. IEEE Trans. Geosci. Remote Sensing 55 (3), 1587–1603.

Xie, S., Tu, Z., 2015. Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1395–1403.

Yang, K., Xia, G., Liu, Z., Du, B., Yang, W., Pelillo, M., 2020. Asymmetric Siamese Networks for Semantic Change Detection, arXiv:2010.05687.

Yang, J., Weisberg, P.J., Bristow, N.A., 2012. Landsat remote sensing approaches for monitoring long-term tree cover dynamics in semi-arid woodlands: comparison of vegetation indices and spectral mixture analysis. Remote Sensing Environ. 119, 62–71.

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

Yu, W., Zhou, W., Qian, Y., Yan, J., 2016. A new approach for land cover classification and change analysis: Integrating backdating and an object-based method. Remote Sensing Environ. 177, 37–47.

Zanetti, M., Bovolo, F., Bruzzone, L., 2015. Rayleigh-rice mixture parameter estimation via EM algorithm for change detection in multispectral images. IEEE Trans. Image Process. 24 (12), 5004–5016.

Zhang, P., Gong, M., Su, L., Liu, J., Li, Z., 2016a. Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images. ISPRS J. Photogram. Remote Sensing 116, 24–41.

Zhang, C., Wei, S., Ji, S., Lu, M., 2019. Detecting large-scale urban land cover changes from very high resolution remote sensing images using cnn-based classification. ISPRS Int. J. Geo-Information 8 (4), 189.

Zhang, L., Zhang, L., Du, B., 2016b. Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Geosci. Remote Sensing Mag. 4 (2), 22–40.

Zhao, Hengshuang, et al., 2017. Pyramid scene parsing network. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.

Zhou, Z., Siddiquee, M.M.R., 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.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci. Remote Sensing Mag. 5 (4), 8–36.