AFSNet: attention-guided full-scale feature aggregation network for high-resolution remote sensing image change detection

Ming Jiang, Xinchang Zhang, Ying Sun, Weiming Feng, Qiao Gan & Yongjian Ruan

To cite this article: Ming Jiang, Xinchang Zhang, Ying Sun, Weiming Feng, Qiao Gan & Yongjian Ruan (2022) AFSNet: attention-guided full-scale feature aggregation network for high-resolution remote sensing image change detection, GIScience & Remote Sensing, 59:1, 1882-1900, DOI: 10.1080/15481603.2022.2142626

To link to this article: https://doi.org/10.1080/15481603.2022.2142626

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 06 Nov 2022.

Submit your article to this journal

View related articles

View Crossmark data
AFSNet: attention-guided full-scale feature aggregation network for high-resolution remote sensing image change detection

Ming Jiang*, Xinchang Zhang*a, Ying Sunc, Weiming Fenga, Qiao Gana and Yongjian Ruan*a

*aSchool of Geography and Remote Sensing, Guangzhou University, Guangzhou, China; bMinistry of Education, Key Laboratory of Geospatial Technology for Middle and Lower Yellow River Areas (Henan University); cSchool of Geography and Planning, Sun Yat-Sen University, Guangzhou, China

ABSTRACT

Using change detection technique precisely analyzes remote sensing images, it has a broad range of applications in resource surveys, surveillance systems, and map updating. In recent years, deep learning-based methods have become a focus area owing to their excellent feature extraction and representation ability. The fusion of multi-scale features is the key to improving change detection performance in fully convolutional network-based structural methods, primarily based on an architecture with skip connections or nested and dense skip connections. However, these methods only fuse features on the same scale and lack sufficient information from multiple scales to generate appropriate results. Traditional feature fusion with redundant and unsupervised information leads to poor model fitting. To solve these problems, we proposed a novel attention-guided full-scale feature aggregation network (AFSNet). The proposed method used a Siamese structure as the backbone network to extract features, which were then aggregated using full-scale skip connections, and an attention mechanism to avoid feature redundancy. Finally, to obtain a highly accurate final change map, a multiple side-outputs fusion strategy was used to fuse the change maps at different scales. To check the reliability of AFSNet, we tested it on two public datasets; the F1-Score/IoU scores improved by 0.57%/0.95% and 1.14%/2.07% in the two datasets, respectively, compared to those obtained using the methods that achieved suboptimal values. The results showed that AFSNet outperforms other mainstream methods while maintaining a good balance between computational costs and model parameters.

1. Introduction

Change detection is a key technology in remote sensing research and is critical in a variety of applications, such as land cover change determination, environmental monitoring, and disaster assessment (Pielke et al. 2011; Liu, Kuffer, and Persello 2019; Liu and Lathrop 2002; Bovolo and Bruzzone 2007). Change detection identifies differences between multitemporal images in the same landscape (Rensink 2002). It combines the features and imaging mechanisms of remote sensing images to analyze changes in the location, state, and characteristics of objects in areas (Singh 1989) and assigns binary attribute labels to image pixels.

Recently, with the development of remote sensing technology, the cost of obtaining remote sensing images with high spatial and temporal resolution has decreased, which provided new ideas for land cover change detection (Rogan and Chen 2004). The remote sensing has become one of the most important means of acquiring information on changes in land use (Khelifi and Mignotte 2020). However, the development of change detection technology lags behind that of high-resolution remote sensing imaging technology. For example, the traditional manual approach relies largely on expert knowledge, requires more time, and is costly, inefficient and inadequate (Zhu et al. 2018). Remote sensing images with higher resolution contain more detailed and complicated information; however, such images demonstrate data interference generated by different image acquisition conditions (Fang et al. 2021), making it difficult to find an effective method for achieving better performance in land cover change detection. The key to the change detection task is to overcome noise interference factors and detect areas of semantic information changes (Kaiyu, Zhe, and Fang 2020). Change detection methods based on various detection
strategies are broadly classified into two categories: traditional and deep learning methodologies.

Pixel-based and object-based methods are two types of traditional change detection methods (Hussain et al. 2013). Typically, pixel-based methods directly compare the feature information of pixels in an image and then define thresholds or clusters to obtain a final difference map (Deng et al. 2008). Pixel-based methods include slow feature analysis (SFA) (Chen et al. 2017), k-means (Liu et al. 2020), change vector analysis (CVA) (Liu et al. 2015), principal component analysis (PCA) (Richards 1984). Pixel-based methods are relatively simple to implement; however, they are less robust and prone to salt-and-pepper noise during processing, rendering it difficult to obtain the desired results (Chen and Shi 2020). Such methods usually require some post-processing techniques to optimize the results, such as parameter mining (Guo and Shihong 2017), fuzzy entropy query (Han, Zhang, and Zhou 2018) and other techniques. Object-based methods first use certain rules to disjoint homogeneous objects from remote sensing images, then extract spectral, textural, and geometric information from the images based on these homogeneous objects, and, finally, analyze the differences between images using those features (Chen et al. 2012). Object-based methods can effectively obtain spatial contextual information from images for change detection (Walter 2004). However, manually designing features is time-consuming and difficult, making it unsuitable for application in remote sensing images with complex spectral and textural features.

In recent years, deep learning techniques have attracted considerable attention and have a variety of applications in the field of remote sensing (Lei et al. 2019). Full convolutional networks (FCN) (Long, Shelhamer, and Darrell 2015) generally consist of encoder-decoder architecture and are often used to solve image segmentation problems. The encoder is similar to a traditional convolutional neural network (CNN) (Krizhevsky et al. 2012) but differs from the CNN in that the fully connected layer is discarded, and the main task is to extract low-level detail features and high-level semantic features via downsampling operations. The FCN adds upsampling convolutional layers to the encoder network, which are restored to the original image size by the decoder. SegNet (Badrinarayanan, Kendall, and Cipolla 2017) and UNet (Ronneberger, Fischer, and Brox 2015) are two of the most common FCNs. SegNet uses the index of the pooling layer in the encoder to guide the corresponding feature graph for the anti-pooling operation. By contrast, UNet fuses low- and high-level features by skipping connections to restore the integrity of high-level semantic information. The attention mechanism (Guo et al. 2022a) was successfully applied to image vision and can extract meaningful information from complicated features and improve the accuracy of remote sensing image change detection using deep learning methods (Liu et al. 2022; Guo et al. 2022b).

Unlike traditional methods, deep learning methods can extract deep discriminative features and automatically learn rich contextual and semantic features from remote sensing images. Generally, deep learning methods are divided into two categories: post-classification (Ji, Wei, and Lu 2019) and direct classification (Gil-Yepes et al. 2016). The post-classification methods individually classify numerous pairs of images before comparing the classification results to determine differences. The results of the change detection are obtained by comparing and analyzing the classification results from remote sensing images (Löw et al. 2022). These methods can only be applied when both change labels and bi-temporal semantic labels are available, but they have the drawback of accumulating classification errors (Tan et al. 2016). Direct classification methods obtain image change maps directly from bi-temporal images. Depending on the image input style, early and late fusion methods can be distinguished (Daudt, Saux, and Boulch 2018). Early fusion methods typically use stacked or differential remote sensing images as input, and the networks are structured to traditional semantic segmentation networks in a similar manner (Peng, Zhang, and Guan 2019). Daudt et al. (Daudt, Saux, and Boulch 2018) proposed a fully convolutional early fusion network (FC-EF), with UNet as the backbone. Alcantarilla et al. (Alcantarilla et al. 2018) proposed a change detection network (CDNet) containing superimposed shrinkage and expansion blocks. Peng et al. (Peng, Zhang, and Guan 2019) proposed a satellite image change detection network (UNet++_MSOF) based on UNet++ (Zhou et al. 2020) that applies UNet++’s nested and dense skip connections to fuse multi-layer features and a multiple side-outputs fusion (MSOF) strategy to obtain a highly accurate final change map. Zheng et al. (Zheng et al. 2021) designed
a change detection network (CLNet) with cross-layer blocks to fuse multi-scale features and multi-level contextual information.

Late fusion methods typically use the Siamese network (Bromley et al. 1993) as the backbone, with a Siamese architecture and weight-sharing strategy, resulting in fewer network parameters and faster convergence (Kaiyu, Zhe, and Fang 2020). Daudt et al. (Daudt, Saux, and Bouch 2018) proposed FC-Siam-Diff and FC-Siam-Conc networks based on differential and connection fusion methods. Chen et al. (Chen and Shi 2020) proposed a Siamese structure based change detection network (STANet), introducing a self-attentive mechanism to obtain more discriminative features. Chen et al. (Chen et al. 2021) proposed a new loss function and captured long-term feature dependencies via a dual attention mechanism (Jun et al. 2019) to improve model performance. Zhang et al. (Zhang et al. 2020) proposed a change detection network (IFN) that refines features in the decoder by using an attention mechanism, including a deeply supervised approach to train the network. Wang et al. (Wang et al. 2021) integrated adaptive attention methods for spatial and channel features while weighting the predictions to optimize the final change detection map. Zhang et al. (Zhang et al. 2021b) proposed HDFNet based on a hierarchical fusion strategy and a dynamic convolutional adaptive learning module, whereas a multi-level supervision strategy multi-scale loss functions was used to supervise the training process. To sum up, while existing deep learning methods have varied degrees of success in remote sensing image change detection, most of them are largely lost some detailed spatial information during downsampling. Detailed spatial information is crucial for detecting boundary details of changed areas. It must be connected to the decoder during upsampling to ensure detection accuracy.

Generally, the network structures of deep learning-based methods are based on UNet and UNet++, and the corresponding change detection methods have achieved good results in remote sensing image change detection tasks (Daudt, Saux, and Bouch 2018; Peng, Zhang, and Guan 2019; Kaiyu, Zhe, and Fang 2020; Zhang et al. 2021a). Features at different scales in neural networks represent different types of information (Lin et al. 2017), with low-level features containing spatial information regarding the changed area, highlighting its boundaries, and high-level semantic features containing category information, both of which contribute differently, with both being critical to the results predicted by the network. UNet employs a single-branch skip connection with the same size encoder and decoder to fuse low-level detailed features with high-level semantic characteristics, whereas UNet++ employs the same size encoder and decoder for different depth networks. These two structures effectively compensate for the detailed information lost because of encoder downsampling and thus improve the network detection accuracy compared to that of some traditional methods. However, these methods do not consider full-scale feature fusion and ignore the interactions between feature maps at different scales, resulting in issues such as blurred change area boundaries and the loss of small-scale areas. Furthermore, a lack of supervision and attention to high-level features merged with low-level features lead to duplicate information, which is unfavorable for model fitting and result in poor change detection.

Based on the above analysis, this research proposes an attention-guided, full-scale feature aggregation change detection network (AFSNet) for high-resolution remote sensing images. To solve the problem of insufficient information fusion caused by the same-scale feature fusion approach in traditional change detection methods, we use the full-scale feature fusion approach to strengthen the connection between multi-scale features, reduce the impact of information loss in change detection performance, and solve the drawbacks of single-scale feature fusion. To address the problem of a single output failing to consider the detection of different size changed scales, we output one result at each scale in the decoder and use the prediction results at multiple scales to reduce the sensitivity to the detection of different changed scales. First, the multi-layer features of the image pairs were extracted using a Siamese network structured as VGG16 (Simonyan and Zisserman 2014). Second, a full-scale feature fusion structure was used to aggregate the features at the full-scale. Third, the convolutional block attention module (CBAM) (Woo et al. 2018) is used to refine the fused features at each scale. Finally, network prediction maps at different scales were fused to obtain the final prediction results. AFSNet was tested on LEVIR-CD and SVCD datasets. Experiments have shown that, compared to other methods, AFSNet achieves excellent performance in both quantitative
metrics and visualization, especially in the detection of changed in small-scale or thin areas.

The main contributions of this research can be summarized as follows:

1. A novel change detection network (AFSNet) is proposed, which is more competitive than other comparison models.
2. A full-scale feature fusion structure is adopted to fuse multi-scale features, reduce information loss, and strengthen the correlation between features at different levels. Also, an attention mechanism is applied before obtaining the network output results to refine the features to obtain highly accurate prediction results.
3. The impact of different supervision methods on the network performance is explored, and a supervision strategy is designed for the proposed AFSNet. We also use a MSOF strategy to combine the predictions at different scales, adapting the method to the detection of different scales of changed areas.

2. Methods

2.1 Overview

As shown in Figure 1(a), AFSNet uses the encoder-decoder architecture for change detection. It primarily consists of feature extraction, feature fusion and refinement, and MSOF modules. The feature extraction module extracts features from bi-temporal images, and the feature fusion and refinement module fuses multi-scale features and refines them using an attention mechanism. The MSOF module is designed to combine the outputs of different sizes to obtain an accurate final result. AFSNet was constructed for change detection in high-resolution remote sensing images.

The feature extraction process adopts a Siamese structure, which ensures independence between the bi-temporal images during feature extraction. The key to improving performance is using a full-scale feature connection structure to aggregate features at multiple scales and fuse adequate information. However, this feature-fusion approach tends to result in information redundancy. To extract useful information, an attention mechanism is employed for feature refining, followed by multi-scale predicted results and a MSOF strategy to produce the final result.

The encoder Siamese network structure uses the VGG16 (Simonyan and Zisserman 2014) network, as shown in Figure 1(b). The modified VGG16 consists of 13 convolutional layers and four maximum pooling layers. The convolutional layers are linked to both batch normalization and the nonlinear activation function of the ReLU layer. With channel numbers (32, 64, 128, 256, 512), the network generated five separate feature maps of various sizes. The encoder uses concatenation to fuse the extracted features,
similar to FC-Siam-Conc (Daudt, Saux, and Boulch 2018). In contrast to the encoder, the decoder network has four upsampling layers. Full-scale skip connections are used for feature fusion. This type of network design can reduce information loss and aggregate features at each scale, which helps obtain a more accurate prediction.

In addition, because the aggregation of full-scale features is rich in feature information and prone to redundancy, we used the CBAM attention module for feature refinement. The refined features at each scale were then passed through a $3 \times 3$ convolutional layer to obtain a prediction map. Finally, the prediction maps at each scale were upsampled to the same size as the label, and the prediction maps at each scale were fused using a MSOF strategy, followed by the final result output through a $3 \times 3$ convolutional layer.

### 2.2 Full-scale skip connection module

Features at different levels can substantially impact the prediction results of deep learning methods, and a proper combination of these features can lead to highly accurate results. In literature, UNet or UNet++ feature connection structures are often used to improve detection results. However, connection structures often lack full-scale feature fusion and, therefore, cannot investigate the relationship between features, resulting in predicted changed areas of blurred locations and boundaries. To solve these problems, the proposed method employs a full-scale skip connection structure (Huang et al. 2020).

Full-scale skip connections change the feature connection of traditional networks and enable interconnections between encoders and decoders and internal connections between decoder networks. Each decoder layer contains feature maps of different sizes from the encoder and decoder, allowing fine-grained details and coarse-grained semantics to be captured at a full-scale.

Taking $D_{AB}^3$ (a feature map) as an example, a full-scale skip connection structure was specified, as shown in Figure 2. To obtain feature maps of various sizes, bi-temporal images were processed using a Siamese structure encoder. A feature map with a larger size was connected and processed via max-pooling to generate a feature map of the same size as the $D_{AB}^3$. Subsequently, we conducted a channel number reduction using a $1 \times 1$ convolution. Feature maps of the same size as $D_{AB}^3$ were concatenated and then operated using a $1 \times 1$ convolution. The smaller feature maps $E_B^3$ and $D_{AB}^1$ were first convolved using a $1 \times 1$ convolution operation, followed by upsampling. The five processed feature maps were concatenated. A feature map $D_{AB}^3$ with full-scale information was obtained using a $1 \times 1$ convolution. The feature maps of the other decoder used the same connection structure as $D_{AB}^3$. For light parameters and fast training, the number of feature channels was reduced through convolution with a kernel size of $1 \times 1$. The number of channels after convolution was one-quarter of the original number. A small feature map size reduces the number of convolution operation parameters and the amount of computation required.

![Figure 2. Architecture of the full-scale skip connection module.](image-url)


2.3 Attention module

The attention module adopted in this study was CBAM, which includes the channel attention module (CAM) and a spatial attention module (SAM). The overall structure is shown in Figure 3. The CBAM attention mechanism can be integrated into the network structure to participate in training. The principle behind the module design was to create an attention feature map by correlating the features in both the channel and spatial dimensions, and using the attention feature map to refine the input features. The attention module employed in this study had the same structure as in the original study (Woo et al. 2018), with relevant semantic information collected via the channel and spatial attention modules.

2.3.1 CAM

The information expressed in the feature maps between the channels in a neural network differs. The general network concatenates channels to fuse semantic information, which has the drawback of missing essential information. The CAM applies channel correlation to provide a one-dimensional weight vector that can be multiplied by the input feature map to focus more on critical semantic information, as shown in Figure 4. First, the input feature map was averaged and maximally pooled to obtain spatial dimensional information to generate two-channel features. The two obtained parts were fed into a shared multilayer perceptron with an implicit layer, outputting the feature vector by element summation. Finally, the sigmoid function was used to calculate the channel attention feature map, which is calculated as follows:

$$W_c(F) = \text{Sigmoid}(\text{MLP}(F^c_{\text{avg}}) + \text{MLP}(F^c_{\text{max}}))$$  \hspace{1cm} (1)

where $W_c(F)$ represents the channel attention feature map for feature $F$.

2.3.2 SAM

Figure 5 illustrates the architecture of the SAM attention module. It primarily uses the positional correlations between features to generate a two-dimensional spatial weight map and then multiplies it with the input feature map to secure more attention to representative features. We used average pooling and max pooling operations along the channel to build two spatial features for the final spatial attention feature map, connect them, and put them through

Figure 3. Architecture of the CBAM attention module.

Figure 4. Architecture of the CAM attention module.

Figure 5. Architecture of the SAM attention module.
a 7 × 7 convolution operation. We then used the resulting feature map to calculate the sigmoid function, which is calculated as follows:

\[ W_s(F) = \text{Sigmoid}(f_7 \times 7 \left( \left[ \frac{F^c_{\text{avg}}; F^c_{\text{max}}} \right] \right) ) \]  

(2)

where, \( W_s(F) \) represents the spatial attentional feature map of feature \( F \) and \( f_7 \times 7 \) represents the convolutional operation using a convolutional kernel size of 7 × 7.

### 2.4 Supervision methods and loss function

The loss in traditional deep learning methods is generally calculated using the final prediction between the result and label, followed by back-propagation to update the parameters. Some studies have combined the influence of deep supervision (Lee et al. 2014) to include loss constraints in the middle of the network to optimize network training. Other studies have combined multiple outputs (Peng, Zhang, and Guan 2019; Zhang et al. 2021b) and enhanced the final network output via a MSOF strategy. Several supervisory approaches were devised to identify the best approach for exploring the characteristics and interactions of areas of change at different scales. As shown in Figure 6, four alternative network architectures were designed by integrating different network supervision strategies.

AFSNetv1 produced only one output using the traditional supervision method in the final stage. AFSNetv2 employed multi-scale supervision with four outputs, in which each decoder’s four stages produced a result and added a constraint loss. AFSNetv3 connected the prediction maps of the four stages to generate the final result using a MSOF strategy. The network output only the final result achieved by MSOF. AFSNetv4 was built on the AFSNetv3 structure by adding an extra constraint loss to the prediction results for the four stages. Five outputs were obtained in the AFSNetv4 structure.

The deep supervision and MSOF strategy involved four decoder scales. First, the attention mechanism was used to refine the features for full-scale aggregation to obtain valuable information. Second, a 3 × 3 convolution layer was used to obtain the prediction at

![Figure 6. Network structures for different network supervision. (a) AFSNetv1. (b) AFSNetv2. (c) AFSNetv3. (d) AFSNetv4.](image-url)
this scale. Third, linear upsampling was used to restore the prediction to the original label size. Fourth, the MSOF strategy was used to concatenate the prediction results from the four scales. Finally, a $3 \times 3$ convolutional layer was used to obtain the final prediction results. The relevant operations are as follows:

$$
\bar{F}_{fsc} = \text{CBAM}(F_{fsc})
$$

(3)

$$
Out_{i} = \text{Conv2D}_{3 \times 3}(\bar{F}_{fsc})
$$

(4)

$$
\tilde{Out}_{i} = \text{Upsample}((H, W), "bilinear")\big(Out_{i}\big)
$$

(5)

$$
Out_{MSOF} = \text{Conv2D}_{3 \times 3}\big(\text{Cat}(\tilde{Out}_{1}, \tilde{Out}_{2}, \tilde{Out}_{3}, \tilde{Out}_{4})\big)
$$

(6)

where $F_{fsc}^i$ represents the feature of stage $i$ of the decoder, $Out_{i}$ represents the prediction result of stage $i$; $H$, $W$ represents the height and width of the image, and $Out_{MSOF}$ represents the prediction result of the MSOF strategy.

The change detection task aimed to find the changed and unchanged areas in bi-temporal images. Sample imbalance was common in images in which the number of pixels in the changed areas was substantially lower than that in the unchanged areas. A hybrid loss approach is often used to mitigate the impact of sample imbalance on accuracy (Fang et al. 2021; Zhang et al. 2020). In this study, a hybrid loss function, including binary cross-entropy loss and dice loss, was used as the final loss function:

$$
l_{bce} = -t_{i}\log(p_{i}) - (1 - t_{i})\log(1 - p_{i})
$$

(7)

$$
l_{dice} = 1 - \frac{2p_{i}t_{i}}{p_{i} + t_{i}}
$$

(8)

where $t_{i}$ represents the ground truth of the pixel $i$, 0 represents unchanged, and 1 represents changed. $p_{i}$ represents the predicted probability that pixel $i$ belongs to the change class. The hybrid loss function is defined as follows:

$$
loss = l_{bce} + l_{dice}
$$

(9)

As some network structures contain multiple output results, the final loss function is the sum of the multiple output losses and is calculated as:

$$
\text{loss}_{total} = \sum_{i=1}^{n} \text{loss}_{i}
$$

(10)

where, $\text{loss}_{total}$ represents the total loss, $\text{loss}_{i}$ represents the loss of output $i$, and $n$ represents the network output.

3. Experiments and discussions

3.1 Dataset description

Two datasets, LEVIR-CD (Chen and Shi 2020) and SVCD (Lebedev et al. 2018), were used to evaluate the methods used in this study. The LEVIR-CD dataset is a building change dataset consisting of 637 pairs of multispectral (R, G, B) images with a size of $1024 \times 1024$ pixels. The dataset was divided into three sets: training, validation, and testing. The images were slide-cropped into small patches of $256 \times 256$ pixels with no overlap owing to the limitation of the computer GPU memory (Figure 7). The training, validation, and test sets included 7120, 1024, and 2048 pairs of images, respectively. The sample images are shown in Figure 7.

The SVCD change detection dataset consists of 11 pairs of multispectral (R, G, B) images obtained by Google Earth, with a resolution ranging from 0.03 m to 1 m. The dataset included different seasonal changes of different sizes and natural objects. By clipping and rotating the original data, Ji et al. (Ji, Wei, and Lu 2019) produced datasets with training sets of 10,000 and 3000 test and validation sets, respectively, which were also used for method evaluation (Figure 8).

3.2 Experimental setup

The proposed AFSNet was implemented using PyTorch. We trained and tested all the models on a server with a single NVIDIA Quadro RTX 6000. Considering the limited computational resources, the batch size was set to 10, and AdamW (Loshchilov and Hutter 2017) was used as the optimizer. The learning rate was initially set to $5e - 4$ with a decay of 0.8 every five epochs for a total of 100 epochs.

3.3 Evaluation metric

In this study, the metrics of Precision, Recall, F1-Score, IoU, mIoU, and OA were used to evaluate the performance of the proposed methods. A higher precision implies that the detected changed pixels detected are
more precise, and a higher recall suggests that the method discovers fewer inaccurate changed pixels. The F1-Score considers precision and recall and provides a more balanced evaluation. The IoU is a measure of the correlation between the ground truth and prediction. The mIoU is calculated by averaging the IoU between unchanged and changed pixels, and OA is an indicator of the overall accuracy of the prediction results. The relevant equations are as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (11)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (12)
\]

\[
F1 - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)
\]

\[
\text{IoU} = \frac{TP}{TP + FP + FN} \quad (14)
\]

\[
mIoU = \frac{1}{k + 1} \sum_{i=0}^{k} \frac{TP}{FN + FN + TP} \quad (15)
\]

\[
\text{OA} = \frac{TP + TN}{TP + FP + TN + FN} \quad (16)
\]

where \(TP, FP, TN,\) and \(FN\) represent the number of true positive, false positive, true negative, and false negative pixels, respectively, and \(k\) represents categories 0 and 1.
3.4 Comparison methods

Eight change detection methods were selected for comparison to evaluate the effectiveness of AFSNet:

(1) FC-EF, FC-Siam-Conc, FC-Siam-Diff (Daudt, Saux, and Bouch 2018): With the three basic full convolutional change detection networks, FC-EF uses early fusion strategy, and FC-Siam-Conc and FC-Siam-Diff use late fusion strategy; the difference between FC-Siam-Conc and FC-Siam-Diff is the feature fusion method – one uses a concatenation method and the other uses a differential method.

(2) UNet++_MSOF (Peng, Zhang, and Guan 2019): An early fusion change detection network based on UNet++, while using MSOF strategy to obtain highly accurate prediction results.

(3) STANet (Chen and Shi 2020): A change detection network that introduces a self-attentive mechanism to obtain more discriminative features.

(4) IFN (Zhang et al. 2020): A change detection network that applies an attention mechanism to refine features during the decoding stage, using deeply supervised methods to train the network.

(5) SNUNet-CD (Fang et al. 2021): A change detection network combined the Siamese network and UNet++. It used the densely connected structure of the UNet++ network to fuse different deep network features. Specifically, it used an ensemble channel attention to strengthen the recognition of features at different semantic levels.

(6) BIT (Chen, Qi, and Shi 2022): A transformer-based change detection network that uses Siamese CNN to extract features, with a transformer to model contextual information and enhance feature representation.

To explore the effects of different supervision methods on the performance of the model, AFSNet, using various methods including AFSNetv1, AFSNetv2, AFSNetv3, and AFSNetv4, was trained and tested under the same conditions.

4. Results

4.1 Results of LEVIR-CD dataset

Table 1 presents the change detection results for the LEVIR-CD dataset. AFSNet’s four network structures achieved a good performance, and the metrics were better than those of the other methods. AFSNetv3 performed the best for almost all metrics, followed by AFSNetv4, AFSNetv1, and AFSNetv2. Compared with other change detection methods, the AFSNetv3 network was significantly superior to the LEVIR-CD dataset. As shown in Table 1, it achieved optimal accuracies in Recall (91.06%), F1-Score (90.90%), IoU (83.31%), mIoU (91.17%), and OA (99.07%). IFN obtained the second-highest ranked accuracy and performed the best in terms of Precision (92.21%). SNUNet-CD also performed well among the remaining methods. Transformer-based BIT performed better than STANet but worse than IFN. Among the three baselines (FC-EF, FC-Siam-Conc, and FC-Siam-Diff), FC-EF and FC-Siam-Conc displayed the worst and best performances, respectively.

Figure 9 shows the visual performances of all methods. Six scenarios, including small-scale and large-scale building changes, were selected to represent the detection effectiveness of each method.

| Method      | Precision (%) | Recall (%) | F1-Score (%) | IoU (%) | mIoU (%) | OA (%) |
|-------------|---------------|------------|--------------|---------|----------|--------|
| FC-EF       | 88.30         | 85.76      | 87.01        | 77.01   | 87.82    | 97.80  |
| FC-Siam-Conc| 89.28         | 88.47      | 88.87        | 79.97   | 89.40    | 98.87  |
| FC-Siam-Diff| 87.24         | 87.77      | 88.00        | 78.57   | 88.64    | 98.77  |
| UNet++_MSOF | 90.46         | 87.80      | 89.11        | 80.37   | 89.61    | 98.91  |
| STANet      | 90.15         | 88.41      | 89.27        | 80.62   | 89.74    | 98.92  |
| IFN         | **92.21**     | 87.73      | 89.91        | 81.50   | 90.31    | 99.00  |
| SNUNet-CD   | 91.44         | 89.24      | 90.33        | 82.36   | 90.67    | 99.03  |
| BIT         | 91.34         | 88.19      | 89.74        | 81.39   | 90.15    | 98.97  |
| AFSNetv1    | 90.80         | 90.52      | 90.66        | 82.92   | 90.96    | 99.05  |
| AFSNetv2    | 90.17         | 90.71      | 90.44        | 82.55   | 90.76    | 99.02  |
| AFSNetv3    | 90.74         | **91.06**  | **90.90**    | **83.31** | 91.17    | **99.07** |
| AFSNetv4    | 92.05         | 89.37      | 90.69        | 82.97   | 90.99    | 99.06  |
AFSNetv3 was used as the final AFSNet structure for the performance comparison. The first, fourth, and fifth rows represent the scenes of small-scale building changes. From the visualization results, it can be seen that the building boundary detection by AFSNet was much clearer, and the gaps between buildings could be detected accurately. The second row represents a narrow building change scene with fewer areas of missed detections by AFSNet, with the other methods showing more serious missed detections. The third row represents a small building change scene in which FC-EF, STANet, SNUNet-CD, and BIT failed to detect the changed area. The sixth row represents the large-scale building change scene; FC-Siam-Conc, FC-Siam-Diff, and AFSNet could detect the buildings more completely, whereas AFSNet detected the building boundary details more precisely and accurately. FC-EF and UNet++_MSOF showed more serious missed detections, whereas STANet, SNUNet-CD, and BIT revealed smaller areas of missed detections.

As can be seen from the above analysis, the AFSNet achieved the best visual performance, with few false and missed detections. It is worth mentioning that the boundary details of the changed buildings were also well detected compared with those detected using other methods. The results of the other methods contained a large number of misclassified pixels and significant missed detections. FC-EF, FC-Siam-Conc, FC-Siam-Diff, UNet++_MSOF, and STANet exhibited poor visual performance in changed building detection. In contrast, IFN, SNUNet-CD, BIT, and AFSNet could detect most changed areas correctly, with few misdetection pixels and fine boundary details. Overall, AFSNet achieved the best visual performance, similar to the ground truth, and could accurately predict the change maps.

### 4.2 Results of SVCD dataset

Table 2 presents the change detection results for the SVCD dataset. As indicated, all four network structures of AFSNet achieved exemplary performance and had better metrics than the other methods. AFSNetv3 achieved the best accuracy out of all the metrics. AFSNetv4 had a worse performance than AFSNetv3, followed by AFSNet2 and AFSNet1. Therefore, AFSNetv3 had the best results and AFSNetv1 performed the worst. Consistent with the results for the LEVIS-CD dataset, AFSNetv3 was adopted as the final structure of the AFSNet. As evident from Table 1, it achieved the optimal accuracies in the Precision (98.44%), Recall (92.85%), F1-Score (95.56%), IoU (91.50%), mIoU (95.15%), and OA (98.94%) indicators. IFN performed the best among other methods, whereas FC-EF performed the worst. The late fusion network was significantly better than the early fusion network in the three baseline networks: FC-EF, FC-Siam-Conc, and FC-Siam-Diff. The difference in the
Table 2. Comparison results of different methods on the SVCD dataset. The metrics are reported in percentage, bold values represent the optimal value, and underlined values represent the suboptimal value achieved by the methods.

| Method          | Precision (%) | Recall (%) | F1-Score (%) | IoU (%) | mIoU (%) | OA (%) |
|-----------------|---------------|------------|--------------|---------|----------|--------|
| FC-EF           | 90.84         | 82.63      | 86.54        | 76.28   | 86.37    | 96.83  |
| FC-Siam-Conc    | 92.55         | 84.14      | 88.15        | 78.80   | 87.84    | 97.21  |
| FC-Siam-Diff    | 92.88         | 86.45      | 89.55        | 81.07   | 89.14    | 97.51  |
| UNet++_MSOF     | 95.12         | 87.01      | 90.88        | 83.29   | 90.44    | 97.85  |
| STANet          | 94.12         | 87.68      | 90.78        | 83.12   | 90.33    | 97.80  |
| IFN             | 96.79         | 92.16      | 94.42        | 89.43   | 93.96    | 98.66  |
| SNUNet-CD       | 97.55         | 90.48      | 93.88        | 88.46   | 93.41    | 98.54  |
| BIT             | 96.77         | 89.94      | 93.23        | 87.32   | 92.75    | 98.39  |
| AFSNetv1        | 97.86         | 91.85      | 94.76        | 90.05   | 94.32    | 98.75  |
| AFSNetv2        | 97.72         | 92.18      | 94.87        | 90.24   | 94.43    | 98.77  |
| AFSNetv3        | 98.44         | 92.85      | 95.56        | 91.50   | 95.15    | 98.94  |
| AFSNetv4        | 97.89         | 92.22      | 94.97        | 90.42   | 94.93    | 98.79  |

Figure 10 shows six typical test areas containing a wide range of typical area variations, from top to bottom, scenes 1 to 6. The AFSNet achieved the best change detection performance by visual comparison, with fine-grained details and few missed and false detections. Compared with other methods, the AFSNet had better detection performance for changes in thin areas, such as detecting road changes in scene 1 and 5, which could maintain the connectivity of the road and accurate boundary details. At the same time, AFSNet could detect changes in small-scale objects, such as the water tower in scene 2 and the cars in scene 6, whereas other methods had more serious omissions in detecting changes in small-scale objects. In addition, the boundary details of the building changes detected by AFSNet were clear, and few voids appeared. Overall, the AFSNet could accurately detect the changed areas and is more accurate in detecting small-scale objects and boundary details, with few false and missed detections of changed areas.

4.3 Analysis of different methods in the LEVIR-CD and SVCD datasets

To verify the efficiency of AFSNet, the F1-Score and IoU of the methods in the LEVIR-CD and SVCD test sets were enumerated. The input image size of the methods was set to $256 \times 256 \times 3$ to calculate the number of floating-point operations (FLOPs) and parameters (Params), where AFSNet used the structure of AFSNetv3, described in the previous section. The evaluation metrics are presented in Table 3. As indicated, AFSNet achieved the highest scores for F1-
Score/IoU in both LEVIR-CD and SVCD datasets, and the F1-Score/IoU metrics reached 90.90%/83.31% and 95.56%/91.50% respectively. In the LEVIR-CD dataset, the F1-Score/IoU scores of SNUNet-CD achieved sub-optimal values, which were 0.57% and 0.95% lower than the F1-Score/IoU scores of AFSNet, respectively. The method that achieved suboptimal values in the SVCD dataset was IFN, with F1-Score/IoU scores 1.14% and 2.07% lower than the F1-Score/IoU scores of AFSNet. Comparing the three baseline models (FC-EF, FC-Siam-Conc, and FC-Siam-Diff), FC-EF had the worst accuracy and FC-Siam-Conc and FC-Siam-Diff achieved optimal values in LEVIR-CD and SVCD, respectively. Owing to the dense connection structure and MSOF strategy, UNet++_MSOF had a higher F1-Score/IoU scores than FC-EF by 2.1%/3.36% and 4.34%/7.01% in the two datasets, respectively. In addition, the numbers of Params and FLOPs of AFSNet were 9.08 M and 13.20 G, respectively, more than the numbers of Params and FLOPs of the three baseline methods (FC-EF, FC-Siam-Conc, FC-Siam-Diff), and less than the UNet++_MSOF, IFN, and SNUNet-CD. STANet had 6.35 G of FLOPs, AFSNet had 6.85 G more FLOPs than STANet, and 3.13 M fewer Params, whereas AFSNet had higher F1-Score/IoU scores than STANet of 2.1%/3.36% and 4.34%/7.01% in the two datasets, respectively. BIT is a lightweight network, with the numbers of FLOPs and Params only 4.24 G and 3.01 M respectively, but its F1-Score/IoU was 89.74%/81.39% and 93.23%/87.32% for the two datasets, 1.16%/1.12% and 2.33%/4.18% less than AFSNet, respectively. Overall, the numbers of Params and FLOPs of AFSNet were lower than those of most typical change detection methods. There was a good balance between computational complexity and model parameters, with the F1-Score/IoU achieving the highest scores, demonstrating the excellent performance of AFSNet.

**Figure 11(a,b)** show the F1-Scores of the training and validation for each training process in the SVCD dataset. **Figure 11(a)** shows that SNUNet-CD, BIT, AFSNet, and STANet fitted rapidly during the training process, with SNUNet-CD having the highest training F1-Score. FC-EF, FC-Siam-Conc, and UNet++_MSOF were the three methods that fitted slowly during the training process, and FC-Siam-Conc had the lowest training F1-Score. AFSNet had the highest F1-Score in the validation, and FC-EF had the lowest F1-Score. The results show that the training process of AFSNet is more stable and effective and that AFSNet had a strong model generalization ability.

### 4.4 Ablation experiments

Relevant ablation experiments were conducted to investigate the effect of each module on the AFSNet. The network with the FSC, CBAM, and MSOF modules was removed as the experimental ablation network. The experimental ablation network with AFSNet and AFSNet_noFSC did not use the full-scale connection structure, but instead used the UNet skip connection structure. These methods were tested on the SVCD dataset and the experimental results are presented in Table 4. From Table 4, we can see that the performances of AFSNet_noFSC, AFSNet_noCBAM, and AFSNet_noMSOF were less accurate than those of AFSNet. The results indicated that the full-scale skip connection module, attention mechanism, and MSOF modules applied in this study can effectively improve the performance of AFSNet. AFSNet_noFSC achieved the worst performance, suggesting that correctly using the full-scale skip connection module is the
Figure 11. Training and validation F1-Score curves for different methods on the SVCD dataset. (a) Training accuracy on SVCD dataset; (b) validation accuracy on SVCD dataset.

Table 4. Quantitative evaluation results of ablation experiments on the SVCD dataset. The metrics are reported in percentage, bold values represent the optimal value, and underlined values represent the suboptimal value achieved by the methods. “_noFSC” denotes the network without the full-scale connection structure; “_noCBAM” denotes the network without the CBAM attention module; and “_noMSOF” denotes the network without the MSOF strategy.

| Method         | Precision (%) | Recall (%) | F1-Score (%) | IoU (%) | mIoU (%) | OA (%) |
|----------------|---------------|------------|--------------|---------|----------|--------|
| AFSNet_noFSC   | 97.00         | 90.98      | 93.89        | 88.49   | 93.42    | 98.54  |
| AFSNet_noCBAM  | 97.74         | 91.61      | 94.58        | 89.72   | 94.13    | 98.70  |
| AFSNet_noMSOF  | 97.86         | 91.85      | 94.76        | 90.05   | 94.32    | 98.75  |
| AFSNet         | **98.44**     | **92.85**  | **95.56**    | **91.50** | **95.15** | **98.94** |

key to improving the performance of the method. The full-scale feature fusion approach enhanced the representation of information and mitigated the impact of information loss on the method performance. AFSNet_noMSOF did not apply the MSOF strategy and adopted the AFSNetv1 structure, outputting only one result at the end of the network. The experimental results of AFSNet_noMSOF achieved suboptimal values in all evaluation metrics, with F1-Score/IoU less than AFSNet by 0.8%/1.45% and mIoU/OA less than AFSNet by 0.83%/0.19%, indicating that the MSOF strategy could improve the performance of the method to a certain extent. The sensitivities of the prediction results on different decoder scales to the changed areas were different. In contrast, the MSOF strategy could combine the output results on multiple scales; this approach can, therefore, produce a highly accurate prediction result.

To understand the influence of each module on the method extraction results, we visualized the results of the experimental ablation network, as shown in Figure 12, which indicates that AFSNet could effectively extract changed areas from remote sensing images, guaranteeing their correctness and integrity. The results of the AFSNet_noFSC extraction revealed a relatively serious phenomenon of missed and false detection; AFSNet_noCBAM was inaccurate in detecting object boundaries, and AFSNet_noMSOF was insensitive to the detection of tiny and thin areas. Nonetheless, in general, AFSNet could effectively improve the above problems, consider the detection test results of complex scene changed areas, and maintain a good performance.

To analyze the impact of the attention module on the network, we visualized the heatmap of the SVCD dataset before and after applying the attention mechanism. The network heatmap before the sigmoid operation and the visualization results are shown in Figure 13. With the inclusion of the attention module, the network could focus more on changed areas and suppress the effects of unchanged areas. As shown in Figure 13(d,e), after adding the attention module, the network exhibited a strong response in the changed areas, whereas the response in the other areas was not evident. The results indicate that the attention mechanism could enhance the feature expression of the relevant regions and make the network focus on the changed areas in the image to improve the change detection performance.
5. Discussion

5.1 The performance of AFSNet

According to the performance of multiple methods in the datasets LEVIR-CD and SVCD, AFSNet had the highest F1-Score/IoU scores, reaching 90.90%/83.31% and 95.56%/91.50%, respectively, while providing a better balance of Recall and Precision. AFSNet can thus be adapted to detect multi-scale changed areas with clear boundary details and few
voids in large-scale changed areas. It exhibited good connectivity in narrow changed areas and detected small-scale changed areas with few missed detections. In addition, AFSNet had a lower number of FLOPs and Params, with F1-Score/IoU scores improving by 0.57%/0.95% and 1.14%/2.07% in the LEVIR-CD and SVCD datasets, respectively, than SNUNet-CD /IFN, which achieved suboptimal values. In this study, four AFSNet architectures were evaluated, with AFSNetv3 achieving the best performance. AFSNetv3 does not supervise the prediction results at each scale and instead uses a MSOF strategy to connect the prediction results at multiple scales to output the final results. The MSOF strategy combines results on multiple scales, each of which is linked to the final output and, to some extent, plays the role of multi-scale supervision. To alleviate the sample imbalance between the changed and unchanged areas, a hybrid of binary cross-entropy loss and dice loss was used in this study. In the future, we will try other loss functions to solve sample imbalance problems.

5.2 The limitations of AFSNet

Although the proposed AFSNet achieved better change detection on different datasets, it still could not obtain perfect results for very challenging scenarios; some failure cases are shown in Figure 14. Figure 14(a) shows the unpredicted changes in lawns, as the differences in lawn changes are minor and cannot be fully detected by AFSNet. This is because the feature information in the bi-temporal images is not significantly different, and the method could not extract the corresponding change features. Figure 14(b,c) show more difficult change detection scenarios, where the size of the change region is too tiny or thin; therefore, our method could not effectively extract the key feature information present in the changed areas, which can render the detection less effective. In addition, shading from trees or the influence of shadows can also make detection difficult (as shown in Figure 14(d)), which may result in the failure of the method to extract key information, leading to poor change detection results. To address this issue, we will devise superior methods for feature fusion and more efficient attention mechanisms to achieve better performance in challenging scenarios.

Deep learning methods can automatically learn features and complete change detection tasks, saving time and labor. In addition, the amount of memory in remote sensing data is considerable, and the image size is large. Owing to the limitations of the computer GPU, we usually crop a large remote sensing image into smaller pieces and feed them into deep learning methods. In the inference stage, the images are fed into the network separately for predictions, and the predictions are then stitched together in cropped order to obtain the changed results of the large remote sensing images. However, the cropping operation leads to insufficient pixel features at the edges of the images, and errors in the predicted edges of the cropped images also affect the accuracy of the final results when stitching. To address this issue, bi-temporal remote sensing images can be cropped with overlap and stitched together after predictions, and the class probability of each overlapping pixel is obtained by averaging the prediction values of that pixel (Huang et al. 2021). The method proposed in this research was oriented toward high-resolution remote sensing images and was a fully supervised change detection method that showed good performance with sufficient labels but also had some drawbacks. First, the proposed method does not solve the
6. Conclusion

This research proposed a change detection method, AFSNet, containing modules of the Siamese structure, full-scale feature aggregation, attention mechanisms, and MSOF structures. The Siamese structure enables the extraction of features, and the full-scale feature aggregation structure effectively obtains full-scale feature information. Subsequently, the attention mechanism further refines these features. The MSOF structure fuses the features at multiple scales, allowing AFSNet to adapt to the detection of changes at different scale areas. The results using two publicly available datasets show that AFSNet achieves impressive results in terms of effectiveness and complexity compared with other methods. AFSNet can detect boundary details of objects more accurately and with few missed and false detections. The proposed method achieved competitive results on change detection in the LEVIR-CD and SVCD datasets, and the F1-Score/IoU scores improved by 0.57%/0.95% and 1.14%/2.07%, respectively, compared with those obtained using the methods that achieved suboptimal values. It is worth noting that the computational cost of the proposed method is 13.20 G and the number of parameters is 9.08 M, both of which are lower than those of most methods. Furthermore, the ground truth dataset is crucial for deep learning methods, and future research could be directed toward semi-supervised or weakly supervised learning.

Acknowledgements

All authors sincerely thank the anonymous reviewers and editors for their constructive comments on improving this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the National Key R&D Program of China (Grant no. 2018YFB2100702) and the National Natural Science Foundation of China (Grant no. 42071441).

Data availability statement

The LEVIR-CD dataset used in this study is available at https://justchenhao.github.io/LEVIR/; the SVCD dataset is available from https://drive.google.com/file/d/1GX656Jqq0yBi_Ef0w65kDGVto-nHrNs9. Both datasets are freely available online.

References

Alcantarilla, P. F., S. Stent, G. Ros, R. Arroyo, and R. Gherardi. 2018. “Street-view Change Detection with Deconvolutional Networks.” *Autonomous Robots* 42 (7): 1301–1322. doi:10.1007/s10514-018-9734-5.

Badrinarayanan, V., A. Kendall, and R. Cipolla. 2017. “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (12): 2481–2495. doi:10.1109/TPAMI.2016.2644615.

Bovolo, F., and L. Bruzzone. 2007. “A Split-Based Approach to Unsupervised Change Detection in Large-Scale Multitemporal Images: Application to Tsunami-Damage Assessment.” *IEEE Transactions on Geoscience and Remote Sensing* 45 (6): 1658–1670. doi:10.1109/TGRS.2007.895835.

Bromley, J., J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah. 1993. “Signature Verification Using a “Siamese” Time Delay Neural Network.” *International Journal of Pattern Recognition and Artificial Intelligence* 7 (4): 669–688. doi:10.1142/S0218001493000339.

Chen, W., B. Du, X. Cui, and L. Zhang. 2017. “A Post-classification Change Detection Method Based on Iterative Slow Feature Analysis and Bayesian Soft Fusion.” *Remote Sensing of Environment* 199: 241–255. doi:10.1016/j.rse.2017.07.009.

Chen, G., G. J. Hay, L. M. T. Carvalho, and M. A. Wulder. 2012. “Object-based Change Detection.” *International Journal of Remote Sensing* 33 (14): 4434–4457. doi:10.1080/01431161.2011.648285.

Chen, H., Z. Qi, and Z. Shi. 2022. “Remote Sensing Image Change Detection With Transformers.” *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–14. doi:10.1109/TGRS.2021.3095166.

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

Chen, J., Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li. 2021. “DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14: 1194–1206. doi:10.1109/JSTARS.2020.303793.

Daudt, R. C., B. L. Saux, and A. Boulch. 2018. “Fully Convolutional Siamese Networks for Change Detection.” 2018 25th IEEE International Conference on Image Processing (ICIP) Athens, Greece, 4063–4067. doi: 10.1109/ICIP.2018.8451652.

Deng, J. S., K. Wang, Y. H. Deng, and G. J. Qi. 2008. “PCA-based Land-use Change Detection and Analysis Using Multitemporal and Multisensor Satellite Data.” International Journal of Remote Sensing 29 (16): 4823–4838. doi:10.1080/01431160801950162.

Fang, S., L. Kaiyu, J. Shao, and L. Zhe. 2021. “SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images.” IEEE Geoscience and Remote Sensing Letters 19: 1–5. doi:10.1109/LGRS.2021.3056416.

Gil-Yepes, J. L., L. A. Ruiz, J. A. Recio, Á. Balaguer-Beser, and T. Hermosilla. 2016. “Description and Validation of a New Set of Object-based Temporal Geostatistical Features for Land-use/Land-cover Change Detection.” ISPRS Journal of Photogrammetry and Remote Sensing 121: 77–91. doi:10.1016/j.isprsjprs.2016.08.010.

Guo, Z., and D. Shihong. 2017. “Mining Parameter Information for Building Extraction and Change Detection with Very High-Resolution Imagery and GIS Data.” GIScience & Remote Sensing 54 (1): 38–63. doi:10.1080/15481603.2016.1250328.

Guo, M.-H., X. Tian-Xing, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, M. Tai-Jiang, S.-H. Zhang, R.-R. Martin, M.-M. Cheng, and H. Shi-Min. 2022a. “Attention Mechanisms in Computer Vision: A Survey.” Computational Visual Media 8 (3): 331–368. doi:10.1007/s41095-022-00271-y.

Guo, Q., J. Zhang, S. Zhu, C. Zhong, and Y. Zhang. 2022b. “Deep Multiscale Siamese Network with Parallel Convolutional Structure and Self-Attention for Change Detection.” IEEE Transactions on Geoscience and Remote Sensing 60: 1–12. doi:10.1109/TGRS.2021.3131993.

Han, M., C. Zhang, and Y. Zhou. 2018. “Object-Wise Joint-Classification Change Detection for Remote Sensing Images Based on Entropy Query-By Fuzzy ARTMAP.” GIScience & Remote Sensing 55 (2): 265–284. doi:10.1080/15481603.2018.1430100.

Huang, H., L. Lin, R. Tong, H. Hongjie, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and W. Jian. 2020. “Unet 3+: A Full-Scale Connected U-Net for Medical Image Segmentation.” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Barcelona, Spain, 1055–1059. doi:10.1109/ICASSP40776.2020.9053405.

Huang, J., X. Zhang, Y. Sun, and Q. Xin. 2021. “Attention-Guided Label Refinement Network for Semantic Segmentation of Very High Resolution Aerial Orthoimages.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14: 4490–4503. doi:10.1109/jstares.2021.3073935.

Hussain, M., D. Chen, A. Cheng, H. Wei, and D. Stanley. 2013. “Change Detection from Remotely Sensed Images: From Pixel-based to Object-based Approaches.” ISPRS Journal of Photogrammetry and Remote Sensing 80: 91–106. doi:10.1016/j.isprsjprs.2013.03.006.

Ji, S., S. Wei, and M. Lu. 2019. “Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set.” IEEE Transactions on Geoscience and Remote Sensing 57 (1): 574–586. doi:10.1109/TGRS.2018.2858817.

Jun, F., J. Liu, H. Tian, L. Yong, Y. Bao, Z. Fang, and L. Hanqing. 2019. “Dual Attention Network for Scene Segmentation.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), California, America, 3141–3149. doi: 10.1109/CVPR.2019.00326.

Kaiyu, L., Z. Zhe, and S. Fang. 2020. “Siamese Nested UNet Networks for Change Detection of High Resolution Satellite Image.” 2020 International Conference on Control, Robotics and Intelligent System, Xiamen, China, 42–48. doi: 10.1145/3437802.3437810.

Khelifi, L., and M. Mignonnet. 2020. “Deep Learning for Change Detection in Remote Sensing Images: Comprehensive Review and Meta-analysis.” IEEE Access 8: 126385–126400. doi:10.1109/ACCESS.2020.3008036.

Krizhevsky, A., I. Sutskever, G. Hinton, and E. P. Simoncelli. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” Advances in Neural Information Processing Systems 25 (2): 3113–3121. doi:10.1145/3065386.

Lebedev, M. A., Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis. 2018. “Change Detection in Remote Sensing Images Using Conditional Adversarial Networks.” ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2: 565–571. doi:10.5194/isprs-archives-XLII-2-565-2018.

Lee, C. Y., S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. 2014. “Deeply-Supervised Nets.” Arxiv Eprint Arxiv 562–570. doi:10.48550/arXiv.1409.5185.

Lei, M., Y. Liu, X. Zhang, Y. Yuanxin, G. Yin, and B. Alan Johnson. 2019. “Deep Learning in Remote Sensing Applications: A Meta-analysis and Review.” ISPRS Journal of Photogrammetry and Remote Sensing 152: 166–177. doi:10.1016/j.isprsjprs.2019.04.015.

Lin, T.-Y., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. 2017. Feature Pyramid Networks for Object Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, America, 936–944. doi:10.1109/CVPR.2017.106.

Liu, S., L. Bruzzone, F. Bovolo, M. Zanetti, and P. Du. 2015. “Sequential Spectral Change Vector Analysis for Iteratively Discovering and Detecting Multiple Changes in Hyperspectral Images.” IEEE Transactions on Geoscience and Remote Sensing 53 (8): 4363–4378. doi:10.1109/TGRS.2015.2396686.

Liu, J., K. Chen, G. Xu, X. Sun, M. Yan, W. Diao, and H. Han. 2020. “Convolutional Neural Network-Based Transfer Learning for Optical Aerial Images Change Detection.” IEEE Geoscience
and Remote Sensing Letters 17 (1): 127–131. doi:10.1109/LGRS.2019.2916601.

Liu, R., M. Kuffer, and C. Persello. 2019. “The Temporal Dynamics of Slums Employing a CNN-Based Change Detection Approach.” Remote Sensing 11 (23): 2844. doi:10.3390/rs11232844.

Liu, X., and R. G. Lathrop Jr. 2002. “Urban Change Detection Based on an Artificial Neural Network.” International Journal of Remote Sensing 23 (12): 2513–2518. doi:10.1080/01431160110097240.

Liu, M., Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang. 2022. “Super-Resolution-Based Change Detection Network with Stacked Attention Module for Images with Different Resolutions.” IEEE Transactions on Geoscience and Remote Sensing 60: 1–18. doi:10.1109/TGRS.2021.3091758.

Long, J., E. Shelhamer, and T. Darrell. 2015. “Fully Convolutional Networks for Semantic Segmentation.” IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (4): 640–651. doi:10.1109/CVPR.2015.7298965.

Loschilov, I., and F. Hutter. 2017. “Decoupled Weight Decay Regularization.” arXiv preprint arXiv:1711.05101. doi:10.48550/arXiv.1711.05101.

Löw, F., D. Dimov, S. Kenjabaev, S. Zaitov, G. Stulina, and V. Dukhovny. 2022. “Land Cover Change Detection in the Aralkum with Multi-Source Satellite Datasets.” GIScience & Remote Sensing 59 (1): 17–35. doi:10.1080/15481603.2021.2009232.

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

Pielke, S., A. Roger, P. Andy, N. Dev, M. Rezaul, M. Clive, H. Faisal, et al. 2011. “Land use/Land Cover Changes and Climate: Modeling Analysis and Observational Evidence.” Wiley Interdisciplinary Reviews: Climate Change 2 (6): 828–850. doi:10.1002/wcc.144.

Rensink, R. A. 2002. “Change Detection.” Annual Review of Psychology 53 (1): 245–277. doi:10.1146/annurev.psych.53.100901.135125.

Richards, J. A. 1984. “Thematic Mapping from Multitemporal Image Data Using the Principal Components Transformation.” Remote Sensing of Environment 16 (1): 35–46. doi:10.1016/0034-4257(84)90025-7.

Rogam, J., and D. Chen. 2004. “Remote Sensing Technology for Mapping and Monitoring Land-cover and Land-use Change.” Progress in Planning 61 (4): 301–325. doi:10.1016/S0305-9006(03)00066-7.

Ronneberger, O., P. Fischer, and T. Brox. 2015. “UNet: Convolutional Networks for Biomedical Image Segmentation.” Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 234–241. doi:10.1007/978-3-319-24574-4_28.

Simonyan, K., and A. Zisserman. 2014. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” arXiv preprint arXiv:1409.1556. doi:10.48550/arXiv.1409.1556.

Singh, A. 1989. “Review Article Digital Change Detection Techniques Using Remotely-Sensed Data.” International Journal of Remote Sensing 10 (6): 989–1003. doi:10.1080/0143169890893939.

Tan, K., X. Jin, A. Plaza, X. Wang, L. Xiao, and P. Du. 2016. “Automatic Change Detection in High-Resolution Remote Sensing Images by Using a Multiple Classifier System and Spectral–Spatial Features.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9 (8): 3439–3451. doi:10.1109/JSTARS.2016.2541678.

Walter, V. 2004. “Object-based Classification of Remote Sensing Data for Change Detection.” ISPRS Journal of Photogrammetry and Remote Sensing 58 (3): 225–238. doi:10.1016/j.isprsjprs.2003.09.007.

Wang, D., X. Chen, M. Jiang, D. Shuhan, X. Bijie, and J. Wang. 2021. “ADS-Net: An Attention-Based Deeply Supervised Network for Remote Sensing Image Change Detection.” International Journal of Applied Earth Observation and Geoinformation 101: 102348. doi:10.1016/j.jag.2021.102348.

Woo, S., J. Park, J.-Y. Lee, and I. So Kweon. 2018. “Cbam: Convolutional Block Attention Module.” Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 3–19. doi:10.1007/978-3-030-01234-2_1.

Zhang, Y., F. Lei, L. Ying, and Y. Zhang. 2021b. “HDFNet: Hierarchical Dynamic Fusion Network for Change Detection in Optical Aerial Images.” Remote Sensing 13 (8): 1440. doi:10.3390/rs13081440.

Zhang, X., Y. Yue, W. Gao, S. Yun, Q. Su, H. Yin, and Y. Zhang. 2021a. “DiFNet++: A Satellite Images Change Detection Network Based on Unet++ and Differential Pyramid.” IEEE Geoscience and Remote Sensing Letters 19: 1–5. doi:10.1109/LGRS.2021.3049370.

Zhang, C., P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu. 2020. “A Deeply Supervised Image Fusion Network for Change Detection in High Resolution Bi-temporal Remote Sensing Images.” ISPRS Journal of Photogrammetry and Remote Sensing 166: 183–200. doi:10.1016/j.isprsjrs.2020.06.003.

Zheng, Z., Y. Wan, Y. Zhang, S. Xiang, D. Peng, and B. Zhang. 2021. “CLNet: Cross-layer Convolutional Neural Network for Change Detection in Optical Remote Sensing Imagery.” ISPRS Journal of Photogrammetry and Remote Sensing 175: 247–267. doi:10.1016/j.isprsjrs.2021.03.005.

Zhou, Z., M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. 2020. “UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation.” IEEE Transactions on Medical Imaging 39 (6): 1856–1867. doi:10.1109/TMI.2019.2959609.

Zhu, X. X., D. Tula, L. Mou, G. S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. 2018. “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources.” IEEE Geoscience & Remote Sensing Magazine 5 (4): 8–36. doi:10.1109/MGRS.2017.2762307.