Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Automated Classification and Change Detection of Remote Sensing Images Based on Convolutional Neural Networks

Nannan Wu[1], Wenxiao Dong[1], Xiangjin Wu[2], Jianxun Li[3], Hongsheng Qian[1,✉]

[1] *School of Tourism and Planning, Pingdingshan University, Pingdingshan, Henan, 467000, China*

[2] *Hefei Urban Planning and Design Institute, Hefei, Anhui, 230000, China*

[3] *Chatone Smart Technology Co., Ltd., Beijing, 100000, China*

ABSTRACT

In order to better realize the automatic classification and change detection of remote sensing images, this paper proposes an automatic remote sensing image classification model based on CNN and migration learning, and constructs a remote sensing image change detection model by combining CNN and Transformer network. In the remote sensing image classification model, DenseNet network and Inception network are used as the backbone network, combined with the new channel attention module to mine the image features of remote sensing images, and then realize the accurate classification of remote sensing images. In the remote sensing image change detection model, the convolution operation of CNN with different sizes of void rate and expansion rate is utilized to better guide the feature map to focus on local information. Combined with the dynamic deformable Transformer to provide more accurate remote sensing image location information and detail information, to reduce the impact of background interference on remote sensing image change detection, and to improve the model's ability to recognize the target of remote sensing images. The parameter count and floating-point computation of the remote sensing image classification model are 7.69MB and 1.89GB, respectively, which are smaller than the parameter count and floating-point computation of the single network model. The RSICD models mF1 and mIoU are 1.66% and 0.58% higher than the optimal ones. Through the effective integration of convolutional neural networks and many different types of deep learning techniques, automated classification and change detection of remote sensing images can be realized.

Keywords: CNN, transfer learning; transformer network, attention mechanism, remote sensing image classification

1. Introduction

Remote sensing image classification refers to the process of distinguishing and categorizing different

✉ Corresponding author.
E-mail address: 13017562013@163.com (H. Qian).
Received 20 February 2024; Revised 15 April 2024; Accepted 20 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-267

objects or features in remote sensing images. In the traditional remote sensing image classification method, the pixel-level classification method is often used, i.e., each pixel is classified into different categories according to its reflectance or radiance value [1-4]. This method is simple and intuitive, but has a large error, especially for the classification of complex features and mixed pixels. Therefore, in recent years, the automation of remote sensing classification has become the main development trend, and this method can more accurately reflect the spatial characteristics and distribution of features and improve the accuracy of classification [5-8].

In addition to image classification, remote sensing technology can also be applied to change detection. Change detection refers to the comparison and analysis of multi-temporal remote sensing images to detect changes in features and environment [9-10]. Change detection in remote sensing images can help people to understand the land use change, the impact of natural disasters, urban expansion and other phenomena, so as to take appropriate measures to adjust and manage in time [11-13]. At present, change detection methods are mainly categorized into image-based and object-based methods [14]. Element-based methods are mainly used to detect changes between elements, i.e., comparing the differences in reflectance or radiance values of elements at the same location in two remote sensing images to determine whether the features have changed [15-18]. This method is simple and intuitive, but the matching requirement between pixels is high, and the effect of change detection is limited for complex features and mixed pixels [19-20].

As an important branch of modern surveying and mapping technology, remote sensing technology is more and more widely used in the field of geographic information [21-22]. Remote sensing image classification and change detection methods play a key role in surveying and mapping technology, which can help people better understand and utilize the information of the earth's surface, and the automated remote sensing image classification and change detection by convolutional neural network is of great significance to the development of this field [23-26].

Literature [27] proposed CE-CNN to classify high-resolution remote sensing images and compared MLP and conventional CNN, emphasizing that the classification accuracy of this method is better than other methods and has less distortion and erroneous results at land cover boundaries. Literature [28] proposed a CNN-based classification method for high-resolution agricultural remote sensing images, which has a very high correct rate in crop classification and provides a reference for the application of CNN in PA remote sensing. Literature [29] proposes a depth-patch based CNN system suitable for medium resolution remote sensing data. It outperforms pixel-based CNNs, patch-based neural networks, etc. in terms of overall classification accuracy. It is believed that combining it with a large amount of medium-resolution remote sensing data can realize a large range of land cover datasets. Literature [30] explored the use of localized description of deep convolutional features and feature encoding. Hellinger kernel and PCA were introduced and two aggregation strategies were proposed. It was revealed that Hellinger kernel, PCA and two aggregation strategies can improve the classification performance. Literature [31] used CNN architecture to classify substances in multispectral remote sensing images. The RUNet model that outperforms other models in remote sensing image classification is proposed. The potential of the method for applications such as land use monitoring is emphasized. Literature [32] proposed an accurate classification method for high resolution remote sensing images based on improved FCN model. The method has significant improvement in accuracy compared with existing methods, and its applicability to multi-resolution image classification is strong. From the above research results, it is easy to see that convolutional neural network has been commonly used in remote sensing image classification, but unfortunately, “automated classification of remote sensing images” has not yet received attention from academics, and the automatic classification of remote sensing images based on convolutional neural network has not been mentioned.

Literature [33] aims to present a review of existing deep learning based change detection algorithms for HR remote sensing images, outlines the issues that must be addressed to improve the performance

of the model by utilizing HR remote sensing images for change detection, and proposes research directions in this area. Literature [34] proposes BIT to efficiently model context in the spatio-temporal domain. The model outperforms several state-of-the-art CD methods, including outperforming attention-based methods in terms of efficiency and accuracy. Literature [35] describes supervised, unsupervised and semi-supervised deep learning techniques for different change detection datasets such as multispectral, hyperspectral, etc. And their challenges are discussed to understand the context of change detection datasets and deep learning model improvement. It is informative for the future development of CD methods. Literature [36] proposed new spatio-temporal attention-based neural network and designed CD self-attention mechanism to model the spatio-temporal relationships of diachronic images, emphasizing that the method is superior to other state-of-the-art methods. Literature [37] launched a systematic review and meta-analysis of deep learning algorithms in remote sensing image change detection. A deep learning based approach for remote sensing image change detection is presented. Promising new directions are identified for future research. Literature [38] proposed NestNet to improve the accuracy of automatic change detection task for remote sensing time series images. In terms of accuracy, the method performs better compared to other methods. It shows that NestNet is a potential method for change detection using high resolution remote sensing images. It can be seen that the application of convolutional neural networks in the detection of changes in remote sensing images is more common, and the results of its application are more obvious as far as the current research results are concerned.

The article takes convolutional neural network as the basis and combines with the theory of transfer learning to construct a deep transfer learning model applied to the automatic classification of remote sensing images. In this model, DenseNet network and Inception network are used as the backbone network, and the improved channel attention module is combined to mine the feature map information of remote sensing images, while realizing the automatic classification and recognition of remote sensing images. In addition, the RSICD model for remote sensing image change detection is constructed by combining convolutional neural network and Transformer, which pays full attention to the edge information of remote sensing images. And the detailed information of remote sensing image is fully extracted by the dynamic deformable Transformer module, so as to enhance the change detection accuracy of remote sensing image. For the practical application of the above two types of models, validation analysis is carried out with different types of datasets, which lays a new research foundation for automatic classification and change detection of remote sensing images.

2. Basic theoretical knowledge

With the rapid development of remote sensing technology, the resolution of remote sensing satellites is getting higher and higher, more and more spectral bands, and the revisit cycle is getting shorter and shorter, and people can get more useful data and information from remote sensing images. The concepts of remote sensing big data, remote sensing base model, and smart city have been put forward one after another in recent years, and the application of massive remote sensing data puts forward higher requirements for the intelligent extraction technology of remote sensing image information. Remote sensing image classification method, as an important part of the intelligent extraction technology of remote sensing image information, has important practical application significance in the fields of land use land cover, land resources survey, natural disaster observation, agricultural yield estimation, forestry protection and so on. With the rapid development of deep learning technology, big data, and big models, remote sensing image scene classification has achieved a series of brand-new results and faces new opportunities and challenges.

2.1. CNN and migration learning

2.1.1. Convolutional Neural Networks. A basic convolutional neural network structure usually contains 3 layers, which are input layer, hidden layer, and output layer. The implicit layer in turn contains 4 parts, which are convolutional layer, pooling layer, fully connected layer, and activation function [39].

1) Convolutional layer is the key to the design of the entire convolutional neural network (CNN), mainly through the linear operation of feature extraction and feature mapping, so it is also called the feature extraction layer, the structure is usually a combination of convolution, the parameters mainly contain the size of the convolution kernel, the number of convolution kernels, the step size, the padding method and so on. The computation of the j feature map of layer l is expressed as:

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} \cdot k_{ij}^l + b_j^l\right) \tag{1}$$

Where l is the number of layers where the convolutional layer is located, k is the convolutional kernel, b is the bias, f is the activation function, M_j is the input feature map, x_i^{l-1} is the ith feature map of the $l-1$th layer, k_{ij}^l is the convolutional kernel connected between x_i^{l-1} and x_j^l, and b_i^l is the bias of x_i^l.

2) The pooling layer usually appears after the convolutional layer, and the main task is to filter the output image of the convolutional layer, downsize the output features after the convolutional layer, reduce the size of the feature maps, simplify the computational complexity, save the computational cost, and at the same time, prevent the occurrence of overfitting. Common pooling methods mainly include maximum pooling and average pooling. The computation of the jth feature map of the lth layer is represented as:

$$x_j^l = f(w_j^l \cdot down(x_j^{l-1}) + b_j^l) \tag{2}$$

Where w_j^l is the pooling parameter of x_j^l, x_j^{l-1} is the jth feature map of layer $l-1$, down() is the pooling function, also known as the downsampling function, and b_i^l is the bias of x_j^l.

3) The fully connected layer is generally in the last layer of the neural network, connecting each neuron to the neurons of the previous layer of the network, and calculating the feature information extracted by integrating the convolutional and pooling layers. And classify the higher order invariant features of the input, output the high level features of the image, and finally use the classifier to calculate and get the final output, which is the probability of the category label corresponding to the input image. It can be expressed as:

$$y_i = f(W_i x_i + b_i) \tag{3}$$

Where, W_i is the weight of the neuron, f is the activation function, x_i is the input of the neuron, b_i is the bias and y_i is the output of the neuron.

4) The role of the activation function is to be able to add some nonlinear factors in the neural network to solve complex problems, and the common activation functions are Sigmoid function, tanh function, ReLU function.

In the choice of activation function, when the gap between the features of the data to be processed is small, the Sigmoid function is preferred. When the gap of data features to be processed is large, tanh function is more applicable. When the Sigmoid function and tanh function are used as the activation function, it is necessary to normalize the last input, otherwise the activation value will enter a fixed interval, and the output of the hidden layer will be close and lose the essential feature expression. The ReLU function, on the other hand, does not have to perform this operation, and the above phenomenon does not occur.

2.1.2. Transfer Learning Theory.In the remote sensing image automated classification scene, the amount of existing data is far from enough, so training a brand new deep neural network model is not suitable for the high-resolution remote sensing image automated classification task, and adopting the strategy of migration learning to build the scene classification model is a better choice. Figure 1 shows the difference between migration learning and traditional deep modeling methods. The migration learning method utilizes the similarity between the data, tasks and models of two task systems, for example, if their inputs are the same as images, speech or others, the source and target domains can be used for knowledge transfer, such as applying the trained model structure and parameters in the source task to the new task. In this case, the source task has a huge amount of data and trained models, and the new task has only a small amount of data [40].

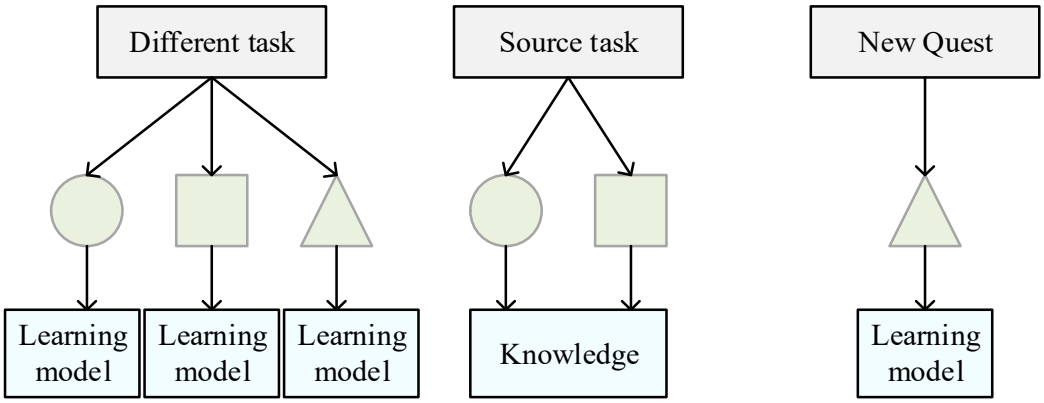

Fig. 1. The migration learning method is different

The basic principle of transfer learning is that deep CNNs are transferable. The image features learned at the first few layers of the deep network model are generally features that can be generalized. For example, image edges and colored patches, etc., while at deeper levels of the network model, image-specific features are learned. If the input data for two research objectives and tasks are closer, similar generic features can be extracted and migration learning will give good results.

There are two general ways to use migration learning, one is Fine-tune, which replaces the randomly initialized network weights of your own model with the parameters of the trained model, and then use the dataset of the project you are working on to re-train the model and fine-tune the parameters of the model. The other is Fixed Feature Extractor, which freezes all the convolutional layers of the pre-trained network, does not participate in parameter updating, and only serves as a feature extractor, and then trains the final fully connected layer. In this paper, we use the migration learning method to train the neural network model, and migrate the model structure and parameters trained with the trained model to the problem of automated classification of remote sensing images, to extract the deep features of remote sensing images.

2.2. Attention Mechanism Theory

2.2.1. Channel Attention.The essence of the channel attention mechanism is to assign trained weight coefficients to the feature maps of each channel of the feature map group by means of spatial compression [41]. The channel attention mechanism can dynamically assign larger weight coefficients to the channel feature maps that contribute more to the recognition task, while assigning smaller weight coefficients to the channel feature maps that contribute less to the recognition task, so as to achieve the effect of highlighting important information and suppressing irrelevant information. For an input feature map $X=\{x_1,x_2,...,x_n\}$, through a network layer with attention, the output feature map X' with attention is calculated as follows:

$$w_i = \frac{\exp(W_w x_i)}{\sum_{i=1}^{n} \exp(W_w x_i)} \tag{4}$$

$$X' = \sum_{i=1}^{n} w_i x_i \tag{5}$$

where w_i is the attention weight and W_w is the parameter in the network layer with attention.

The input feature map is first compressed in the spatial dimension using global average pooling, from the C*H*W data dimension is compressed into a C*1*1 one-dimensional vector, which is trained on each element of the vector through an excitation structure consisting of 2 fully connected layers. The trained one-dimensional vector is multiplied by the Sigmoid function to compress the vector elements to between (0,1) and the original feature map in the channel dimension to realize the weights assigned to each channel feature map.

2.2.2. Spatial attention.The spatial attention mechanism is able to model the importance of the spatial dimensions of the feature map, and to assign attention in the width and height dimensions of the data [42]. Specifically, for inputs of dimension (C,H,W), where C is the channel dimension, H is the height of the feature map, and W is the width of the feature map. Spatial attention is concerned with the importance of the data in the H and W dimensions, while for different channel features in the same spatial region, spatial attention treats them as equally important. The computational expression for the spatial attention mechanism for each channel dimension feature map is:

$$F_{out} = Broadcast(Attn \odot F_{in}) \tag{6}$$

Where $F_{in} \in \mathbb{R}^{H \times W \times C}$ and $F_{out} \in \mathbb{R}^{H \times W \times C}$ denote the input and the output feature maps weighted by spatial domain attention, respectively, C denotes the size of the channel dimension, $Atm \in \mathbb{R}^{H \times W}$ denotes the spatial domain attention mask weight, each element denotes the importance of a spatial location, $\odot$ denotes the Hadamard product operation of the matrix, and Broadcast denotes the broadcast mechanism. It can be seen that the dimensions of $Atm \in \mathbb{R}^{H \times W}$ and $F_m \in \mathbb{R}^{H \times W \times 1}$ are inconsistent and cannot perform the Hadamard product operation, so the broadcast mechanism should be applied when realizing the spatial self-attention, thus realizing the assignment of the spatial domain of the feature map.

2.2.3. Mixed Attention. Channel attention generally forms a hybrid attention module with spatial attention. With spatial attention, the important information in the original image can be converted into a spatial representation. This approach focuses attention on focused spatial regions and ignores unfocused irrelevant regions by generating weight values. The spatial interactions of features can be obtained and can be adaptively calibrated according to the spatial dimensions. The hybrid attention module that combines channel attention and spatial attention in a cascading combination in a channel-first-spatial order makes it easier to obtain cues that are most favorable to the output, but its pixel-level multiplication operation greatly increases the number of parameters in the network. This attention process can be represented as:

$$F' = M_C(F) \otimes F \tag{7}$$

$$F'' = M_S(F') \otimes F' \tag{8}$$

where F is the original input of the feature, M_C denotes the one-dimensional channel feature, F' is the feature outputted by the channel attention mechanism, M_S denotes the two-dimensional spatial feature, $\otimes$ is the pixel multiplication operation, and F'' is the final input feature. Where M_C and M_S are shown below:

$$M_C(F) = f(mlp(AvgPool(F)) + mlp(MaxPool(F))) = f(W_1(W_0(F^c_{ang})) + W_1(W_0(F_{\max^c}))) \quad (9)$$

$$M_s(F) = f(f_{conv}([AvgPool(F); MaxPool(F)])) = f(f_{conv}([F^s_{avg}; F_{\max^s}])) \quad (10)$$

Where, F_{avg} is the average pooling operation, $F_{\max}$ is the maximum pooling operation, W is the weights, f_{conv} is the convolution process and f is the activation function.

2.3. Transformer Network

CNNs are commonly used to extract deep features from images, and because the size of its convolution kernel is fixed, CNNs can extract local detail features well, but CNNs lack the ability to extract global contextual features and cannot establish long distance dependencies. In order to solve this problem, Transformer network is proposed, which is mainly used in the field of natural language processing to deal with the task from sequence to sequence.Transformer network is mainly divided into three parts, namely, input module, encoder and decoder.In the input module, the sequence data is encoded in position, which is because Transformer does not have the position information of the sequence data. So additional position coding information needs to be introduced. In the encoder, a coding block is formed by a multi-head self-attention and a feed-forward neural network using residual connections, and N coding blocks are used to extract the feature information of the input data. In the decoder, a decoding unit is composed of a multi-head mask self-attention, a multi-head self-attention and a feed-forward neural network utilizing residual connections, and the decoder is composed using N decoding units, linear layers, and SoftMax functions [43].

1) Location coding module. Transformer network model is a sequence-to-sequence end-to-end model, its input data is serialized data, but its network structure does not have the operation of obtaining the location information of the sequence data, so it is necessary to additionally obtain the location information of the sequence data through the constructed location encoding module.

2) Multihead Attention Module. Multihead attention module is the core module of Transformer network model, and the multihead attention mechanism is composed of multiple self-attention mechanisms.

In the multi-head attention module Q, K, V are query vector, key vector and value vector respectively, which are usually the input of the model or the output of the previous encoding module in Transformer, and it can be expressed as:

$$Attention(Q, K, V) = softmax(\frac{QK}{\sqrt{d_k}})V \quad (11)$$

The multi-headed attention mechanisms Q, K, and V first pass through multiple linear operations, then through h self-attention mechanisms, and then through a linear layer after cascading them together again. Where h denotes the multiple heads of attention, that is, the number of self-attention. The multiple heads of attention are denoted as:

$$MultiHead(Q, K, V) = Concat(head_i, head_h)W^O \quad (12)$$

Parameters are not shared between modules in the Transformer network model, so rich global features can be learned.

3. Automatic remote sensing image classification model

With the gradual maturation of remote sensing imaging technology, remote sensing image scene data with higher resolution continue to appear, and the research on remote sensing image scene

classification methods has received extensive attention. As remote sensing scene images are characterized by large intra-class diversity and high inter-class similarity, the task of high-resolution remote sensing image scene classification faces great challenges. In addition, high-resolution remote sensing image scene classification methods are of great significance in urban planning, military target identification, and disaster assessment. Therefore, remote sensing image scene classification has important theoretical research significance and application value. Currently, deep learning and CNN are widely used in the field of computer vision, which makes the accuracy of target detection, semantic segmentation and image classification greatly improved. Therefore, introducing CNN into the automated classification of remote sensing images is a hot spot in the research of land scene classification.

3.1. Remote Sensing Image Classification Model

3.1.1. Backbone network selection

1) DenseNet network. The basic structure of DenseNet is a dense block Dense Block and a transition layer Transition Layer.The basic structure of Dense Block is BN-ReLU-1×1 Conv-BN-ReLU-3×3 Conv, and each Dense Block is connected to each other by Transition Layer. The basic structure of Transition Layer is BN-ReLU-1×1 Conv-Average Pooling (2×2).

Dense Block is an important part of DenseNet, which is a continuous layer composed of multiple convolutional layers. In DenseNet, each Dense Block is connected to all the previous Dense Blocks, thus enabling very tight feature extraction and information transfer.The main purpose of Dense Block is to increase the depth and width of the network while using the idea of dense connectivity to enhance feature reuse and gradient flow.

The Dense Block in DenseNet consists of multiple densely connected layers, and the output of each densely connected layer is passed to the next densely connected layer, resulting in a network with strong feature extraction and feature reuse capabilities. Assume that the input image feature is X_0, and as it passes through the n-layer DenseNet network, each layer performs a nonlinear transformation $H_n(.)$. Layer n consists of the feature mappings of all the previous convolutional layers, assuming that the input feature mappings of layers 0 to $n-1$ are connected as $X_0 \ldots X_{n-1}$. Thus, the n-layer DenseNet network has $n\times(n+1)/2$ connections, and the output of layer n is characterized as $X_n = H_n([X_0 \cdots\cdots X_n - 1])$. where X_n is the current layer n, $[X_0 \cdots\cdots X_{n-1}]$ is the concatenation of the mappings obtained from layers 0 to $n-1$, and $F_n(.)$ is a composite function of the batch normalization (BN) and the corrected linear unit (ReLU).

In DenseNet, Transition Layer is an intermediate layer connecting two dense blocks. It has two main roles, one is to reduce the dimensionality of the feature maps, due to the large number of channels in the feature maps in the dense blocks, this can lead to a large computational burden and memory consumption in the later stages of the network. Therefore, the transition layer can reduce the dimension of the feature map by using pooling layer or 1×1 convolution to reduce the computational complexity and memory consumption. Secondly, to control the complexity of the model, the size of the feature graph and the number of channels may keep increasing during the training process, which will lead to the increasing complexity of the model. The transition layer can control the complexity of the model to make it easier to train and optimize.

2) Inception Network. The Inception network performs parallel operations on different convolutional operations on the same mapping and then connects all the results to a single output. That is, each layer of the model performs a 3*3 convolution operation, a 5*5 convolution operation, and a 3*3 maximum pooling operation, and it is up to the next layer of the model to decide how and whether to apply each piece of information. In this way, the computational cost of the Inception network will increase dramatically. Because not only the 5*5 convolutional operation itself has a high computational cost, but also the different filters stacked side by side will increase the number of feature maps in each layer accordingly, in addition, the output of the pooling operation is equal to the number of feature maps

inputted from the previous layer is also an important factor to increase the computational cost. Therefore, the model used in this paper is the improved Inception network, which adds a 1*1 convolution operation for dimensionality reduction where the computational load is large, so that the neurons of the input feature maps are fully connected to the newly added 1*1 convolution layer. The computationally heavy 3*3 and 5*5 convolution filters are then accessed so that the number of connections is reduced to a manageable number of 1*1 convolutions.

3.1.2. New channel attention.In order to make the features extracted by the DenseNet121 model more discriminative, a new channel attention module (MECANet) is introduced into the DenseNet121 model. The design of this module draws on the previous design ideas of channel attention and the CAM module in CBAM. That is, the ECANet module adopts a one-dimensional convolutional layer to replace the fully connected network composed of two fully connected layers to capture the nonlinear dependencies between different channel feature vectors, which reduces the computational complexity of the channel attention module and realizes more effective utilization of different channel feature vectors.The CAM module takes the global average pooling value and the global maximum pooling value as the global spatial information descriptor on each channel feature vector. It is conducive to capturing channel feature vectors that are more effective than using a single pooling value as the global spatial information descriptor for the automated classification of remote sensing images, thus facilitating the subsequent automated classification of remote sensing images.

Global pooling is performed for feature vector X of input size $H\times W\times C$ in spatial dimension. In order to extract the valuable channel feature vectors more perfectly and to reduce the loss of spatial feature information in the compression process, three spatial information descriptors $X_{avg}, X_{\max}, X_{sto}$ of size $1\times 1\times C$ are obtained using global average pooling, global maximum pooling, and global random pooling, respectively. i.e:

$$X_c^{avg}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}X_c(i,j),\quad c=0,1,\ldots.C \tag{13}$$

$$X_c^{\max}=\max_{1\le i\le H;}X_c(i,j),\quad c=0,1,\ldots.C1\le j\le W \tag{14}$$

$$X_c^{sto}=multi\left(\frac{X_c(1,1)}{\sum_{i=1}^{H}\sum_{j=1}^{W}X_c(i,j)},\ldots,\frac{X_c(i,j)}{\sum_{i=1}^{H}\sum_{j=1}^{W}X_c(i,j)},\ldots,\frac{X_c(H,W)}{\sum_{i=1}^{H}\sum_{j=1}^{W}X_c(i,j)}\right),c=0,1,\ldots.C \tag{15}$$

where H,W,C denotes the height, width, and number of channels of feature vector X, respectively, $X_c^{avg}, X_c^{\max}, X_c^{sto}$ denotes the global average pooling value, the global maximum pooling value, and the global random pooling value of the feature vector on the c th channel, respectively, and multi(·) denotes polynomially distributed randomness sampling.

In order to avoid the problems of the fully connected network's dimensionality reduction operation affecting the channel attention prediction and the inefficiency of capturing the nonlinear dependencies among all channel features. A one-dimensional convolutional layer is used instead of the fully connected network to obtain nonlinear dependencies between feature vectors across channels without dimensionality reduction. Then:

$$X_{avg}^{n}=conv1D_k(W,X_{avg}) \tag{16}$$

$$X_{\max}^{n}=conv1D_k(W,X_{avg}) \tag{17}$$

$$X_{sto}^{n}=conv1D_k(W,X_{avg}) \tag{18}$$

Where $X_{avg}^{n}, X_{\max}^{n}, X_{sto}^{n}$ represents the feature vectors of different spatial feature information descriptors after passing through the one-dimensional convolutional layer, respectively, $conv1D(*)$ represents the one-dimensional convolutional operation, k represents the number of channels realizing the cross-connection, and W represents the one-dimensional convolutional kernel parameters.

The three feature vectors extracted by fusing the one-dimensional convolutional layer containing the nonlinear dependencies of the channels are then obtained by the Sigmoid activation function to obtain the normalized channel weight coefficient matrix w_{MECA}, and then the weight coefficient matrix is convolved with the feature vectors x to obtain the channel attention-weighted feature vectors X_{MECA}. Then:

$$w_{MECA} = \sigma(X_{avg}^{n} + X_{max}^{n} + X_{sto}^{n}) \tag{19}$$

$$X_{MECA} = conv(w_{MECA}, X) \tag{20}$$

where $\sigma(\cdot)$ denotes the formula for the Sigmoid activation function and $conv(*)$ denotes the two-dimensional convolution operation.

3.1.3. Classification model framework.In migration learning, instead of training the model in the target domain from scratch, the network is first trained using a large-scale training dataset, i.e., within the source domain. Secondly, a part of the network trained on the source domain is transferred as a part of the new network designed in the target domain. In this paper firstly the feature extraction module of the network is trained on the source domain i.e. ImageNet dataset to obtain the weights, then the weights obtained from the pre-training are loaded into the new network instead of the random initialization of the weights, the classification layer of the original network is removed and replaced with a new classification layer. Finally, the remote sensing dataset is input for training.

Subsequently, for the automated classification of remote sensing images in the source domain, this paper constructs a remote sensing image classification network model (DI) based on the new channel attention, whose specific structure is shown in Fig. 2, consisting of a feature extraction module, a global pooling layer, a splicing layer and a classifier.

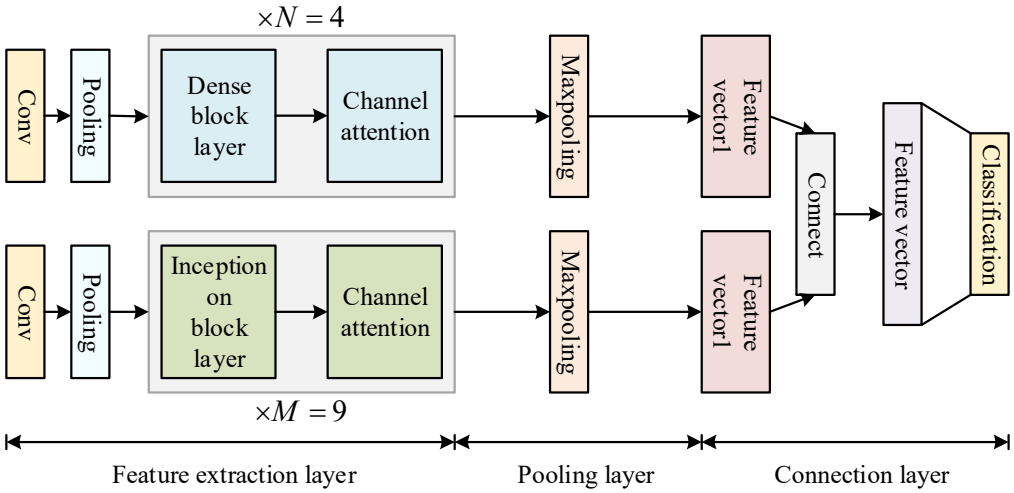

Fig. 2. Overall architecture of the DI network

The feature extraction modules include DenseNet feature extraction module, which consists of multiple Dense Blocks plus channel attention.Inception feature extraction module, which consists of multiple Inception Blocks plus channel attention. The channel attention layer is added to add weights to the output feature maps. The global pooling layer takes the feature maps extracted by the DenseNet

feature extraction module and the Inception feature extraction module and performs global maximum pooling to obtain feature vectors with dimensions of 1024 and 2048, respectively. The channel splicing layer splices the two feature vectors into a new feature vector of 3072. The classifier puts the obtained 3072-dimensional feature vector into the SoftMax classifier for classification. The feature extraction module of the model is pre-trained using the ImageNet natural dataset and the obtained weights are migrated over before the subsequent training on the remote sensing dataset.

3.2. Experimental environment and evaluation indicators

3.2.1. Experimental environment design.The experimental data used in this chapter are from the remote sensing image dataset RSSCN7, which is a commonly used dataset for remote sensing image scene categorization, in which there are seven categories of data, including Grass, Field, Industry, RiverLake, Forest, Resident and Parking, and there are Grass, Field, RiverLake, Forest, and also Industry, Resident and Parking, which represent the natural factors with a wide coverage, RiverLake and Forest, and also Industry, Resident and Parking, which represent the scenes of human production and life, and are characterized by wide coverage. The pixel of each image is 500*500, and there are 500 images for each category, totaling 3000 images. The experiment in this chapter is to set up this dataset according to the ratio of 8:2 for the training set and the test set, that is, there are 7 categories totaling 2400 images in the training set, and 7 categories totaling 600 images in the test set as well.

The experiments in this chapter are built on the basis of the Windows 10 flagship edition system with the TensorFlow deep learning framework, with the GPU processing environment of CUDA9 and CUDNN7, and the experiments in this chapter are conducted under the cooperation of various hardware and software conditions, and its specific hardware and software environment configuration is shown in Table 1.

Table 1. Experimental hardware and software environment configuration

Experimental environment		Concrete configuration
Hardware environment	CPU	i3-10560k CPU, 4.28HZ
	GPU	NVIDIA Tesla K20c GPU
	Memory	64G
Software environment	Operating system	Windows 10 64 bit
	Algorithm integration framework	TensorFlow-GPU
	Programming language	Python
	GPU processing frame	CUDA9 and CUDNN7

3.2.2. Selection of evaluation indicators.Four commonly used metrics, namely Overall Accuracy (OA), Average Accuracy (AA), Kappa Coefficient (Kappa), and Confusion Matrix (CM), are used to quantitatively assess the performance of automated classification of remote sensing images.

1) Overall accuracy. Defined as the number of all correctly classified images divided by the number of the entire test dataset. It is a direct measure that reveals the classification performance on the entire dataset. It can be expressed as:

$$OA = \frac{N_t}{N_t + N_f} \tag{21}$$

Where N_t and N_f denote the number of correctly and incorrectly categorized samples. oa denotes the performance of the model in predicting image categories, ranging from 0 to 1.

2) Average accuracy. It means the average prediction accuracy per category and can be described as:

$$AA = \frac{1}{n}\sum_{i=1}^{n}\frac{C_i}{N_i} \tag{22}$$

where C_i is the number of correctly categorized samples in category i, N_i is the number of total samples in category i, and n is the number of categories.

3) The Kappa coefficient is a metric used to measure classification accuracy. Its calculation formula is:

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{23}$$

where p_0 is the sum of the number of correctly classified samples in each category divided by the total number of samples, i.e., the overall classification accuracy.

Assuming that the true number of samples in each category is $a_1, a_2, ...a_c$ respectively, and the predicted number of samples in each category is $b_1, b_2, ...b_c$ respectively, and the total number of samples is n, we have:

$$p_e = \frac{a_1 \times b_1 + a_2 \times b_2 + \cdots + a_c \times b_c}{n \times n} \tag{24}$$

4) The confusion matrix is a specific table layout to visualize the classification results for each remotely sensed scene class. Each row in the CM represents the predicted category, the total number of data in each row indicates the number of data predicted for that category, and the value in each row indicates the number of real data predicted for that category. Each column represents the true class to which the data belongs, and the total number of data in each column represents the number of data instances in that class.

3.3. Remote Sensing Image Classification Experiment

3.3.1. Classification Model Training.In order to realize the automated classification of remote sensing images, this paper constructs a remote sensing image classification network model based on CNN and migration learning, combined with DenseNet121 network and new channel attention mechanism. Before carrying out the validation of the classification model, the model is trained based on the selected dataset as a way to ensure that the model obtains better classification results when performing remote sensing image classification. Figure 3 shows the training accuracy and loss curves of the DI model.

From the figure, it can be seen that since the parameters of the model are randomly initialized by PyTorch, the accuracy in the first iteration cycle is very low, 23.46% and 27.58% in the training and validation sets, respectively. In the next 100 iteration cycles, the accuracy of the training set increases rapidly and the loss value decreases rapidly, and the accuracy and loss value of the test set oscillate up and down. After the 100th iteration cycle, the change trend of the model slows down, and the classification accuracy on the training set stabilizes at about 95.54%, while the test set fluctuates up and down with a large magnitude in the 75th and 165th iteration cycles, and the loss value of the DI model on the test and validation sets stabilizes after the 200th iteration cycle, at which time the model has converged. As can be seen from the figure, the difference between the average accuracy of the DI model in the training set and the test set is more than 22%, and the difference between the loss values is more than 0.5. Overfitting refers to the phenomenon that the model predicts the known data very well but the unknown data very poorly when the amount of training data is small. The DI model has a large difference between the classification results on the training set and the test set, indicating that the model has entered a state of overfitting.

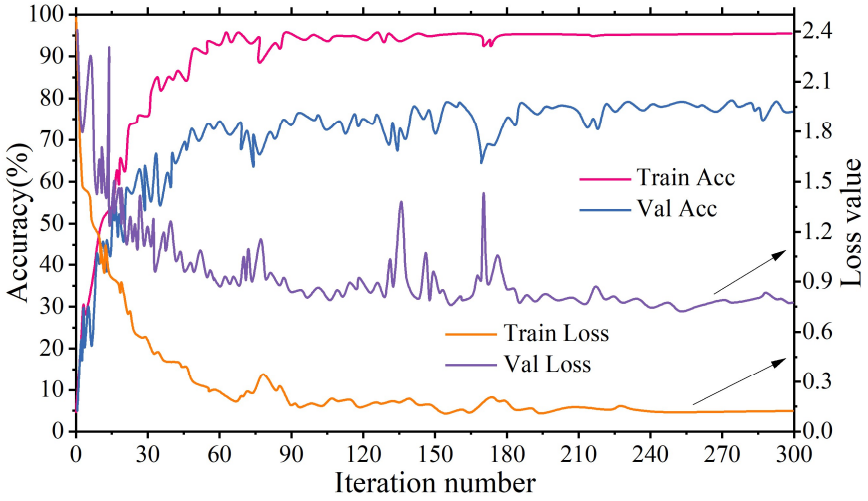

Fig.3. Training accuracy and loss curve of DI model

Taking the trained DI model as the basis for classifying remote sensing images of different categories in the RSSCN7 dataset, and using the overall accuracy (OA) as the evaluation index, the specific classification accuracies of the model on each category of the validator after the training is completed are obtained as shown in Fig. 4.

From the figure, it can be seen that after the training is completed, the DI model has a high classification effect on the validation set for Grass, Industry, and Parking, which are three more homogeneous types of remotely sensed features, and the accuracy of the DI model reaches 92.73%, 91.98%, and 90.89% for these three types of scenes, respectively. The model has lower classification accuracy for Resident and Forest with OA values of 85.12% and 88.54%, respectively. The class inclusions of these two classes of images are highly complex and both include features such as house

buildings, trees and roads, and are highly similar between classes. They are only categorized into different classes due to different sparsity of features, indicating that the model has poor ability to characterize the sparsity of features within classes. However, on the whole, the DI model designed in this paper can realize a more accurate classification effect of remote sensing images, which can provide reliable identification and classification results for urban planning and disaster assessment.

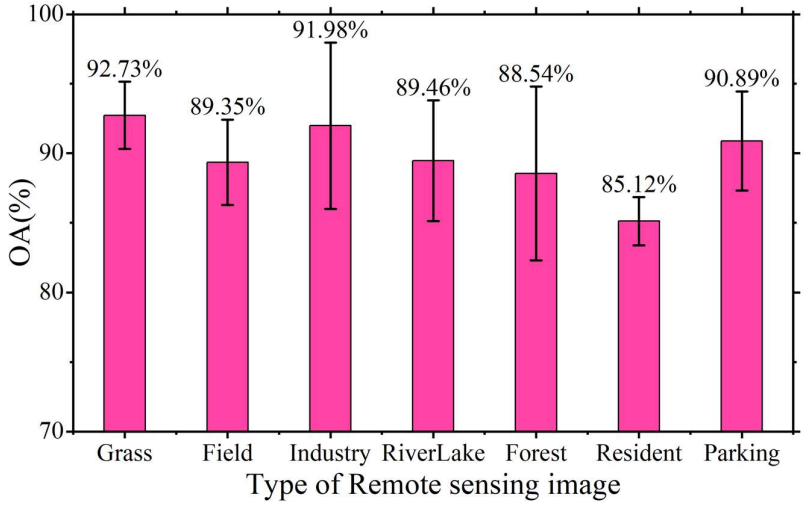

Fig. 4. Specific classification accuracy of the DI model

3.3.2. Model Comparison Validation.In order to further verify the effectiveness of the model in this paper in the automated classification of remote sensing images, this paper carries out model comparison experiments. GLCM, LBP, GLCM-LBP, AlexNet80, LBP-CNN, and ResNet50 are selected as the comparison models, and the classification accuracy (Pre), average accuracy (AA), and Kappa coefficient are used as the evaluation indexes, and the comparison results of remote sensing image classification of different models are obtained as shown in Table 2.

As can be seen from the table, the DI model on RSSCN7 dataset improves 25.23% and 18.27% in classification accuracy and 44.79% and 46.22% in AA and Kappa coefficient, respectively, compared with GLCM and LBP individual features. The DI model designed in this paper gets the optimal result among the comparison results of various models, which is due to the fact that the model in this paper first utilizes DenseNet network and Inception network for remote sensing image feature extraction, and then combines the MECANet attention to obtain the nonlinear dependency between different channel feature vectors. The ability to express image information after fusion of features is better than single feature, and different shallow features can play the role of complementary information, thus enhancing the generalization ability and classification performance of the model.

Table 2. Remote sensing image classification of different models

Model	Pre (%)	AA	Kappa
GLCM	65.42	0.853	0.614
LBP	72.38	0.805	0.608
GLCM-LBP	78.94	0.869	0.787
AlexNet80	76.53	0.827	0.742
LBP-CNN	84.81	0.898	0.816
ResNet50	76.86	0.831	0.753
Ours	90.65	0.924	0.889

In addition, the classification effect for the model in this paper is also visualized by the confusion matrix, and Figure 5 shows the confusion matrix of the DI model on the RSSCN7 dataset. From the confusion matrix, it can be seen that the classification accuracies of grassland, industry and parking

lot are over 90%, except for the four categories of remote sensing images, namely, field, river and lake, forest and residential area, whose classification accuracies do not exceed 90%. Among them, 5.89% of the images misclassified in the field category were misclassified into the grass category, which is due to the fact that the overall remote sensing image of fields is closer to grass, thus leading to the classification error. In addition, 11.85% of the images misclassified in the residential area category were misclassified into the parking lot category, which also affected the overall classification accuracy of the DI model to some extent. The above results show that the traditional handmade features cannot fully express the information of complex feature images, and cannot effectively solve the problem of high inter-class similarity and low intra-class similarity. Moreover, after the introduction of migration learning and new channel attention mechanism in this paper, the extraction ability of feature information of complex remote sensing images can be significantly enhanced, which helps to better realize the accurate classification and recognition of remote sensing features.

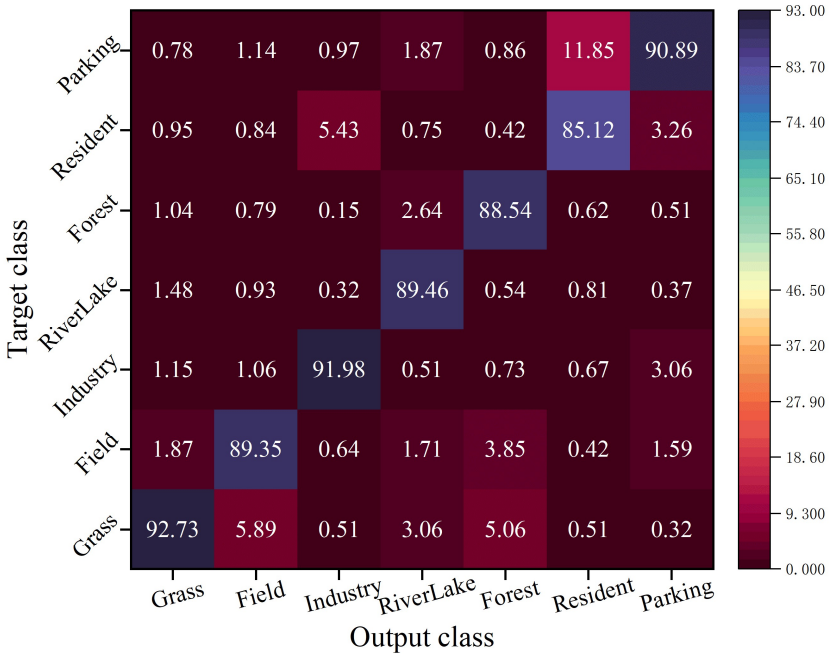

Fig. 5. The confusion matrix of the DI model

3.3.3. Robustness of models.In this paper, the combination of Densenet121 network and Inception network is selected in the backbone network of the DI model, which aims to further improve the extraction of image features from remote sensing images, as well as effectively improve the classification accuracy of remote sensing images. In order to verify the contribution of this network to the model, this paper compares the performance of different deep networks. The average accuracy (AA), the computational amount of the network (FLOP) and the number of parameters (Params) of different network models on the RSSCN7 dataset are statistically analyzed, and their specific results are shown in Table 3.

As can be seen from the table, the average accuracy of the DI model fused with Densenet121 network and Inception network on the dataset reaches 92.89%, which is 1.65% and 2.04% higher than the average accuracy of the single Densenet121 network and Inception network, respectively. And the number of model parameters and floating-point computation volume of the fusion network constructed in this paper are 7.69MB and 1.89GB, respectively, and the number of model parameters is reduced by 2.04% and 4.59% compared with the single Densenet121 network and Inception network, respectively, and the floating-point computation volume is reduced by 33.68% and 35.05% compared with the single Densenet121 network and Inception network, respectively. 33.68% and 35.05%. Although the number of parameters and floating-point computation of the fusion network in

this paper have a slight disadvantage compared with the Shufflenet network, based on the consideration of the average accuracy of remote sensing image classification, it can be considered that the performance of the fusion network designed in this paper is relatively better, and it can meet the automated classification and recognition of remote sensing images.

Table 3. Different networks are compared to the index

Network	AA (%)	Params (MB)	FLOP (GB)
Densenet121	91.24	7.85	2.85
Inception	90.85	8.06	2.91
ResNet	91.16	25.43	4.17
ResNeXt	91.27	23.61	4.36
ConvNeXt	91.38	24.87	4.28
VGG16	82.32	136.38	15.64
Shufflenet	91.91	7.51	0.72
Ours	92.89	7.69	1.89

4. Remote sensing image change detection model

Change detection determines the different processing of the same object or phenomenon according to its observation at different times. Remote sensing image change detection is the use of multi-source remote sensing images and related geospatial data covering the same surface area at different times, combined with the characteristics of the corresponding features and remote sensing imaging mechanism, and the use of image and graphic processing theories and mathematical modeling methods, to identify and analyze the changes in the features in the geographic area, including changes in the location and scope of the features, and the changes in the nature and state of the features. The purpose of its research is to identify the change information of interest and filter out irrelevant change information as interference factors. With the rapid development and application of remote sensing earth observation technology, the system of change detection technology is also constantly developing and evolving.

4.1. Remote sensing change detection model

4.1.1. Overall framework of the model.Effective fusion of local and global features helps to improve the change detection accuracy of remote sensing images, CNN captures local features in remote sensing images through convolutional operations and hierarchical feature representation, while Transformer network realizes the extraction of global features in images through cascading self-attention mechanism. In order to fully utilize the local details and global semantic features in remote sensing images, this paper proposes a remote sensing image change detection network (RSICD) based on Edge Information Guidance Module (EIG) and Dynamic Deformable Transformer (DDaT). The overall architecture of the network is shown in Fig. 6, and the RSICD model inputs the dual time-phase image to the weight-sharing encoder to get the difference image of each phase, and feeds it to the decoder to finally get the change detection result.

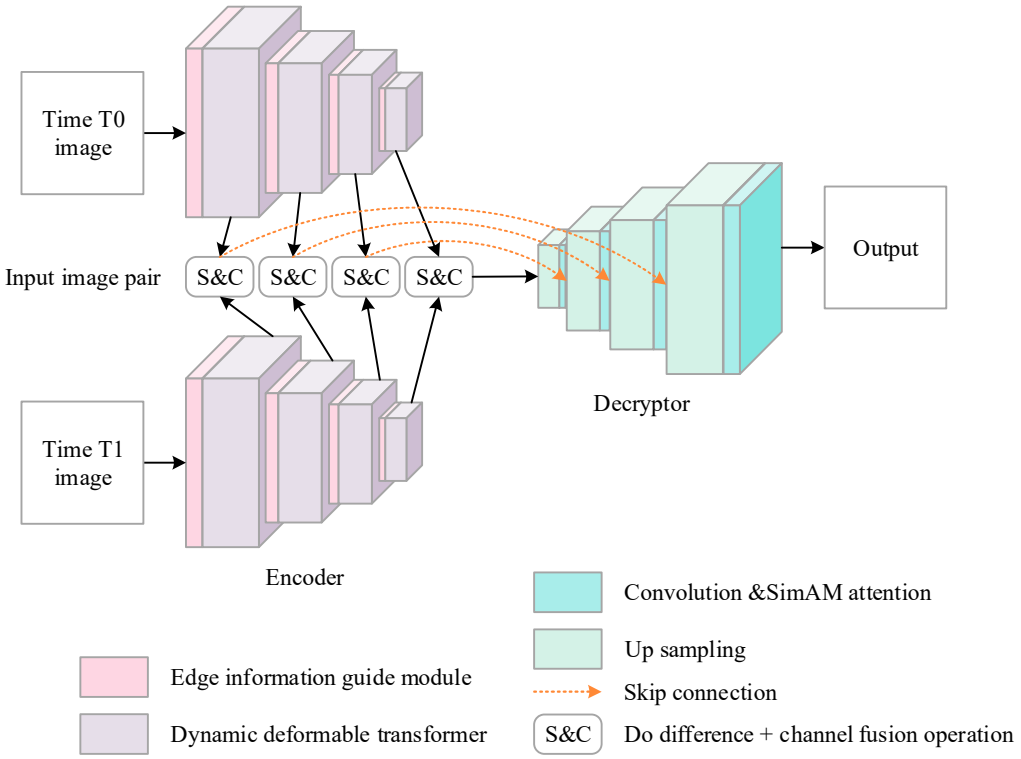

Fig. 6. The overall architecture of the RSICD network

RSICD feeds the local detail information extracted by the CNN into the Transformer to enhance the global information features, thus obtaining a more excellent feature representation.RSICD adopts a standard coding and decoding structure with shared encoder weights, which consists of two modules cascaded by EIG and DDaT. In the coding stage, the edge integrity of the image change region is firstly enhanced by EIG, and then the change target region is better matched by DDaT, which reduces the negative impact of background pixels on the network accuracy. In the jump-connected difference fusion module, the bitemporal feature maps obtained from each coding layer are first channel fused and downscaled, and then subsequently added with the difference maps obtained through the difference operation. This module can provide richer dual-time-phase information and solve the problem of information loss caused by the traditional difference operation. In the decoding stage, the difference map with rich semantic information is up-sampled and decoded by bilinear interpolation. It is then fed into a convolutional decoder, which consists of a convolutional layer and SimAM attention. Different from the traditional attention mechanism, SimAM attention focuses on both the spatial and channel dimensions of the target region, and explores the importance of each neuron by optimizing the energy function, which is a simple and efficient attention mechanism that does not require the introduction of additional network parameters to further enhance the semantic information. Finally, the shallow feature information obtained from the coding layer is fused into each output layer to enrich the detail information, which finally forms the codec structure change detection network.

4.1.2. Edge information guidance module.In recent years, much progress has been made in effectively integrating edge information into CNNs, but little literature has investigated the effective combination of Transformer and boundary information. For the change detection task, the detailed description of boundary features can effectively improve the accuracy of change results. For this reason, this study embeds an Edge Information Guidance Module (EIG) in the front-end of each DDaT to extract the features of remote sensing images.

The designed EIG has convolutional operations with different sizes of nulling and swelling rates. The purpose is to capture features with more receptive fields and thus refine the edge features. For feature map $X \in \mathbb{R}^{L\times h\times w}$, L denotes the channel dimension.

First, the EIG needs to perform convolution operations with different nulling and expansion rates for X. Batch normalization and nonlinear activation functions are also needed after each convolution operation, which are intended to make the model more stable to avoid gradient explosion or gradient vanishing. The formulas are described below:

$$x_i = \mathrm{R}\left(\mathrm{B}(\mathrm{Conv}(k,d,p)(X))\right) \tag{25}$$

where i denotes the number of layers, k denotes the kernel size, d denotes the null rate, and p denotes the expansion rate. B denotes batch normalization. R denotes the ReLU activation function. When $i=0$, only the convolution operation with convolution kernel 1 is performed, i.e., $k=1$. And at this time, the null rate and the expansion rate are not set, i.e., $d=0, p=0$. When $i=[1,2,3,4]$, $k=3$ is set uniformly, and $d=[1,2,4,6]$ as well as $p=[1,2,4,6]$ are set respectively. Then, the feature maps with different receptive fields can be obtained in this study $x_i \in \mathbb{R}^{\frac{L}{r}\times h\times w}, i\in[0,1,2,3,4]$, respectively. x_i is able to effectively capture the building change regions at different scales. r is the channel attenuation rate, which can reduce the number of parameters of EAM and reduce the computational burden.

Next, the feature maps need to be aggregated by performing channel dimension concatenation on x_i. After performing 1×1 convolution operation and linear activation function on the aggregated feature maps, the attention confidence w can be obtained. w The weight redistribution of the feature maps can be realized, thus effectively guiding the feature maps to focus on local information. The enhanced feature map is able to recognize significant change regions and suppress irrelevant background. The specific formula is described as follows:

$$w = \xi(Cat(x_i)) \tag{26}$$

$$X^{'} = X \oplus (X \otimes w) \tag{27}$$

where $Cat(\cdot)$ denotes concatenate. $Conv_{1\times1}$ denotes convolution operation with convolution kernel 1. ξ denotes Sigmoid activation function. $\otimes$ denotes element multiplication. $\oplus$ denotes elementwise addition. The feature map $X^{'}$ with boundary enhancement can be obtained by using v to perform multiplication and addition operations on the original feature map X.

4.1.3. Dynamic deformable Transformer.After completing the feature extraction, this paper, in order to further suppress the interference information in the complex background and improve the accuracy of foreign object target detection, this paper designs the Dynamic Deformable Transformer decoder, which dynamically generates the position query to provide more accurate positional and detail information, reduce the impact of the background interference on the target detection, and improve the model's ability to recognize the target.

The cross-attention layer in the DDaT module is a basic module for updating queries by aggregating the image context. Attention weights are obtained by calculating the similarity between the query Q and the keys K and values V generated from the image features, and the formula for cross attention is as follows:

$$Crs\text{-}Attention(Q,K,V) = soft\max\left(\frac{QK^{r}}{\sqrt{D}}\right)V \tag{28}$$

where $Q \in \mathbb{R}^{N\times D}$ denotes the query, $\mathbb{R}$ denotes the real number space, N is the number of queries, and D is the dimension of the vector. Each query Q consists of a content query Q_e and a location query Q_p. The image features K_e and its location encoding K_p together form the key

$K \in \square^{N\times D}$. value $V \in \square^{N\times D}$ contains the feature information in the image. Both the content part and the location part of Q and K help to compute the attention score.

In order to enhance the model's focus on important regions, this paper introduces dynamic focus-aware location queries in the cross-attention computation process, which dynamically generates queries containing location information by aggregating the location encoding K_p of the previous decoder block, which is denoted as:

$$Q_p^t = h(A^{t-1}K_p^{t-1} + B) \tag{29}$$

where t is the index of the t nd Transformer decoder block, A^t is the cross-attention score from the $t-1$ th Transformer decoder block, $B \in \square^{N\times D}$ is the learnable network parameters, and $h(\cdot)$ denotes the mapping module, consisting of a two-layer multilayer perceptron. The bias term B represents the original randomly initialized learnable location query. By generating the location query in this dynamic way, more accurate location information and local details can be provided to the query.

In order to make the model better able to detect and segment these fine targets, this paper also uses a high-resolution cross-attention HRCA layer to mine the detail information in the high-resolution feature map and optimize the segmentation results. In the high-resolution cross-attention layer, it is first necessary to obtain the cross-attention score at low resolution A_t. Then, its corresponding cross-attention score at high resolution is computed by the bilinear up-sampling operation $f(\cdot)$. $A_h = f(A_l)$ Finally, the top k pixels with the highest scores in A_h are merged into the set Ω for the high-resolution cross-attention, which is computed as follows:

$$HRCA(Q,K,V,\Omega) = soft\max\left(\frac{QK^T}{\sqrt{D}}\right)V \tag{30}$$

where $K' = g(K,\Omega), V' = g(V,\Omega)$, g are indexing operations to generate K' and V' based on Ω selecting vectors from K and V respectively.

4.2. Data sets and evaluation indicators

4.2.1. Data set selection.In this paper, the algorithm is applied to the remote sensing image change detection task, and two open-source high-resolution (building) remote sensing image change detection datasets are used as the experimental datasets to validate the effectiveness of the algorithm designed in this paper, which are the LEVIR-CD dataset and the S2Looking dataset.

1) LEVIR-CD dataset. The LEVIR-CD dataset is a high-resolution remote sensing image change detection dataset, which is mainly satellite data of urban areas obtained from Google Earth, mainly through mapping satellites to ensure high resolution and geometric accuracy. The dataset contains 31,352 independent examples of changing buildings, captured from 2000 to 2019, all from different areas of different cities. The size of each of these images is 1024×1024, the spatial resolution is 0.5m/pixel, the image format is PNG image, and each set contains three images, which are the image of time-phase I, the image of time-phase II, and the real change map.

2) S2Looking data set. The S2Looking dataset is a building change detection dataset containing large-scale side-looking satellite images. This dataset mainly targets rural areas captured by surveillance or reconnaissance satellites through different off-zero angles, mainly from satellites such as GF, SV, BJ-2, etc. The use of vortex imaging can obtain better imaging range and shorter period, representing more challenging change detection problems. The dataset contains 5000 dual time-phase image pairs with more than 65,980 building change instances. Among them, each image size is 1024×1024 with a spatial resolution of 0.5-0.8 m/pixel. The time span is from 1 to 5 years, and the image format is 8-bit PNG map.

4.2.2. Selection of evaluation indicators.In this paper, five common change detection metrics are used, which are precision (P), recall (R), mean F1 score (mF1), and mean intersection and merger ratio (mIoU). These five metrics are calculated as:

$$\begin{cases} P = (P_c + P_{uc}) / 2 \\ R = (R_c + R_{uc}) / 2 \\ mF1 = P_c R_c / (P_c + R_c) + P_{uc} R_{uc} / (P_{uc} + R_{uc}) \\ mIoU = TP / (TP + FP + FN) + TN / (TN + FN + FP) \end{cases} \tag{31}$$

where, where P_c and P_{uc} represent the accuracy of detecting changing and unchanging regions, respectively.TP, TN, FP and FN represent the number of true positives, true negatives, false positives and false negatives, respectively. For all metrics, larger values indicate better performance.

4.3. Remote Sensing Image Change Detection

4.3.1. Model comparison experiments.In order to evaluate the performance of RSICD model in performing remote sensing image change detection, this paper compares it with FC-EF, FC-Siam-Diff, FC-Siam-Conc, STA-Net, IFN, SNU-Net, BIT, ChangeFormer on LEVIR-CD and S2Looking datasets. Table 4 shows the performance comparison results of different models on different datasets.

The quantitative results show that mF1 and mIoU outperform the optimal model by 1.66% and 0.58% on the LEVIR-CD dataset. On the S2Looking dataset P, mF1, and mIoU are 5.11%, 2.32%, and 3.91% higher than the optimal model, respectively. The CNN-based method is unable to obtain the precise localization of the change target due to its limited sensory field, which leads to the crippling of the extracted change region. Although the Transformer-based method can localize the change region, it does not have enough edge detail information, resulting in unclear edges to the extracted change region. The RSICD model designed in this paper has both the global feature extraction capability of CNN and the local feature extraction effect of Transformer. Because of the combination of global and local information, it can accurately localize some irregular shapes and large scales while retaining enough edge information, and the RSICD model realizes primary-advanced feature alignment, so it can accurately extract the features of “region of interest” even in the background of the big difference in the style of the two-time-phase remote sensing images. “The RSICD model realizes primary-advanced feature alignment. In the LEVIR-CD dataset, the change samples are all regular shaped houses, but other methods have the problem of incomplete extraction of some irregular shaped change samples. The RSICD model can well avoid this problem, and its unique global-local feature representation can not only accurately locate the position of the change samples, but also retain enough edge information. In the S2Looking dataset, the change samples are usually irregular and widely varying scales of house buildings, and other methods can well recognize the change samples with smaller targets, but the extraction of some large-size change samples is incomplete. The RSICD model can well avoid these problems and can accurately identify not only small-size change samples but also large-size change samples in a complete way.

Table 4. Performance comparison results of different models (%)

Model	LEVIR-CD				S2Looking			
	P	R	mF1	mIOU	P	R	mF1	mIOU
FC-EF	86.92	80.15	83.41	71.82	72.63	82.74	61.05	43.99
FC-Siam-Diff	89.54	83.36	86.32	75.93	59.65	65.73	62.57	45.51
FC-Siam-Conc	91.98	76.79	83.69	71.98	66.41	54.26	59.78	42.56
STA-Net	83.82	91.04	87.25	77.45	67.84	61.69	64.51	47.68
IFN	94.06	82.95	88.14	78.72	67.89	53.98	60.16	42.94
SNU-Net	89.19	87.21	88.19	78.72	60.65	72.81	66.21	49.47
BIT	92.06	88.84	90.42	82.46	88.43	84.96	86.72	76.53
Change Former	89.27	89.35	89.34	80.69	68.37	70.19	69.35	52.95
RSICD	93.31	89.94	90.57	82.94	92.95	83.74	88.73	79.52

4.3.2. Model efficiency analysis.The purpose of this paper is to achieve high accuracy detection of remote sensing image changes and reduce the computational complexity. Therefore, this paper analyzes and compares the network on the LEVIR-CD dataset from multiple perspectives, including floating-point operations (FLOPs), the number of parameters (Params) and the F1-score, and still chooses the previous models as a comparison, and obtains the results of the computational efficiency of the different models as shown in Table 5.

It can be seen that the FLOPs and Params of the RSICD model are lower than the number of parameters and computational complexity of the Change Former network that replaces the EIG module with the traditional self-attentive SA, and the F1 of the RSICD model is the highest at 90.63%, which verifies the validity of the proposed EIG module.FC-EF, FC-Siam-Conc and FC-Siam-Diff use fewer feature extraction layers and the number of channels in the last layer is less, resulting in a smaller model. However, when confronted with more complex remote sensing image change detection problems, their feature extraction capabilities are weak, resulting in low accuracy and poor robustness. The other networks can obtain relatively good accuracy, but the computation and model size are large. Compared with these networks, the RSICD model not only greatly reduces the number of parameters and model size, but also achieves the best remote sensing image change detection performance. Therefore, combining CNN with Transformer can better extract the feature changes of remote sensing images and provide a reliable model support for realizing the change detection accuracy of remote sensing images.

Table 5. Comparison results of the calculation efficiency of different models

Model	Params (MB)	FLOPs (GB)	F1-score (%)
FC-EF	0.86	2.68	83.59
FC-Siam-Diff	1.04	4.05	82.26
FC-Siam-Conc	0.89	3.42	85.61
STA-Net	7.45	18.53	87.43
IFN	43.93	50.37	87.84
SNU-Net	7.52	18.59	87.45
BIT	6.95	8.46	89.16
Change Former	6.31	9.18	89.87
RSICD	4.87	5.21	90.63

4.3.3. Model ablation experiments.To validate the effectiveness of the modules, this paper conducts ablation experiments on the LEVIR-CD dataset to explore in depth the application of different module

combinations. The underlying network adopts the Siam-UNet codec structure and omits the EIG (E), DDaT (D) and SimAM (S) modules to simplify the network. When evaluating the model performance, the F1 score is used as a comprehensive evaluation index, with higher values indicating better model performance and reflecting the overall performance more accurately. The experimental results show that the proposed module is effective and robust to the change detection algorithm, providing strong support for the practical application of remote sensing image change detection. Table 6 shows the results of the model ablation experiments.

The experimental results on the LEVIR-CD dataset fully verify the effectiveness of the proposed network module. By introducing the Edge Information Guided (EIG) module, the network receptive field is significantly enlarged and the ability to capture multi-scale context information is enhanced, which improves the F1 score of the model by 4.31 percentage points compared to the Base network. The addition of the Dynamic Deformable Transformer (DDaT) module further improves the model performance by learning the global context and extracting deep features, which improves the F1 score by 2.32 percentage points compared to the Base network, while the introduction of the SimAM module significantly improves the model accuracy and accelerates the convergence process, which results in an increase of 0.74 percentage points in the F1 score. In addition, adding the DDaT or SimAM module to EIG improves the F1 score by 4.37 percentage points and 3.98 percentage points, respectively. Finally, the combination of the three modules has a higher F1 score of 6.48 percentage points compared to the Base network, which fully demonstrates the superiority of the proposed network modules and their combinations. These improvements not only enhance the model performance, but also provide more reliable technical support for remote sensing image change detection.

Table 6. Results of the experiment of the model ablation (%)

	EIG	DDaT	SimAM	IOU	F1	P	R
Base	×	×	×	72.75	84.15	90.84	78.54
Base+E	√	×	×	79.23	88.46	89.45	87.42
Base+D	×	√	×	73.68	86.47	88.97	81.06
Base+S	×	×	√	76.11	84.89	89.75	83.37
Base+E+D	√	√	×	79.48	88.52	89.23	87.93
Base+E+S	√	×	√	78.76	88.13	85.87	90.53
Base+D+S	×	√	√	75.68	86.16	87.64	84.64
Ours	√	√	√	82.69	90.63	88.96	92.28

5. Conclusion

The article proposes an automatic remote sensing image classification model based on CNN and migration learning, and constructs a remote sensing image change detection model by combining CNN and Transformer network. Different types of datasets are used to carry out validation analysis for the effectiveness of the remote sensing image automatic classification and change detection model.

1) The DI model on the RSSCN7 dataset improves the classification accuracy by 25.23% and 18.27%, and the AA and Kappa coefficients by 44.79% and 46.22%, respectively, compared with the GLCM and LBP individual features. And the number of parameters and floating-point computation volume of the fusion network model are 7.69MB and 1.89GB, respectively, which are smaller than the number of parameters and floating-point computation volume of the single network model, and the model computation efficiency is faster and has better robustness.

2) The RSICD models mF1 and mIoU designed in this paper are 1.66% and 0.58% higher than the optimal models on the LEVIR-CD dataset. On the S2Looking dataset P, mF1, and mIoU are 5.11%, 2.32%, and 3.91% higher than the optimal model, respectively, and the RSICD model has the highest F1 of 90.63%. The combination of edge information guidance and dynamic deformable Transformer

can better explore the change characteristics of remote sensing images and provide reliable technical support for the accurate realization of remote sensing image change detection.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 41601022).

References

[1] Song, J., Gao, S., Zhu, Y., & Ma, C. (2019). A survey of remote sensing image classification based on CNNs. Big earth data, 3(3), 232-254.

[2] Mehmood, M., Shahzad, A., Zafar, B., Shabbir, A., & Ali, N. (2022). Remote sensing image classification: A comprehensive review and applications. Mathematical Problems in Engineering, 2022(1), 5880959.

[3] Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., & Ajlan, N. A. (2021). Vision transformers for remote sensing image classification. Remote Sensing, 13(3), 516.

[4] Cai, W., & Wei, Z. (2020). Remote sensing image classification based on a cross-attention mechanism and graph convolution. IEEE Geoscience and Remote Sensing Letters, 19, 1-5.

[5] Salzano, R., Salvatori, R., Valt, M., Giuliani, G., Chatenoux, B., & Ioppi, L. (2019). Automated classification of terrestrial images: The contribution to the remote sensing of snow cover. Geosciences, 9(2), 97.

[6] Matci, D. K., & Avdan, U. (2020). Optimization-based automated unsupervised classification method: A novel approach. Expert Systems with Applications, 160, 113735.

[7] Huang, Q., Wu, G., Chen, J., & Chu, H. (2012, June). Automated remote sensing image classification method based on FCM and SVM. In 2012 2nd International Conference on Remote Sensing, Environment and Transportation Engineering (pp. 1-4). IEEE.

[8] Wang, T., Thomasson, J. A., Yang, C., Isakeit, T., & Nichols, R. L. (2020). Automatic classification of cotton root rot disease based on UAV remote sensing. Remote Sensing, 12(8), 1310.

[9] Sicong, L. I. U., Kecheng, D. U., Yongjie, Z. H. E. N. G., Jin, C. H. E. N., Peijun, D. U., & Xiaohua, T. O. N. G. (2023). Remote sensing change detection technology in the Era of artificial intelligence: Inheritance, development and challenges. National Remote Sensing Bulletin, 27(9), 1975-1987.

[10] Zhu, Q., Guo, X., Li, Z., & Li, D. (2024). A review of multi-class change detection for satellite remote sensing imagery. Geo-spatial Information Science, 27(1), 1-15.

[11] Afaq, Y., & Manocha, A. (2021). Analysis on change detection techniques for remote sensing applications: A review. Ecological Informatics, 63, 101310.

[12] Bai, T., Wang, L., Yin, D., Sun, K., Chen, Y., Li, W., & Li, D. (2023). Deep learning for change detection in remote sensing: a review. Geo-spatial Information Science, 26(3), 262-288.

[13] Panuju, D. R., Paull, D. J., & Griffin, A. L. (2020). Change detection techniques based on multispectral images for investigating land cover dynamics. Remote Sensing, 12(11), 1781.

[14] Hecheltjen, A., Thonfeld, F., & Menz, G. (2014). Recent advances in remote sensing change detection–a review. Land Use and Land Cover Mapping in Europe: Practices & Trends, 145-178.

[15] 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 Journal of Photogrammetry and Remote Sensing, 119, 402-414.

[16] Liu, Z., Li, G., Mercier, G., He, Y., & Pan, Q. (2017). Change detection in heterogenous remote sensing images via homogeneous pixel transformation. IEEE Transactions on Image Processing, 27(4), 1822-1834.

[17] Tewkesbury, A. P., Comber, A. J., Tate, N. J., Lamb, A., & Fisher, P. F. (2015). A critical synthesis of remotely sensed optical image change detection techniques. Remote Sensing of Environment, 160, 1-14.

[18] Cheng, G., Huang, Y., Li, X., Lyu, S., Xu, Z., Zhao, H., ... & Xiang, S. (2024). Change detection methods for remote sensing in the last decade: A comprehensive review. Remote Sensing, 16(13), 2355.

[19] Han, Y., Javed, A., Jung, S., & Liu, S. (2020). Object-based change detection of very high resolution images by fusing pixel-based change detection results using weighted Dempster–Shafer theory. Remote Sensing, 12(6), 983.

[20] 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. IEEE Geoscience and Remote Sensing Letters, 14(10), 1845-1849.

[21] Favorskaya, M. N., & Jain, L. C. (2017). Handbook on advances in remote sensing and geographic information systems. New York: Springer.

[22] Mani, P. K., Mandal, A., Biswas, S., Sarkar, B., Mitran, T., & Meena, R. S. (2021). Remote sensing and geographic information system: a tool for precision farming. Geospatial technologies for crops and soils, 49-111.

[23] Canty, M. J. (2019). Image analysis, classification and change detection in remote sensing: with algorithms for Python. Crc Press.

[24] Yu, S., Tao, C., Zhang, G., Xuan, Y., & Wang, X. (2024). Remote Sensing Image Change Detection Based on Deep Learning: Multi-Level Feature Cross-Fusion with 3D-Convolutional Neural Networks. Applied Sciences (2076-3417), 14(14).

[25] Asokan, A., & Anitha, J. J. E. S. I. (2019). Change detection techniques for remote sensing applications: A survey. Earth Science Informatics, 12, 143-160.

[26] Ye, F., Su, Y., Xiao, H., Zhao, X., & Min, W. (2018). Remote sensing image registration using convolutional neural network features. IEEE Geoscience and Remote Sensing Letters, 15(2), 232-236.

[27] Pan, X., & Zhao, J. (2017). A central-point-enhanced convolutional neural network for high-resolution remote-sensing image classification. International Journal of Remote Sensing, 38(23), 6554-6581.

[28] Yao, C., Zhang, Y., & Liu, H. (2017). Application of convolutional neural network in classification of high resolution agricultural remote sensing images. The international archives of the photogrammetry, remote sensing and spatial information sciences, 42, 989-992.

[29] Sharma, A., Liu, X., Yang, X., & Shi, D. (2017). A patch-based convolutional neural network for remote sensing image classification. Neural Networks, 95, 19-28.

[30] Liu, N., Wan, L., Zhang, Y., Zhou, T., Huo, H., & Fang, T. (2018). Exploiting convolutional neural networks with deeply local description for remote sensing image classification. IEEE access, 6, 11215-11228.

[31] Lin, C. H., & Wang, T. Y. (2021). A novel convolutional neural network architecture of multispectral remote sensing images for automatic material classification. Signal Processing: Image Communication, 97, 116329.

[32] Fu, G., Liu, C., Zhou, R., Sun, T., & Zhang, Q. (2017). Classification for high resolution remote sensing imagery using a fully convolutional network. Remote Sensing, 9(5), 498.

[33] Jiang, H., Peng, M., Zhong, Y., Xie, H., Hao, Z., Lin, J., ... & Hu, X. (2022). A survey on deep learning-based change detection from high-resolution remote sensing images. Remote Sensing, 14(7), 1552.

[34] Chen, H., Qi, Z., & Shi, Z. (2021). Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-14.

[35] Shafique, A., Cao, G., Khan, Z., Asad, M., & Aslam, M. (2022). Deep learning-based change detection in remote sensing images: A review. Remote Sensing, 14(4), 871.

[36] Chen, H., & Shi, Z. (2020). A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sensing, 12(10), 1662.

[37] Khelifi, L., & Mignotte, M. (2020). Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis. Ieee Access, 8, 126385-126400.

[38] Yu, X., Fan, J., Chen, J., Zhang, P., Zhou, Y., & Han, L. (2021). NestNet: A multiscale convolutional neural network for remote sensing image change detection. International Journal of Remote Sensing, 42(13), 4898-4921.

[39] Baoyi Zhang,Jiacheng Tang,Yuke Huan,Lei Song,Syed Yasir Ali Shah & Lifang Wang. (2024). Multi-scale convolutional neural networks (CNNs) for landslide inventory mapping from remote sensing imagery and landslide susceptibility mapping (LSM). Geomatics, Natural Hazards and Risk(1).

[40] Amal Alshardan,Nuha Alruwais,Hamed Alqahtani,Asma Alshuhail,Wafa Sulaiman Almukadi & Ahmed Sayed. (2024). Leveraging transfer learning-driven convolutional neural network-based semantic segmentation model for medical image analysis using MRI images. Scientific reports(1),30549.

[41] Haowen Shi,Yubin Su,Yuan Pan,Weihan Zhang,Zhong Chen & Ruiquan Liao. (2025). Research on failure diagnosis analysis of plunger gas lift system using convolutional neural network with multi-scale channel attention mechanism based on wavelet transform. Chemical Engineering Science121031-121031.

[42] Xiaohui Li,Yuheng Chen,Feng Yuan,Simon M. Jowitt,Mingming Zhang,Can Ge... & Yufeng Deng. (2024). 3D mineral prospectivity modeling using multi-scale 3D convolution neural network and spatial attention approaches. Geochemistry(4),126125-126125.

[43] Jingliu He,Yuqi Ma,Mingyue Yang,Wensong Yang,Chunming Wu & Shanxiong Chen. (2024). TAC-UNet: transformer-assisted convolutional neural network for medical image segmentation. Quantitative imaging in medicine and surgery(12),8824-8839.