J. Meas. Sci. Instrum., 2025, 16(3): 371-383. http://jmsi.publish.founderss.cn/
DOI: 10.62756/jmsi.1674-8042.2025036 jmsi@nuc.edu.cn

FPCNet-based change detection for remote sensing images

LI Jiying*, WANG Qi, SHI Hongping

School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730000, China

***Corresponding author:** LI Jiying (ljy7609@126.com)

Received: December 3, 2023 **Revised:** January 11, 2024 **Accepted:** January 21, 2024

Abstract: The objective of this study is to address semantic misalignment and insufficient accuracy in edge detail and discrimination detection, which are common issues in deep learning-based change detection methods relying on encoding and decoding frameworks. In response to this, we propose a model called FlowDual-PixelClsObjectMec (FPCNet), which innovatively incorporates dual flow alignment technology in the decoding stage to rectify semantic discrepancies through streamlined feature correction fusion. Furthermore, the model employs an object-level similarity measurement coupled with pixel-level classification in the PixelClsObjectMec (PCOM) module during the final discrimination stage, significantly enhancing edge detail detection and overall accuracy. Experimental evaluations on the change detection dataset (CDD) and building CDD demonstrate superior performance, with $F1$ scores of 95.1% and 92.8%, respectively. Our findings indicate that the FPCNet outperforms the existing algorithms in stability, robustness, and other key metrics.

Key words: remote sensing image change detection; semantic misalignment; dual flow alignment; deep supervised discrimination

0 Introduction

Traditional change detection algorithms primarily identify changed regions through low-level or mid-level feature analysis. They can be categorized into direct comparison method and analysis-based method. The former relies on pixel or various features to construct a difference map, which is then used to extract the changed regions. The latter includes image differencing, image ratioing, and change vector analysis[1,2]. For example, we propose a PCA-k-means algorithm that leverages principal component analysis (PCA) to extract orthogonal feature vectors from the difference maps of two images[2], and subsequently k-means clustering is employed to yield the change detection results. Unlike the direct comparison method that imposes stringent data preprocessing requirements and may yield noisy results if preprocessing is suboptimal, the analysis-based method involves the examination of target objects in images and decision-making based on the analysis outcomes. It encompasses such techniques as decision trees[3], support vector machines, and random forests. Traditional algorithms require manual feature extraction and classifier design, while the analysis-based method is influenced by the feature extraction and generalization capabilities of the classifier, making it difficult to detect complex change scenarios in remote sensing images.

In recent years, the algorithms based on convolutional neural networks (CNNs) have achieved remarkable outcomes in change detection. Early works such as fully convolutional early fusion (FC-EF)[4], fully convolutional Siamese-concatenation (FC-Siam-conc)[4], fully convolutional Siamese-difference (FC-Siam-diff)[4], and convolutional-wavelet neural network (CWNN)[5] have achieved end-to-end detection processes, with the latter two networks applying Siamese architecture and conducting in-depth research on classification problems.

The algorithm based on information-fuzzy network (IFN)[6] introduces attention mechanism and deep supervision method to improve the integrity and compactness of change region. Furthermore, the SNUNet[7] designs Siamese sub-networks as encoders and employs a nested U-Net[8] as the decoder to fully utilize low-level detail features for intensive information transfer between networks to ensure accurate deep-level feature localization. Meanwhile, to obtain more discriminative feature maps, STANet[9] and DASNet[10] introduce a self-attentive mechanism to capture remote dependency in the features.

Although some CNN-based methods have achieved good performance, there are still some limitations: 1) Semantic misalignment often occurs when fusing feature

maps with different resolutions, leading to a decrease in detection accuracy[11]; 2) Bi-temporal images contain rich contextual and channel information, which is intertwinded and complex, therefore, it is necessary to address how to extract useful features to improve detection accuracy, and how to suppress irrelevant information to avoid interference; and 3) Single recognition methods have the problems of edge details and insufficient detection accuracy, pixel-based classification methods provide good prediction results but are susceptible to noise interference, and object-level detection methods based on similarity measures can quickly detect results, but with poor accuracy in contour edges[12,13].

In order to tackle the previously mentioned challenges, we introduce a novel deep supervised discriminative detection approach that seamlessly integrates the dual-stream alignment technique[12,13].

To begin with, a ResNet50-based model serves as the backbone network for comprehensive feature extraction from individual images. To address the diverse scale of target objects within high-resolution remote sensing images, we incorporate an atrous spatial pyramid pooling (ASPP) module, thereby bolstering the capability to detect changes across different scales.

Subsequently, the extracted features from the backbone network are channelled into the feature fusion-interaction module. In this module, feature fusion is executed using the dual flow alignment module called FD to rectify semantic misalignment. Simultaneously, feature interaction is facilitated through the dual attention mechanism (DAM), which aggregates features and assigns weights to effectively amalgamate rich contextual information, discern change-related features, and suppress irrelevant data, ultimately elevating detection accuracy.

Ultimately, the fused and interacted features are input into the PixelClsObjectMec (PCOM) module for the detection of change regions. It is important to highlight the innovative aspects of our proposed method, which specifically reside in the PCOM and FD modules. Collectively, we refer to these modules as FPCNet.

1 Network structure

1.1 General framework

The overall network architecture is shown in Fig.1.

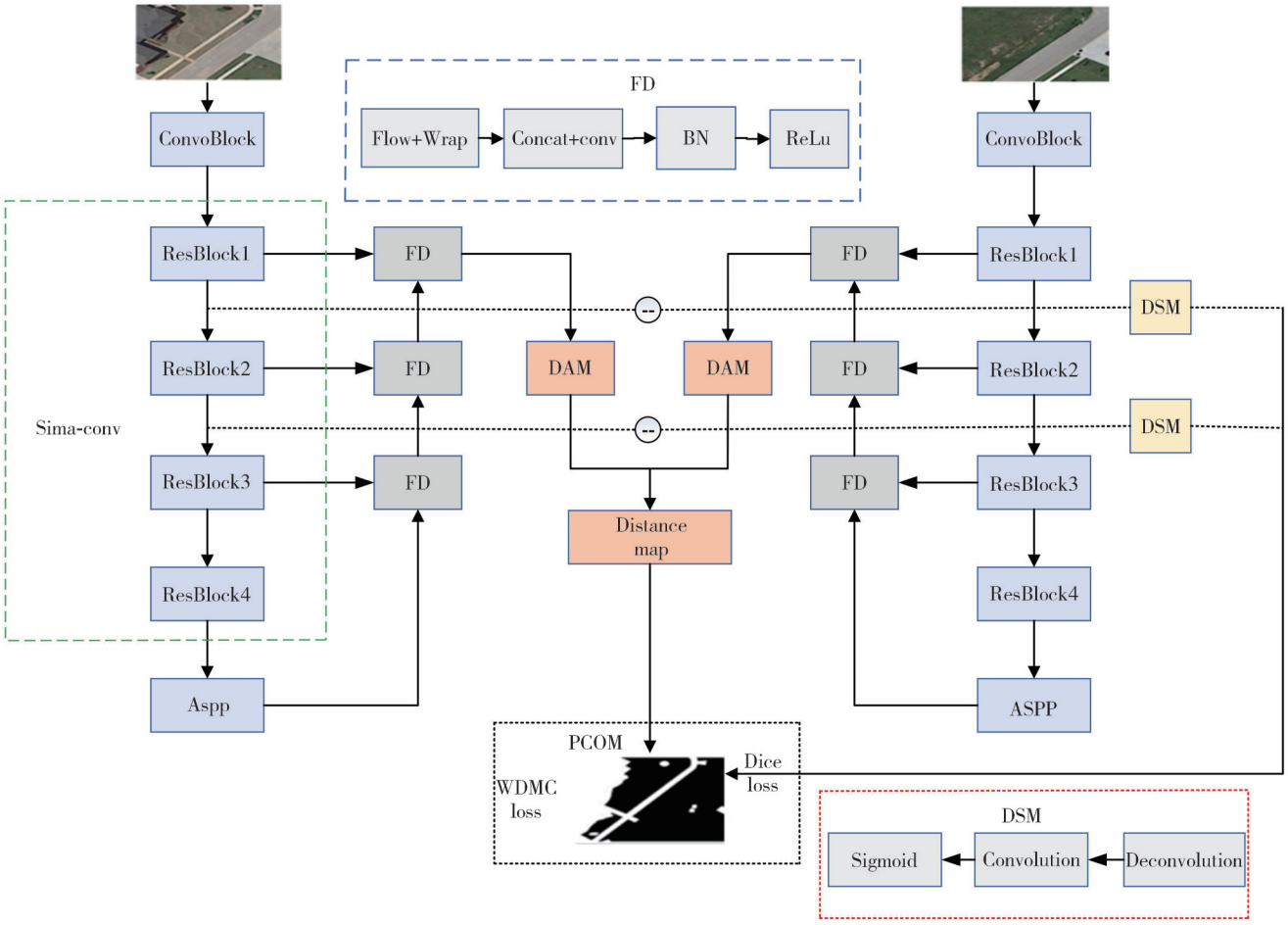

Fig. 1 Schematic diagram of FPCNet

The proposed FPCNet method, based on the encoding-decoding framework, consists of three main modules: deep feature extraction, feature fusion-interaction, and change region discrimination. At the encoding stage, to extract features effectively from individual images, a ResNet50-based backbone network is constructed. Considering the objects at different scales in high-resolution remote sensing images, an ASPP[14] module is embedded in the last layer of the backbone network to enhance the feature extraction capability for different scales of objects. The decoder adopts a structure similar to U-Net and consists of two steps: feature fusion-interaction and change region discrimination. The features obtained from the encoder are fed into the feature fusion-interaction module, where the FD module is used to fuse different resolution feature maps to address the misalignment issue during feature fusion. Then, the DAM is applied for feature interaction. Finally, the fused and interacted features are fed into the PCOM module, which combines object-level similarity measurement[15] and pixel-level classification to achieve deep supervision discrimination.

The deep feature extraction module leverages the initial four layers of ResNet50 as its feature extraction backbone network. Initially, both input images undergo feature extraction within the network. The final layer incorporates the ASPP module to produce feature maps at varying scales, thereby enhancing the model's ability of multi-scale object detection.

Subsequently, commencing with the multi-scale feature maps, a flow-guided wrap operation progressively aligns them with adjacent higher-level feature maps, replacing linear interpolation for upsampling. This alignment process is followed by dimensional concatenation, succeeded by 3×3 convolution, ReLU and batch normalization (BN) operations to fuse features. These fused and aligned feature maps are then input into the DAM module to facilitate feature interaction, thereby enhancing the model's semantic understanding capabilities. Ultimately, the fused and interacted features are input into the PCOM change region detection module to identify altered regions.

1.2 Deep feature extraction

Compared to general optical remote sensing images, high-resolution remote sensing images contain richer information and require higher demands for feature extraction. In this study, the FPCNet, a weight-sharing Siam-Conv network framework based on ResNet50 is designed, which is renowned in image processing tasks like image classification and object detection for its strong feature extraction capabilities. The ResNet50 basic network structure[16] is applied due to its balanced depth—neither too shallow nor overly complex—providing excellent feature representation without excessive computational demands. Deeper models introduce more parameters, computations, memory requirements, and training time, potentially limiting their utility in resource-constrained scenarios. In contrast, ResNet-50, being relatively shallow, operates efficiently under lower memory constraints, and trains quickly. To implement FPCNet, only the first four layers of ResNet50 are utilized, discarding subsequent downsampling pooling operations and fully connected layers. Additionally, the pretrained ResNet50's weight parameters are fine-tuned for these initial layers. Furthermore, the dilated convolutions in the last two ResNet blocks are used to enhance the receptive field and capture more comprehensive features. To complete our deep feature extraction network, Siam-Conv is integrated with ASPP, which leverages dilated convolutions with varying dilation rates, executing multiple dilated convolution operations at different scales on the input feature map to obtain a diverse range of receptive fields. The resulting feature maps at different scales are then concatenated or merged to obtain a more comprehensive feature representation. These feature maps include global contextual information, multi-scale object information, and fine-grained edge information, effectively enhancing the model's receptive field and accuracy. The Siam-Conv module takes the input of dual-temporal images for deep feature extraction, generating feature vectors $(\boldsymbol{F}_{t0}^{i},\boldsymbol{F}_{t1}^{i})$, where $i\in(1,2,3,4)$, $t\in(t0,t1)$, $t0$ represents the period before the change, and $t1$ represents the period after the change.

1.3 Feature fusion and interaction

The feature fusion-interaction[17] consists of feature fusion FD and feature interaction DAM. Through FD, the features $(\boldsymbol{F}_{t0}^{i},\boldsymbol{F}_{t1}^{i})$ are combined to generate fused features $\boldsymbol{F}_{t0}^{C\times H\times W}$ and $\boldsymbol{F}_{t1}^{C\times H\times W}$, where C is the pixel channel, H is the pixel height, and W is the pixel width. Then, the DAM is utilized to construct discriminative feature pair $(\bar{\boldsymbol{F}}_{t0},\bar{\boldsymbol{F}}_{t1})$.

In the features $(\boldsymbol{F}_{t0}^{i},\boldsymbol{F}_{t1}^{i})$ extracted from the dual-temporal images by Siam-Conv, the low-level features contain rich spatial details, while the high-level features aggregate semantic information. The fusion of these features $(\boldsymbol{F}_{t0}^{i},\boldsymbol{F}_{t1}^{i})$ is beneficial for change area recognition. However, when fusing the features with different resolutions, it is often necessary to upsample

the low-resolution features before merging them with high-resolution features. This process may lead to semantic misalignment. Bilinear interpolation is a common upsampling method that generates upsampled pixel values by linearly interpolation between pixel positions in the input image. Bilinear interpolation leads to not only blurring or loss of details when dealing with large-scale upsampling, but also misalignment of information between different feature channels, which may affect the semantic consistency of the features. To address this problem, the approach of FD is proposed, which can be viewed as a very localized alignment operation when using flow-guided wrap operation in the FD module. The main difference between this approach and direct bilinear interpolation upsampling is that it takes into account the semantic flow information to better align the features, where the semantic flow describes the offsets of each pixel from the low-resolution feature maps to the high-resolution feature maps. The flow-guided wrap operation involves not only applying a semantic flow field matrix to the feature map, which describes the offset of each pixel from the low-resolution feature map to the high-resolution feature map, but also performing a positional transformation of the feature map based on the predicted offsets, which maps each pixel in the low-resolution feature map to a position in the high-resolution feature map, thus more accurately corresponding to the transfer of semantic information and more efficiently transferring coarse features to high-resolution features to enhance the feature representation capability of the model. It takes (F_{t0}^{i},F_{t1}^{i}) as input and generates fused features $(F_{t0}^{C\times H\times W}, F_{t1}^{C\times H\times W})$ through feature fusion. High-resolution remote sensing images inherently contain abundant contextual information. By aggregating these contextual features, facilitating interaction among features, and redistributing the features, the model can learn more useful features, with a focus on changing areas and suppressing unchanged regions to improve detection accuracy. Subsequently, DAM is utilized to obtain feature pair $(\bar{F}_{t0}, \bar{F}_{t1})$ for further processing.

1.3.1 Feature fusion by FD

When fusing high-level and low-level features (F_{t0}^{i},F_{t1}^{i}), there often exist some spatial position differences known as semantic misalignment, whose impact on the accuracy of detection results can be neglected. Inspired by Li's work in 2020[18] on flow alignment, we propose an improved flow alignment algorithm called FD. It learns semantic flows between the features with different resolutions to guide the transfer of semantic information from low-resolution coarse features to high-resolution refined features, thereby eliminating the issue of semantic misalignment during feature fusion. As shown in the blue dashed box in Fig.1, (F_{t}^{i-1},F_{t}^{i}) undergoes a 1×1 convolution to change the channel number, which helps reduce the number of features. Then, channel concatenation is performed, followed by two 3×3 convolutions to learn a semantic flow field with a size of $2\times H\times W$, as depicted in Fig.2. This flow field contains the semantic flows between the features with different resolutions and guides the wrap operation to perform upsampling more effectively than regular bilinear interpolation, as illustrated in Fig.3. This reduces spatial position differences and eliminates semantic misalignment. Next, the wrapped channels F_{t}^{i} and F_{t}^{i-1} are concatenated, and a 3×3 convolution, BN and ReLU activation are unilized to further fuse the features and eliminate the differences. By feeding (F_{t0}^{i},F_{t1}^{i}) into FD, features $F_{t0}^{C\times H\times W}$ and $F_{t1}^{C\times H\times W}$ are generated, which exhibits more structural characteristics compared to general upsampling methods.

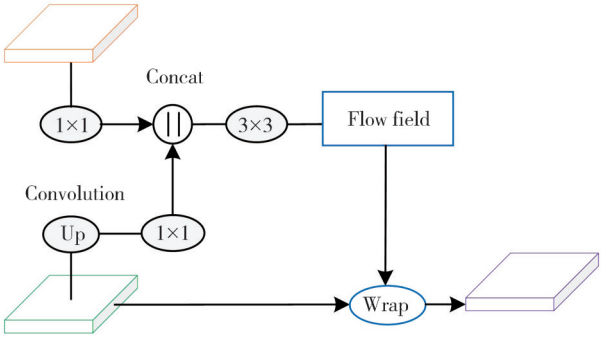

Fig. 2 Flow generation diagram

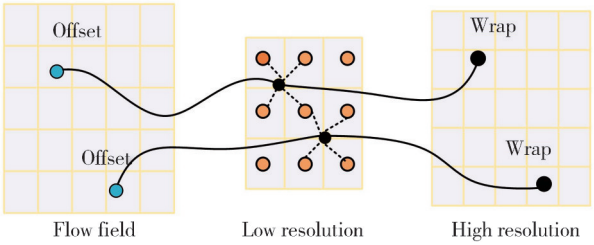

Fig. 3 Diagram of wrap operation

In the FD, the channels of the upsampled and dimensional processed $F_{i}^{C_i\times H_i\times W_i}$ and $F_{i-1}^{C_{i-1}\times H_{i-1}\times W_{i-1}}$ are spliced, and then the optical flow matrix Δ with the size of $2\times H_{i-1}\times W_{i-1}$ via two convolution operations with 3×3 size is obtained, namely

$$\Delta_{i-1}=conv_1(cat(F_i,F_{i-1})). \tag{1}$$

The optical flow field represents the displacement or motion of each pixel in the x and y directions, providing guidance for obtaining accurate positional information during the upsampling of low-resolution features.

Then, it is mapped to the adjacent higher-level spatial

location by a simple algebraic operation to obtain a new spatial coordinate location. The positional transformation can be expressed as

$$p_i = \frac{p_{i-1} + \Delta_{i-1}(p_{i-1})}{2}, \tag{2}$$

where p_i is the underlying spatial location, and p_{i-1} is the higher-level spatial location.

Subsequently, a differentiable bilinear upsampling operation[19] is performed, which linearly interpolates the values of the four neighborhoods of around p_i: top-left, top-right, bottom-left, and bottom-right positions as

$$F_i(p_{i-1}) = \sum_{p \in N(p_i)} w_p F_i(p), \tag{3}$$

where w_p is the weight of the bilinear interpolation kernel, $F_i(p)$ is the value of the low-resolution feature map at position p_i, and $N(p_i)$ is set of the four neighborhood positions. This is used to compute the final output result $F_i(p_{i-1})$ of the FD module.

1.3.2 Feature interaction by DAM

High-resolution remote sensing images are rich in background information and image features, and the fusion of multiple features is a major research area for improving detection accuracy. Extensive research[20,21] has underscored the limitations of relying solely on features generated by traditional fully convolutional networks (FCNs) for discrimination, potentially resulting in misclassification. To address this issue and enhance detection accuracy, we introduce DAM, a dual-channel attention mechanism. DAM comprises two components: spatial attention and channel attention. Spatial attention directs its focus to diverse spatial locations within the feature map, capturing inter-region relationships. This mechanism guides the model to prioritize the regions that harbor crucial information, enhancing its comprehension of the semantic significance across image locations. Conversely, channel attention centers on distinct channels within the feature map, determining their relevance to the specific task at hand. It amplifies responses to task-relevant features while suppressing those irrelevant to the task. Spatial attention enhances the precision of channel attention by guiding it to concentrate more accurately on vital features within a given region, mitigating interference from irrelevant features. In turn, channel attention complements spatial attention by enabling it to target and emphasize the most pertinent feature channels across various regions. By uniting these components, the model gains an improved ability to focus its attention on specific locations while dynamically strengthening or attenuating different channels as needed. DAM performs feature aggregation and redistribution in both spatial and channel dimensions, allowing for adaptive refinement of features in each dimension. Subsequently, the spatially and channel-wise refined features are added pixel-wise, followed by weighted fusion of contextual information, resulting in features $\bar{F}_{t0}$ and $\bar{F}_{t1}$. These features are then used for subsequent change area discrimination. Specifically, for the spatial dimension, we introduce spatial attention, which calculates a weight coefficient for each pixel based on its position in the image. This weight coefficient is multiplied with the corresponding pixel to obtain the weighted feature representation. By assigning different weights to the features at different spatial positions, the network can focus more on important feature information. The spatial attention mechanism is illustrated in Fig.4, where $F_{t0}^{C\times H\times W}$ and $F_{t1}^{C\times H\times W}$ are input into three convolutional layers with the same structure, resulting in three new features F_a, F_b, F_c, $\{F_a, F_b, F_c\} \in \mathbf{R}^{C\times H\times W}$; and C, H, W represent the number of channels, height, and width, respectively. Then, we reshape F_a and F_b to $\mathbf{R}^{C\times N}$, where $N = H \times W$. Subsequently, we perform matrix multiplication between the transposed F_b and F_a, and obtain the spatial attention map $F_s \in \mathbf{R}^{N\times N}$ via softmax operaton. The value of $F_{s_{ji}}$ can be used to measure the effectiveness of feature at position i for feature at position j. A stronger connection between two features results in a larger value of $F_{s_{ji}}$. F_c is reshaped as $\mathbf{R}^{C\times N}$ and multiplied with F_s through matrix multiplication, resulting in an output with a size of $\mathbf{R}^{C\times H\times W}$. The result is then element-wise added to F to obtain the final output F_{sa_j}.

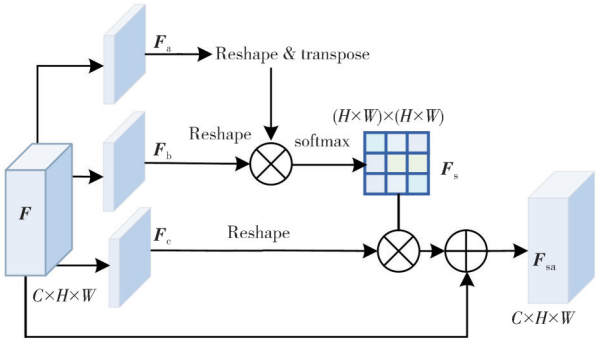

Fig. 4 Spatial attention schematics

The feature obtained at each position is the weighted sum of all positions' features, with the weights determined by $F_{s_{ji}}$, representing a selective aggregation of context based on the spatial attention map. The entire spatial attention feature map can be expressed as

$$F_{sa_j} = \eta \sum_{i=1}^{N} (F_{s_{ji}} F_{c_j}) + F_j, \tag{4}$$

where η is a scale parameter initialized as 0 and gradually learning to allocate more weights.

For the channel dimension, we introduce the channel attention module to compute the attention coefficients for each channel. These coefficients are then multiplied with the feature vectors of each channel to obtain the weighted feature representation, establishing relationships among channels. By utilizing the correlations between channel mappings, we can enhance interdependent feature maps and improve the feature representation with specified semantics for better differentiation of the changing channels. The channel attention is illustrated in Fig.5. It can be seen that the feature $F \in \mathbf{R}^{C\times H\times W}$ is reshaped as $\mathbf{R}^{C\times N}$, where $N = H \times W$. Then, the transpose of F is multiplied with F through matrix multiplication, and the softmax operation is performed to obtain the channel attention map $F_c \in \mathbf{R}^{N\times N}$. The value $F_{c_{ji}}$ can be used to measure the influence of the channel i on the channel j. A larger value of $F_{c_{ji}}$ indicates a stronger connection between the two channels. Subsequently, the feathure F is reshaped as $\mathbf{R}^{C\times N}$ and multiplied with F_c through matrix multiplication, and the result is then element-wise added with F to obtain the final output F_{ca_j}.

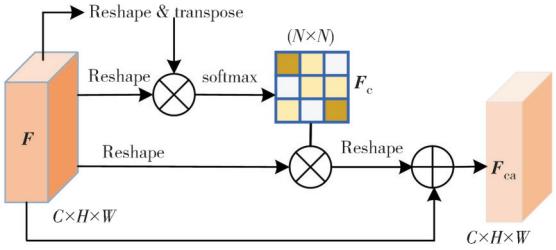

Fig. 5 Channel attention schematics

we can conclude that the final feature for each channel is the weighted sum of all channels' features and the original feature, This modeling captures long-term semantic dependencies among feature maps, enhancing the discriminability of features and highlighting the representation of changing regions. The entire channel attention feature map can be expressed as

$$F_{ca_j} = \gamma \sum_{i=1}^{C} (F_{c_{ji}} F_i) + F_j, \tag{5}$$

where γ is a scale parameter, initialized as 0 and gradually learning to allocate more weights.

Finally, the feature maps F_{sa} and F_{ca} are element-wise added to obtain the final feature interaction map F_{se}, as shown in Eq. (6). Each pixel point in the feature map is weightedly fused with rich contextual information.

$$F_{se} = F_{sa} + F_{ca}. \tag{6}$$

After $F_{t0}^{C\times H\times W}$ and $F_{t1}^{C\times H\times W}$ pass through the DAM module separately, the feature $\bar{F}_{t0}$ and $\bar{F}_{t1}$ are obtained for subsequent discriminations of change regions.

1.4 Change area discrimination by PCOM

1.4.1 Loss functions and optimizers

Due to the inherent imbalance between positive and negative samples in change detection tasks, where changed regions tend to be considerably smaller than unchanged regions, training complex models can inadvertently introduce a bias towards learning unchanged feature pairs while neglecting those associated with changes. This bias can significantly undermine the model's detection capabilities. Therefore, in this study, we leverage the characteristics of the change detection dataset (CDD) to introduce an enhanced loss function termed the dual supervision loss (DSL), which combines the Dice loss[22] (DL) and contrastive loss. The weighted distance map contrast loss (WDMCL) function is a loss function commonly used in change detection tasks to balance the contribution of two classes of samples to the loss value during the training process.

DL quantifies the degree of overlap in segmentation results, producing values within the range from 0 to 1. A higher DL value signifies a superior prediction, facilitating the model in capturing target boundaries and shapes with greater precision. Compared to the traditional cross-entropy loss function, DL exhibits advantages when addressing sample imbalance issues by focusing more on critical segments of the target during computation. To harness the combined strengths of WDMCL and DL, we propose their weighted fusion in the training process. This approach enables the network model to rapidly and accurately identify change regions, ultimately yielding superior detection results.

1) WDMCL

WDMCL function is a loss function used in remote sensing change detection tasks to balance the contributions of the two classes of samples to the loss value during the training process. The formula can be expressed as

$$L_{\mathrm{WDMC}} = \frac{1}{2}\left[\omega_1 (1 - y_{ij}) \max(d_{ij} - m_1, 0)^2 + \omega_2 y_{ij} \max(d_{ij} - m_2, 0)^2\right], \tag{7}$$

where ij is the pixel position; y_{ij} is the label of the pixel, which can take values of 0 or 1, indicating unchanged and changed regions, respectively; d_{ij} is the feature distance at pixels i and j; ω_1 and ω_2 are the weights for unchanged and changed feature pairs, respectively, balancing the contributions of the two classes of samples to the loss value, here $\omega_1 = \frac{1}{p_u}$ and $\omega_2 = \frac{1}{p_c}$ indicating

that the feature distances for unchanged and changed pairs are the same, and P_u and P_c are the frequencies of unchanged and changed pixel pairs, respectively.

This loss function consists of two parts as follows:

For unchanged feature pairs, the loss is computed using the $(d_{ij}-m_1,0)^2$ term, and m_1 is the threshold for unchanged sample pairs.

For changed feature pairs, the $(d_{ij}-m_2,0)^2$ term is used, where m_2 represents the threshold for changed sample pairs. This term contributes to the loss only in case of $d_{ij}>m_2$.

2) DL

DL is named after the Dice coefficient, which is a measure to evaluate the similarity between two samples. A higher Dice coefficient indicates a higher similarity between the samples. Here DL is expressed as

$$L_{\text{Dice}}=1-\frac{1}{m}\sum_{j=1}^{m}\frac{2\sum_{i=1}^{N}y_{ij}\hat{y}_{ij}}{\sum_{i=1}^{N}y_{ij}+\sum_{i=1}^{N}\hat{y}_{ij}}, \tag{8}$$

where y and $\hat{y}$ are the prediction graph and the target label, respectively. Finally, we get

$$L_{\text{DS}}=L_{\text{WDMC}}+\lambda L_{\text{Dice}}, \tag{9}$$

where λ is the weight of the DL. This loss function improves the performance of the network in identifying changes by balancing the contribution of invariant and changing feature pairs to the loss values during the training and by using a deep supervision strategy to enhance the feature representation performance of the network. The Adam optimizer is employed with a learning rate of 10^{-4}.

1.4.2 PCOM module

In Fig.1, the black dashed line represents the PCOM module, an integral part of the overall network architecture. While similarity-based methods often fail to accurately delineate changed region contours, pixel-level change detection methods excel in contour accuracy. However, they are prone to noise interference and exhibit extended prediction times due to their heavy reliance on individual pixel pair judgments, disregarding local feature information and potentially misclassifying noisy points as changes.

To address these limitations in existing detection algorithms, the PCOM module combines similarity-based discrimination with pixel-level classification and introduces an auxiliary discrimination method, the deep supervised module (DSM). During the process of network training, the similarity feature maps, based on similarity metrics, contribute to the training of the WDMCL calculation. Simultaneously, the shallow feature difference maps, derived from pixel classification, undergo pixel-by-pixel classification and participate in DL calculation training. This approach effectively merges and synergizes both aspects, as highlighted within the red dashed box in Fig.1.

The DSM comprises two 3×3 deconvolution layers, two convolution layers, and a sigmoid layer. It calculates a feature difference map by subtracting the first and second layer feature maps from the feature extractor. This map is then upsampled via deconvolution to match the input image's size. After passing through convolution layers, it undergoes a sigmoid transformation, generating intermediate change results. These intermediate results, in conjunction with the outputs of the similarity discrimination module, guide the network's optimization process, enhancing the model's feature learning capabilities and ensuring the acquisition of more effective features.

Specifically, after obtaining the fused and interacted features $\bar{F}_{t0}$ and $\bar{F}_{t1}$ through the feature fusion-interaction process, they are fed into the distance module to calculate the Euclidean distance between feature pairs, which measures the similarity between images. Generally, if the pixel values at corresponding positions in the two image periods change, the distance will be larger, indicating lower similarity. Conversely, if the distance is smaller, the similarity is higher. The distance map is then calculated by comparing the distance with the labels to obtain the WDMCL. In the deep supervised part, the two sets of feature maps (F_{t0}^{i},F_{t1}^{i}), $i\in(1,2,3,4)$ extracted by the first two layers of the network are pixel-wise subtracted to obtain feature difference maps $(\bar{F}_{t0}^{1},\bar{F}_{t1}^{1})$ and $(\bar{F}_{t0}^{2},\bar{F}_{t1}^{2})$. These difference maps then pass through the DSM to generate intermediate change results, which are involved in the computation of the DL. Finally, DSL is calculated as shown in Eq. (9). The PCOM module enables the training of the network model to learn more effective features, thereby improving the detection accuracy of the model.

2 Experiment and analysis

2.1 Datasets and experimental procedure

2.1.1 Datasets

The CDD consists of multiple source remote sensing images and includes a total of 11 pairs of original images. Among them, there are 7 pairs of seasonal change images with a size of 4 725×2 200 pixels, and 4 pairs of seasonal change images with a size of 1 900×1 000 pixels. In Ref.[23], Ji et al. cropped and rotated the raw data to obtain a training set of 10 000 pairs, a validation set of 3 000 pairs, and a test set of 3 000 pairs.

The partial scene images are shown in Fig.6.

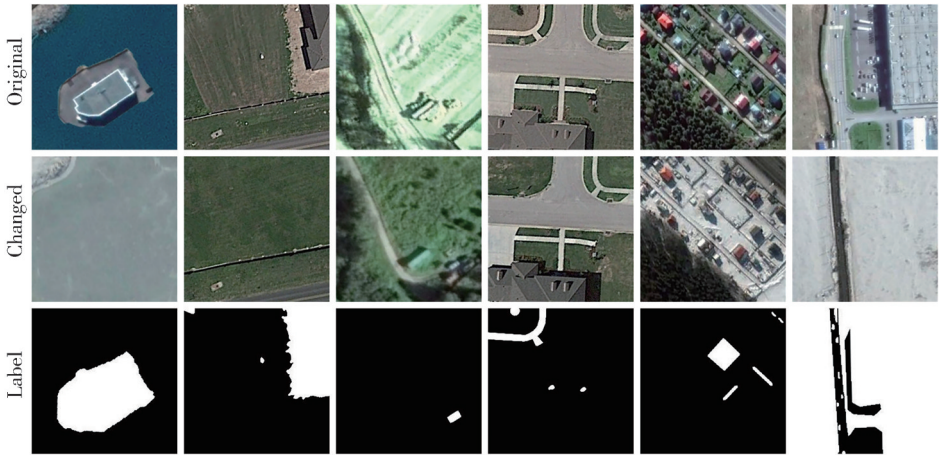

Fig. 6 Partial scene images of CDD

The WHU building change detection dataset (BCDD) contains two image scenes captured at the same location in 2012 and 2016, along with the semantic labels and change detection labels of buildings. Since the size of each image is large (32 507 × 15 354 pixels), the images are cropped into 256×256 patches by data enhancement methods such as image cropping, rotation, etc., resulting in a training set of 6 408 pairs, a validation set of 505 pairs, and a test set of 521 pairs. The partial scene images are shown in Fig.7.

Fig. 7 Partial scene images of BCDD

2.1.2 Model training and evaluation

During the model training process, appropriate hyperparameters, such as learning rate, batch size, and optimizer, must be carefully selected and fine-tuned based on validation set outcomes. In this experiment, we employed the PyTorch deep learning framework, with a single 2080 11G GPU. The initial learning rate for the Adam optimizer was set to 10^{-4} and adjusted every 20 epochs. Specifically, we set the learning rate to 5×10^{-5} for epochs 20 — 40, 10^{-5} for epochs 40 — 60, 5×10^{-6} for epochs 60 — 70, and 10^{-6} for epochs 70 — 80. The learning rate directly impacts training speed and model convergence.

Regarding batch size, we conducted several experiments and determined the batch size to be 4. Along with a validation batch size of 1, it yields the best model performance. It is noteworthy that when the batch size was set to 8 or 16, the model's accuracy did not reach optimal levels. The model exhibits the most effective convergence around the 67th epoch, as depicted in Fig.8.

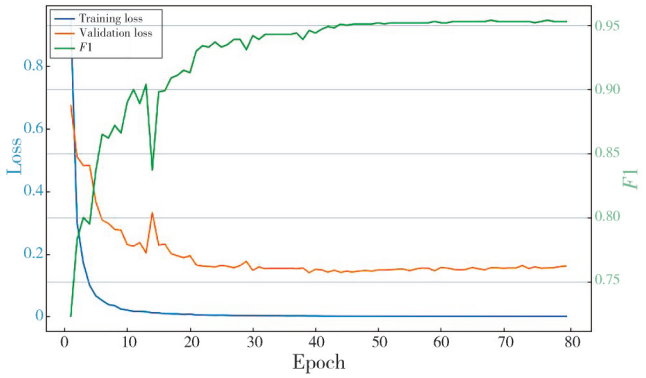

Fig. 8 Loss-accuracy curve of training process

To evaluate the performance of the proposed FPCNet, three evaluation metrics are used, including precision (P), recall (R), and $F1$. In the context of change detection task, a high precision value indicates fewer false detections in the prediction results, while a high recall value means fewer missed detections. $F1$ score is a comprehensive evaluation metric of the prediction results, with higher values indicating better prediction performance. The metrics can be defined as

$$P=\frac{TP}{TP+FP}, \tag{10}$$

$$R=\frac{TP}{TP+FN}, \tag{11}$$

$$F1=\frac{2PR}{P+R}, \tag{12}$$

where TP is the true positive number, FP is the false positive number, TN is the true negative number, and FN is false negative number.

2.2 Experimental results

2.2.1 Comparative test

To validate the performance of the model, comparative experiments were conducted on the CDD and BCDD test sets against FC-EF, FC-Siam-Diff, CDnet, BiDateNet, and ESCNet[24]. The evaluation metrics used include precision, recall, and $F1$-score.

The test results on the CDD dataset are illustrated in Fig.9, where results for FC-EF, FC-Siam-Diff, CDnet, BiDateNet, ESCNet, and FPCNet are presented from left to right. Significant differences in detection results are highlighted within red boxes, while green boxes indicate the ground truth labels. Notably, FPCNet outperforms other models, exhibiting superior performance in detecting change areas with more accurate localization and complete edge information.

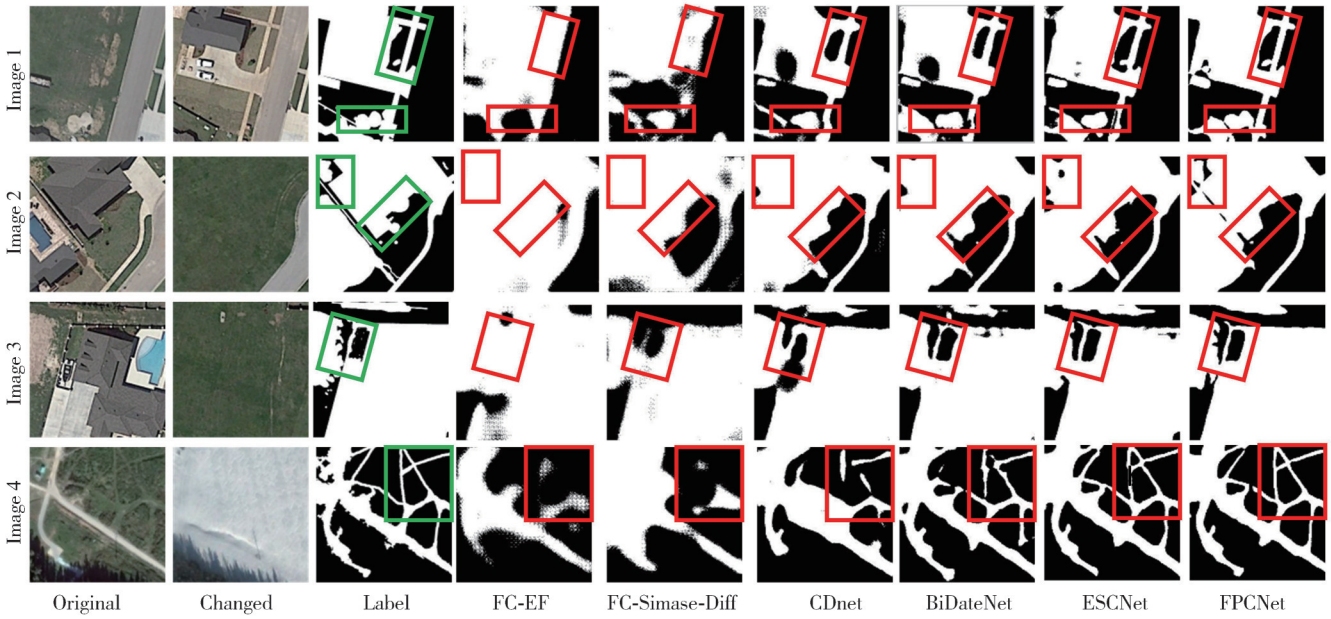

Fig. 9 Experimental results on CDD

Table 1 underscores the outstanding performance of the FPCNet model, boasting precision, recall, and $F1$ scores of 0.946, 0.958, and 0.951, respectively, surpassing all other models. This superiority is especially pronounced in terms of model robustness and detection accuracy.

Table 1 Test results on CDD

Model	R	P	$F1$
FC-EF	0.528	0.684	0.596
FC-Siam-Diff	0.703	0.677	0.689
CDnet	0.817	0.827	0.822
BiDateNet	0.894	0.901	0.898
ESCNet	0.937	0.941	0.938
FPCNet	0.958	0.946	0.951

The test results on BCDD are shown in Fig.10, It can be observed that the results obtained by FPCNet are comparatively better than the other models.

Table 2 shows the test results of different models on the BCDD dataset. The results show that the FPCNet model performs best among all the models, with precision, recall, and $F1$ of 0.952, 0.905, and 0.928, respectively.

Table 2 Test results on BCDD

Model	R	P	$F1$
FC-EF	0.746	0.841	0.791
FC-Siam-Diff	0.723	0.776	0.749
CDnet	0.821	0.908	0.862
BiDateNet	0.819	0.889	0.852
ESCNet	0.870	0.934	0.901
FPCNet	0.905	0.952	0.928

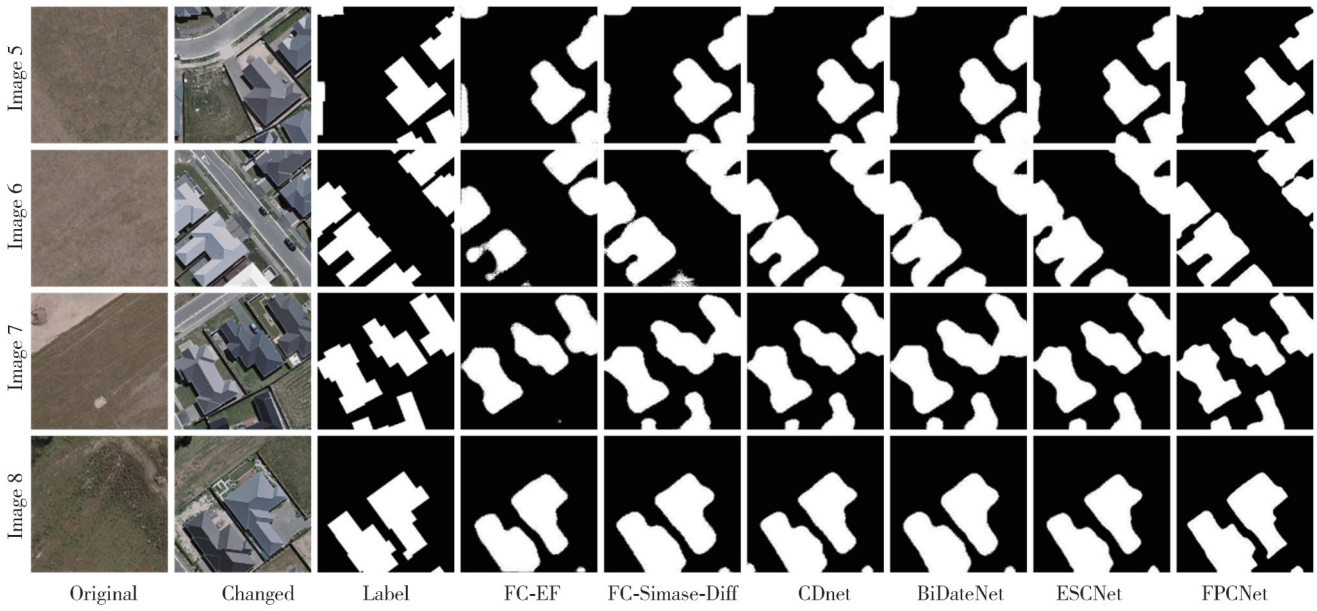

Fig. 10 Experimental results on BCDD

2.2.2 Ablation experiments

To evaluate the impact of each module on detection performance, ablation experiments were conducted on the CDD and BCDD test sets. As depicted in Fig.11, ablation experiments were performed on the CDD with identical hyperparameter settings for all experiments.

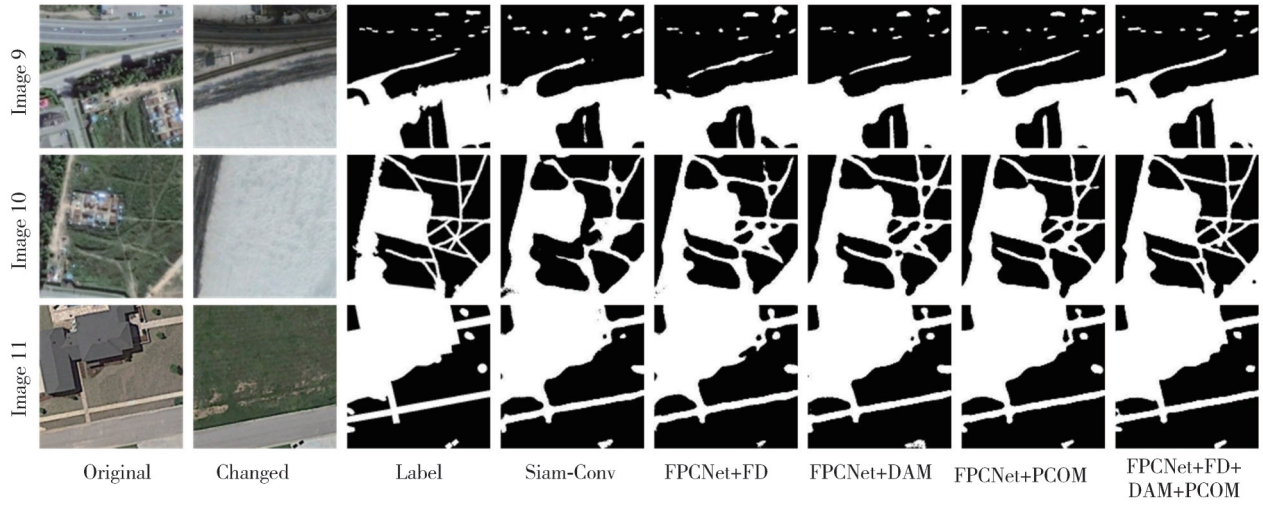

Fig. 11 Ablation experimental results on CDD

The first row in Table 3 represents the baseline model, relying directly on similarity measurement results after feature extraction via Siam-Conv.

Table 3 Accuracy comparison of ablation experiments on CDD

Model	FD	DAM	PCOM	*R*	*P*	*F*1
Baseline				0.918	0.894	0.906
FPCNet	√			0.921	0.935	0.927
FPCNet		√		0.923	0.916	0.919
FPCNet			√	0.935	0.928	0.931
FPCNet	√	√	√	0.958	0.946	0.951

It can be seen that metrics in this row exhibit relatively low values. In the second row, the FD module replaces conventional linear interpolation during the processes of feature fusion and upsampling, resulting in a 0.003 improvement in recall, a 0.041 improvement in precision, and a 0.021 improvement in $F1$. This indicates the FD module effectively eliminates semantic misalignment, generating more structured features.

Introducing the DAM module for feature interaction in the third row results in a 0.005 improvement in recall, a 0.022 improvement in precision, and a 0.013 improvement in $F1$. This demonstrates DAM's effectiveness in aggregating and redistributing features, enabling adaptive feature refinement in various dimensions. The fourth row employs PCOM to modify area discrimination, leading to a 0.017 improvement in recall, a 0.034 improvement in precision, and a 0.025 improvement in $F1$, illustrating that PCOM combined with similarity measurement and pixel-level

classification characteristics enhances detection accuracy. Finally, the combination of all modules to create the complete FPCNet in the last row yields favorable experimental results, underscoring its robustness and strong performance in remote sensing image change detection.

The results of the ablation experiments for each module on the BCDD are shown in Fig.12.

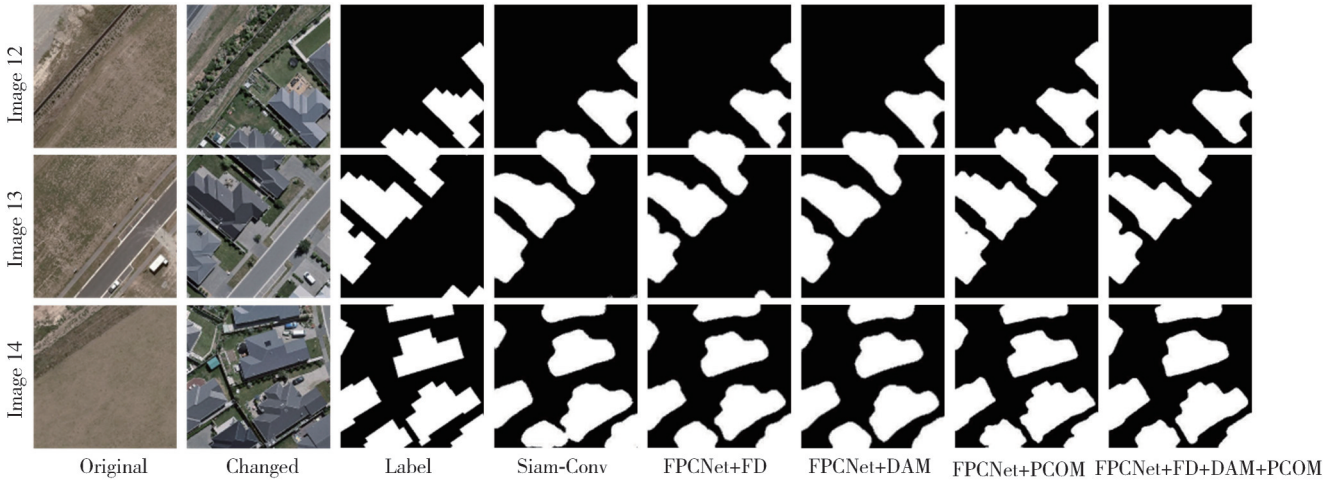

Fig. 12 Ablation experimental results on BCDD

As the experimental results shown in Table 4, the first row represents the baseline model, which demonstrates relatively low values for all the performance metrics. In the second row, the FD module is utilized, resulting in a 0.023 improvement in recall, a 0.011 improvement in precision, and a 0.017 improvement in $F1$. The third row incorporates the DAM module for feature interaction, leading to a 0.020 improvement in precision and a 0.005 improvement in $F1$. The fourth row utilizes the PCOM module for change area discrimination, resulting in a 0.027 improvement in recall, a 0.015 improvement in precision, and a 0.021 improvement in $F1$.

Table 4 Accuracy comparison of ablation experiments on BCDD

Model	FD	DAM	PCOM	R	P	$F1$
Baseline				0.869	0.924	0.896
FPCNet	√			0.892	0.935	0.913
FPCNet		√		0.861	0.944	0.901
FPCNet			√	0.896	0.939	0.917
FPCNet	√	√	√	0.905	0.952	0.928

2.3 Discussion

The experimental results demonstrate the outstanding performance and robustness of the FPCNet model in remote sensing image change detection, surpassing other models significantly. On the CDD, FPCNet achieves a 0.5% increase in precision, a 2.2% boost in recall, and a 1.3% improvement in $F1$ compared with ESCNet. Likewise, on the BCDD, $F1$ of FPCNet outperforms ESCNet with a 2.2% precision increase, a 4.5% recall improvement, and a 3.4% $F1$ enhancement. These enhancements are attributed to the FD module, which replaces standard linear interpolation to rectify feature misalignment during the fusion. Furthermore, the DAM mechanism introduced in the feature interaction module enhances contextual information capture and feature refinement. Lastly, the PCOM module elevates the model's attention to change areas, mitigating background noise and heightening accuracy and stability.

Moreover, hyperparameter settings such as learning rate, batch size, and segmentation threshold exert a pivotal influence on model performance. Notably, the threshold value profoundly impacts performance and recognition outcomes. In our work, the code automatically generates a threshold list with multiple values, computes each of them, and ultimately selects the optimal threshold. Learning rate also affects model convergence and training speed. To balance the training speed and the convergence speed, an initial learning rate of 10^{-4} is set, with subsequent automatic adjustments: 5×10^{-5} for epochs 20—40, 1×10^{-5} for epochs 40—60, 5×10^{-6} for epochs 60—70, and 1×10^{-6} for epochs 70 — 80. This automated hyperparameter tuning eliminates manual adjustments, resulting in the optimal-performing FPCNet model.

In summary, this paper presents FPCNet, a highly effective deep learning-based approach for remote sensing image change detection. Extensively evaluated on public datasets, FPCNet consistently exhibits exceptional performance and robustness in this challenging task. By proficiently identifying change areas in remote sensing images, FPCNet provides an efficient solution with broad applications in the field.

3 Conclusions

In this study, an FPCNet for change detection in remote sensing image is proposed. The model achieves fast and accurate predictions by directly relying on the similarity recognition results between the two temporal images. With the FD and DAM, the model effectively fuses and interacts features, enhancing the detection accuracy. An improved loss function DSL is also employed to balance the impact of imbalanced sample distribution between change and unchanged regions on the network. Compared with the baseline method and some compared algorithms, the proposed network performs well on both CDD and BCDD datasets, showing high accuracy and reliability in detecting change areas in remote sensing images. In the future, further research will be conducted to improve the mobility and robustness of change detection, focusing on small-sample scenarios and noisy environments.

Acknowledgement

I would like to express my gratitude to the reviewers and editors.

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] CHEN J, CHEN X H, CUI X H, et al. Change vector analysis in posterior probability space: a new method for land cover change detection. IEEE Geoscience and Remote Sensing Letters, 2011, 8(2): 317-321.

[2] LI Z Y, LIU T F, SHI C, et al. Novel land cover change detection method based on *k*-means clustering and adaptive majority voting using bitemporal remote sensing images. IEEE Access, 2019, 7: 34425-34437.

[3] QIN R J, HUANG X, GRUEN A, et al. Object-based 3-D building change detection on multitemporal stereo images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015, 8(5): 2125-2137.

[4] CAYE DAUDT R, LE SAUX B, BOULCH A. Fully convolutional Siamese networks for change detection// 2018 25th IEEE International Conference on Image Processing, October 7-10, 2018, Athens, Greece. New York: IEEE, 2018: 4063-4067.

[5] YIN Y H, ZHANG Y P, ZHAO T B, et al. Sea ice change detection algorithm based on multi-scale reconstruction and constrained clustering. Journal of Test and Measurement Technology, 2023, 37(3): 199-207.

[6] CHEN P, LI C, ZHANG B, et al. A region-based feature fusion network for VHR image change detection. Remote Sensing, 2022, 14(21): 5577.

[7] FANG S, LI K Y, SHAO J Y, et al. SNUNet-CD: a densely connected Siamese network for change detection of VHR images. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 8007805.

[8] ZHOU Z W, RAHMAN SIDDIQUEE M M, TAJBAKHSH N, et al. UNet++: A nested U-Net architecture for medical image segmentation//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Cham: Springer International Publishing, 2018: 3-11.

[9] CHEN H, SHI Z W. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sensing, 2020, 12(10): 1662.

[10] CHEN J, YUAN Z Y, PENG J, et al. 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, 2020, 14: 1194-1206.

[11] FU W F. Research on remote sensing image change detection algorithm based on deep learning. Chengdu: University of Electronic Science and Technology, 2021.

[12] GU L, XU S Q, ZHU L Q. Remote sensing image building change detection based on FlowS-Unet. Journal of Automation, 2020, 46(6): 1291-1300.

[13] WANG H, SONG H X, ZHANG Z. Two-stream face image restoration algorithm based on texture and structure. Journal of Test and Measurement Technology, 2024, 38(3): 274-280.

[14] JI S P, TIAN S Q, ZHANG C. Urban land cover classification and change detection using full null convolutional neuron network. Journal of Wuhan University (Information Science Edition), 2020, 45(2): 233-241.

[15] BAO T F. High-resolution remote sensing image change detection based on convolutional neural network. Shanghai: Shanghai Jiaotong University, 2020.

[16] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[17] FANG S, LI K Y, LI Z. Changer: feature interaction is what you need for change detection. 2022: 2209.08290. https://arxiv.org/abs/2209.08290v1.

[18] LI X T, YOU A S, ZHU Z, et al. Semantic flow for fast and accurate scene parsing//European Conference on Computer Vision, October 23 – 28, 2020, Glasgow, Scotland, UK.. Cham: Springer International Publishing, 2020: 775-793.

[19] ZHANG C, CHEN X P, HAN G Q, et al. Spatial transformer network on skeleton-based gait recognition. 2022: 2204.03873. https://arxiv.org/abs/2204.03873v1.

[20] PENG C, ZHANG X Y, YU G, et al. Large kernel matters: improve semantic segmentation by global convolutional network//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1743-1751.

[21] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6230-6239.

[22] LU C, GUAN S. Research on the change detection method of urban high-resolution remote sensing images based on deep learning. Computer Application Research, 2020, 37(S1): 320-323.

[23] JI S P, WEI S Q, LU M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(1): 574-586.

[24] ZHANG H Y, LIN M H, YANG G Y, et al. ESCNet: an end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(1): 28-42.

基于FPCNet的遥感图像变化检测

李积英*, 王　奇, 石红萍

兰州交通大学 电子与信息工程学院, 甘肃 兰州 730000

摘　要： 目前，很多基于编码-解码深度学习的变化检测方法，在解码阶段进行不同分辨率特征融合时，往往存在语义不对齐现象以及判别结果中边缘细节和检测精度不够精确的问题。为此，基于编码解码思想，提出了双流像素分类相融合模块（FlowDual-PixelClsObjectMec，FPCNet）模型。在解码阶段，利用双流对齐技术进行特征矫正融合，以消除不同分辨率特征图融合时的语义不对齐现象。在判别阶段，基于对象级相似性度量和像素级分类思想，采用像素分类相融合模块（PixelClsObjectMec，PCOM）进行深度监督判别，解决了识别结果中存在的边缘细节和检测精度不够精确的问题。此算法在变化检测数据集（Change detection dataset，CDD）上，$F1$达到95.1%，在建筑物变化检测数据集（Building change detection dataset，BCDD）上，$F1$达到92.8%。和已知的一些算法相比，该算法具有更好的稳定性和鲁棒性，显示出一定的性能优越性。

关键词： 遥感图像变化检测；语义不对齐；双流对齐；深度监督判别

引用格式： LI Jiying, WANG Qi, SHI Hongping. FPCNet-based change detection for remote sensing images. Journal of Measurement Science and Instrumentation, 2025, 16(3): 371-383. DOI: 10.62756/jmsi.1674-8042.2025036