International Journal of Applied Earth Observations and Geoinformation 105 (2021) 102591

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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

DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images

Qing Ding [a], Zhenfeng Shao [a,*], Xiao Huang [b], Orhan Altan [c]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[b] Department of Geosciences, University of Arkansas, Fayetteville, AR 72701, USA
[c] Department of Geomatics Engineering, Istanbul Technical University, Istanbul 36626, Turkey

ARTICLE INFO

Keywords:
Building change detection
Deep learning
DSA-Net
CLA-Con-SAM
Deep supervision

ABSTRACT

Building change detection (BCD) plays a crucial role in urban planning and development and has received extensive attention. However, existing deep learning-based change detection methods suffer from limited accuracy, mainly due to the information loss and inadequate capability in feature extraction. To overcome these shortcomings, we propose a novel deeply supervised attention-guided network (DSA-Net) for BCD tasks in high-resolution images. In the DSA-Net, we innovatively introduce a spatial attention mechanism-guided cross-layer addition and skip-connection (CLA-Con-SAM) module to aggregate multi-level contextual information, weaken the heterogeneity between raw image features and difference features, and direct the network's attention to changed regions. We also introduce an atrous spatial pyramid pooling (ASPP) module to extract multi-scale features. To further improve detection performance, we implement a new deep supervision module to enhance the ability of middle layers to extract more distinctive features. We conduct quantitative and qualitative experiments on the two publicly available datasets, i.e., the LEVIR-CD and the WHU Building datasets. Compared with the competing methods, the proposed DSA-Net achieves the best performance in all evaluation metrics. The efficiency analysis reveals that the proposed DSA-Net achieves a great balance between BCD performance and complexity/efficiency, with faster convergence and higher robustness.

1. Introduction

Change detection (CD), as a hot topic in remote sensing applications, aims to identify the differences from bi-temporal images that cover the same area. Existing CD studies are more focused on identifying large-scale changing phenomena, such as forest and water changes (Rokni et al., 2015). At a micro-scale, building change detection (BCD) also plays an important role in various applications that include illegal building identification, urban disaster assessment, to list a few (Awrangjeb et al., 2018; Gong et al., 2016).

With the continuous progress of earth observation technology, sensors with improved spatial resolution have greatly enriched the available data for remote sensing applications (Chen et al., 2020; Pan et al., 2021). Compared with medium and low-resolution images, high-resolution (HR) images contain more detailed information, thus facilitating the monitoring of small urban objects such as buildings. Traditional BCD methods require intensive manual interpretation, which is time- and labor-consuming, with low data utilization. Therefore, it is urgent to develop effective BCD algorithms that take advantage of massive remote sensing data.

According to the scale of the detection unit, traditional CD methods can be divided into pixel-based CD (Bruzzone and Prieto, 2000; Celik, 2009) and object-based CD (Bouziani et al., 2010; Im et al., 2008). In pixel-based CD, a difference map is usually obtained by directly comparing the corresponding pixels of bi-temporal images. Then pixels are classified into changed and unchanged categories via thresholding segmentation or cluster analysis (Bovolo et al., 2012; Nielsen, 2007). However, pixel-wise analysis largely ignores the contextual information, inevitably leading to salt-and-pepper noises. In contrast, the object-based CD can take advantage of structural and geometric information (Cai and Liu, 2013; Johansen et al., 2008). Nonetheless, the uncertainty of object segmentation and the sophistication of spectral and textural features in HR images bring new challenges to object-based CD.

In recent years, deep learning has shown notable advantages in remote sensing tasks such as semantic segmentation (Dai et al., 2016) and CD (Zhang et al., 2018; Hou et al., 2021). Generally, deep learning-

* Corresponding author at: No. 129 Luoyu Road, Hongshan District, Wuhan 430079, Hubei, China.
E-mail addresses: dingqing@whu.edu.cn (Q. Ding), shaozhenfeng@whu.edu.cn (Z. Shao), xh010@uark.edu (X. Huang), oaltan@itu.edu.tr (O. Altan).

https://doi.org/10.1016/j.jag.2021.102591
Received 28 May 2021; Received in revised form 23 September 2021; Accepted 15 October 2021
Available online 22 October 2021

Fig. 1. Representative samples of changed buildings in the LEVIR-CD dataset.

Fig. 2. Representative samples of changed buildings in WHU Building dataset.

based CD methods can be divided into three categories: (1) feature-based, (2) patch-based, and (3) image-based methods (Peng et al., 2019). Feature-based methods use networks to extract robust deep-level features, then realize CD based on the difference map generated by these features (Abdi and Jabari, 2021). Patch-based methods feed image patches into the network to determine whether the central pixels have changed (Gong et al., 2017; Yu et al., 2019). However, the feature-based and patch-based CD methods still have notable error propagation issues. To overcome this problem, image-based CD methods have been made to adopt the fully convolutional network (FCN), aiming to achieve high-precision results (Chen et al., 2021; Daudt et al., 2018a, 2018b; Liu et al., 2020). These image-based CD methods integrate feature extraction and difference discrimination operations, and the results are generated via an end-to-end manner, thereby minimizing error transmission.

However, most FCNs in CD studies are modified from single-image semantic segmentation networks. Concatenating bi-temporal images as one input makes the early layers of the network unable to extract informative features of raw images. In addition, the repeated usage of pooling layers (for image downsampling) leads to notable information loss, thus affecting the detection accuracy. Furthermore, an excessively deep architecture may cause gradient vanishing, slow convergence, and overfitting.

To solve the problems in FCNs, we propose a novel deeply supervised attention-guided network (DSA-Net) for BCD tasks. DSA-Net is a dual-branch network that adopts an encoder-decoder architecture. The encoders of different inputs have independent weights so as to effectively extract the deep features of raw images. The main contributions of this article are summarized as follows:

(1) We propose a spatial attention mechanism-guided cross-layer addition and skip-connection (CLA-Con-SAM) module. The

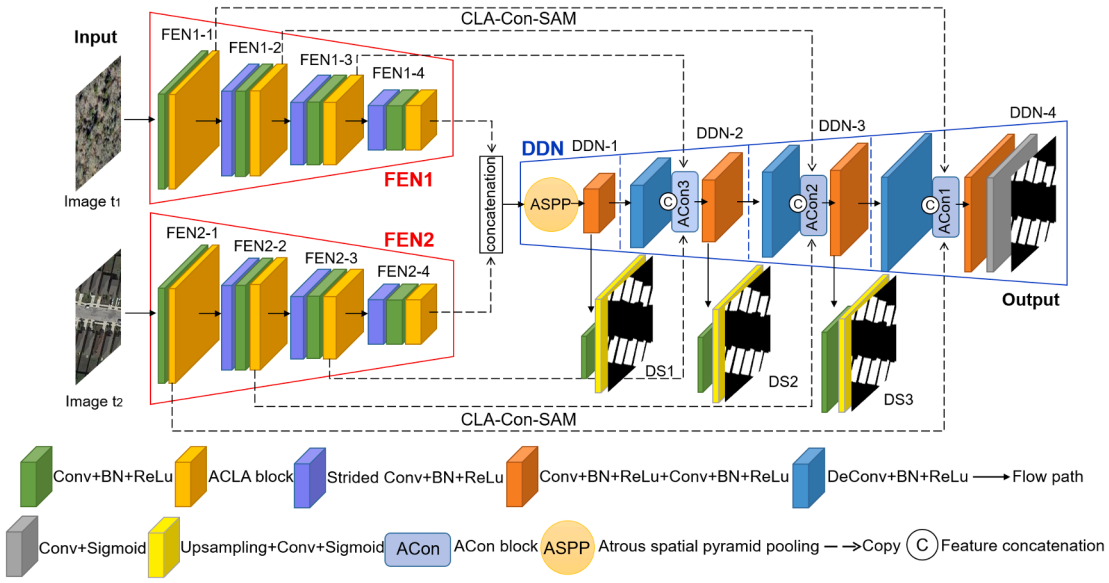

Fig. 3. The architecture of the proposed DSA-Net.

combination of parallel structures and the spatial attention mechanism (SAM) leads to a better expression of spatial details. The introduction of SAM in skip-connection facilitates the integration of raw image features and difference features, achieving better information retention.

(2) We implement a novel deep supervision module. Auxiliary classifiers are added in the middle layers of the decoder to assist in training network parameters, thus improving the generalization ability and detection accuracy.

(3) We evaluate DSA-Net on two different BCD datasets. By comparing DSA-Net with other existing methods, both qualitative and quantitative results confirm that the proposed DSA-Net achieves a better performance.

2. Materials and methods

2.1. Dataset description

LEVIR-CD (Chen and Shi, 2020) dataset focus on the changes in small and dense buildings (Fig. 1), which contains 637 pairs of HR bi-temporal image (0.5 m) patches covering various U.S. cities with a size of 1,024 × 1,024 pixels captured from 2002 to 2018. Small non-overlapping patches with a size of 128 × 128 pixels are cropped randomly, forming training (4,800 pairs), validation (1,200 pairs), and test sets (1,200 pairs).

In contrast, the WHU Building dataset (Ji et al., 2019) focuses more on changes in large and sparse buildings with diversified types and shapes (Fig. 2). The bi-temporal images (0.075 m) were acquired in 2012 (with 12,796 buildings) and 2016 (with 16,077 buildings), respectively. To verify the robustness of DSA-Net, non-overlapping patches with a size of 128 × 128 pixels are cropped, forming training (4,000 pairs), validation (1,600 pairs) and test sets (1,600 pairs).

2.2. Basic architecture of DSA-Net

Different from FCNs used for image semantic segmentation, DSA-Net (Fig. 3) is a dual-branch end-to-end network. Each branch with independent weights is a feature extraction network (FEN) that extracts individual features from a raw image. Then, the extracted features are transmitted to the difference detection network (DDN) to detect building changes. Notably, we use strided convolutions rather than poolings to achieve the downsampling of feature maps, aiming to facilitates the preservation of details (Isola et al., 2016; Radford et al., 2015).

Except for the first one, the other three stages of each FEN contain a strided convolutional layer, a convolutional layer, and an attention-guided cross-layer addition (ACLA) block. ACLA block aims to fuse cross-layer features and leads to increased attention on building changes (described in Section 2.3). For the convenience of description, we use $F^l_{t_1}$, $F^l_{t_2} \in \mathbb{R}^{C_l \times W_l \times H_l}$ to represent the features respectively for images t_1 and t_2 extracted by ACLA block at the l-th FEN stage.

As HR images have low interclass variability and high intraclass variability, concatenating the raw image features extracted by different FENs is expected to achieve better information retention (Ding and Huo, 2014). Therefore, features $F^4_{t_1}$ and $F^4_{t_2}$ with large receptive field and compact global information are concatenated as the input of ASPP module (Chen et al., 2017) at DDN-1 stage. The ASPP module can sample features in parallel and capture the multi-scale information (described in Section 2.4). Then, two convolutional layers are sequentially applied over ASPP's output to generate difference features with compact size.

DDN-2 starts with the difference features extracted by DDN-1. First, a deconvolutional layer is applied for upsampling purposes. Then, features $F^3_{t_1}$ and $F^3_{t_2}$ extracted by FEN1-3 and FEN2-3 are fed into an attention-guided skip-connection (ACon) block to generate a redefined joint feature (described in Section 2.3). The joint feature is concatenated with the features upsampled by the deconvolution. Further, the concatenated features are fed into two sequential convolutional layers to extract deep features. DDN-3 has exactly the same structure as DDN-2. In contrast, one 1 × 1 convolutional layer and one sigmoid function are added at the end of DDN-4 to generate change maps.

Besides the supervision on the output layer of the backbone network, DSA-Net also introduces three deep supervision branches (DS_1, DS_2, and DS_3) with the same structure. Note that change maps generated by deep supervision branches are only used for auxiliary training networks, and the final results of DSA-Net are obtained at the DDN-4 stage.

Taking DS_1 as an example, a convolutional layer is applied over the output of DDN-1 to reduce the number of feature channels. Then, features are restored to the same size as raw images by upsampling operations. Finally, the sigmoid function is used to obtain the output of DS_1. And the output is compared with the ground truth to enhance the network's generalization ability. Detailed parameters of DSA-Net can be found in Table 1.

Table 1
Detailed structure and parameters of DSA-Net.

Stage	Layer	Kernel size	Stride	Repeat	Output size
Image t_1, Image t_2	–	–	–	–	128 × 128 × 3
FEN1-1, FEN2-1	Conv ACLA block	3 × 3 –	1 –	1 1	128 × 128 × 64 128 × 128 × 64
FEN1-2, FEN2-2	Strided Conv Conv Conv ACLA block	2 × 2 3 × 3 –	2 1 –	1 1 1	64 × 64 × 128 64 × 64 × 128 64 × 64 × 128
FEN1-3, FEN2-3	Strided Conv Conv Conv ACLA block	2 × 2 3 × 3 –	2 1 –	1 1 1	32 × 32 × 256 32 × 32 × 256 32 × 32 × 256
FEN1-4, FEN2-4	Strided Conv Conv Conv ACLA block	2 × 2 3 × 3 –	2 1 –	1 1 1	16 × 16 × 512 16 × 16 × 512 16 × 16 × 512
DDN-1	ASPP Conv	– 3 × 3	– 1	1 2	16 × 16 × 1024 16 × 16 × 512
DDN-2	DeConv ACon3 block Conv	2 × 2 Concatenation 3 × 3	2 – 1	1 1 2	32 × 32 × 256 32 × 32×(256 + 256 + 256) 32 × 32 × 256
DDN-3	DeConv ACon2 block Conv	2 × 2 Concatenation 3 × 3	2 – 1	1 1 2	64 × 64 × 128 64 × 64×(128 + 128 + 128) 64 × 64 × 128
DDN-4	DeConv ACon1 block Conv Conv Sigmoid	2 × 2 Concatenation 3 × 3 1 × 1 –	2 – 1 1 –	1 1 2 1 1	128 × 128 × 64 128 × 128×(64 + 64 + 64) 128 × 128 × 64 128 × 128 × 2 128 × 128 × 2
DS_1	Conv Upsampling Conv Sigmoid	3 × 3 – 1 × 1 –	1 – 1 –	1 1 1 1	16 × 16 × 2 128 × 128 × 2 128 × 128 × 2 128 × 128 × 2
DS_2	Conv Upsampling Conv Sigmoid	3 × 3 – 1 × 1 –	1 – 1 –	1 1 1 1	32 × 32 × 2 128 × 128 × 2 128 × 128 × 2 128 × 128 × 2
DS_3	Conv Upsampling Conv Sigmoid	3 × 3 – 1 × 1 –	1 – 1 –	1 1 1 1	64 × 64 × 2 128 × 128 × 2 128 × 128 × 2 128 × 128 × 2

2.3. CLA-Con-SAM module

Existing deep learning-based CD studies have shown that multi-level and multi-scale information can enlarge receptive fields and aggregate spatial details for better parsing the changing scenes (Zhao et al., 2017; Zheng et al., 2021). To enhance the detection ability, the CLA-Con-SAM and ASPP modules are introduced into DSA-Net to extract multi-level and multi-scale features, respectively.

The CLA-Con-SAM module (Fig. 4) contains two ACLA blocks located at FEN1 and FEN2, respectively. In the ACLA block, features extracted from the previous convolutional layer are used as its input. In its first branch, two sequential convolutional layers are applied to extract the deep features. In the second branch, a convolutional layer is used to extract shallow features, and then a SAM is applied to optimize features in the spatial domain. Further, cross-layer outputs are added to obtain the multi-level features that benefit accurate detection by increasing the diversity of features.

Then, the extracted features of the corresponding ACLA blocks in FEN1 and FEN2 are concatenated as the input of the ACon block that uses SAM to optimize the concatenated features. After that, the redefined features are concatenated with features extracted by DDN, achieving information retention and reducing the heterogeneity of features.

The main role of SAM in CLA-Con-SAM module (Fig. 5) is to code the importance of each pixel location so as to direct the network's attention to changes (Woo et al., 2018). Specifically, the input feature $T \in \mathbb{R}^{C\times W\times H}$ is first processed through channel-based max pooling and average pooling to obtain features T_{max} and T_{avg}. Further, T_{max} and T_{avg} are concatenated and fed into the 7 × 7 convolutional layer to obtain the initial weight matrix G. Then, a sigmoid function is used to obtain the final weight matrix G_s. Finally, the input feature T is multiplied by the weight matrix G_s to obtain the output feature T_{sa}:

$$T_{sa} = T \times \left(\sigma\left(f^{7\times7}([max\ \ pool(T); avgpool(T)])\right)\right) \tag{1}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2}$$

where σ donates sigmoid function, $f^{7\times7}$ represents 7 × 7 convolutional operation, and $[;]$ represents feature concatenation.

2.4. ASPP module

The ASPP module conducts parallel sampling on input feature P using atrous convolutional layers at different rates to capture multi-scale features (Fig. 6). Notably, atrous convolution can increase the receptive field of convolution kernels without reducing the resolution of feature maps (Chen et al., 2017).

ASPP module consists of four parallel branches. The first branch contains one 1 × 1 convolutional layer, while the second and the third branches contain one 3 × 3 atrous convolutional layer with a rate of 2 and 4, respectively. The fourth branch contains one global pooling layer and one upsampling operation to integrate global information. Features P_1, P_2, P_3, and P_4 extracted by four different branches have the same size. The fusion feature P_{con} is obtained by channel-wise concatenation. Furthermore, one 1 × 1 convolutional layer is used to obtain the output feature P_{aspp} that reflects the multi-scale contextual information:

$$P_{aspp} = f^{1\times1}([f^{1\times1}(P); f^{3\times3}_{r=2}(P); f^{3\times3}_{r=4}(P); up(glo\ \ pool(P))]) \tag{3}$$

where $f^{1\times1}$ represents the 1 × 1 convolutional operation, $f^{3\times3}_{r=2}$ and $f^{3\times3}_{r=4}$ represent the atrous convolutional operations with rates of 2 and 4, respectively.

2.5. Deep supervision and loss function

The essence of a neural network is the utilization of the back-propagation algorithm to optimize the parameters. In general, increasing the depth can improve the representation ability of the network. However, as the network deepens, the multiplicative effect in the process of gradient back-propagation may lead to unstable weight updates, leading to unsatisfactory performance in CD tasks (Liu et al., 2019; Mao et al., 2018).

To solve the above problems and improve network performance, the deep supervision module is introduced in DSA-Net (Fig. 7). This module adds some auxiliary classifiers to realize gradient back-propagation by assisting middle layers in extracting features with higher discriminative information. As shown in Figs. 3 and 7, the three branches of the deep supervision module are denoted as DS_1, DS_2, and DS_3 respectively. Auxiliary branch DS_i is employed as follows:

$$DS_i = \sigma(f^{1\times1}(up(f^{3\times3}(D_i)))) \tag{4}$$

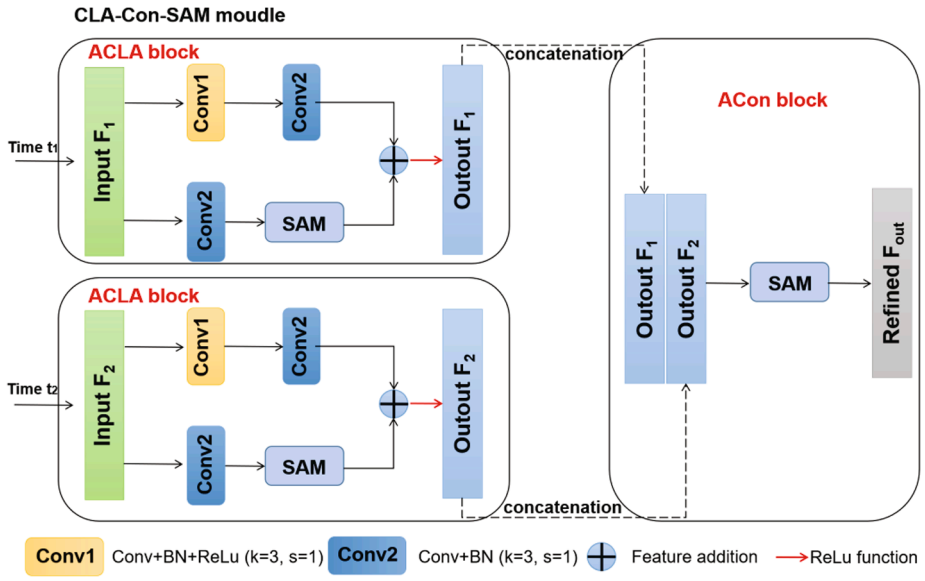

Fig. 4. The structure of the CLA-Con-SAM module.

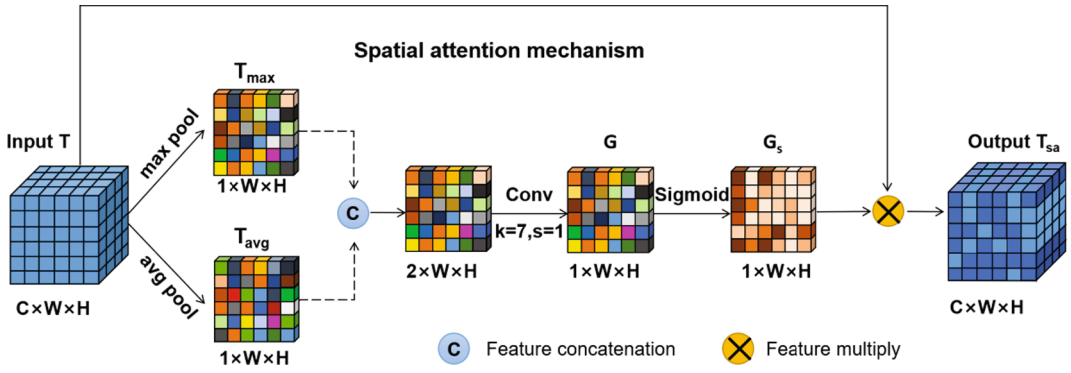

Fig. 5. The spatial attention mechanism.

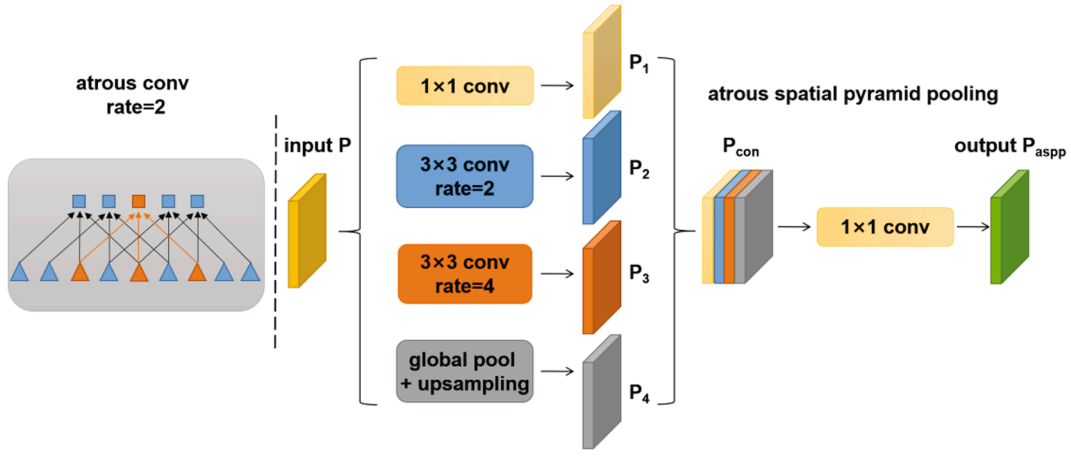

Fig. 6. The atrous convolution and atrous spatial pyramid pooling (ASPP).

where σ donates sigmoid function, $f^{1\times1}$ represents the 1 × 1 convolutional operation, *up* represents the upsampling operation, $f^{3\times3}$ represents the 3 × 3 convolutional operation to reduce the channel number, and D_i represents the feature maps extracted at *i*-th DDN stage.

Since image-based CD can also be regarded as a binary classification task, the binary cross-entropy (BCE) loss is selected as the loss function of the DSA-Net, as follows:

$$L_{bce} = -\frac{1}{m^{*}n}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[y_{(i,j)}\log\left(p_{(i,j)}\right) + \left(1-y_{(i,j)}\right)\log\left(1-p_{(i,j)}\right)\right] \quad (5)$$

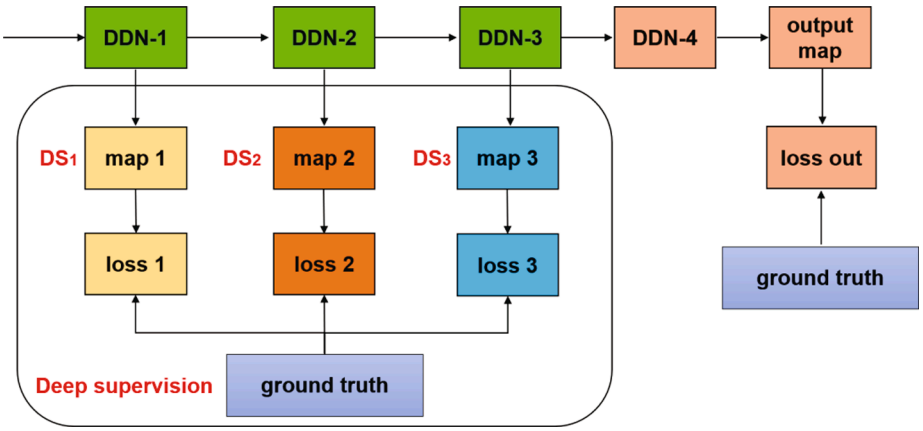

Fig. 7. The deep supervision module.

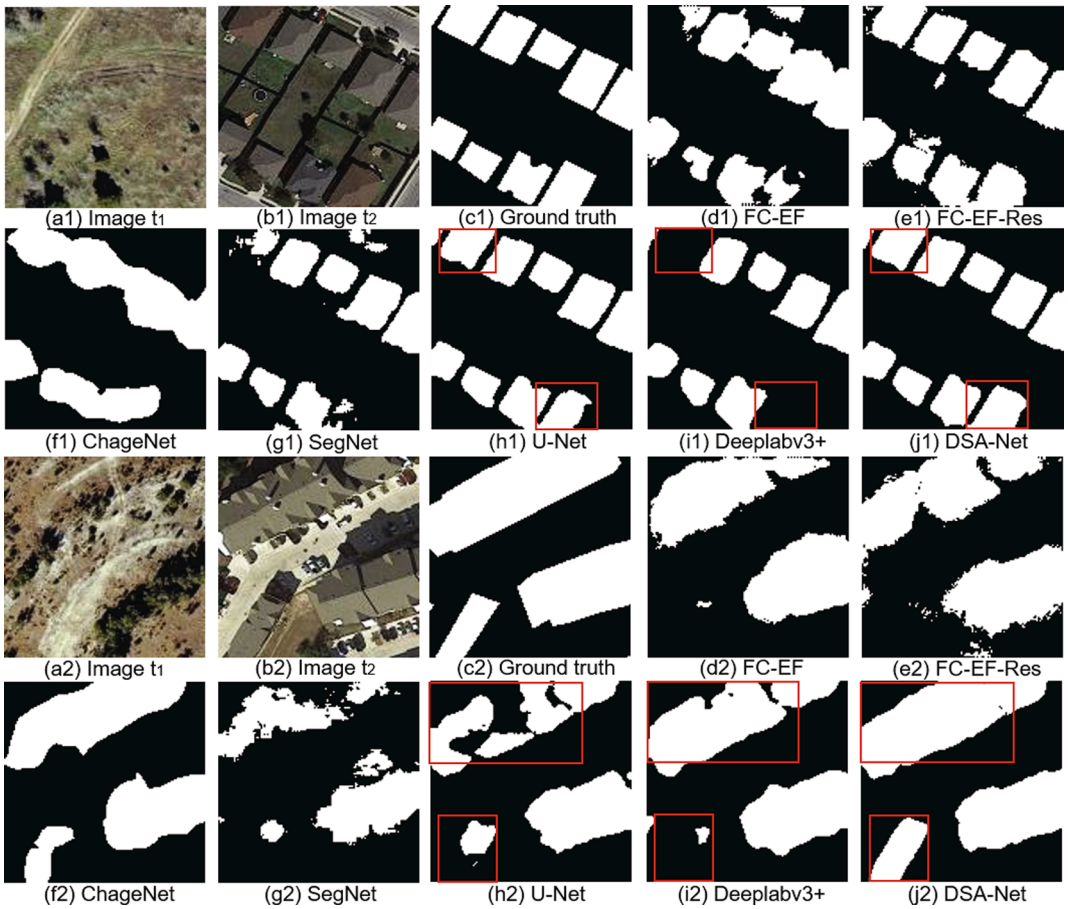

Fig. 8. Visual comparison of CD results on small and dense buildings from the LEVIR-CD dataset.

where $y_{(i,j)}$ represents the true value of pixel j in i-th layer of ground truth map after one-hot coding. $y_{(1,j)} = 0$ and $y_{(2,j)} = 1$ represent that pixel j belongs to the changed category. $p_{(i,j)}$ represents the score of the pixel j in i-th layer of the predicted map. If $p_{(1,j)}$ is smaller than $p_{(2,j)}$, pixel j is divided into the changed category in the predicted results.

Supposing that the corresponding weights of the four predicted maps are λ_1, λ_2, λ_3, and λ_{out} respectively, the total loss L_{total} can be expressed as:

$$L_{total} = \lambda_1 L_{bce}^1 + \lambda_2 L_{bce}^2 + \lambda_3 L_{bce}^3 + \lambda_{out} L_{bce}^{out} \tag{6}$$

where L_{bce}^1, L_{bce}^2, and L_{bce}^3 represent the BCE loss of the branches DS_1, DS_2 and DS_3, respectively. L_{bce}^{out} represents the BCE loss obtained at the DDN-4 stage.

2.6. Performance assessment

Precision (P), recall (R), intersection over union (IOU) are selected to evaluate the network performance. In BCD tasks, the higher the precision and recall, the lower the false detections and omission of the changed pixels. IOU represents the ratio of the intersection and union between the changed pixels in the predicted results and ground truths.

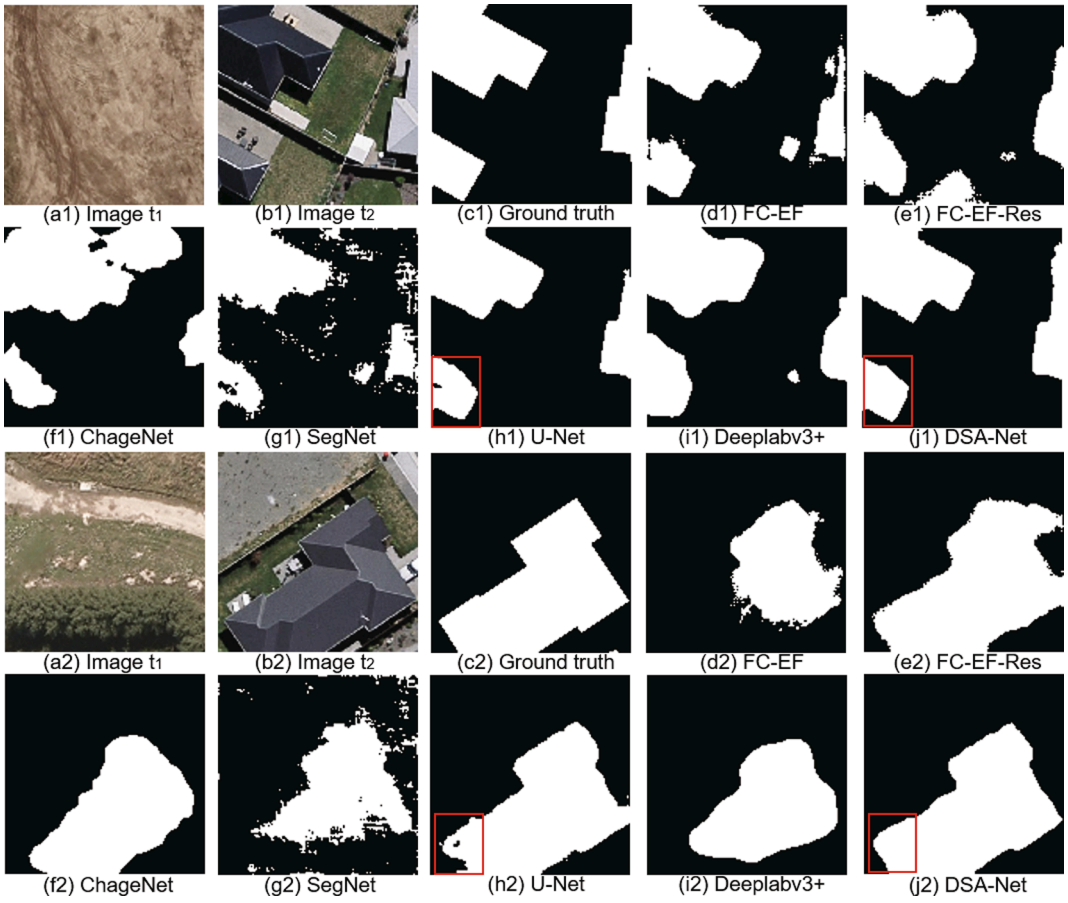

Fig. 9. Visual comparison of CD results on large and sparse buildings from WHU Building dataset.

Table 2
Quantitative performances of different methods on the LEVIR-CD dataset.

Methods	P	R	IOU	F_1	Kappa	OA
FC-EF	0.7103	0.7805	0.5920	0.7437	0.7302	0.9742
FC-EF-Res	0.7712	0.7922	0.6414	0.7815	0.7704	0.9788
ChangeNet	0.8218	0.7139	0.6182	0.7641	0.7530	0.9788
SegNet	0.8811	0.5966	0.5522	0.7115	0.6998	0.9768
U-Net	0.8971	0.8366	0.7633	0.8658	0.8593	0.9876
Deeplabv3+	0.8963	0.7806	0.7159	0.8344	0.8267	0.9851
DSA-Net	**0.8974**	**0.8812**	**0.8005**	**0.8892**	**0.8837**	**0.9895**

Table 3
Quantitative performances of different methods on the WHU Building dataset.

Methods	P	R	IOU	F_1	Kappa	OA
FC-EF	0.6460	0.7413	0.5272	0.6904	0.6747	0.9699
FC-EF-Res	0.7467	0.8040	0.6317	0.7743	0.7632	0.9788
ChangeNet	0.7847	0.6666	0.5635	0.7208	0.7087	0.9766
SegNet	0.7335	0.6906	0.5520	0.7114	0.6981	0.9746
U-Net	0.8920	0.8156	0.7423	0.8521	0.8454	0.9872
Deeplabv3+	0.8047	0.7781	0.6545	0.7912	0.7814	0.9814
DSA-Net	**0.8935**	**0.8763**	**0.7935**	**0.8848**	**0.8794**	**0.9897**

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$IOU = \frac{TP}{TP + FP + FN} \tag{9}$$

where TP represents the number of positive samples classified correctly, FP represents the number of negative samples classified incorrectly, and FN represents the number of positive samples classified incorrectly.

Besides, F_1 score, Kappa coefficient ($Kappa$), and overall accuracy (OA) are selected to evaluate the overall quality of the results. The greater the value of these metrics, the higher the consistency between predicted results and ground truths. These metrics are calculated as follows:

$$F_1 = \frac{2PR}{P + R} \tag{10}$$

$$Q = \frac{(TP + FN)(TP + FP) + (TN + FN)(TN + FP)}{TP + TN + FP + FN} \tag{11}$$

$$Kappa = \frac{TP + TN - Q}{TP + TN + FP + FN - Q} \tag{12}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

where TN represents the number of negative samples classified correctly; Q represents the intermediate variable in the calculation of the $Kappa$.

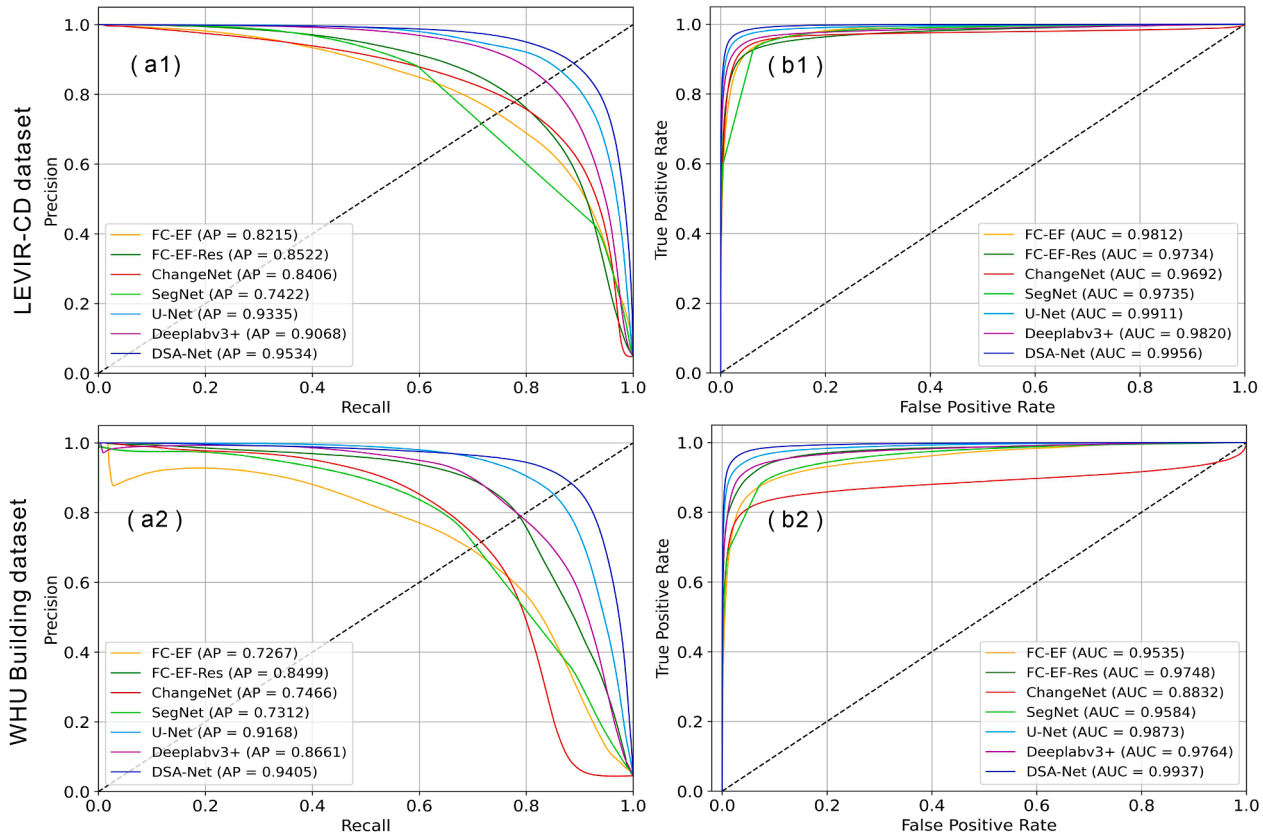

Fig. 10. The PR and ROC curves of BCD results.

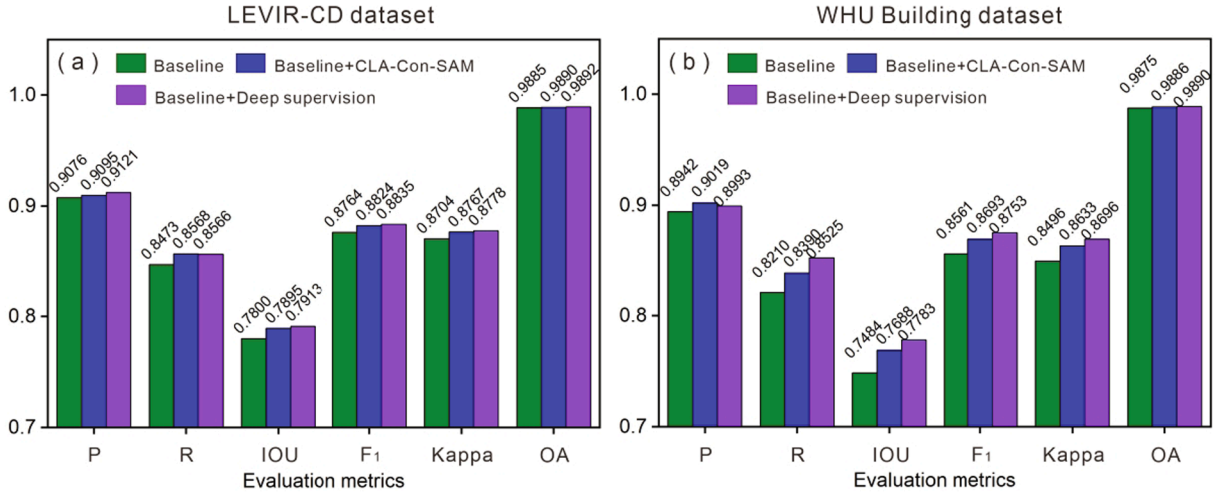

Fig. 11. The effect of the CLA-Con-SAM and deep supervision modules on the BCD results.

To evaluate the superiority and effectiveness of the DSA-Net, several state-of-the-art FCN-based methods are selected as competing methods:

(1) Fully convolutional-early fusion (FC-EF) (Daudt et al., 2018a) stacks bi-temporal images as its input, uses skip-connection to fuse shallow and deep features, then obtains the change maps.
(2) FC-EF with residual blocks (FC-EF-Res) (Daudt et al., 2018b) is an extension architecture of the FC-EF that introduces residual blocks to improve detection accuracy.
(3) ChangeNet (Varghese et al., 2019) is an image CD network that combines siamese structure and deconvolution.
(4) SegNet (Badrinarayanan et al., 2017) is an encoder-decoder network applied to semantic segmentation, whose encoder is the same as the first 13 layers of the VGG16 network.
(5) U-Net (Ronneberger et al., 2015) is a classic semantic segmentation network that introduces skip-connection between encoder and decoder to reduce information loss.
(6) Deeplabv3+ (Chen et al., 2018) is a semantic segmentation network that realizes the extraction of multi-scale features by integrating spatial pyramid pooling and encoder-decoder.

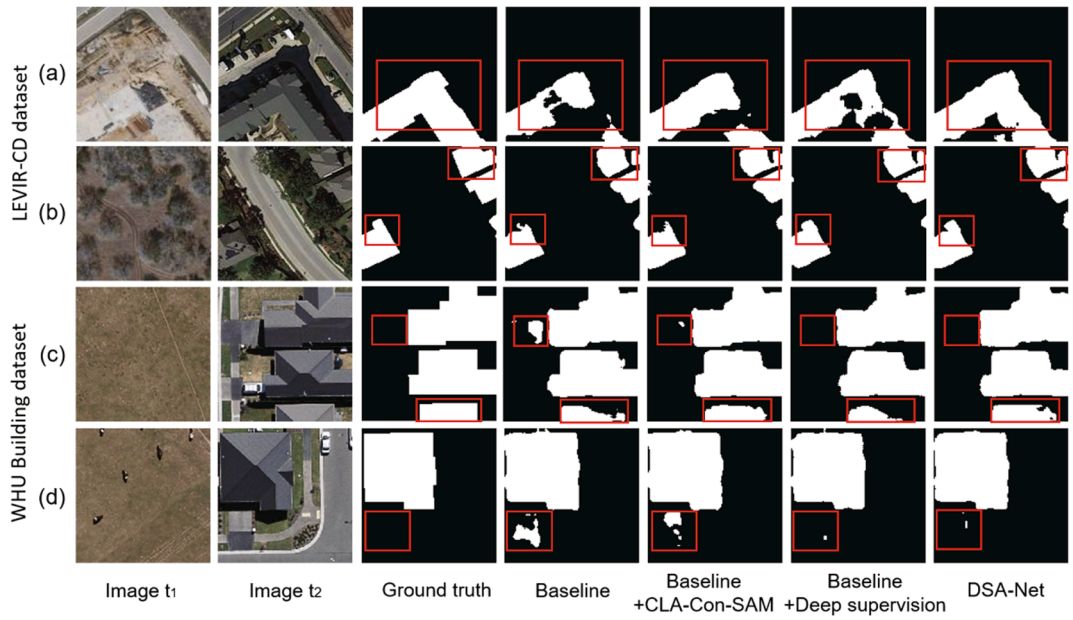

Fig. 12. Selected BCD results of the baseline, baseline + CLA-Con-SAM and baseline + deep supervision methods.

Table 4
Performance comparison among different methods.

Methods	Parameter (M)	Training epoch (s)	Validation (s)	Minimum loss	F_1
FC-EF	~1.35	~54	~20	0.0864	0.7437
FC-EF-Res	~1.10	~90	~37	0.0781	0.7815
ChangeNet	~51.30	~193	~96	0.0642	0.7641
SegNet	~24.95	~66	~30	0.4564	0.7115
U-Net	~31.04	~67	~20	0.0372	0.8658
Deeplabv3+	~59.35	~172	~84	0.0449	0.8344
DSA-Net	~38.53	~152	~45	0.0298	0.8892

3. Results and discussions

3.1. Experimental setup

The proposed DSA-Net is constructed with the PyTorch framework as the backend. The model training is implemented on a server with 16 GB RAM memory, Intel Xeon(R) E5-2687 W v4 cores at 3.00 GHz, and NVIDIA RTX8000-8Q GPU with 8 GB memory. During the training process, the Adam optimizer with a weight decay of $1 \times e^{-4}$ is selected as the optimization algorithm. The batch size and epoch are set to 16 and 60, respectively. The initial learning rate is set to $1 \times e^{-4}$ and multiplied by 10% after 30 epochs. For the proposed loss function, λ_{out} is set to 0.4, while λ_1, λ_2 and λ_3 are set to 0.2 to highlight the importance of the output layer. The competitive methods have the same hyperparameter setup and training strategy as DSA-Net.

3.2. Building change detection results

We conduct BCD experiments on the LEVIR-CD and WHU Building datasets to verify the superiority of DSA-Net over competing methods. To evaluate the BCD results more intuitively, we select some representative examples for visualization. Figs. 8 and 9 present two groups of change maps obtained from these two datasets, respectively.

Compared with the competing methods, DSA-Net achieves the best visual results with less false detection and misdetection (Fig. 8). Besides, results from DSA-Net are consistent with the ground truths, evidenced by accurately detected boundaries. Among the competing methods, the results of FC-EF, FC-EF-Res, ChangeNet, and SegNet contain a large number of misclassified unchanged pixels, leading to inaccurate

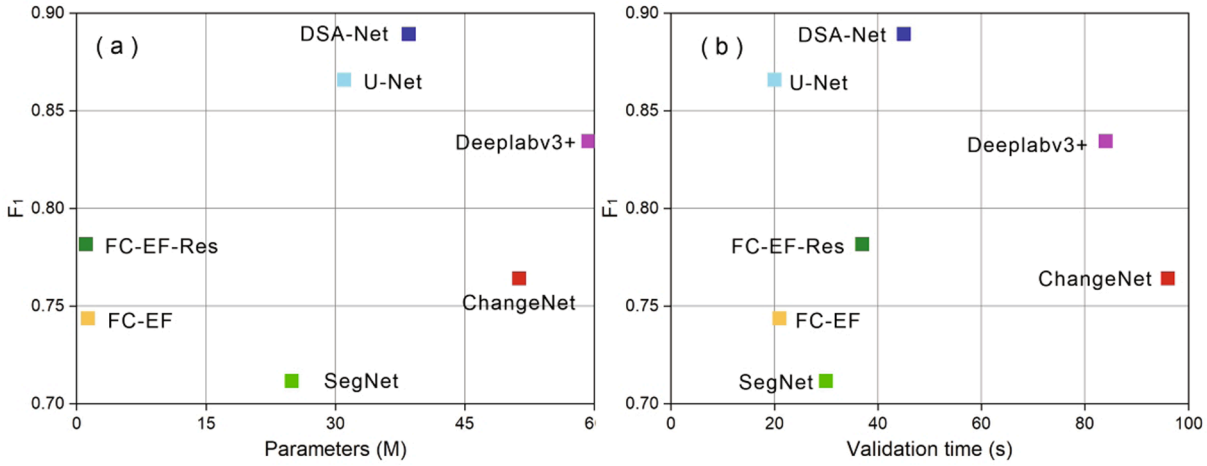

Fig. 13. Models' overall accuracy against (a) complexity and (b) efficiency.

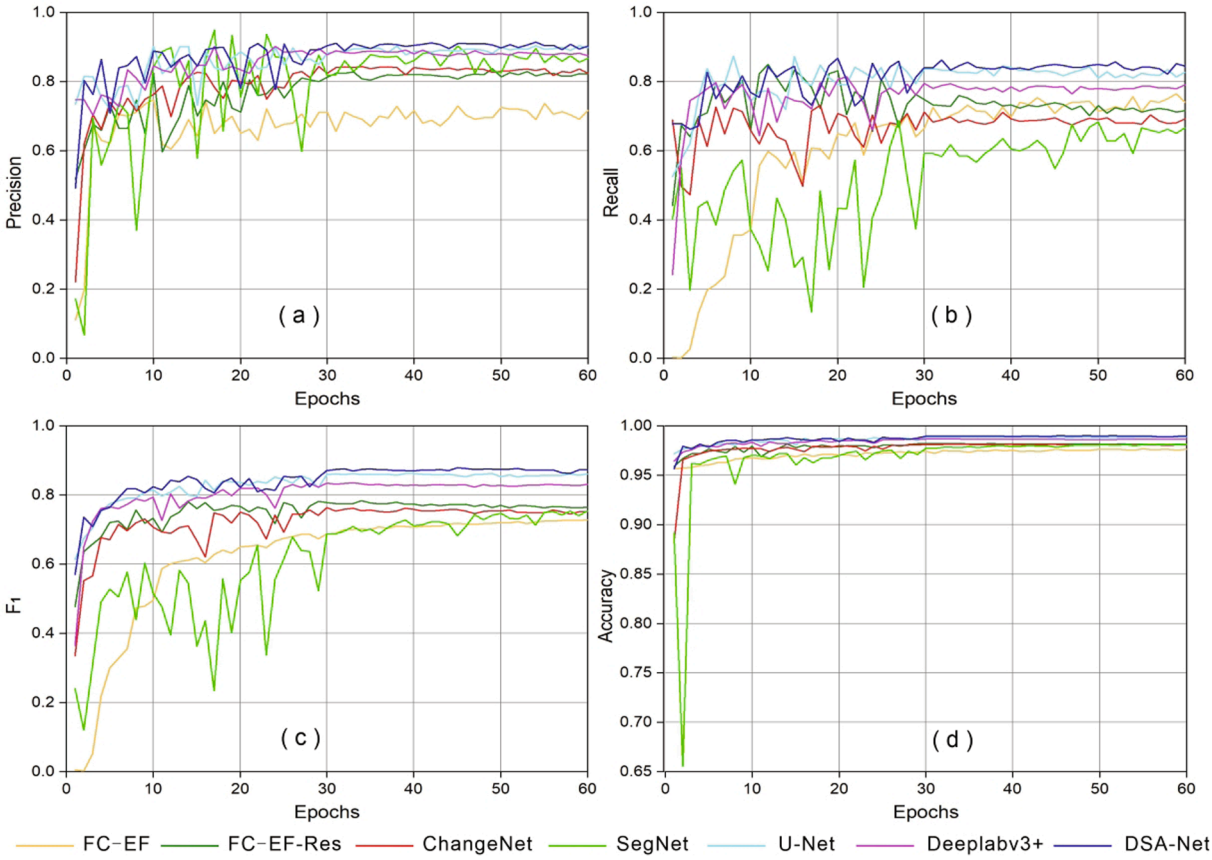

Fig. 14. Comparison of validation learning curves.

boundaries of changed buildings. The results from U-Net and Deeplabv3+ contain many missed detections.

We notice that the DSA-Net obtains the best detection results, with clear building boundaries, high internal compactness, and fewer missing/false detections (Fig. 9). In contrast, a large number of pixels are mistakenly detected in the results of FC-EF, FC-EF-Res, ChangeNet, and SegNet. The results of ChangeNet contain fuzzy building boundaries, while results of SegNet contain notable salt-and-pepper noises. U-Net obtains satisfactory BCD detection results, but with incomplete building structures, notable omissions and misclassification.

As can be seen from Tables 2 and 3, DSA-Net leads in all evaluation metrics, among which F_1 reaches 0.8892 and 0.8848 on LEVIR-CD dataset and WHU Building dataset, respectively. Compared with other competing methods, DSA-Net realizes the maximum improvement of 17.77% and 19.44% in F_1, respectively. The improvement can be attributed to its capability in extracting multi-level and multi-scale features, which greatly reduces the missing rate of changed pixels.

Precision-recall (PR) and receiver operating characteristic (ROC) curves are presented in Fig. 10 to further compare models' performance. Average precision (AP) and area under the curve (AUC) suggest the area under the PR curve and ROC curve, respectively. Higher AP and AUC values suggest better model performance and stronger generalization capability. The PR curves (Fig. 10(a1) and (a2)) of DSA-Net have the largest areas under curves with APs of 0.9534 and 0.9405, respectively. Compared with the competing methods, the ROC curves of DSA-Net have the largest deviation from the 45° diagonals, with the largest AUCs of 0.9956 and 0.9937, respectively (Fig. 10(b1) and (b2)). The above results prove again that the proposed DSA-Net outperforms other competing methods on both datasets.

3.3. Effect of the CLA-Con-SAM and deep supervision modules

We design and conduct ablation experiments to verify the effectiveness of CLA-Con-SAM and deep supervision modules proposed in DSA-Net. In this session, DSA-Net without CLA-Con-SAM and deep supervision modules is selected as the baseline method, where one 3×3 convolutional layer and skip-connection without attention mechanism are used to replace the ACLA block and ACon block, respectively.

The quantitative comparisons in Fig. 11 show that the designed CLA-Con-SAM and deep supervision modules are effective in improving the BCD accuracy. The introduction of the CLA-Con-SAM module increases F_1 and *Kappa* by 0.60%, 0.63% on the LEVIR-CD dataset, and 1.32%, 1.37% on the WHU Building dataset, respectively. Similarly, when the deep supervision module is introduced, the increases of F_1 and *Kappa* are 0.71%, 0.74% on the LEVIR-CD dataset, and 1.92%, 2.00% on the WHU Building dataset, respectively.

The introduction of CLA-Con-SAM and deep supervision modules increases the recall of changed pixels (see red boxes in Fig. 12(a) and (b)) and reduces false detections (see red boxes in Fig. 12(c) and (d)) with varying degrees. Notably, DSA-Net with both CLA-Con-SAM and deep supervision modules achieves the best performance, indicating their essential and intertwining role in extracting meaningful features and retaining complete information.

3.4. Comparison of network performance

On the LEVIR-CD dataset, the performance of different methods is evaluated from four aspects (Table. 4). Among them, the number of network parameters represents the model complexity. The average time cost of each training epoch and the whole validation process (the batch size is set to 1) represent the computational efficiency. In addition, the output loss of the validation set is selected to represent the convergence

ability, and F_1 is selected to reflect the overall accuracy.

Among all methods, FC-EF with lower complexity has the fastest training speed (about 54 s per epoch). SegNet that ignores the skip-connection has the worst convergence ability, with the lowest F_1 of 0.7115. The F_1 of FC-EF-Res with skip-connection and residual blocks is 7.00% higher than that of SegNet. U-Net achieves the best detection accuracy among the competing methods, with F_1 reaching 0.8658. Deeplabv3+ has high network complexity, with 59.35 M parameters. Notably, the DSA-Net achieves the best BCD performance, with minimum output loss (0.0298) and the highest F_1 (0.8892). In addition, the number of parameters of DSA-Net is 38.53M, about the same as that of U-Net and much lower than that of ChangeNet and Deeplabv3+. This comparison proves that the DSA-Net achieves a great balance between BCD performance and network complexity.

From Fig. 13, we notice that U-Net, SegNet, and DSA-Net have similar complexity. In contrast, the proposed DSA-Net achieves the best accuracy by sacrificing an acceptable range of efficiency.

Fig. 14 shows that the DSA-Net achieves faster convergence, evidenced by the rapid rise of evaluation metrics when training epochs increase, suggesting the great learning and generalization ability of the proposed DSA-Net. In comparison, the evaluation metrics of SegNet oscillate the most with the increase of epochs, indicating its poor robustness and difficulty in converging. The evaluation metrics of FC-EF are generally low when the model converges; as a result, the detection effect of FC-EF is relatively poor.

4. Conclusions

In this study, we propose a novel dual-branch end-to-end DSA-Net for bi-temporal BCD in HR remote sensing images. The FENs with independent weights benefit the extraction of the deep features of individual raw images. To reduce the information loss, we propose a CLA-Con-SAM module to effectively aggregate multi-level contextual features, weakening the heterogeneity between raw image features and difference features, and directing the network's attention to building changes. We also introduce an ASPP module into the decoder to realize the extraction and fusion of multi-scale features. In addition, we implement a novel deep supervision module to facilitate the middle layers in extracting more distinctive features, enhancing the model's performance and generalization ability, and avoiding gradient vanishing and slow convergence.

Experiments on the LEVIR-CD and WHU Building datasets confirm the effectiveness and robustness of the DSA-Net in different types of BCD tasks. Compared with the competing methods, DSA-Net has the highest accuracy and best BCD results, with greatly reduced missing/false detections. Ablation experiments show that the proposed CLA-Con-SAM and deep supervision modules significantly boost model performance. In addition, the proposed DSA-Net achieves a better balance between CD performance and network complexity/efficiency. The efficiency analysis suggests that the DSA-Net achieves faster convergence with stronger convergence ability. Further works will improve network's performance while simplifying the structure, and apply DSA-Net to other remote sensing tasks.

CRediT authorship contribution statement

Qing Ding: Methodology, Software, Writing – original draft. **Zhenfeng Shao:** Methodology, Data curation, Visualization. **Xiao Huang:** Writing – review & editing, Writing – review & editing. **Orhan Altan:** Validation, Conceptualization.

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

This work was supported in part by the National Natural Science Foundation of China under Grants 42090012, 41771452, 41771454; in part by the Key R&D Program of Yunnan Province in China under Grant 2018IB023.

Appendix. List of abbreviations

ACLA	attention-guided cross-layer addition
ACon	attention-guided skip-connection
AP	average precision
ASPP	atrous spatial pyramid pooling
AUC	area under the curve
BCD	building change detection
BCE	binary cross-entropy
CD	change detection
CLA-Con-SAM	spatial attention mechanism-guided cross-layer addition and skip-connection
DDN	difference detection network
DS	deep supervision
DSA-Net	deeply supervised attention-guided network
FCN	fully convolutional network
FEN	feature extraction network
HR	high-resolution
IOU	intersection over union
Kappa	kappa coefficient
OA	overall accuracy
P	precision
PR	precision-recall
R	recall
ROC	receiver operating characteristic
SAM	spatial attention mechanism

References

Abdi, G., Jabari, S., 2021. A Multi-Feature Fusion Using Deep Transfer Learning for Earthquake Building Damage Detection. Can. J. Remote Sens. 47 (2), 337–352.

Awrangjeb, M., Gilani, S.A.N., Siddiqui, F.U., 2018. An Effective Data-Driven Method for 3-D Building Roof Reconstruction and Robust Change Detection. Remote Sens. 10 (10), 1512.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Bouziani, M., Goïta, K., He, D.-C., 2010. Automatic change detection of buildings in urban environment from very high spatial resolution images using existing geodatabase and prior knowledge. ISPRS-J. Photogramm. Remote Sens. 65 (1), 143–153.

Bovolo, F., Marchesi, S., Bruzzone, L., 2012. A framework for automatic and unsupervised detection of multiple changes in multitemporal images. IEEE Trans. Geosci. Remote Sens. 50 (6), 2196–2212.

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.

Cai, S., Liu, D., 2013. A comparison of object-based and contextual pixel-based classifications using high and medium spatial resolution images. Remote Sens. Lett. 4 (10), 998–1007.

Celik, T., 2009. Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and k-Means Clustering. IEEE Geosci. Remote Sens. Lett. 6 (4), 772–776.

Chen, H., Shi, Z., 2020. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. Remote Sens. 12 (10), 1662. https://doi.org/10.3390/rs12101662.

Chen, H., Wu, C., Du, B.o., Zhang, L., Wang, L.e., 2020. Change Detection in Multisource VHR Images via Deep Siamese Convolutional Multiple-Layers Recurrent Neural Network. IEEE Trans. Geosci. Remote Sensing 58 (4), 2848–2864.

Chen, J., Yuan, Z., Peng, J., Chen, L.i., Huang, H., Zhu, J., Liu, Y.u., Li, H., 2021. Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 14, 1194–1206.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv 1706.05587.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Computer Vision – ECCV 2018, Proceedings, Part VII, Lecture Notes in Computer Science, pp. 833–851.

Dai, J., He, K., Sun, J., 2016. Instance-Aware Semantic Segmentation via Multi-task Network Cascades. In: IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 3150–3158.

Daudt, R.C., Saux, B.L., Boulch, A., 2018a. Fully Convolutional Siamese Networks for Change Detection. In: 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece 4063–4067.

Daudt, R.C., Saux, B.L., Boulch, A., Gousseau, Y., 2018b. High Resolution Semantic Change Detection. arXiv 1810.08452v1.

Ding, K., Huo, C., 2014. Sparse Hierarchical Clustering for VHR Image Change Detection. IEEE Geosci. Remote Sens. Lett. 12, 577–581.

Gong, M., Zhang, P., Su, L., Liu, J., 2016. Coupled Dictionary Learning for Change Detection From Multisource Data. IEEE Trans. Geosci. Remote Sensing 54 (12), 7077–7091.

Gong, M., Zhan, T., Zhang, P., Miao, Q., 2017. Superpixel-based difference representation learning for change detection in multispectral remote sensing images. IEEE Trans. Geosci. Remote Sensing 55 (5), 2658–2673.

Hou, X., Bai, Y., Li, Y., Shang, C., Shen, Q., 2021. High-resolution triplet network with dynamic multiscale feature for change detection on satellite images. ISPRS-J. Photogramm. Remote Sens. 177, 103–115.

Im, J., Jensen, J.R., Tullis, J.A., 2008. Object-based change detection using correlation image analysis and image segmentation. Int. J. Remote Sens. 29 (2), 399–423.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2016. Image-to-Image Translation with Conditional Adversarial Networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134.

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 57 (1), 574–586.

Johansen, K., Roelfsema, C., Phinn, S., 2008. High spatial resolution remote sensing for environmental monitoring and management preface. J. Spat. Sci. 53 (1), 43–47.

Liu, J., Gong, M., Qin, A.K., Tan, K.C., 2020. Bipartite Differential Neural Network for Unsupervised Image Change Detection. IEEE Trans. Neural Netw. Learn. Syst. 31 (3), 876–890.

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

Mao, T., Liu, W., Zhao, Y., Huang, J., 2018. Change Detection in Semantic Level for SAR Images. In: IEEE 3rd International Conference on Image, Vision and Computing.

Nielsen, A.A., 2007. The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data. IEEE Trans. Image Process. 16 (2), 463–478.

Pan, D., Zhang, M., Zhang, B.o., 2021. A Generic FCN-Based Approach for the Road-Network Extraction from VHR Remote Sensing Images – Using OpenStreetMap as Benchmarks. IEEE J Sel. Top. Appl. Earth Observ. Remote Sens. 14, 2662–2673.

Peng, D., Zhang, Y., Guan, H., 2019. End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++. Remote Sens. 11 (11), 1382. https://doi.org/10.3390/rs11111382.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 1511.06434.

Rokni, K., Ahmad, A., Solaimani, K., Hazini, S., 2015. A new approach for surface water change detection: Integration of pixel level image fusion and image classification techniques. Int. J. Appl. Earth Obs. Geoinf. 34, 226–234.

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.

Varghese, A., Gubbi, J., Ramaswamy, A., Balamuralidhar, P., 2019. ChangeNet: A Deep Learning Architecture for Visual Change Detection. LNCS. Springer Verlag. doi: 10.1007/978-3-030-11012-3_10.

Woo, S., Park, J., Lee, J. Y., Kweon, I. S., 2018. CBAM: Convolutional Block Attention Module. European Conference on Computer Vision. Springer, Cham.

Yu, L., Liu, X., Jiao, S., Chao, L., Jing, W., 2019. Multiscale Superpixel Segmentation With Deep Features for Change Detection. IEEE Access 7, 36600–36616.

Zhang, Z., Vosselman, G., Gerke, M., Tuia, D., Yang, M.Y., 2018. Change Detection between Multimodal Remote Sensing Data Using Siamese CNN. In: Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. In: IEEE Conference on Computer Vision and Pattern Recognition.

Zheng, Z., Wan, Y.i., 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.