Engineering Applications of Artificial Intelligence 133 (2024) 108446

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Change detection on multi-sensor imagery using mixed interleaved group convolutional network

Kun Tan [a,b], Moyang Wang [a,b], Xue Wang [a,b,*], Jianwei Ding [c], Zhaoxian Liu [c], Chen Pan [d], Yong Mei [e]

[a] Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai, 200241, China
[b] Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), East China Normal University, Shanghai, 200241, China
[c] The Second Surveying and Mapping Institute of Hebei, Shijiazhuang, 050037, China
[d] Shanghai Municipal Institute of Surveying and Mapping, Shanghai, 200063, China
[e] Institute of Defense Engineering, AMS, Beijing, 100036, China

ARTICLE INFO

Keywords:
Change detection algorithms
Multi-loss supervision
Mixed convolution
Interleaved group convolution
Multi sensor imagery

ABSTRACT

The difference of the spatio-spectral features of multi-sensor image causes big difficulty in change detection because of the difficulties of the feature extraction. Unlike the traditional approaches that mainly relying on manually feature design, the advances of deep learning-based methods in deep feature extraction provide new alternatives for multi-sensor imagery change detection. Specifically, the incorporation of multi-scale information from remote sensing images holds paramount importance in change detection, consistently applied in the design of various deep learning models. This study investigated a change detection approach utilizing a mixed interleaved group convolutional network (MIGCNet) on multi-sensor remote sensing imagery, with a specific focus on fine-grained kernel space and multi-scale feature analysis within convolution operations. The proposed MIGCNet, with parallel branches as the fundamental architecture, can distinguish the change information effectively by the proposed mixed interleaved group convolution (MIGC) module, which combined mixed convolution with interleaved group convolution. Meanwhile, multi-loss supervision is utilized to promote the performance of the proposed MIGCNet. Experimental results demonstrate the outperformance of the MIGCNet to handle change detection with multi-sensor images on urban area. Considering different datasets, the Overall Accuracy and Kappa Coefficient are reaching 0.97 and 80.67%, respectively, and the miss detection rate and the false alarm rate are as low as 0.17 and 0.18, respectively.

1. Introduction

Remote sensing is a technology that enables the characterization of the Earth's surface through both active and passive methods, allowing observation without direct contact (Hemati et al., 2021). The unparalleled and rapid progress in sensor technology has provided a great impetus for the use of remote sensing applications in various fields (Ban and Yousif, 2016). As an important research direction, change detection (CD) aims to extract differences in land cover features through multi-temporal observations (Lu et al., 2004; Seydi and Hasanlou, 2021; Singh, 1989). Numerous applications have widely applied based on CD technology, such as nature disaster assessment (Brunner et al., 2010), urban growth monitoring (Xiao et al., 2016), and land-cover investigation (Mubea and Menz, 2012).

Most of the investigations have been focused on remote sensing CD from single data source, however, the ability for a single sensor to acquire the target scene periodically is tightly bound by the satellite revisiting period and imaging quality (Wang et al., 2019; Wu et al., 2013), while collaborative observation using multiple sensors provides high-frequency, multi-modal remote sensing imagery, which fulfills the requirement for massive high-quality data in land-cover change monitoring. The use of different sensors increases the probability of being able to interpret the study area under cloudless conditions, which ensures the possibility and accuracy of CD (Zhao et al., 2017).

Decades of research have been dedicated to change detection. Traditional CD methods can be divided into three main categories

* Corresponding author. Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai, 200241, China.
E-mail address: wx_ecnu@yeah.net (X. Wang).

https://doi.org/10.1016/j.engappai.2024.108446
Received 10 February 2023; Received in revised form 5 February 2024; Accepted 11 April 2024
Available online 18 April 2024
0952-1976/

(Zhang et al., 2020): 1) image arithmetic operation-based approaches, such as change vector analysis (Chen and Chen, 2016), refer to separate changed pixels from unchanged pixels by determining an appropriate threshold; 2) image transformation-based methods, for instance, principal component analysis (PCA) (Celik, 2009), transform the image spectral combinations into a specific feature space to identify the changed pixels; and 3) image classification based-methods. In particular, with the widespread use of machine learning methods, the accuracy of CD methods based on classification algorithms has been considerably improved (Li et al., 2018). Early methods mainly use independent pixels as the detection units for capturing the different characteristics by exploring the pixel spectral differences. With the development of high spatial resolution remote sensing imagery, object-oriented image analysis (OBIA) technology has been introduced into high-resolution remote sensing image change detection (Lv et al., 2020). Compared to the traditional pixel-based CD methods, OBIA enhances the detection process by incorporating information from image objects. It takes into account spectral, textural, and geometrical features of pixels, effectively suppress the influence of salt-and-pepper noise (Chen et al., 2014; Hussain et al., 2013; Lv et al., 2020; Tan et al., 2019). For example, an object-oriented method with uncertainty analysis was investigated by (Hao et al., 2016) to detect the changes in high-resolution images.

However, due to the different data distributions, it is difficult to compare the multi-sensor images directly in the original low-dimensional feature space. As a result, change detection on multi-sensor remote sensing images is more challenging than single-sensor images change detection. In recent years, deep learning-based methods have been developed in various fields, such as image analysis (Hong et al., 2023), point cloud registration (Wu et al. 2022a, 2022b), especially in CD of remote sensing images (Wang et al., 2023). DL-based CD methods can learn the latent relationships between single or multi sensor images with the powerful capability of deep feature extraction (Habibollahi et al., 2022; Wang et al., 2021), such as multi-dimensional convolution neural network (CNN) (Seydi et al., 2020), ACE-Net (Luppino et al., 2021), and Y-Net (Wang et al., 2022a). The deep learning model unifies the images from originally different domains based on their depth characteristics, facilitating convenient comparisons (Andresini et al., 2023; Wu et al., 2021b; Yuan et al., 2021; Zhang et al., 2018). Recently, a novel network architecture known as the Transformer has been introduced and applied to the change detection task, primarily owing to its attention mechanism (Bandara and Patel, 2022; Wang et al., 2022b). Furthermore, the graph neural network (GNN), characterized by its unique graph structure, proves to be well-suited for handling data with intricate spatial relationships and has also been explored for processing remote sensing data (Wang et al., 2024; Zhou et al., 2023).

For multi-sensor image CD, Liu et al. adopted a deep convolutional coupling network for the change detection using radar and optical images (Liu et al., 2016), while Wang et al. (2020) proposed a hybrid convolutional module to detect changes on multi-sensor optical images. These proposed network architectures were implemented by leveraging the concept of using blocks of layers as structural units and incorporating multi-path information processing. Seydi et al. (2020) designed a CNN-based CD network which composed of parallel channels for exploiting the spatial and channel information. These channels contain three parts: the first and second channels extract deep feature on two temporal images, and the third channel obtains change information on differencing and staking imagery. Recently, building change detection on multi-sensor remote sensing images was implemented using a deep learning-based framework that incorporates multi-feature fusion (Li et al., 2023).

In fact, one of the main challenges in processing multi-sensor images comes from the different feature representation of ground objects in different types of images, increasing the difficulty of obtaining difference maps. The extraction and application on multi-scale features have been proved to be one of the effective ways to obtain representative differences (He et al., 2016; Khan et al., 2020; Szegedy et al., 2015). The concept of branching within a layer was first utilized in the Inception module for extracting multi-scale features (Srivastava et al., 2015; Szegedy et al., 2015). Examples include the use of an asymmetric Siamese neural network for learning semantic changes (Yang et al., 2020) and the design of an unit for extracting multi-scale features in the same layer (Chen et al., 2019). Additionally, Wu et al. (2021a) adopted the graph convolutional network for CD combining with the multiscale object-based technique. This method also employed the idea of multi-scale feature extraction for improving accuracy on multi-sensor images CD.

Despite the numerous successes and contributions on the multi-sensor images CD, there are limitations requiring more attention. Firstly, in the most of existing CNN-based CD methods, the extraction of multi-scale features depend on multiple convolution layers with distinct convolution kernel sizes (Tan and Le, 2019). However, this approach often tends to increase the size of the network, and finer-grained kernel spaces within the same convolution are frequently overlooked. As a result, it is essential to reduce the computational costs without accuracy reduction. Consequently, the recent CD methods tend to lose crucial change information, including the diversity of the scale features and the difference characteristics which will cause the low change detection accuracy. Lastly, the back-propagation training method can give rise to issues such as vanishing or exploding gradients in neural network learning. In general, the longer the error back-propagation distance is, the smaller the gradient in the early layers is, which potentially causes an unstable gradient problem.

Regarding the widely application of land resource monitoring in the era of multi-source data, the utilization of a deep convolutional neural network (DCNN) is explored in this paper to for change detection with multi-sensor high-resolution remote sensing imagery. To handling the multi-scale feature learning and the optimization of the convolution, an innovative Siamese network architecture and supervised training method are also introduced. On this basis, the deep learning-based framework called MIGCNet is proposed to detect the differences with the two periods images acquired by multi-sensor. By combining mixed convolution (Tan and Le, 2019) with interleaved group convolution (Zhang et al., 2017), the mixed interleaved group convolution (MIGC) module can extract abundant image difference features for determining the changes. Meanwhile, multi-loss supervision is utilized to promote the network performance. The main contributions of this work are given as follows.

(1) We investigate a novel MIGCNet for multi-sensor images CD. The network obtains different scale information at a fine-grained level and integrate the feature representation with multi scales, and the detailed change information is well kept for the final detection.
(2) We propose a novel MIGC module, employed by mixing multiple convolution kernel sizes in one convolution operation, which contributes to the multi-scale convolutional features learning without any expanding.
(3) We devise a multi-loss supervision training strategy using different parameter optimizers to handling the hard training issues of deep neural network.
(4) From comprehensive comparisons among the multi-sensor images change detection approaches, our proposed method can achieve state-of-the-art performance.

2. Materials and methods

To addressing two time-domain images collected by different sensors, this study initiates optimization at the fine-grained convolution level to attain multi-scale feature extraction within a unified convolutional framework. A novel deep learning network called mixed interleaved group convolutional network (MIGCNet) has been proposed to

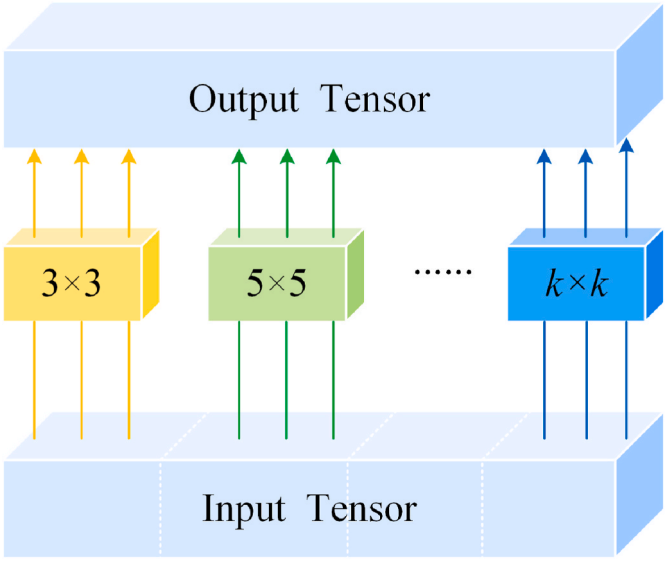

Fig. 1. The framework of mixed depthwise convolution.

extract the changes from multi-sensor images. The network comprises two key components: feature extraction and multi-loss supervision. In this section, the proposed model architecture will be illustrated firstly.

2.1. Model structure

2.1.1. Mixed interleaved group convolution module

Convolution kernels are the main way that feature extraction is implemented in CNNs. Conventional convolution kernels usually use the same kernel size to extract the feature map. We let $X^{(h,w,c)}$ denote the input image data with shape (h, w, c), where c, w and h refer to the channel size, the width and height, respectively. We then assume that the convolution kernel has the same height and width k. $W^{(k,c,m)}$ denotes a convolution kernel with input channel size c. $k \times k$ represents the kernel size, and m is the channel multiplier. Then, each output feature map value can be calculated as:

$$Y_{a,b,t} = \sum_{-\frac{k}{2}\le i\le\frac{k}{2},-\frac{k}{2}\le j\le\frac{k}{2}} X_{a+i,b+j,\frac{t}{m}} \bullet W_{i,j,t}, \tag{1}$$

$$\forall t = 1, \cdots, c \bullet m$$

where $c \bullet m$ denotes output size of the output tensor $Y^{(h,w,c\bullet m)}$, which has the same shape (h, w) as the input.

Differing from conventional convolution, mixed convolution (Tan and Le, 2019) partitions the input into several groups, and applies convolution kernels of different sizes to each group. Fig. 1 shows the mixed depthwise convolution. Specifically, the input tensor is partitioned into l groups of independent tensors according to the channel, which are denoted as $\{X^{(h,w,c_1)}, \cdots, X^{(h,w,c_l)}\}$. The sum of channel sizes is equal to the original input tensor, i.e., $c = c_1 + \cdots + c_l$ .We also partition the original convolution kernels into l groups, expressed as $\{W^{(k_1,c_1,m)}, \cdots, W^{(k_l,c_l,m)}\}$, where each kernel for the different groups has a different size. The output of the input tensor in s-th group is calculated as:

$$Y^s_{a,b,t} = \sum_{-\frac{k}{2}\le i\le\frac{k}{2},-\frac{k}{2}\le j\le\frac{k}{2}} X^s_{a+i,b+j,t/m} \bullet W^s_{i,j,t}, \tag{2}$$

$$\forall t = 1, \cdots, c_s$$

By concatenating all of the output tensor, the final output tensor can be calculated as:

$$Y_{a,b,t_0} = Concat\left(Y^1_{a,b,t_1}, \cdots, Y^l_{a,b,t_l}\right) \tag{3}$$

where $t_0 = t_1 + t_2 + \cdots + t_l$ is the final output channel size.

The MixConv operation can easily obtain different scale information at a fine-grained level by mixing multiple convolution kernels in a single convolution. However, the large convolution kernel size requires more parameters and has a higher computational cost. Thus, we introduced interleaved group convolution (IGC) (Zhang et al., 2017) to mixed convolution in this study to solve this problem. In IGC, the input channels consist of several partitions. This can not only guarantee the increase of the width under the same depth, but also solves the problem of the computational cost and data redundancy.

The MIGC module is composed of primary mixed convolution part and secondary group convolution part. The input tenser shown in Fig. 2 is first equally divided into two groups, i.e., $l = 2$, and convolution kernels of size [3 × 3, 5 × 5] are performed over each group. Then, the output tensors are concatenated and input into next convolution. The secondary group convolution is performed to shuffle the cross-partition channels output by the primary mixed convolution. After that, the partition x_1 is then connected with y_2, the partition of output tensor, through the MIGC module, as shown by the red line in Fig. 2. There also has a similar path connecting each channel of the secondary group convolution output with each channel of the primary mixed convolution input.

Under the same computational complexity, the MIGC module has the bigger receptive field comparing with a conventional convolution module. We assume that 1 and $S = k \times k$ is the group convolution and the kernel size in the primary mixed convolution, respectively.

$$T_{IGC} = L \times M \times M \times S + M \times L \times L \tag{4}$$

where L denotes the count of partitions in the primary group convolution. M is the count of channels in each partition. We let $G = M \times L$ denotes the width of an IGC block. Equation (4) can then be calculated as:

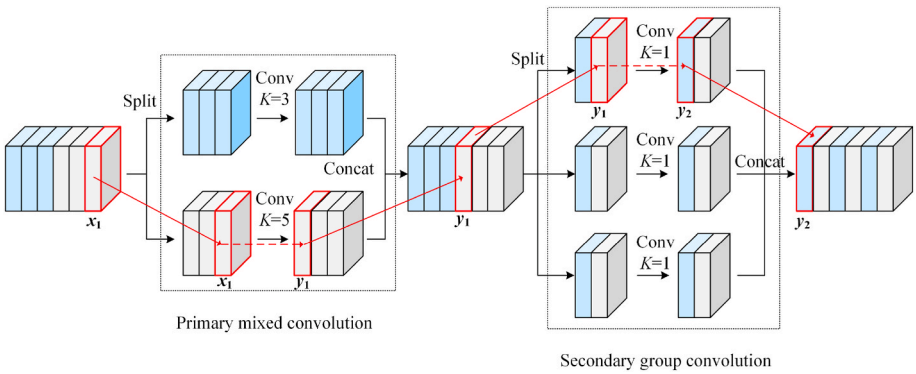

Fig. 2. The framework of mixed interleaved group convolution.

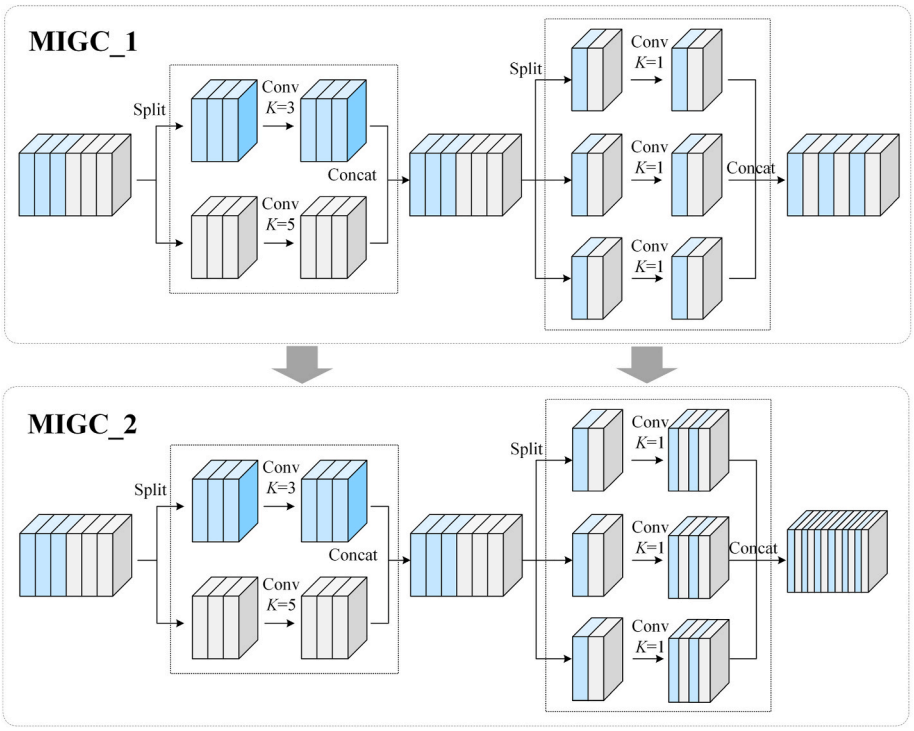

Fig. 3. The framework of the MIGC module.

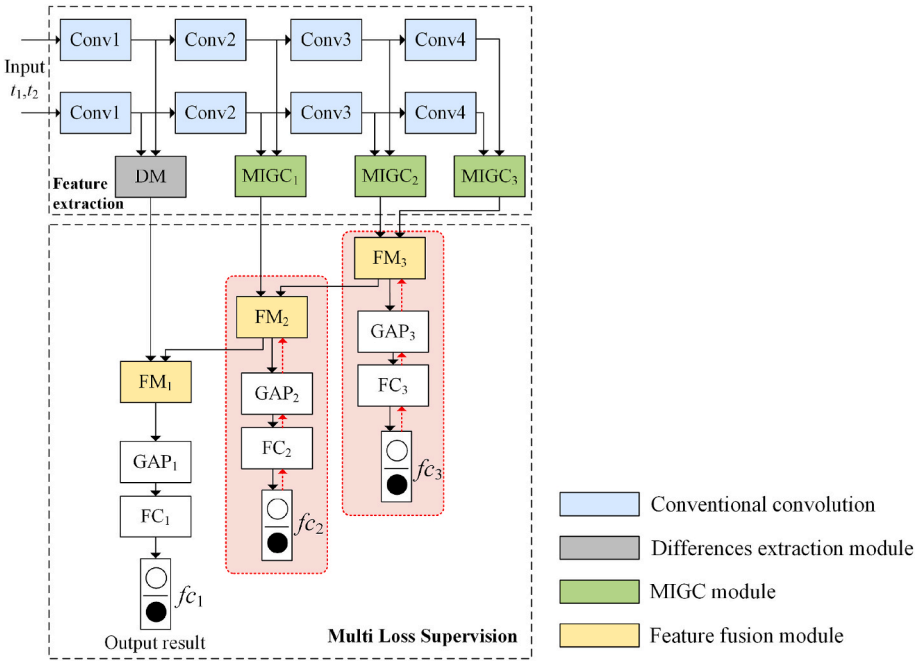

Fig. 4. The proposed MIGCNet architecture. The different modules are marked by different colors. Block color legend: blue represents conventional convolution with kernel size 3 × 3, gray represents the difference extraction module (DM), green represents the proposed MIGC module, and yellow represents the feature fusion module (FM).

$$T_{IGC} = G^2 \times (S/L + 1/M) \tag{5}$$

The number of parameters of a traditional convolution for a single spatial position is:

$$T_{norm} = C \times C \times S \tag{6}$$

where C represents the count of channels. Given the same size of parameters, i.e., $T_{IGC} = T_{norm} = T$, we have $G^2 = T/(S/L + 1/M)$, $C^2 = T/S$. It is therefore easy to obtain:

$$G > C, when\ L/L - 1 < M \times S \tag{7}$$

In a typical case, e.g., $S = 3 \times 3$, we have $G > C$ when $L > 1$. As for the MIGC block, i.e., $L = 2$, $S_1 = 3 \times 3$, and $S_2 = 5 \times 5$. when $L > 1$, we can easily obtain $G > C$. As shown in Fig. 3, we adopt two successive MIGC blocks to compose a complete MIGC module.

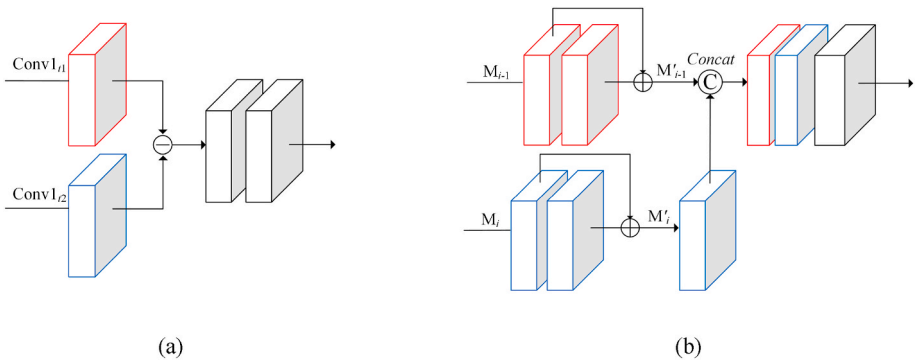

Fig. 5. The framework map of (a) the difference extraction module (DM), (b) the feature fusion module (FM).

2.1.2. Network architecture

Based on the MIGC module, MIGCNet has been proposed to tackle with the change detection with multi-sensor high-resolution images. Fig. 4 illustrated the structure of the MIGCNet framework. MIGCNet had two inputs, and each input patch was fed into an equal stream in the first half of the network. Each stream consists of four groups of conventional convolutions (colored in blue in Fig. 4). After progressive abstraction through stacked convolutional layers, the deepest layer in streams T1 and T2 obtained compact local information.

The outputs of the last three convolution groups (i.e., conv2, conv3, conv4) in streams T1 and T2 are then fed into the MIGC module (colored in green in Fig. 4), in turn, to extract abundant features. Meanwhile, the output of the first convolution group is input into the difference extraction module (DM, colored in gray in Fig. 4, and shown in Fig. 5), to obtain the initial difference features for fusion with the other high-dimensional features in later modules. The DM module utilizes convolution after subtraction to extract the low-level difference information. The detailed components of the proposed feature fusion module (FM) are shown in Fig. 5(b). Given the features of two adjacent MIGC modules, M_{i-1} and M_i, the FM first reduced the dimension of each feature by stacking the convolutions to facilitate efficient training. $M_i^{'}$ is then concatenated with $M_{i-1}^{'}$ after residual connection, and applies subsequent additional convolutions. Combining the two features in different modalities through FM complementarily is crucial for the feature extraction of the network.

MIGCNet applies a global average pooling layer (GAP) after the FM module for avoiding overfitting. The results are then generated after the following fully connected layer. Moreover, multi-loss supervision is included to enhance the network performance and guarantee that the

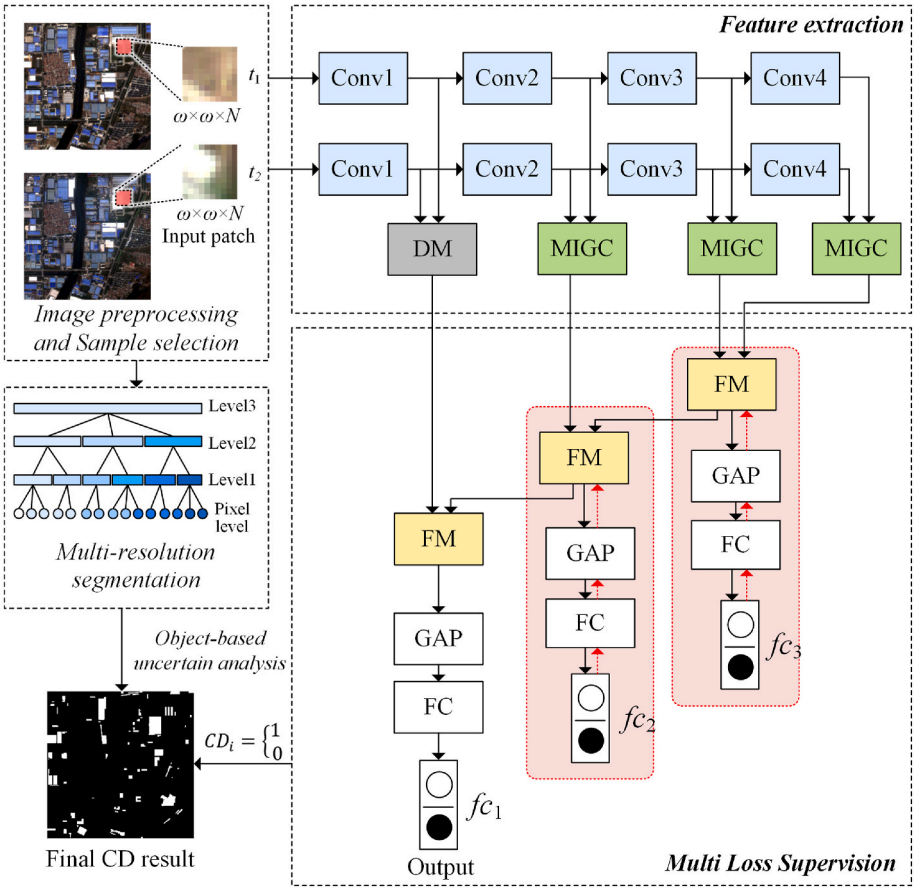

Fig. 6. Flowchart of the change detection framework.

network is well trained. Details of the multi-loss supervision process are presented in the next section.

2.2. Multi-loss supervision

In the field of neural network learning, the discriminative deep features are learned through defining and minimizing the loss function, and then features are used to train the classifier (Cheng et al., 2020). Neural networks tend to achieve architecture innovation with increased depth and width, and the parameter updating depends on the back-propagation algorithm (LeCun et al., 2015). Therefore, the model performance is dependent on the following: 1) the update rate of the different layers having variation; and 2) the update speed of a layer close to the output being faster than that of a layer close to the input. w_{jk}^{l} denotes the weight between layer $l-1$ and next layer l. C denotes the loss function. b_j^l denotes the bias in layer l. The output of the j-th neuron is defined as:

$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right) \tag{8}$$

where σ represents the activation function, and a_k^{l-1} denotes the output from the k-th neuron in layer $l-1$. By rewriting (8) to matrix form, we can obtain the formula: $a^l = \sigma(w^l a^{l-1} + b^l)$. As z_j^l is the weight input of neuron j in layer l, i.e., $z_j^l = w^l a^{l-1} + b^l$, (8) can be calculated as:

$$a_j^l = \sigma\left(z_j^l\right) \tag{9}$$

By using the chain rule to calculate the partial derivative, the error of the output layer is as follows:

$$\delta_j^l = \frac{\partial C}{\partial z_j^l} = \frac{\partial C}{\partial a_j^l} \bullet \frac{\partial a_j^l}{\partial z_j^l} = \frac{\partial C}{\partial a_j^l} \bullet \sigma'\left(z_j^l\right) \tag{10}$$

where $\sigma'(z_j^l)$ represents the partial derivative of z_j^l by activation function σ. The error of w_{jk}^l can be calculated as follows:

$$\frac{\partial C}{\partial w_{jk}^l} = \frac{\partial C}{\partial z_j^l} \bullet \frac{\partial z_j^l}{\partial w_{jk}^l} = \frac{\partial C}{\partial z_j^l} \bullet a_k^{l-1} = \delta_j^l \bullet a_k^{l-1} \tag{11}$$

As can be seen in (11), when the activation output of the upper layer approaches zero, no matter how large the error is, $\partial C/\partial w$ has smaller value which will create a smaller gradient. Therefore, the back-propagation training method may lead to the problem of vanishing or exploding gradients (Zhang et al., 2020).

To handling the problem of gradient vanishment of MIGCNet, we introduced the multi-loss supervision method to train the difference identification layers effectively. In this context, the middle layer of the network does not solely depend on gradients gradually backpropagating from the output layer; instead, it is supervised by distinct parameter optimizers. This direct feedback from the change outcome enables the middle layer to generate features that exhibit greater differentiation within the region of change. As illustrated in the red box in Fig. 4, throughout the training process, the loss for each depth supervision is independently calculated and directly backpropagated to the middle layer.

2.3. The proposed change detection framework

As shown in Fig. 6, the proposed change detection framework includes three main steps: 1) training sample acquisition; 2) network training; and 3) object-based uncertainty analysis. A pair of bi-temporal images (i.e., the pre-change image T1 and the post-change image T2) is fed into the two parallel streams separately, which allows the original features of each individual bi-temporal image to be preserved as much as possible. The introduction of multi-loss supervision enhances the performance of this network. Meanwhile, the object-based uncertainty analysis is applied to refine the results to the object level. The three steps are described as follows.

2.3.1. Training sample acquisition

The training samples are selected in combination with an automatic analysis process, based on the differences in the multi-feature images. By combining the individual detection results from the spectral and texture features, initial selection of the changed and unchanged pixels is achieved.

Firstly, the Gabor features are constructed in the 0°, 45°, 90°, and 135° directions, with kernel sizes of [7,9,11,13,15,17], for the transform-based texture features. The multi-kernel Gabor features are generated as follows:

$$\boldsymbol{G}_{direction}^{e} = \sum\nolimits_k g_k^e, k \in [7, 9, 11, 13, 15, 17], \tag{12}$$

where k denotes kernel size, g_k^e represents the Gabor features on the e-th spectral band with k, and the original images have E spectral bands. After Equation (8), the $4 \times E$ Gabor texture features are obtained to generate the difference map.

The difference image $\boldsymbol{D}$ is generated from the two temporal images, with the dataset consisting of the spectral features and the Gabor texture features. For the images with r spectral bands at times $\boldsymbol{T}^1$ and $\boldsymbol{T}^2$, $\boldsymbol{D}$ is calculated as follows:

$$\boldsymbol{D} = \left|\boldsymbol{T}^1 - \boldsymbol{T}^2\right|, \tag{13}$$

Each dimension of $\boldsymbol{D}$ must be normalized in the range [0, 1], and the data in the b-th dimensional $\boldsymbol{D}_b$ are normalized as follows:

$$\boldsymbol{D}_b = \frac{\boldsymbol{D}_b - D_{min}}{D_{max} - D_{min}}, b = 1, 2, \ldots, r, \tag{14}$$

Then, Equation (15) is applied to generate the pixel-based results CD^b on each band by the threshold T_b, calculated according to the expectation maximization (EM) algorithm (Bruzzone and Prieto, 2000).

$$cd_{i,j}^b = \begin{cases} 0, if\ d_{i,j}^b < T_b \\ 1, if\ d_{i,j}^b \geq T_b \end{cases}, \tag{15}$$

where $cd_{i,j}^b$ indicates whether the pixel at position (i,j) in CD^b belongs to the unchanged or changed part. In order to select reliable train and valid samples, the uncertainty analysis on each band of CD is considered, and a conservative decision is made as follows:

$$L_{i,j} = \begin{cases} 0, p \leq \lfloor 0.3 \times b \rfloor \\ 1, p \geq \lfloor 0.7 \times b \rfloor \end{cases}, \tag{16}$$

$$p = \sum_{r=1}^{b} cd_{i,j}^b, \tag{17}$$

where $cd_{i,j}^b$ indicates the category of the pixel at (i,j) in CD^b. p is the score that a pixel at position (i,j) is regard as changed in all dimensions. If the score p on the position (i,j) is greater than the threshold $\lfloor 0.7 \times b \rfloor$, then the pixel is labeled as the "changed". Likewise, if p is less than $\lfloor 0.3 \times b \rfloor$, then the pixel is labeled as the "unchanged". Training samples are selected from these augmented samples randomly. Therefore, the inputs in this study are $[patch_1, patch_2, label]$, where $patch_1$ and $patch_2$ represent the patches of a fixed size ω on the two temporal images.

2) Network Training

Patches of the multi-sensor remote sensing images are utilized in MIGCNet to extract the high-level features. We randomly collected 2000

Table 1
Parameter settings of multi-loss supervision.

	Optimization	Layers	Learning Rate
OP_1	Adam	All	$1e-03$
OP_2	Adam	FM_2, GAP_2, FC_2	$1e-03$
OP_3	Adam	FM_3, GAP_3, FC_3	$1e-04$

bi-temporal image pairs for each dataset. Among them, 70% of the sample dataset was used for model training, and the other 30% of the sample dataset was used for the model performance assessment. Adam optimizer (Kingma and Ba, 2014) and cross-entropy loss function was utilized in the network training. Three optimizations, i.e., OP_1, OP_2, and OP_3, were utilized to optimize the different parts parameters of the MIGCNet. Table 1 lists the parameter settings of optimizations.

2.1.3. Object-based uncertainty analysis

Due to the different imaging conditions, images collected by multiple sensors always show great diversity (Wang et al., 2020). Object-oriented change detection (OBCD) can effectively suppress the influence of noise on change detection. In the proposed approach, two temporal images are stacked into a single image by simple band stacking. The fractal net evolution approach (FNEA) is then utilized to over segment the stacked image (Hay et al., 2003; Tan et al., 2019). According to the heterogeneity of the segmented objects, the segmented objects are merged into multiple scales.

The optimal segmentation scale S_l is first obtained according to the global score (GS) value (Espindola et al., 2006; Lu et al., 2017; Tan et al., 2019), and then five segmentation scales [S_{l-2}, S_{l-1}, S_l, S_{l+1}, S_{l+2}] are selected. We combine the pixel-based result obtained by MIGCNet with the segmentation with different scales, and extract the enhanced spatial features of the multi-sensor images and obtain optimized object-based results.

For a segmented object O_i, the number of pixels n_c^i for the changed class is counted. The percentage of object O_i belonging to changed class C is calculated as follows:

$$p_c = \frac{n_c^i}{n_i}, \tag{18}$$

where n_i is the total number of pixels in object O_i. A threshold T is then set and compared with p_c to classify the object O_i of the current segmentation scale. The final change detection result is calculated as follows:

$$CD_i = \begin{cases} 1, if\ p_c > T \\ 0, others \end{cases}, \tag{19}$$

If CD_i satisfies $P_c > T$, the object O_i is labeled as a changed object.

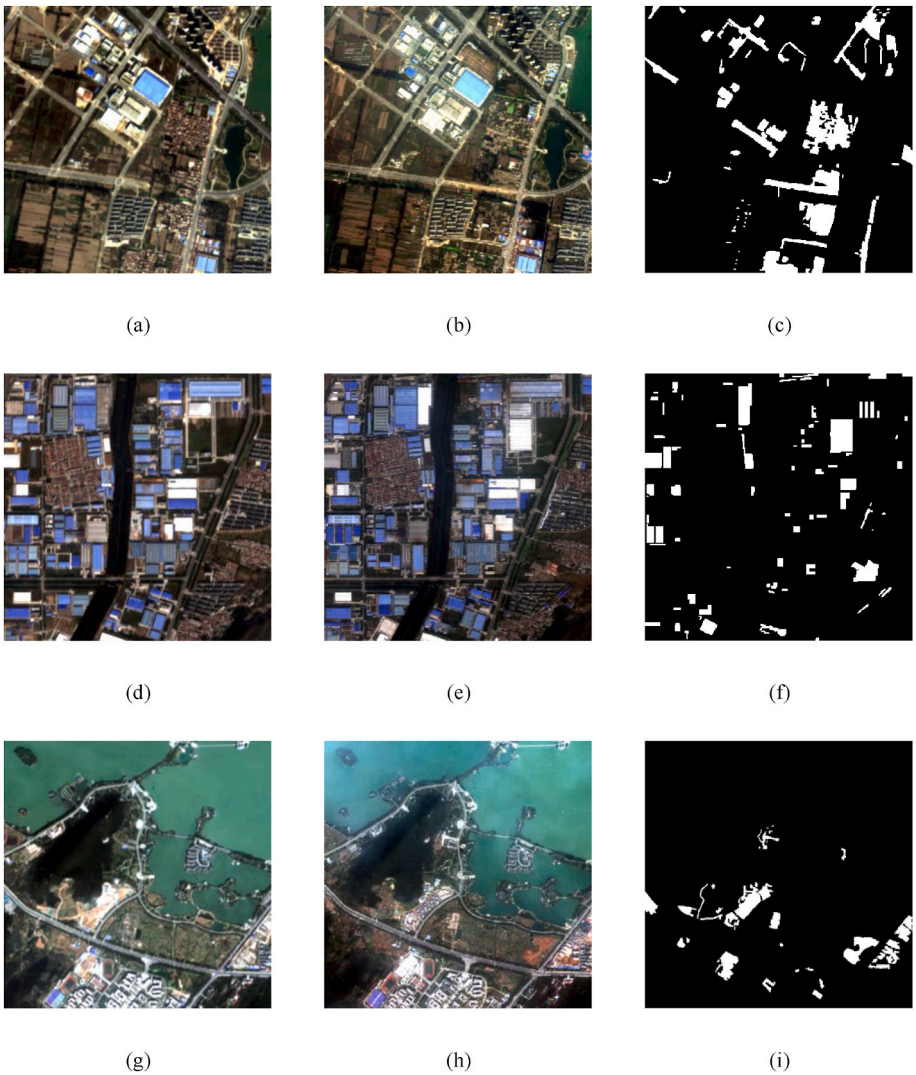

Fig. 7. The image and the corresponding labeled maps for the three datasets.

Table 2
Details of the GF-2 and ZY-3 imagers.

Satellite	Sensor	Band	Spectrum (μm)	Spatial resolution (m)	Time
ZY-3	MUX	Blue	0.45–0.52	5.8	2014.10.14
		Green	0.52–0.59		
		Red	0.63–0.69		
		Nir	0.77–0.89		
GF-2	PMS	Blue	0.45–0.52	4.0	2016.10.05
		Green	0.52–0.59		
		Red	0.63–0.69		
		Nir	0.77–0.89		

$CD_i = 0, 1$ indicates that R_i belongs to the unchanged and changed classes, respectively. The change map is obtained by integrating the segmentation maps and the pixel-wise change detection result through the uncertainty analysis, according to the accuracy evaluation. The pixel-based result obtained by MIGCNet can be refined by an additional constraint on the same object, so as to make better use of the spatial information of the multi-sensor images.

3. Experiments and results

3.1. Dataset description

We evaluated MIGCNet using three urban area datasets. Dataset I covers part of Big Dragon Lake, Xuzhou, China, and the details of this dataset are depicted in Fig. 7(a) and (b). Dataset II covers Tongshan District, Xuzhou China, which is presented in Fig. 7(d) and (e). Dataset III, depicted by Fig. 7(g) and (h), covers Cloud Dragon Lake, Xuzhou, China. The main types of changes in the three multi-sensor datasets are mainly additional building in urban area. The first temporal image was acquired by the ZY-3 satellite taken on October 14, 2014, and the second temporal image was obtained by the GF-2 satellite on October 5, 2016. The key characteristics of these two satellites, e.g., band combination and resolution, are listed in Table 2. Each dataset consisted of 350 × 350 pixels, and all of the images were unified transformed to the same spatial resolution. In the stage of image preprocessing, the geometric registration root-mean-square error (RMSE) was found to be lower than 0.5 pixel. The corresponding labeled maps for the three datasets were generated via manual expert knowledge based on prior knowledge and field investigation, and are shown in Fig. 7(c), (f), and (i).

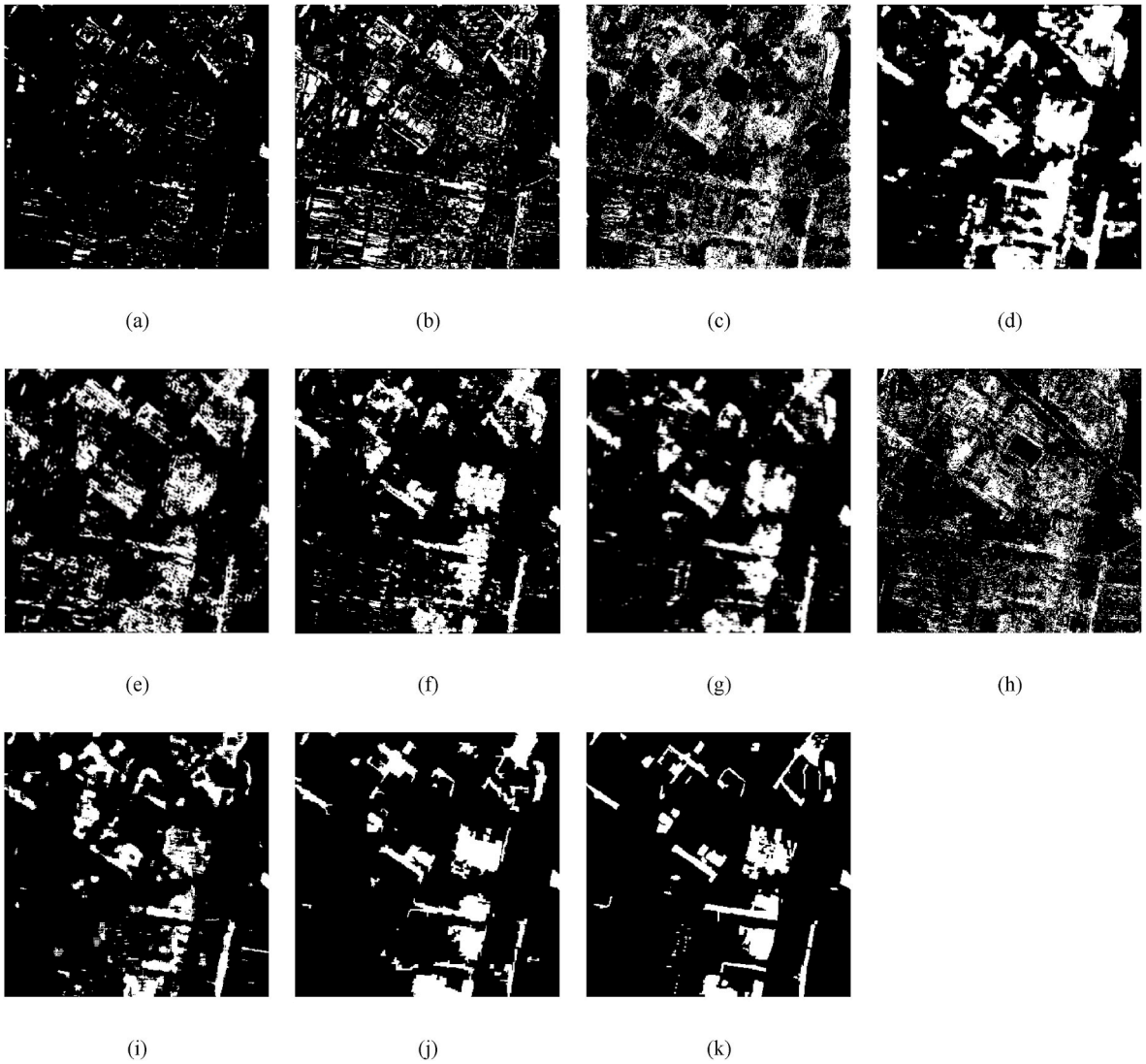

Fig. 8. Detection map on Dataset I by: (a) SVM, (b) MLP, (c) TSCNN, (d) DSMS-CN, (e) DCNN, (f) MixNet, (g) DSCNH (h) ResViT (i) PASSNet and (j) MIGCNet with multi-loss supervision ($l = 45$, $\omega = 11$). (k) Ground truth.

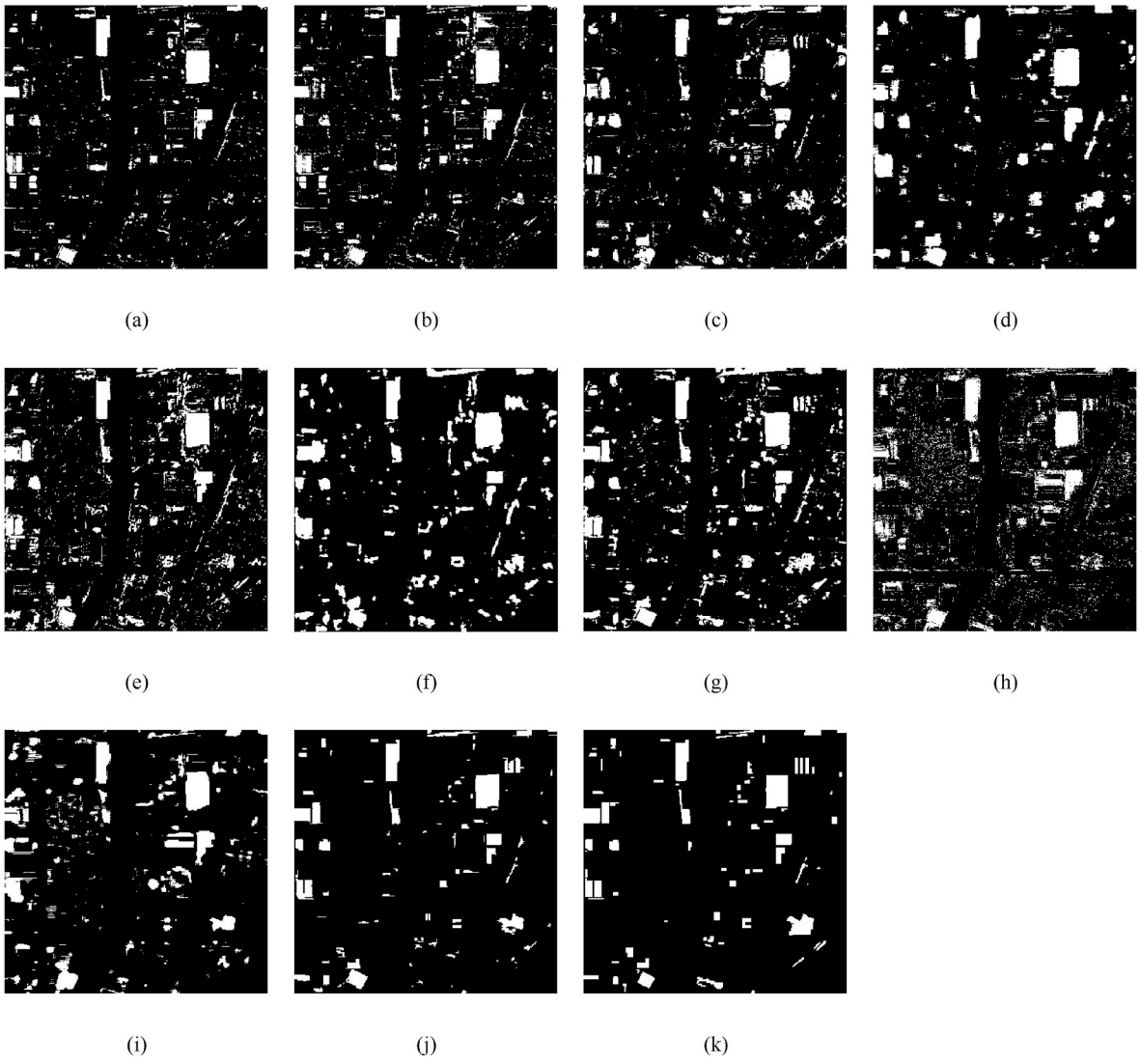

Fig. 9. Detection map on Dataset II by: (a) SVM, (b) MLP, (c) TSCNN, (d) DSMS-CN, (e) DCNN, (f) MixNet, (g) DSCNH (h) ResViT (i) PASSNet and (j) MIGCNet with multi-loss supervision ($l = 30$, $\omega = 9$). (k) Ground truth.

3.2. Benchmark methods

To fairly evaluate the proposed method, the following methods were selected to conduct compared experiments.

1) Supervised pixel-wise change detection methods, i.e., multi-layer perceptron (MLP) and SVM-based method.
2) Change detection based on DCNN. This network is made up of five conventional convolution groups and three fully connected layers. Each conventional convolution group is made up of two convolutional layers, following with batch normalization and activation function. The input is the absolute difference of the paired samples.
3) The approach using the traditional Siamese CNN (TSCNN) (Zhan et al., 2017).
4) The approach using the deep Siamese multi-scale CNN (DSMS-CN) (Chen et al., 2019) which consists of multi-scale feature extraction module.
5) The network is composed of MixConv, as per the method introduced by (Tan and Le, 2019).
6) Change detection based on the deep Siamese convolutional network with hybrid convolutional feature extraction module (DSCNH) (Wang et al., 2020).
7) Change detection based on the visual transformers (ResViT) (Wu et al., 2020).
8) Change detection on a spatial-spectral feature extraction network with patch attention module (PASSNet) (Ji et al., 2023).

In addition, the patch size ω used in the comparison experiments of deep learning methods were settled as same as MIGCNet.

3.3. Accuracy evaluation metrics

To assess the performance of the proposed approach, six indicators are adopted for comparing the detection results with the ground truth: 1) overall accuracy (OA); 2) kappa coefficient; 3) the missing alarm rate (MAR); 4) the false alarm rate (FAR); 5) intersection over union (IoU); 6) F1 Score. Furthermore, we calculated the mean intersection over union (MIoU), which is the average of individual IoU values. These metrics are obtained as follows:

$$OA = \frac{(N_{11} + N_{00})}{(N_{11} + N_{00} + N_{01} + N_{10})} \tag{20}$$

$$Kappa = \frac{N \times (N_{11} + N_{00}) - ((N_{11} + N_{10}) \times (N_{11} + N_{01}) + (N_{01} + N_{00}) \times (N_{10} + N_{00}))}{N^2 - ((N_{11} + N_{10}) \times (N_{11} + N_{01}) + (N_{01} + N_{00}) * (N_{10} + N_{00}))} \quad (21)$$

$$MAR = \frac{N_{01}}{(N_{01} + N_{11})} \quad (22)$$

$$FAR = \frac{N_{10}}{(N_{10} + N_{00})} \quad (23)$$

$$IoU = \frac{N_{11}}{(N_{11} + N_{10} + N_{01})} \quad (24)$$

$$Precision = \frac{N_{11}}{(N_{11} + N_{10})} \quad (25)$$

$$Recall = \frac{N_{11}}{(N_{11} + N_{01})} \quad (26)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (27)$$

where N_{11} presents the count of correctly identified "changed" labels, N_{00} is the count of pixels correctly detected which are unchanged, N_{10} represents the number of missed changed pixels; N_{01} is the number of pixels which are identified as changed in change map while are unchanged in ground reference; and N is the count of all labeled pixels.

3.4. Experimental results

Figs. 8–10 provide a visual representation of the results obtained across the three datasets. For Dataset I, the changed regions mainly comprise new constructions and high-complexity features. As shown in Fig. 8(a), the detection result obtained by SVM on Dataset I contains a considerable number of undetected pixels, which demonstrates the inadequacy of the SVM classifier on multi-sensor images. In change detection, accurately labeling an area as unchanged is imperative, especially when covered by crops. However, change maps generated by

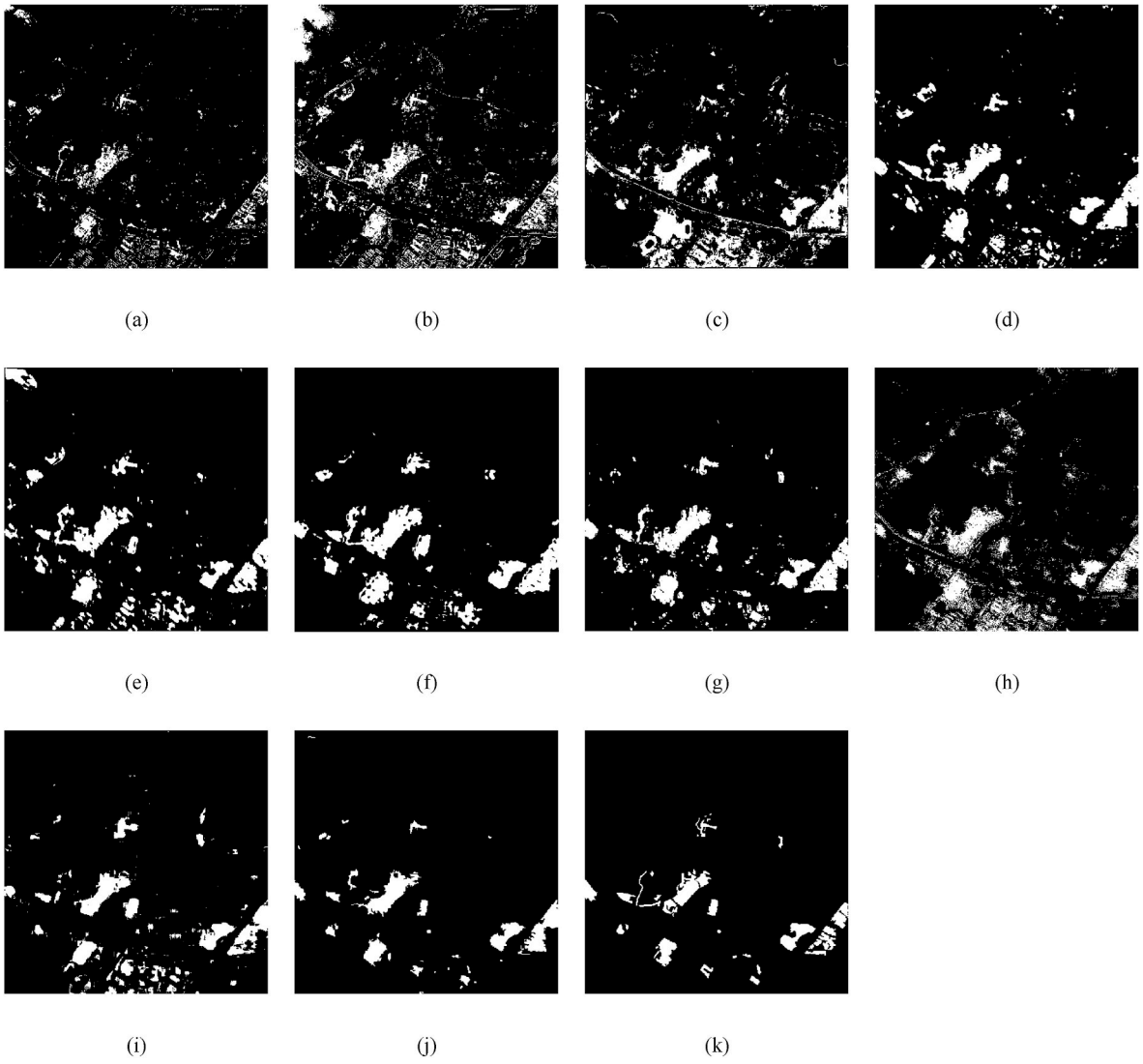

Fig. 10. Detection map on Dataset III by: (a) SVM, (b) MLP, (c) TSCNN, (d) DSMS-CN, (e) DCNN, (f) MixNet, (g) DSCNH (h) ResViT (i) PASSNet and (j) MIGCNet with multi-loss supervision ($\omega = 9$, $l = 30$). (k) Ground truth.

Table 3
Quantitative analysis of the Different Approaches on Dataset I.

Method		SVM	ANN	DCNN	DSCNH	MixNet	DSMS-CN	TSCNN	PASSNet	ResVIT	MIGCNet ($\omega = 11, l = 45$)
IoU	C	0.1584	0.2481	0.4238	0.6124	0.6055	0.5650	0.2892	0.3888	0.2470	**0.6269**
	U	0.8888	0.8329	0.8454	0.9428	0.9321	0.9165	0.7939	0.8850	0.8087	**0.9340**
MIoU		0.5236	0.5404	0.6346	0.7776	0.7688	0.7407	0.5416	0.6369	0.5279	**0.7804**
F1-score		0.2735	0.3976	0.5953	0.7548	0.7544	0.7221	0.4486	0.5600	0.3962	**0.7655**
OA		0.8803	0.8416	0.8783	0.9167	0.9138	0.8959	0.8898	0.8929	0.8200	**0.9419**
Kappa		0.3167	0.3085	0.5074	0.6545	0.6436	0.4925	0.4201	0.4994	0.2978	**0.7303**
MAR		0.5440	0.6523	0.5263	0.4114	0.4208	0.5122	0.5421	0.4789	0.6815	**0.2551**
FAR		0.6729	0.5362	0.2697	**0.1340**	0.1419	0.2625	0.3130	0.3948	0.4754	0.1581

Table 4
Quantitative analysis of the Different Approaches on Dataset II.

Method		SVM	ANN	DCNN	DSCNH	MixNet	DSMS-CN	TSCNN	PASSNet	ResVIT	MIGCNet ($\omega = 9, l = 30$)
IoU	C	0.4687	0.4474	0.4621	0.6332	0.4793	0.4554	0.4684	0.4048	0.3317	**0.6866**
	U	0.9443	0.9344	0.9168	0.9663	0.9284	0.9316	0.9359	0.9234	0.9022	**0.9721**
MIoU		0.7065	0.6909	0.6895	0.7997	0.7039	0.6933	0.7022	0.6641	0.6170	**0.8294**
F1-score		0.6382	0.6182	0.6202	0.7754	0.6480	0.6255	0.6380	0.5763	0.4982	**0.8142**
OA		0.9470	0.9378	0.9268	0.9335	0.9295	0.9353	0.9393	0.9272	0.9067	**0.9748**
Kappa		0.6097	0.5850	0.5805	0.6124	0.6033	0.5913	0.6058	0.5378	0.4500	**0.8067**
MAR		0.3817	0.4528	0.5094	0.4810	0.4982	0.4693	0.4468	0.5091	0.5970	**0.1713**
FAR		0.3405	0.2895	0.1655	0.1425	**0.1202**	0.2385	0.2465	0.3021	0.3473	0.1880

Table 5
Quantitative analysis of the Different Approaches on Dataset III.

Method		SVM	ANN	DCNN	DSCNH	MixNet	DSMS-CN	TSCNN	PASSNet	ResVIT	MIGCNet ($\omega = 9, l = 25$)
IoU	C	0.3468	0.2702	0.4275	0.4995	0.4509	0.9482	0.3645	0.4383	0.3237	**0.6048**
	U	0.9571	0.9116	0.9508	0.9603	0.9534	0.4410	0.9526	0.9562	0.9296	**0.9772**
MIoU		0.6520	0.5909	0.6892	0.7299	0.7021	0.6946	0.6585	0.6973	0.6267	**0.7910**
F1-score		0.5150	0.4254	0.5989	0.6663	0.6215	0.6121	0.5343	0.6095	0.4891	**0.7538**
OA		0.9581	0.9145	0.9491	0.9619	0.9524	0.9502	0.9539	0.9576	0.9319	**0.9780**
Kappa		0.4932	0.3886	0.5692	0.6476	0.5983	0.5889	0.5107	0.5883	0.4579	**0.7423**
MAR		0.5006	0.7041	0.5542	0.4740	0.5343	0.5457	0.5374	0.5034	0.6433	**0.2912**
FAR		0.4683	0.2434	0.1139	0.0915	0.0696	**0.0621**	0.3676	0.2108	0.2221	0.1951

MLP (Fig. 8(b)) and TSCNN (Fig. 8 (c)) exhibit a significant number of falsely detected pixels in the southwest of the image, corresponding to cultivated land. Additionally, in Fig. 8(h), there is a significant presence of misdetected pixels in the results of ResViT. Fig. 8(d) and (g) demonstrate that the siamese convolutional neural network (Siamese CNN) with multi-scale features outperforms in capturing both spatial and spectral information. Fig. 8(e) and (f) present the result maps of the traditional DCNN and MixNet, and it can be seen that the method combined with multi-scale convolution has a good performance which successfully restrains most of the salt-and-pepper noise. The detection performance of the PASSNet (Fig. 8(i)) is commendable; the transform-based method effectively captures differences between images without introducing confusion. However, it is noted that the detection result still falls short compared to the method employed in this study.

The change area within Dataset II comprises new bare land, roads, and demolished buildings. In Fig. 9(a) and (b), the results from machine learning methods exhibit increased noise and simultaneous missed detections. This may be attributed to the lack of consideration for spatial context. In contrast, the block-based deep learning approach, illustrated in Fig. 9(c), (d), (e), and (f), successfully extracts features beneficial for the detection task. Among the transformer-based detection models, as depicted in Fig. 9 (h) and (i), the performance of PASSNet is better than that of ResViT. In our proposed model, multi-scale information is integrated into the fine-grained space, and a multi-loss supervised training approach is employed to enhance the consistent of the change result map with the ground truth.

Similarly, MIGCNet also achieves superior results on Dataset III, where the change area primarily involves new buildings and vegetation reduction—a distinct scenario from the first two datasets. As shown in Fig. 10(a) and (b), there are many falsely detected pixels in the water area when employing SVM and MLP. Conversely, Fig. 10(f) and (g) showcase the adept detection of invariant information in the results obtained through the deep learning methods incorporating multi-scale features. Notably, the southern region of the data exhibits a tendency toward false detections illustrating in Figures (h) and (i), and the proposed method effectively mitigates this phenomenon. Furthermore, the segmented features significantly suppress salt-and-pepper noise within the image.

On the quantitative analysis (i.e., Tables 3–5), MIGCNet obtained the superior results with the highest kappa values and OA, and the lowest MAR across all datasets. In the table, “C” and “U” represent the Intersection over Union (IoU) for the changed and unchanged categories, respectively.

MIGCNet outperform other counter parts, with the OA of 0.9419, 0.9748, and 0.9780 on these three datasets, respectively. For Dataset I, MIGCNet surpasses TSCNN by 5.21% in OA and 31.02% in kappa. In comparison with DSMS-CN, MIGCNet achieves a 20% higher kappa and an OA 3.95% superior in Dataset II. Moving to Dataset III, MIGCNet’s OA surpasses neural network-based methods by over 2.41%, and its kappa is higher by 15.34% compared to DSMS-CN. Moving to Dataset II, MIGCNet achieves a 20% higher kappa compared to DSM-CN, with an OA superior by 3.95%. On Dataset III, MIGCNet enhances OA by more than 2.41% compared to neural network-based methods, and its kappa surpasses DSM-CN by 15.34%.

While MIGCNet may not achieve the lowest FAR on all three datasets, its overall effectiveness stands out. Additionally, MIGCNet attains the highest mean intersection over union (mIoU) and F1-score across the three datasets. In summary, MIGCNet exhibits robustness and superior

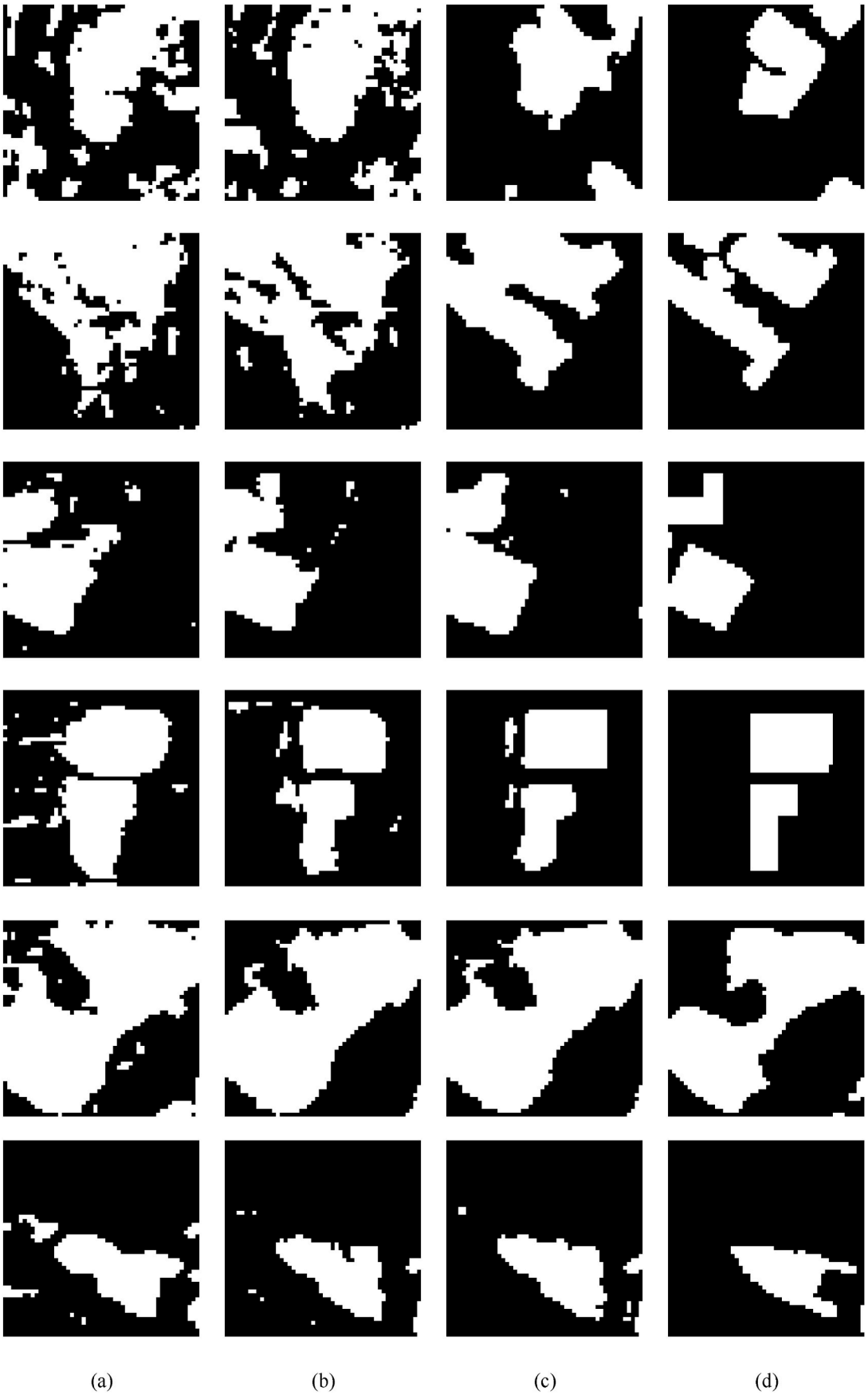

Fig. 11. Several presentations of the details of change detection results. Results generated by means of (a) fc_3, (b) fc_2, (c) fc_1 (the final output of MIGCNet), (d) Reference map.

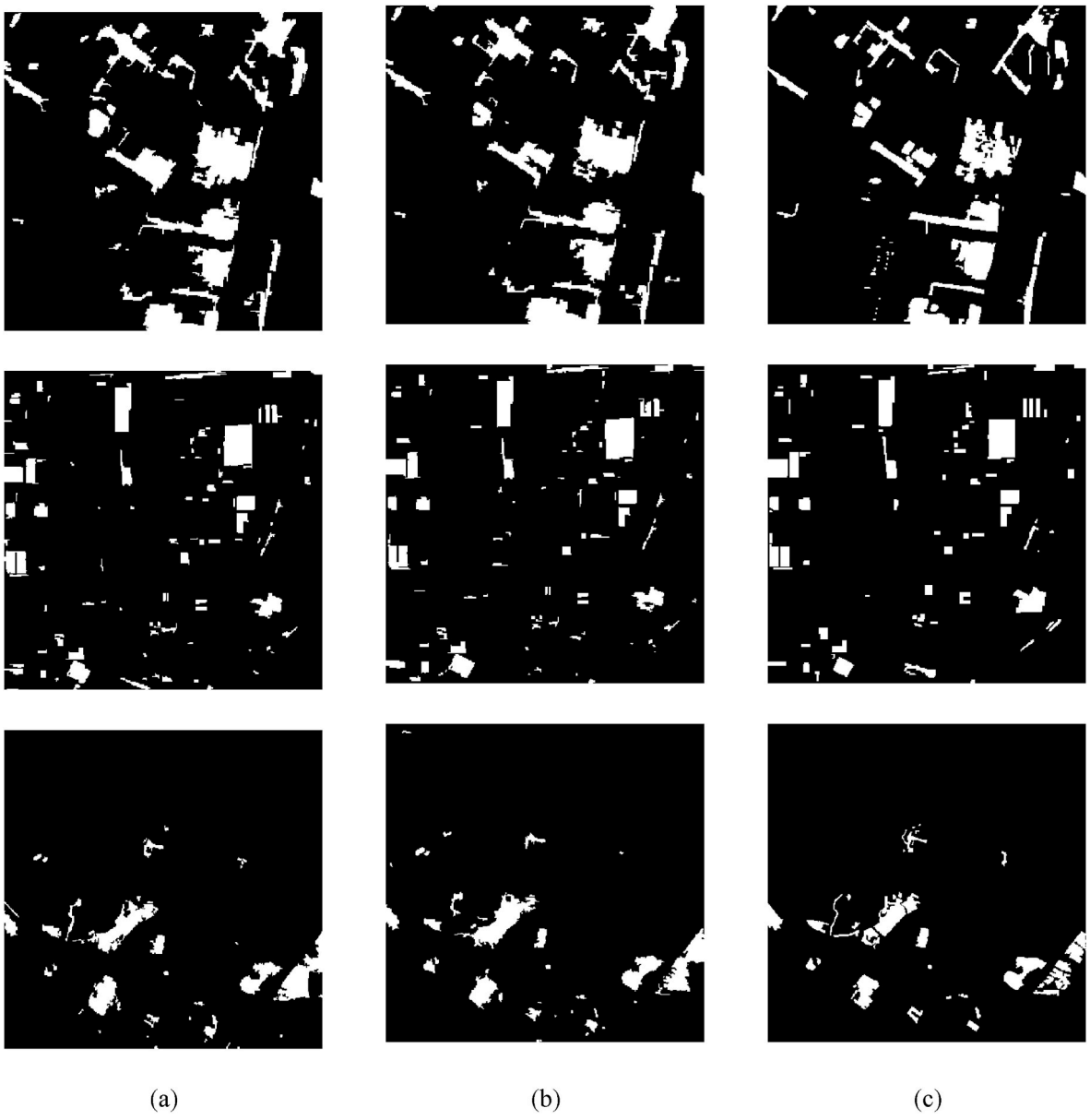

Fig. 12. Change detection results obtained on the three datasets by: (a) MIGCNet without multi-loss supervision, and (b) MIGCNet with multi-loss supervision. (c) Reference map.

performance in the comparative experiments across the three datasets. The proposed method shows a superior detection capability in most evaluation indictors comparing with other advanced change detection methods, which demonstrated that the designed neural network structure results in effective utilization of the multi-scale information, and the multi-loss supervision enhances the performance of the network.

4. Discussion

4.1. Validation of the multi-loss supervision

The proposed method introduced multi-loss supervision in the training process after each feature fusion operation. To explore the effect of the multi-loss supervision, groups of results for the three datasets were selected for analysis. Fig. 11(a), (b), and (c) are the results produced by fc_3, fc_2, and fc_1, respectively, and (d) is the corresponding reference map. From Fig. 11, fc_3 produces the poorest change results, which have broken boundaries and lower compactness. The results obtained by fc_2 are somewhat better. The change maps generated by fc_1 shows the boundaries of the recognition results are becoming clearer and the internal compactness of the objects is improved. We also compared the results of the network without multi-loss supervision with those of the network with multi-loss supervision.

Fig. 12 depicts the results obtained by the network with these two

Table 6
Quantitative results of the different training approaches.

Dataset	Method	OA	Kappa	Commission	Omission
I	−	0.9385	0.7196	0.3141	0.1620
	+	0.9419	0.7303	0.2551	0.1581
II	−	0.9737	0.7987	0.1819	0.1897
	+	0.9748	0.8067	0.1713	0.1880
III	−	0.9767	0.7273	0.3055	0.2094
	+	0.9780	0.7423	0.2912	0.1951

* − : MIGCNet without multi-loss supervision; +: MIGCNet with multi-loss supervision.

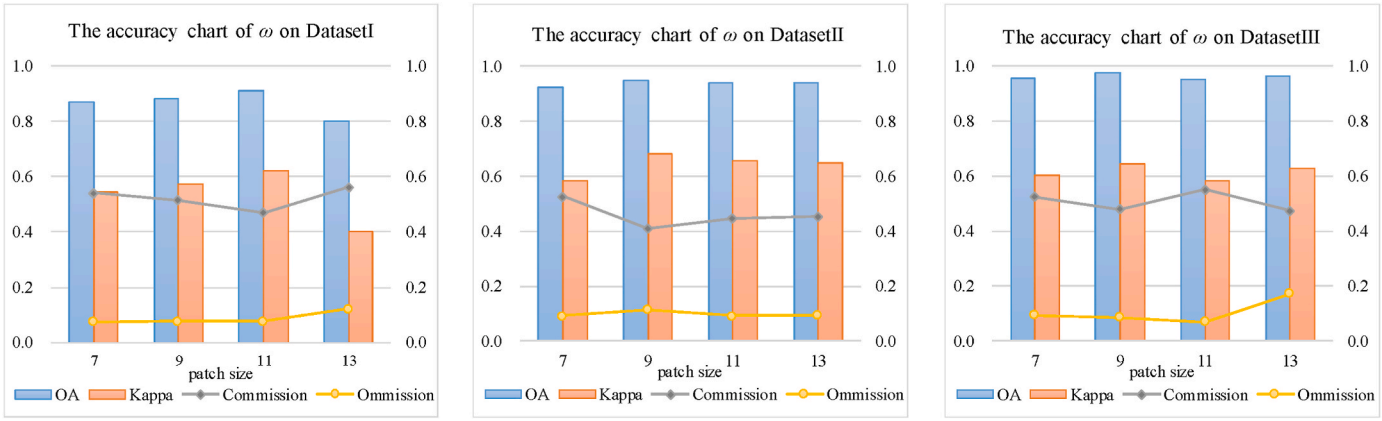

Fig. 13. Charts of the accuracy curves with different patch sizes on (a) Dataset I, (b) Dataset II, and (c) Dataset III.

Table 7
Quantitative results of the ablation experiments.

Dataset	Method	OA	Kappa	Commission	Omission	mIoU	F1 score
I	M-2	0.9348	0.6608	0.2689	0.3334	0.7323	0.6972
	M-1	0.9411	0.7204	0.2685	0.2027	0.7765	0.7245
	MIGCNet	**0.9419**	**0.7303**	**0.2551**	**0.1581**	**0.7804**	**0.7655**
II	M-2	0.9680	0.7436	0.1891	0.2835	0.7900	0.7607
	M-1	0.9691	0.7525	0.1720	0.2694	0.7931	0.7849
	MIGCNet	**0.9748**	**0.8067**	**0.1713**	**0.1880**	**0.8294**	**0.8142**
III	M-2	0.9328	0.5665	0.4379	0.2068	0.6323	0.4972
	M-1	0.9629	0.6935	0.3022	0.1955	0.7484	0.6047
	MIGCNet	**0.9780**	**0.7423**	**0.2912**	**0.1951**	**0.7910**	**0.7538**

Table 8
The comparison of Params and FLOPs for networks.

Network	#Params	#FLOPs
DSCNH	5.7M	0.46G
MixNet	21.9M	0.54G
TSCNN	20.4M	0.89G
PASSNet	0.23M	18.35M
ResVIT	1.06M	11.48M
MIGCNet	9.7M	0.23G

Table 9
Running and testing time on each dataset.

Network	Time	Dataset I	Dataset II	Dataset III
DSCNH	Train	498.26s	481.98s	495.21s
	Test	120.42s	121.57s	120.61s
DCNN	Train	821.92s	807.53s	903.01s
	Test	125.39s	136.01s	125.78s
MixNet	Train	308.50s	299.18s	321.71s
	Test	108.43s	115.24s	120.70s
DSMS-CN	Train	96.21s	107.19s	105.11s
	Test	107.66s	101.57s	104.01s
TSCNN	Train	384.20s	379.54s	391.08s
	Test	157.21s	151.09s	147.33s
PASSNet	Train	179.01s	176.25s	175.47s
	Test	50.99s	50.10s	55.02s
ResViT	Train	188.84s	195.65s	187.38s
	Test	90.71s	90.61s	95.83s
MIGCNet	Train	85.62s	89.86s	84.47s
	Test	102.26s	101.42s	102.83s

training approaches, and the corresponding accuracy evaluation is presented in Table 5. Fig. 12(a) presents the change maps produced by MIGCNet without multi-loss supervision (i.e., $-$) on the three datasets, and Fig. 12(b) shows the results from MIGCNet with multi-loss supervision (i.e., $+$). The accuracy of the results obtained without multi-loss supervision is generally lower. As is shown in Table 6, higher OA and kappa values are obtained by introducing the multi-loss supervision, and the values of the commission and omission are also reduced. The improvement in accuracy is most significant in Dataset I. The complex change categories and irregular change shapes are likely the main reason why the multi-loss supervision has a significant impact on Dataset I. This confirms that the multi-loss supervision is beneficial to the performance improvement of the MIGCNet.

4.2. Influence of the patch size ω

The image patches with free size is fed into the proposed MIGCNet. We have chosen four sizes of input patch, i.e., 7, 9, 11, and 13, for analyzing the influence of patch size on detection accuracy. Fig. 13 illustrated the performance of the MIGCNet using different patch sizes are illustrated in Fig. 13. For Dataset I, the model with 11 patch size obtains the highest OA, where the commission is lower than for the other patch sizes. With the patch size 9, MIGCNet yields the best results on both Dataset II and Dataset III.

In Dataset II, buildings contribute the most of changes. In other words, this dataset has relatively simple change categories and regular change shapes. As a result, input patch with large size may lead to too much spatial neighborhood information, resulting in a decrease of the accuracy. As shown in Fig. 13(a), the accuracy is significantly improved from patch size 7 to patch size 11 because of the complexity of the land surface in Dataset I. Compared with Dataset II, the change scenarios in Dataset I are more intricacy. For example, many buildings have been turned into bare land. The network is lack of sensitivity for change features based on the small patch size, which results in the detection flaw. Moreover, the situation of Dataset III is similar to that of Dataset II. Except that there is a large area of water coverage, the information of change areas is not as complex as that of Dataset I, thus small size will be more suitable on Dataset III.

Fig. 14. The image and the corresponding labeled maps for the two datasets.

4.3. Ablation study

To further validate the efficacy of the proposed modules, we performed ablation experiments by systematically removing one or two layers of MIGC blocks and scrutinizing their impact on model performance. Specifically, M-1 denotes the network configuration with one block of MIGC removed, while M-2 represents the network with two blocks of MIGC removed. As depicted in Table 7, the detection accuracy of the network with a single layer of MIGC module (i.e., M-1) surpasses that without the inclusion of the MIGC module (i.e., M-2), yet falls short of the performance achieved by the newly proposed network, i.e., MIGCNet. Further incorporation of MIGC block may lead to overfitting and increased computing resources. To trade off the accuracy and model complexity, we opted for a configuration with two layers of MIGC modules to build the current detection network.

4.4. Model complexity and time efficiency

In this section, we conducted supplementary experiments to compare the complexity, training time, and testing time of each deep learning model, as presented in Table 8 and Table 9, respectively. As indicated in Table 8, we calculated the floating-point operations (FLOPs) and parameter size (Params) to assess the model complexity of each model. While the number of parameters in MIGCNet is relatively modest, it still surpasses that of DSCNH and TSCNN. Notably, PASSNet involves a relatively lower count of FLOPs, whereas TSCNN has the highest number of it. MIGCNet demonstrates a moderate performance in terms of FLOPs. It is evident that MIGCNet shows a balanced performance compared to other models, considering both the Params and FLOPs. This could explain its relatively efficient performance while demanding fewer computational resources.

As indicated in Table 9, MIGCNet demonstrates a significant superiority over MixNet with TSCNN during the training phase, manifesting a notably reduced training time. In the testing phase, our proposed method also exhibits exceptional efficiency, showing a time reduction of approximately 18s compared to DSCNH. In a holistic assessment, the change detection model should not only prioritize model efficiency but also consider accuracy and robustness as important indicators. The proposed research has yielded optimal results on three distinct datasets with a comprehensive evaluation that accounts for time, complexity, and accuracy.

4.5. Robustness study

To assess the robustness of our proposed model, we incorporate a publicly available dataset, namely the SZTAKI Air Change benchmark (Benedek and Szirányi, 2008), for our research. This dataset comprises images with a resolution of 1.5 m/pixel and dimensions of 952 × 640. Furthermore, we have selected two sets of images for experimentation, and the true color image and ground truth of the datasets are shown in Fig. 14.

Tables 10 and 11 present the detection accuracy of various algorithms applied to those datasets, and the detection maps are shown in Figs. 15 and 16. Examining both the result figures and the accuracy evaluation table, our method consistently outperforms on the public datasets, underscoring the robustness inherent in the proposed approach. Specifically, when applied to datasets A and B, MIGCNet excels in accurately detecting change regions and exhibits a strong capability to effectively suppress noise. These results affirm the resilience and efficacy of our proposed method in various scenarios.

5. Conclusion

MIGCNet has been proposed as a supervised change detection

Table 10
Quantitative analysis of the Different Approaches on Dataset A.

Method		SVM	ANN	DCNN	DSCNH	MixNet	DSMS-CN	TSCNN	PASSNet	ResVIT	MIGCNet
IoU	C	0.3267	0.2932	0.3841	0.3304	0.3521	0.2864	0.2722	0.4491	0.3380	**0.4601**
	U	0.9266	0.9112	0.9493	0.8933	0.9019	0.8819	0.8536	**0.9310**	0.8976	0.9281
MIoU		0.6266	0.6022	0.6667	0.6119	0.6270	0.5841	0.5629	0.6900	0.6178	**0.6991**
F1-score		0.4925	0.4534	0.5551	0.4967	0.5208	0.4452	0.4279	0.6198	0.5052	**0.6302**
OA		0.9291	0.9144	0.9507	0.8987	0.9069	0.8872	0.8612	0.9346	0.9026	**0.9412**
Kappa		0.4556	0.4104	0.5291	0.4516	0.4785	0.3952	0.3727	0.5882	0.4612	**0.6008**
MAR		0.5798	0.6407	0.4178	0.6515	0.6293	0.6889	0.7191	0.5330	0.6422	**0.4950**
FAR		0.4048	0.3854	0.4696	0.1346	0.1242	0.2168	0.1013	**0.0782**	0.1400	0.1326

Table 11
Quantitative analysis of the Different Approaches on Dataset B.

Method		SVM	ANN	DCNN	DSCNH	MixNet	DSMS-CN	TSCNN	PASSNet	ResVIT	MIGCNet
IoU	C	0.3738	0.3687	0.7522	0.8028	0.8079	0.8069	0.4121	0.7716	0.8010	**0.8265**
	U	0.8563	0.8500	0.9738	0.9627	0.9615	0.9640	0.7890	0.8761	0.9599	**0.9656**
MIoU		0.6151	0.6094	0.8630	0.8827	0.8847	0.8854	0.6006	0.8216	0.8804	**0.8960**
F1-score		0.5442	0.5387	0.9115	0.8906	0.8938	0.8931	0.5837	0.8314	0.8895	**0.9050**
OA		0.8677	0.8621	0.9701	0.9676	0.9668	0.9687	0.8162	0.9794	0.9655	**0.9704**
Kappa		0.4669	0.4578	0.8486	0.8716	0.8743	0.8748	0.4829	0.8193	0.8692	**0.8876**
MAR		0.4555	0.4761	0.2228	0.3258	0.1635	**0.1137**	0.5651	0.1964	0.2680	0.1517
FAR		0.4558	0.4454	0.0725	0.0922	0.0404	0.0998	0.1125	0.0389	0.0443	**0.0299**

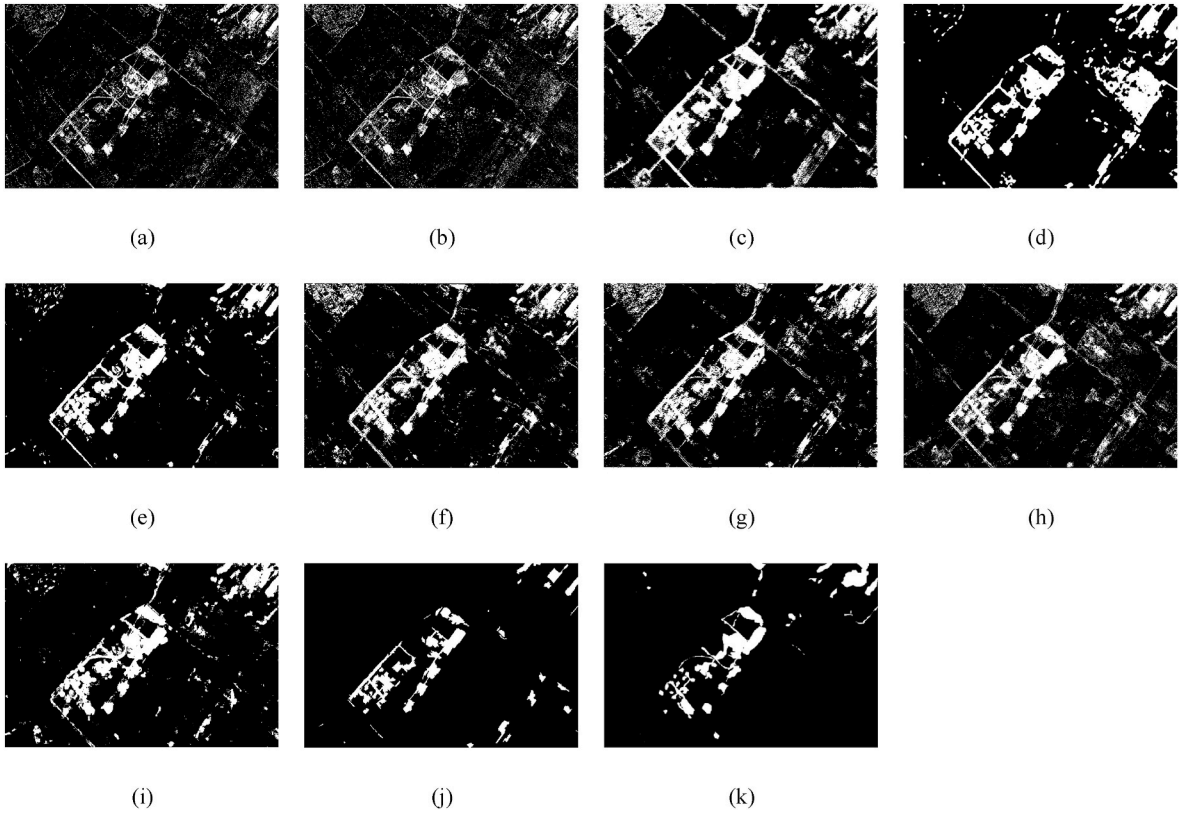

Fig. 15. Detection map on Dataset A by: (a) SVM, (b) MLP, (c) TSCNN, (d) DSMS-CN, (e) DCNN, (f) MixNet, (g) DSCNH (h) ResViT (i) PASSNet and (j) MIGCNet. (k) Ground truth.

network in this work, which takes multiple kernel sizes into account in the establishment of the CNN. The proposed MIGC module, which is based on mixed convolution and interleaved group convolution, was found to be of learning the effective multiscale features. MIGCNet was also found to be capable of being applied to tackle with multi-sensor images. Meanwhile, we conduct the multi-loss supervision to overcome the problem of insufficient training. The results obtained for the three multi-sensor datasets demonstrated that the MIGCNet is superiority compared with the mainstream methods. The introduction of multi-loss supervision can improve the performance of MIGCNet. Although the proposed MIGCNet has obtained satisfactory results, still there are some limitations requiring more attention in future.

1) Firstly, although the model has obtained a certain degree of practicality, there are still some defects for the large-scale remote sensing image. Therefore, in the future, we need to consider how to combine the new model to carry out efficient and low-cost network learning of remote sensing big data.
2) Secondly, supervised learning is easily affected by the difficulty of the sample acquisition and the insufficient number of samples. The sample imbalance problem also poses a great challenge during the application of supervised learning. The unsupervised representation learning methods will also be considered during the detection process.
3) Moreover, our feature work is to extend the change detection framework to more types of heterogeneous images, such as SAR and optical images.

CRediT authorship contribution statement

Kun Tan: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing. **Moyang Wang:** Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision. **Xue Wang:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jianwei Ding:** Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft. **Zhaoxian Liu:**

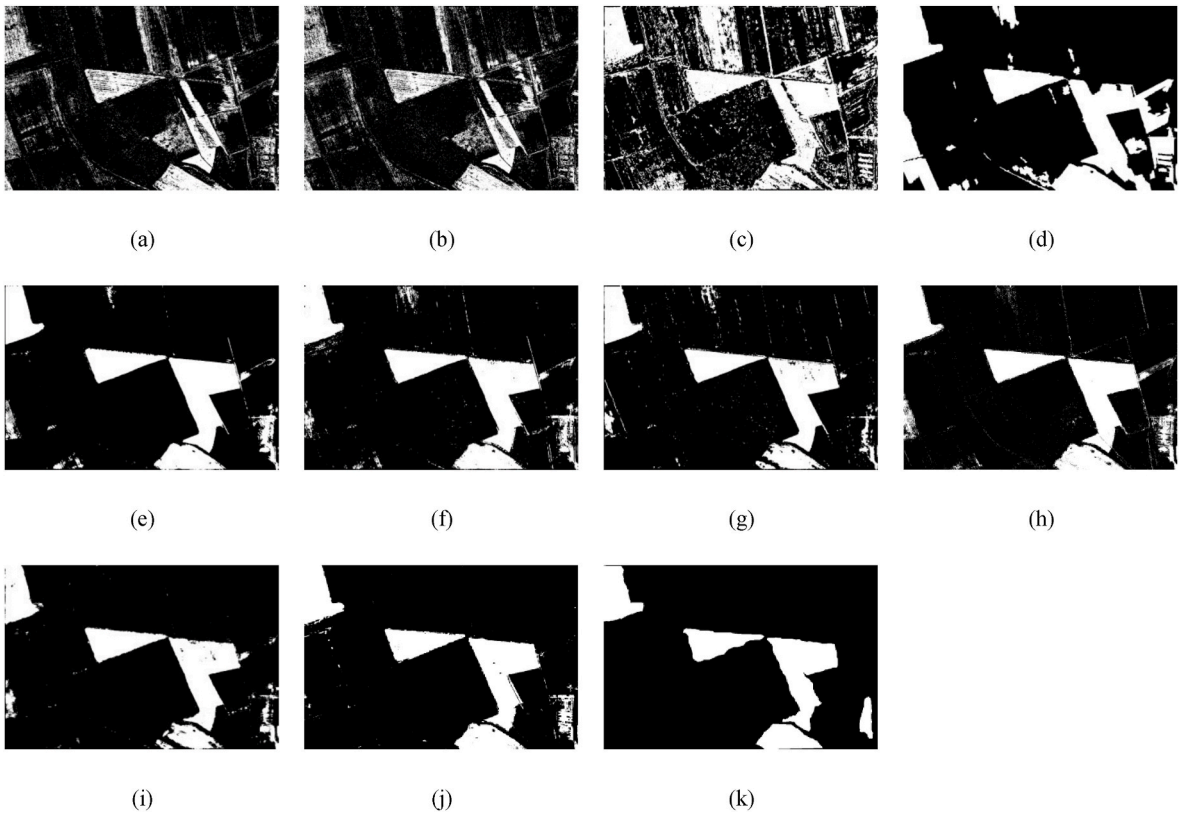

Fig. 16. Detection map on Dataset B by: (a) SVM, (b) MLP, (c) TSCNN, (d) DSMS-CN, (e) DCNN, (f) MixNet, (g) DSCNH (h) ResViT (i) PASSNet and (j) MIGCNet. (k) Ground truth.

Investigation, Resources, Visualization, Writing – original draft, Writing – review & editing. **Chen Pan:** Methodology, Software, Visualization, Writing – review & editing. **Yong Mei:** Validation, Writing – review & editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgements

This work is jointly supported by the Natural Science Foundation of China (grant no. 42171335), the National Civil Aerospace Project of China (No. D040102) and the Open Fundations of Jiangsu Province Engineering Research Center of Airborne Detecting and Intelligent Perceptive Technology (JSECF2023-10).

References

Andresini, G., Appice, A., Ienco, D., Malerba, D., 2023. SENECA: change detection in optical imagery using Siamese networks with Active-Transfer Learning. Expert Syst. Appl. 214, 119123.
Ban, Y., Yousif, O., 2016. Change Detection Techniques: A Review. Springer.
Bandara, W.G.C., Patel, V.M., 2022. A transformer-based siamese network for change detection. In: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 207–210.
Benedek, C., Szirányi, T., 2008. A mixed Markov model for change detection in aerial photos with large time differences. In: 2008 19th International Conference on Pattern Recognition. IEEE, pp. 1–4.
Brunner, D., Lemoine, G., Bruzzone, L., 2010. Earthquake damage assessment of buildings using VHR optical and SAR imagery. IEEE Trans. Geosci. Rem. Sens. 48, 2403–2420.
Bruzzone, L., Prieto, D.F., 2000. Automatic analysis of the difference image for unsupervised change detection. IEEE Trans. Geosci. Rem. Sens. 38, 1171–1182.
Celik, T., 2009. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. Geosci. Rem. Sens. Lett. IEEE 6, 772–776.
Chen, G., Zhao, K., Powers, R., 2014. Assessment of the image misregistration effects on object-based change detection. ISPRS J. Photogrammetry Remote Sens. 87, 19–27.
Chen, H., Wu, C., Du, B., Zhang, L., 2019. Deep siamese multi-scale convolutional network for change detection in multi-temporal VHR images. In: 2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp). IEEE, pp. 1–4.
Chen, Q., Chen, Y., 2016. Multi-feature object-based change detection using self-adaptive weight change vector analysis. Rem. Sens. 8, 549.
Cheng, Q., Li, H., Wu, Q., Ngan, K.N., 2020. Hybrid-loss supervision for deep neural network. Neurocomputing 388, 78–89.
Espindola, G., Câmara, G., Reis, I., Bins, L., Monteiro, A., 2006. Parameter selection for region-growing image segmentation algorithms using spatial autocorrelation. Int. J. Rem. Sens. 27, 3035–3040.
Habibollahi, R., Seydi, S.T., Hasanlou, M., Mahdianpari, M., 2022. TCD-Net: a novel deep learning framework for fully polarimetric change detection using transfer learning. Rem. Sens. 14, 438.
Hao, M., Shi, W., Deng, K., Zhang, H., He, P., 2016. An object-based change detection approach using uncertainty analysis for VHR images. J. Sens. 2016.
Hay, G.J., Blaschke, T., Marceau, D.J., Bouchard, A., 2003. A comparison of three image-object methods for the multiscale analysis of landscape structure. ISPRS J. Photogrammetry Remote Sens. 57, 327–345.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
Hemati, M., Hasanlou, M., Mahdianpari, M., Mohammadimanesh, F., 2021. A systematic review of landsat data for change detection applications: 50 years of monitoring the earth. Rem. Sens. 13, 2869.
Hong, D., Zhang, B., Li, H., Li, Y., Yao, J., Li, C., Werner, M., Chanussot, J., Zipf, A., Zhu, X.X., 2023. Cross-city matters: a multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. Rem. Sens. Environ. 299, 113856.
Hussain, M., Chen, D., Cheng, A., Wei, H., Stanley, D., 2013. Change detection from remotely sensed images: from pixel-based to object-based approaches. ISPRS J. Photogrammetry Remote Sens. 80, 91–106.

Ji, R., Tan, K., Wang, X., Pan, C., Xin, L., 2023. PASSNet: a spatial-spectral feature extraction network with patch attention module for hyperspectral image classification. Geosci. Rem. Sens. Lett. IEEE.
Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S., 2020. A survey of the recent architectures of deep convolutional neural networks. Artif. Intell. Rev. 53, 5455–5516.
Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Li, H., Gong, M., Wang, C., Miao, Q., 2018. Self-paced stacked denoising autoencoders based on differential evolution for change detection. Appl. Soft Comput. 71, 698–714.
Li, S., Wang, Y., Cai, H., Lin, Y., Wang, M., Teng, F., 2023. MF-SRCDNet: multi-feature fusion super-resolution building change detection framework for multi-sensor high-resolution remote sensing imagery. Int. J. Appl. Earth Obs. Geoinf. 119, 103303.
Liu, J., Gong, M., Qin, K., Zhang, P., 2016. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. IEEE Transact. Neural Networks Learn. Syst. 29, 545–559.
Lu, D., Mausel, P., Brondizio, E., Moran, E., 2004. Change detection techniques. Int. J. Rem. Sens. 25, 2365–2401.
Lu, Q., Ma, Y., Xia, G.-S., 2017. Active learning for training sample selection in remote sensing image classification using spatial information. Remote Sensing Letters 8, 1210–1219.
Luppino, L.T., Kampffmeyer, M., Bianchi, F.M., Moser, G., Serpico, S.B., Jenssen, R., Anfinsen, S.N., 2021. Deep image translation with an affinity-based change prior for unsupervised multimodal change detection. IEEE Trans. Geosci. Rem. Sens. 60, 1–22.
Lv, Z., Liu, T., Benediktsson, J.A., 2020. Object-oriented key point vector distance for binary land cover change detection using VHR remote sensing images. IEEE Trans. Geosci. Rem. Sens. 58, 6524–6533.
Mubea, K., Menz, G., 2012. Monitoring land-use change in Nakuru (Kenya) using multi-sensor satellite data. Adv. Rem. Sens. (1).
Seydi, S.T., Hasanlou, M., 2021. A new structure for binary and multiple hyperspectral change detection based on spectral unmixing and convolutional neural network. Measurement 186, 110137.
Seydi, S.T., Hasanlou, M., Amani, M., 2020. A new end-to-end multi-dimensional CNN framework for land cover/land use change detection in multi-source remote sensing datasets. Rem. Sens. 12, 2010.
Singh, A., 1989. Review article digital change detection techniques using remotely-sensed data. Int. J. Rem. Sens. 10, 989–1003.
Srivastava, R.K., Greff, K., Schmidhuber, J., 2015. Highway Networks arXiv preprint arXiv:1505.00387.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.
Tan, K., Zhang, Y., Wang, X., Chen, Y., 2019. Object-based change detection using multiple classifiers and multi-scale uncertainty analysis. Rem. Sens. 11, 359.
Tan, M., Le, Q.V., 2019. Mixconv: Mixed Depthwise Convolutional Kernels. *arXiv preprint arXiv:1907.09595*.
Wang, D., Zhao, F., Wang, C., Wang, H., Zheng, F., Chen, X., 2022a. Y-Net: a multiclass change detection network for bi-temporal remote sensing images. Int. J. Rem. Sens. 43, 565–592.
Wang, M., Li, X., Tan, K., Mango, J., Pan, C., Zhang, D., 2024. Position-aware graph-CNN fusion network: an integrated approach combining geospatial information and graph attention network for multi-class change detection. IEEE Trans. Geosci. Rem. Sens.
Wang, M., Tan, K., Jia, X., Wang, X., Chen, Y., 2020. A deep siamese network with hybrid convolutional feature extraction module for change detection based on multi-sensor remote sensing images. Rem. Sens. 12, 205.
Wang, X., Du, J., Tan, K., Ding, J., Liu, Z., Pan, C., Han, B., 2022b. A high-resolution feature difference attention network for the application of building change detection. Int. J. Appl. Earth Obs. Geoinf. 112, 102950.
Wang, X., Tan, K., Du, Q., Chen, Y., Du, P., 2019. Caps-TripleGAN: GAN-assisted CapsNet for hyperspectral image classification. IEEE Trans. Geosci. Rem. Sens. 57, 7232–7245.
Wang, X., Yan, X., Tan, K., Pan, C., Ding, J., Liu, Z., Dong, X., 2023. Double U-Net (W-Net): a change detection network with two heads for remote sensing imagery. Int. J. Appl. Earth Obs. Geoinf. 122, 103456.
Wang, Y., Bashir, S., Khan, M., Ullah, Q., Wang, R., Song, Y., Guo, Z., Niu, Y., 2021. Remote Sensing Image Super-resolution and Object Detection: Benchmark and State of the Art.
Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P., 2020. Visual Transformers: Token-Based Image Representation and Processing for Computer Vision. *arXiv preprint arXiv:2006.03677*.
Wu, J., Li, B., Qin, Y., Ni, W., Zhang, H., Fu, R., Sun, Y., 2021a. A multiscale graph convolutional network for change detection in homogeneous and heterogeneous remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 105, 102615.
Wu, X., Zhu, X., Wu, G.-Q., Ding, W., 2013. Data mining with big data. IEEE Trans. Knowl. Data Eng. 26, 97–107.
Wu, Y., Ding, H., Gong, M., Qin, A., Ma, W., Miao, Q., Tan, K.C., 2022a. Evolutionary multiform optimization with two-stage bidirectional knowledge transfer strategy for point cloud registration. IEEE Trans. Evol. Comput.
Wu, Y., Li, J., Yuan, Y., Qin, A., Miao, Q.-G., Gong, M.-G., 2021b. Commonality autoencoder: learning common features for change detection from heterogeneous images. IEEE Transact. Neural Networks Learn. Syst. 33, 4257–4270.
Wu, Y., Zhang, Y., Fan, X., Gong, M., Miao, Q., Ma, W., 2022b. Inenet: inliers estimation network with similarity learning for partial overlapping registration. IEEE Trans. Circ. Syst. Video Technol. 33, 1413–1426.
Xiao, P., Zhang, X., Wang, D., Yuan, M., Feng, X., Kelly, M., 2016. Change detection of built-up land: a framework of combining pixel-based detection and object-based recognition. ISPRS J. Photogrammetry Remote Sens. 119, 402–414.
Yang, K., Xia, G.-S., Liu, Z., Du, B., Yang, W., Pelillo, M., 2020. Asymmetric Siamese Networks for Semantic Change Detection. *arXiv preprint arXiv:2010.05687*.
Yuan, X., Shi, J., Gu, L., 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Syst. Appl. 169, 114417.
Zhan, Y., Fu, K., Yan, M., Sun, X., Wang, H., Qiu, X., 2017. Change detection based on deep siamese convolutional network for optical aerial images. Geosci. Rem. Sens. Lett. IEEE 14, 1845–1849.
Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., Liu, G., 2020. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS J. Photogrammetry Remote Sens. 166, 183–200.
Zhang, M., Xu, G., Chen, K., Yan, M., Sun, X., 2018. Triplet-based semantic relation learning for aerial remote sensing image change detection. Geosci. Rem. Sens. Lett. IEEE 16, 266–270.
Zhang, T., Qi, G.-J., Xiao, B., Wang, J., 2017. Interleaved group convolutions. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4373–4382.
Zhao, W., Wang, Z., Gong, M., Liu, J., 2017. Discriminative feature learning for unsupervised change detection in heterogeneous images based on a coupled neural network. IEEE Trans. Geosci. Rem. Sens. 55, 7066–7080.
Zhou, Y., Wang, J., Ding, J., Liu, B., Weng, N., Xiao, H., 2023. SIGNet: a siamese graph convolutional network for multi-class urban change detection. Rem. Sens. 15, 2464.