Aircraft Detection in Remote Sensing Image based on Multi-scale Convolution Neural Network with Attention Mechanism

Xianfeng Wang

XiJing University

Changqing Yu (✉ xaycq@163.com)

XiJing University

Lei Huang

CNPC Tubular Goods Research Institute

Shanwen Zhang

XiJing University

Research Article

Keywords: Aircraft Detection in remote sensing images, Convolution neural networks (CNN), Multi-scale CNN (MS-CNN), Global attention mechanism (GAM), MS-CNN with attention (MSCNNA).

Posted Date: December 30th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1179245/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Aircraft Detection in Remote Sensing Image based on Multi-scale Convolution Neural Network with Attention Mechanism

Xianfeng Wang¹, Changqing Yu^{1*}, Lei Huang², Shanwen Zhang¹

1. College of Information Engineering, XiJing University, Xi'an 710123, China

2. CNPC Tubular Goods Research Institute, Xi'an 710077, China

*email: xaycq@163.com

Abstract: Detecting aircraft from remote sensing image (RSI) is an important but challenging task due to the variations of aircraft type, size, pose, angle, complex background and small size of aircraft in RSIs. An aircraft detection method is proposed based on multi-scale convolution neural network with attention (MSCNNA), consisting of encoder, decoder, attention and classification. In MSCNNA, the multiple convolutional and pooling kernels with different sizes are utilized to learn the multi-scale discriminant features, and the global attention mechanism (GAM) is employed to capture the spatial and channel dependencies and adaptively preserve the relationships of the entire image. Compared with the standard deep CNN, multi-scale convolution neural networks (CNN) and GAM are integrated to learn the multi-scale features for aircraft detection, especially small aircrafts. Experiment results on the aircraft image dataset of the public EORSSD dataset show that the proposed method outperforms the state-of-the-art method on the same dataset and the obtained multi-size aircraft edge is clearer.

Keywords: Aircraft Detection in remote sensing images; Convolution neural networks (CNN); Multi-scale CNN (MS-CNN); Global attention mechanism (GAM); MS-CNN with attention (MSCNNA).

1. Introduction

Automatic aircraft detection from remote sensing image (RSI) is an important research due to its great value in civilian and military and high application value in dynamic monitoring and military surveillance. However, it is a very difficult and challenging task because the aircraft objects occupy a relatively small and blurry proportion in RSIs, and have various sizes, shapes, poses, shadows, illuminations, and complex background as shown in Fig.1.

Fig.1 Samples of aircrafts in RSIs

With the development of computer vision and image processing technologies, many aircraft detection methods have been proposed in recent years [1,2]. Zhao et al.[3] proposed a multiscale sliding-window framework for aircraft detection based on aggregate channel features and well-designed features. In the framework, the features are trained by Cascade AdaBoost including multiple rounds of bootstrapping, and a two-step nonmaximum suppression algorithm is designed based on a given detection set. Yan et al.[4] designed a aircraft detection method using centre-based proposal regions and invariant features, consisting of three steps, firstly extracting

proposal regions from each RSI, secondly training the ensemble learning classifier with the extracted rotation-invariant features for aircraft classification, finally detecting aircraft from RSIs by the trained well classifier. The detecting performance of the above methods mainly relies on the extracted features, but it is difficult to extract the proper discriminant features for various aircraft objects in the RSIs. Recently, with the development of big data technology and computing ability, many convolution neural networks (CNN) based aircraft detection methods have been presented and achieved significant advances due to the prominent capability of CNN on feature extraction. Bin et al.[5] proposed a cascade CNN (CCNN) framework based on transfer-learning and geometric feature constraints to implement the aircraft detection in large-area RSIs, and achieved high detection accuracy with relatively few samples. Guo et al. [6] proposed an end-to-end aircraft detection method, which completely eliminates proposal generation and encapsulates all calculations in a network, making it simple and efficient. Hu et al.[7] proposed an aircraft detection in RSIs approach based on saliency and CNN, where the region of interests (RoIs) are obtained using the saliency algorithm, and the feature vector of the position information is obtained by deep CNN (DCNN) from the whole image. Yang et al.[8] constructed an effective airplane detection framework called Markov random field-fully convolutional network (M-FCN) by using a cascade strategy that consists of a fully convolutional network (FCN) based coarse candidate extraction stage, a multi-Markov random field (multi-MRF)-based region proposal (RP) generation stage, and a final classification stage. Zhang et al.[9] constructed an effective aircraft detection framework based on CNN to detect multi-scale targets in extremely large and complicated scenes, designed a constrained Edge-Boxes approach to generate a modest number of target candidates quickly and precisely, and presented a modified GoogLeNet based on Fast Region-based CNN (R-CNN) to detect various kinds of multi-scale aircrafts. Zhong et al.[10] collected RSIs of airports from Google Earth and took full advantage of data augmentation, transfer learning, DCNN and limited training samples to implement end-to-end trainable airplane detection. Li et al.[11] collected a multi-resolution aircraft remote sensing dataset from Google Earth and proposed a R-CNN based aircraft detection model. They trained the detection model by end-to-end fine-tuning on the collected dataset and realized automatic aircraft detection and positioning. Wang et al.[12] proposed an aircraft detection algorithm which can detect weak and small aircraft objects of varying sizes. Fu et al. [13] proposed a feature-fusion algorithm to generate a multi-scale features, which can form a powerful feature representation for multi-scale aircraft detection. They constructed a rotation-aware object detector to localize objects in RSIs, utilized the oriented proposal boxes to enclose objects rather than horizontal proposals that can coarsely locate the oriented objects, and adopted the orientation RoI pooling operation to extract the feature maps of oriented proposals for the following R-CNN sub-network. Liu et al.[14] proposed an aircraft detection scheme based on corner clustering and CNN, and compared with three methods, namely selective search (SS)+CNN, Edge-boxes+CNN and histogram of oriented gradient (HOG)+support vector machine (SVM). The proposed scheme contains two main steps: region proposal and classification, where candidate regions are generated by utilizing mean-shift clustering algorithm, and feature extraction and classification of candidate regions possibly containing aircrafts are realized by CNN.

Although CNN and its improved models have achieved promising results on various image datasets, they require a large amount of labeled image data to learn, while labeling RSIs requires high labor costs, and many standard CNN models focus only on the local features while ignore the global region

features, then some important global features are lost inevitably. In fact, both local and global features are important for complex image classification and recognition, but it is difficult to effectively extract the global and local features by CNN and its improved models from RSIs with different scales of aircrafts. As for the small aircrafts in RSIs, pooling layer of CNN may further reduce the amount of the classification information. Suppose a 24×24 difficult only has about 1 pixel after 4-layer pooling, making the dimension too low to be distinguished. That is to say, it is difficult to detect the small aircrafts. To overcome these shortages and enhance the robustness and classification performance, a lot of multi-scale CNN (MS-CNN) models have been proposed and successfully applied to complex image segmentation and multi-scale object detection tasks. Some MS-CNN models adopt multi-scale input blocks of the original image instead of the directly inputting the original images [15-17], and some MS-CNN models use the multi-scale input images of different channels to convolve with multi-scale filters [18,19]. Deng et al.[20] proposed a unified and effective method for multi-class object detection in RSIs with large scales variability, where the detection is performed by two sub-networks: a multi-scale object proposal network for object-like region generation from several intermediate layers, and an accurate object detection network for object detection based on fused feature maps. Wang et al.[21] proposed a compact multi-scale dense CNN (MS-DenseNet) model for aircraft detection in RSIs, where DenseNet is utilized for feature extraction, the feature pyramid network is combined with DenseNet to form a MS-DenseNet for learning multi-scale features, and three compact architectures namely MS-DenseNet-41, MS-DenseNet-65 and MS-DenseNet-77 are designed by compressing some of the unnecessary convolution layers of each dense block. Recently, various attention mechanisms have been presented and widely introduced into CNN to allow the detection and classification networks to concentrate on the most salient features [22,23], but they are difficult to directly applied these attention mechanisms to aircraft object detection in RSIs due to the aircraft objects usually vary and are relatively small in RSIs [24]. To overcome this problem, Dayananda et al.[25] designed a global attention module between the encoder and decoder parts with a multi-scale guided input.

The traditional deep learning based image segmentation and detection models such as FCN [8], SegNet, U-Net and PSP-Net apply bilinear interpolation to recover image resolution during deconvolution [26]. Multi-scale multi-column CNN (MSMC-CNN) adopts the bicubic convolution as a deconvolution for the decoder structure to restore the original image size [27]. The results validate that MSMC-CNN is superior to the other segmentation methods. Since many aircraft in RSIs are small targets (dozens or even a few pixels) containing little target information, RSIs are taken from overhead, then the direction of the target is uncertain, and RSI has a lot of interference, such as noise, complex background, and illumination changes, the aircraft detection in RSIs is still a challenging topic. There are a lot of shortcomings in the existing CNN and MS-CNN based aircraft detection algorithms, such as low recognition accuracy and time-consuming. Inspired by CNN, MS-CNN and MSMC-CNN and attention mechanism [28], a modified MS-CNN with attention (MSCNNA) is constructed to automatically detect the aircraft objects from RSIs with different scales. It is also an encoder-decoder network. It adopts the multi-scale input images into MS-CNN instead of directly inputting the original image. In the traditional CNNs model, the pooling operation can compress the feature maps and simplify reduce CNN computation complexity, but it results in the loss of features. Comparing with CNN and MS-CNN models, MSCNNA is able to extract multi-scale features for improving the accuracy and robustness of aircraft detection. The main contributions of this paper are indicated as follows:

- ✓ A modified MS-CNN model namely MSCNNA is constructed to deal with scale variation

problem of RSIs in aircraft detection.

- ✓ Global attention mechanism (GAM) is introduced to MSCNNA to capture global and local features of multi-scale aircrafts and to improve the aircraft detection accuracy.
- ✓ MSCNNA is validated on a public aircraft RSI dataset.

The rest of the paper is organized as follows. Section 2 provides the review of related methods, including MS-CNN and attention mechanism and global attention mechanism (GAM). The details of MSCNNA for aircraft detection are described in Section 3. Then, a lot of experiments and results are given in Section 4. Finally, some conclusion remarks are drawn in Section 5.

2. Related works

2.1 Multi-scale CNN (MS-CNN)

To solve the problem of object scale inconsistency in object detection, many MS-CNN models have been proposed. There are two common MS-CNN architectures as shown in Fig.2, where Fig.2B is the convolutional operation comparing with the traditional CNN model as seen in Fig.2A, Fig.2C is a multi-channel input scheme using multi-scale convolutional kernels to extract the multi-scale features from the input images, Fig.2D is a multi-scale input scheme by changing the original input to multi-scale images as different channel input. Two kinds of MS-CNN models can be implemented by utilizing multiple convolutions to form multi-scale features for objective image detection with different sizes in the convolution layer. Meanwhile, to ensure that the spatial feature information between the input and output of the multi-scale convolutional kernels remains unchanged, the edge of the feature graph is filled according to the size of the convolution kernel.

(A) Convolutional operation with one convolution (B) Convolutional operation with multi-scale convolution

(C) Architecture of MS-CNN with different size kernels

(D) Architecture of MS-CNN with different size input images

Fig.2 Architecture of MS-CNN

As shown in Fig.2D, each original input image is reshaped to three images with three sizes. When the multi-scale input scheme is adopted, three images are input into three channels with different sizes of convolution and pooling kernels, respectively, and are carried out separately. Finally, the extracted feature graphs of different channels are combined or integrated in the fully-connected layers. The features extracted by the convolved kernels of large size are more global, while those of small size are more reflective of local characteristics. So MS-CNN can extract both global and local features.

2.2 Attention mechanism and global attention mechanism (GAM)

Attention mechanism makes full use of the local and global features of CNN to reduce noise and improve the classification accuracy by giving higher weight to the important features [23,24]. It is widely used to obtain the better feature representation for CNN. It can learn the weight between the extracted feature graphs, and then multiplies back to the original feature graphs, so that the feature graphs can be better represented. The channel attention module firstly conducts global average-pooling and global max-pooling respectively for the original feature graphs, then connects the two pooling feature vectors to one fully-connected layer, adds point-wise, and then activates. The spatial attention module uses 1×1 pooling including average-pooling and max-pooling instead of 1×1 convolution. Both pooling operations are concatenated followed by a layer of 7×7 convolution, and then activated [25].

In image segmentation and classification, attention model is responsible for learning the attention weight a_{ij} , which can automatically capture the correlation between h_i (encoder hidden state namely candidate state) and s_j (decoder hidden state namely query state). These attention weights are then used to build the content vector C , which as input is passed to the decoder. At each decoding position j , the content vector c_j is the weighted sum of all the hidden states of the encoder and their corresponding attention weights, calculated as follows,

$$c_j = \sum_{i=1}^T a_{ij} h_i \quad (1)$$

Attention weights are learned by adding an additional feed-forward neural network to the architecture. The feed-forward network learns a special attention weight a_{ij} as a function of two states h_i and s_{j-1} . That is about using these two states as inputs to CNN, and learning about a_{ij} as their

function.

Global attention mechanism (GAM) can be used to refine the feature extraction and improve the perception ability of CNN [28]. The multi-scale information and the image classification module output and are fed as input to the GAM at each encoding layer as follows,

$$\begin{aligned} \text{GAM}_l &= \text{GAM}(ms_l, \text{expan}_l) \\ ms_l &= \text{Concat}[f(m_l, w_l^{1 \times 1}), f(m_l, w_l^{3 \times 3})] \end{aligned} \quad (2)$$

where expan_l is the fire module output of the l -th network layer, ms_l is the input of GAM, $w_l^{1 \times 1}$ and $w_l^{3 \times 3}$ are 1×1 and 3×3 convolution filters, $f()$ is activation function such as the standard Rectified

Linear Unit (ReLU), $\text{Concat}()$ is a concatenate function, m_l is the input of ms_l as

$m_l = \text{maxpool}(x_{l-1}, 2)$, that is the input x_l is the down-sampled using max-pooling with a stride of 2×2 .

The output of GAM is input into the max-pooling layer to reduce the dimensionality and focus on the fine details of the feature map, as follows,

$$\text{encoder}(l) = \text{Maxpool}(\text{GAM}_l, 2) \quad (3)$$

In the process of GAM in RSI segmentation, pooling indices are stored at each encoder layer so that the decoder uses the information to up-sample the feature maps. The output at each encoder layer is referred to down-sampling unit obtained by Eq.(3) [25].

3. MS-CNN with attention for aircraft detection

The above experiment results show that it is difficult to extract the detection and classification features from aircraft RSIs of different sizes, because there are various aircrafts in RSIs with large difference in size, location and number of aircrafts in RSIs [10-15]. Different image scales can provide multi-scale features in the same RSI. In fact, the aircraft targets in RSIs can be roughly divided into three different sizes: small, medium and large. Then RSIs of three different scales are used for training the detection network, which can well provide the aircraft features of the three different sizes in a RSI. The features extracted from the multi-scale images by network model are not all the same, where large scale images are better to extract small target features, medium scale images are better to extract medium target features, and small scale images are better to extract large area targets. Fusing the features of the multi-scale images can obtain more comprehensive classification features and achieve more accurate detection. Because the attention mechanism can assign different attention weights to multi-scale objects, it is convenient to obtain more discriminant features [25,28].

From the above analysis, a modified MS-CNN with attention model namely MSCNNA is constructed as shown in Fig.3, consisting of 4 modules of encoder, attention, decoder and classification, where the encoder module contains 3 parallel sub-CNN models, which utilize different size and number of convolution kernels and pooling operations to extract the multi-scale features. Each sub-CNN model is based on LeNet5, which is a relatively simple CNN. The sizes of kernels and feature graphs of convolution, pooling and deconvolution are denoted in Fig.3, respectively.

Fig.3 Architecture of MSCNNA

The main process of MSCNNA is introduced as follows:

(1) Input data. The original image is resized into three input images with three different sizes, which are input into three CNNs to extract multi-scale features of the aircraft image, respectively. Three CNNs are carried out separately, and the extracted feature graphs are integrated by global attention mechanism (GAM) in attention module. This integrating process of GAM is similar to the existing attention mechanism in MS-CNNs [25]. In each CNN of MSCNNA, multi-scale connections are made to the feature maps of different layers to adjust the scale and viewing angle of the sonar image.

(2) Encoder. In encoder module, the multi-scale input features at each encoding layer can encode both the global and local features. After the original input image is convolved with the kernel, an activation function ReLU is adopted to obtain the first-layer output feature maps, and the specific steps of convolution and pooling are the same as CNN, introduced as follows,

Convolutional operation is used to extract feature maps from the input images. The output x^l of the l -th layer in CNN is calculated as follows,

$$x^l = f(x^{l-1} \otimes W^l + b^l) \quad (4)$$

where x^{l-1} is output in $(l-1)$ -th and W^l is a weight, $f(\cdot)$ is an activation function, ' $\otimes$ ' is convolutional operation, b^l is an additive bias in the convolution layer, in default $f(x) = 1/[1 + \exp(-x)]$.

Pooling operation can reduce the computing time of next layers, make the network less prone to over-fitting and extract the invariance features to translation. Average-pooling and max-pooling are two widely used operations. A pooling layer takes an $h \times w \times n$ input and performs the average (max) pooling operation with a small pooling size $m \times m$ to produce the so-called pooled maps a $\frac{h}{m} \times \frac{w}{m} \times n$

sub-sampled output. The output of the l -th pooling layer is calculated as,

$$x^l = f(\text{down}(x^{l-1}) + b^l) \quad (5)$$

where x^{l-1} is the input of the l -th layer, $\text{down}(\cdot)$ is a sub-sampling function to shrink a feature map by using a pooling operation, and b^l is a bias.

(3) Attention. In attention mechanism module, GAM between the encoder and decoder focus on the most relevant details needed for the aircraft detection task, meanwhile filters the irrelevant information. GAM is implemented by Eqs.(2) and (3).

(4) Decoder. Decoder module uses several deconvolutions to up-sampling the feature graphs integrated in the attention mechanism. It consists of deconvolution and upsampling operations, where deconvolution with a parameter forms the encoder information similar to that of a convolutional layer, and upsampling is implemented on the input feature maps using the memorized max-pooling indices

from the corresponding encoder feature maps. Max-pooling index from the encoder feature mapping is often used for upsampling of the input feature mapping in the decoder block. To effectively preserve the information from the original image for edge details, the entire feature maps are transferred from the encoder block to the corresponding decoder block, and are connected to the upsampling decoder feature map using the pooled indexes. Decoder can use fewer parameters and obtain better image features to recover the object information from the original image.

In decoder, the different convolution blocks of the sampling layer can be connected by bilinear and bicubic interpolation [26], while the interpolated surface by bicubic interpolation is smoother than that by bilinear interpolation or nearest-neighbor interpolation [29]. Then in MSCNNA, bicubic interpolation is applied to deconvolution, which can be implemented by Lagrange polynomials, cubic splines, and cubic convolution algorithm. Bicubic convolution interpolation function is defined as follows,

$$u(s) = \frac{c_{i+2} - c_{i+1} + c_i}{2} s^2 + \frac{-c_{i+2} + 4c_{i+1} - 3c_i}{2} s + c_i \quad (6)$$

where $c_i = f(x_i)$, $c_m = 3c_{m-1} - 3c_{m-2} + c_{m-3}$, $c_{m+1} = 3c_m - 3c_{m-1} + c_{m-2}$, $f(x)$ is the input data.

By Eq.(6), two-dimensional bicubic convolution interpolation can be calculated as follows,

$$g(s_1, s_2) = \sum_{a=-1}^2 \sum_{b=-1}^2 c_{i+a, j+b} u(s_1 - a) u(s_2 - b) \quad (7)$$

where $s_1 = \frac{x}{Vx} - i$, $i = \left\lfloor \frac{x}{Vx} \right\rfloor$; $s_2 = \frac{y}{Vy} - j$, $j = \left\lfloor \frac{y}{Vy} \right\rfloor$, $g(s_1, s_2)$ is regarded as deconvolution graph.

By Eqs.(6) and (7), the calculation coefficient relation of the two-dimensional cubic convolution interpolation is as follows,

$$\begin{cases} c_{i,j} = f(x_i, y_j) \\ c_{i+a, j-1} = 3c_{i+a, j} - 3c_{i+a, j+1} + c_{i+a, j+2} \\ c_{i+a, j+2} = 3c_{i+a, j+1} - 3c_{i+a, j} + c_{i+a, j-1} \\ c_{i-1, j+b} = 3c_{i, j+b} - 3c_{i+1, j+b} + c_{i+2, j+b} \\ c_{i+2, j+b} = 3c_{i+1, j+b} - 3c_{i, j+b} + c_{i-1, j+b} \end{cases} \quad (8)$$

where $f(x, y)$ is the two-dimensional integrated feature graph by GAM.

(5) Aircraft detecting and model training. The training dataset is divided into a positive sample set containing aircrafts and a negative sample set without aircraft. MSCNNA is used for a binary aircraft classification task, i.e., determining whether a RSI contains aircraft or not. When training MSCNNA is complete, a test RSI is input into the trained MSCNNA for forward propagation. Then, the weight corresponding to the image classification result of having aircraft in the last deconvolution layer is extracted, weighted, and averaged with the characteristic map output. The highlight area is the basis for MSCNNA to detect the test image using the aircraft category. The main difference between two RSIs of having aircraft and without aircraft is that the former contains an aircraft target, so these areas correspond to the position of the aircraft in the test RSI. Then, adaptive threshold based detection is used to segment a binary graph. Finally, according to the threshold, the detection of aircraft in the test RSI is completed [30].

For a RSI with size of $M \times N$, the foreground and background are segmented according to the optimal grayscale threshold, which can be evaluated by the OTSU algorithm as follows [31]:

$$T = \arg \max_{T \in \{1, 2, \dots, 255\}} (\varpi(T)) \quad (9)$$

where $\varpi(T)$ is the variance between the foreground and background, and ϖ is calculated as,

$$\varpi = \frac{N_f}{M \times N} \frac{N_b}{M \times N} (G_f - G_b)^2 \quad (10)$$

where N_f and N_b are the pixel numbers in the foreground and background, respectively, G_f and G_b are the average gray values in the foreground and background, respectively.

In RSI, the pixels larger than the threshold are marked as the foreground and the rest is the background. Then, the foreground and background are segmented and binarized using the optimal threshold. The training strategies mostly follow LeNet5, including multi-scale training, data augmentation, convolutional with anchor boxes, and loss function. For a fair comparison, LeNet5 is trained in the same way as that of MSCNNA.

(6) Model evaluation. Intersection over Union (IoU) is often used in object detection to measure the correlation between real and predicted. The higher the correlation, the higher the IoU. It is calculated by the overlap rate of generated predicted images and bounding boxes, that is, the ratio of the intersection and union of two boxes. It is defined as follows,

$$IoU = \frac{area(C) \cap area(G)}{area(C) \cup area(G)} \quad (11)$$

In semantic segmentation, pixel regions are used to replace the marked boundary boxes to calculate the intersection ratio of image segmentation. The average intersection ratio is $\frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i - n_{ii}}$, where

the number of classes is n_{cl} , n_{ji} is the pixel number of the i th class predicted to be j th class, and

$t_i = \sum_j n_{ij}$ is the total pixel number of the i th class. IoU is often used to measure the difference

between the detection results and the actual labeled images. It can be used as the evaluation index of the aircraft class, that is, the number of pixels marked as built in the forecast and reference is divided by the number of pixels marked in the forecast or reference. When a fractional intersection of the detected aircraft overlaps with a test aircraft greater than 0.5, it is taken as a true detected aircraft. On the other side, it is regarded as a missed aircraft.

To evaluate the detection effect of aircraft more objectively, the detection precision rate *Precision*, recall rate *Recall* and average *F*-value is introduced as three evaluation indexes to measure the difference between the detection results and the actual labeled images. They are calculated as follows:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \quad F = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (12)$$

where TP is the overlap between the detection aircraft region and the original aircraft region, FP is the part of the detection result that does not belong to the aircraft region, and FN is the part that does not belong to the normal region.

In Eq. (12), *Precision* is the consistency between the reflected aircraft area and the real aircraft area, *Recall* is the proportion of correctly segmented aircrafts to the total number of aircrafts, and *F* reflects the evaluation index of the overall accuracy by combining *Precision* and *Recall*. Since the detection time is a key technical index to evaluate the practicability of the model, the detection time of a single image is used as a time index to measure the detection speed.

4. Experiments and analysis

To validate MSCNNA based aircraft detection method, a lot of experiments are conducted on the aircraft RSI dataset, and compared with five state-of-art methods, i.e., Deep CNN (DCNN) [6], saliency and CNN (SCNN) [7], Markov random field-FCN (M-FCN) [8], CNN based weakly supervised learning (CNNWSL)[31], CNN based semantic segmentation (CNNSS) [32]. All models are trained on the EORSSD dataset, and the main experimental configuration is listed in Table 1.

Table 1 Experimental configurations

Hard-software	Profile
CPU	Intel Xeon E5-2643v3 @3.40GHz, 64GB RAM, 8GB of video memory, CUDA Toolkit 9.0, CUDNN V7.0
GPU	NVidia Quadro M4000
Operating system	Ubuntu 16.04
Programming language	Python 3.5.2
Algorithm framework	Tensorflow-GPU 1.8.0

4.1 Data and Preprocessing

The EORSSD dataset (<https://github.com/rmcong/EORSSD-dataset>) contains 2,000 images corresponding pixel-level annotated true images, including 258 aircraft RSIs and most of them are complex and challenging having one or several aircraft targets with small sizes, arbitrary position and angle, diverse resolution and blurred background, as shown in Fig.1. In this Section, all aircraft RSIs are used to construct an aircraft RSI dataset to carry on the aircraft detection experiments.

Since the sizes of the original aircraft RSIs are various with different resolutions ranging from 973×760 to 242×239, each image is resized to size of 128×128 to relieve the computation time and GPU memory. LabelMe Matlab toolbox is used to label RSIs for aircraft detection and share the annotations with the rest of the community. The number of the aircraft RSIs of EORSSD is relatively small for training MSCNNA. To increase the image diversity, decrease the overfitting problem, and improve the aircraft detection ability, each original aircraft RSI is augmented by a rotation method (90^0 , 180^0 , 270^0) and a randomly cropping method with an overlap of 100 pixels, and then data cleaning is adopted to reduce sub-images without objects [27]. Each image is augmented to 4 images including aircraft targets with 3 orientations and a cropping, and then the augmented dataset contains 1548 images in total for aircraft detection task. There are no any image preprocessing besides of resizing, augmentation and annotating. Four-fold validation experiments are conducted on the augmented dataset and the average detection results are regarded as the accuracy rates.

4.2 Experiments

MSCNNA is initialized by being pertained on the EORSSD dataset with various resolutions ranging from 0.1 to 1 m [33,34]. It is a benchmark dataset for evaluating the performance on salient object detection in optical RSIs. During the fine-tuning training [6,7], the initial learning rate of each layer is set to 0.001 and reduced to the previous 10% after 500 iterations. The momentum and weight decay are set to 0.9 and 0.0005, respectively. The max training iteration is 3,000, and the batch size of iteration is set to 40, and the total number of iterations depends on the convergence of the network model. During testing, the proposed MSCNNA is utilized to detect the aircraft in the test RSIs.

4.3 Results

In network training, the convolutional layers are used to extract features from the three-size input images, and the extracted features by different convolutional kernels are different. The shallow convolution kernel mainly extracts the low-level features of the image, including color and contour, etc. Deep convolution kernel extracts more semantic features from images, including texture and detail features. To observe the different features extracted by different convolution kernels, all convolution feature maps of C1 and C3 layers are visualized in Fig.4. From Fig.4, all feature graphs obtained by convolution C1 and C3 layers can display the contour features of aircraft images.

Fig.4 Feature graphs

To verify the feasibility of MSCNNA in aircraft detection, MSCNNA is compared with two classical models for complex image segmentation, i.e., FCN [35] and SegNet [36]. In general, the last one to three layers of CNN adopts full connection to classify the whole image. FCN replaces the final fully-connection of CNN by convolution operation, classifies each pixel of image, utilizes deconvolution operation at the end of the network to up-sample the final extracted feature graphs of CNN, and reshapes the output to the same size as the segmentation result. SegNet applies decoder to up-sampling and directly uses the pooling index retained during maximum pooling in the corresponding Encoder stage for nonlinear up-sampling. In upsample, the input values are directly assigned to the corresponding positions, and other positions are set to zero, finally a classification layer is used for pixel-wise. The difference between FCN and SegNet is that the feature graph of FCN is obtained by deconvolution operation in decoding and then added with the corresponding encoder's feature graph to get the output, while SegNet does not need to save the feature graphs of the whole Encoder part, but only makes use of the pooled index to save memory space. The training process and

detection accuracy versus iterations of different methods are shown in Fig.5.

Fig.5 Training process and detection accuracy versus iterations of FCN, SegNet and MSCNNA

It is seen from Fig.5 that MSCNNA has excellent stability in the aircraft detection task, and during the training process MSCNNA converges faster and the loss value is less fluctuating than that of FCN and SegNet, which converge slowly during training and their loss values are still high when training is completed, and FCN converges faster during the initial training, but its loss value gradually increases and greatly fluctuates when it continues to train. The training process of SegNet is relatively stable after 1000 iterations, and its loss value is higher than MSCNNA when the train is completed, which indicates that the training effect of SegNet is poor under the same epoch times. The reason is that it employs the pooling index retained during maximum pooling in the corresponding Encoder stage for nonlinear up-sampling, which can reduce the parameter calculation of the network model and accelerates the convergence speed in the training process. Fig.5B shows that MSCNNA achieves the highest accuracy than FCN and SegNet in the aircraft detection, indicating that MSCNNA is effective for aircraft detection task.

Fig.6 shows the detection results by three classical algorithms and three deep learning methods, i.e., K-Means clustering (KMC) [37], Fuzzy C-means clustering (FCMC) [38] and Fuzzy Region Competition & Gaussian Mixture Model (FRC-GMM) [39].

(A) Two original images, one single-aircraft and multi-aircraft

(B) Labeled aircrafts

(B) Segmented aircrafts by three classical segmentation methods

(C) Segmented aircrafts by three segmentation methods

Fig.6 Aircraft detection

From Fig.6, it is obviously found that the results of the deep learning models are very better than that of three classical algorithms. The reason is that deep learning can extract rich semantic features from complex RSIs to detect the contour region of aircrafts, while the classical algorithms utilize the manually designed low-level features which cannot effectively describe the background-complex RSIs. From Fig.6, the detection results indicated that MSCNNA is the best than FCN and SegNet, because MSCNNA can produce the multi-resolution features for multi-scale RSIs, while FCN and SegNet cannot sufficiently extract the detailed information from the multi-scale RSIs for aircraft detection.

To validate the robustness of MSCNNA for aircraft detection in the noise background, the Gaussian noise and Salt&pepper noise with a signal-to-noise of 40 are added to two original RSIs. The aircraft detection experiments with noise are performed by MSCNNA and SegNet, and the results are shown in Fig.7. IoU is adopted to measure the similarity between the segmentation results and the labeled areas of aircraft, as listed in Table 2.

(A) Gaussian

(B) Salt&pepper

(C) Aircraft detection by MSCNNA

(D) Aircraft detection by SegNet

Fig.7 Aircraft detection by MSCNNA and SegNet with Gaussian Salt&pepper noise

Table 2 IoU of MSCNNA and SegNet with different noise types

Noise type	Model	
	MSCNNA	SegNet
Gaussian noise	88.16	81.36
Salt&pepper noise	82.57	74.24

It can be seen from Fig.6 and Table 2 that MSCNNA still has a favorable detection results effect under the noise background, in which Gaussian noise has lower influence on the aircraft detection effect and IoU reaches 88.16%, while the detection effect is poor under Salt&pepper noise, but IoU is also reaches 82.57%. The detection results under the noise background show that MSCNNA is robust to noises.

From Fig.1, it is seen that the aircraft RSIs are various and the aircraft size is relatively small. To demonstrate the advantages of MSCNNA, a lot of detection experiments are implemented using several RSIs and further compared with FCN, SegNet and DeepLab [40]. The detected aircraft images are shown in Fig.8. These RSIs have complex background, wide space scope, low resolution, small and scattered, and one or several multi-directional aircrafts, as shown in Fig.8A.

(A) Original RSI samples

(B) Labeled aircrafts

Fig.8 Aircraft detection by four deep learning based image detection methods

In Fig.8, by comparing the detected aircrafts with the labeled aircrafts, it can be found that MSCNNA has a good detecting effect, and the aircraft detection regions are closest to the labeled aircraft regions, which can meet the high precision requirements of aircraft detection from the natural RSIs. FCN and SegNet are difficult to accurately detect multi-scale aircrafts, especially small aircrafts, and the edges of the detected aircrafts are a little fuzzy with a lot noise. The reason is that MSCNNA can integrate the aircraft features of the multi-scale RSIs.

Four-fold validation experiments of aircraft detection are performed on the augmented dataset of aircraft RSI and compared with DCNN [6], SCNN [7], M-FCN [8], CNNWSL [31] and CNNSS [32]. The average detection results are regarded as the aircraft detection results, as shown in Table 3.

Table 3 Performance comparison of different methods for aircraft detection results

Methods	<i>Precision</i> /%	<i>Recall</i> /%	<i>F</i> /%	Detection time /s
DCNN	71.12	69.51	70.31	0.71
SCNN	69.75	67.68	68.35	0.65
FCN	65.92	68.25	66.94	0.73
M-FCN	76.27	68.44	72.14	0.64
CNNWSL	75.38	69.17	72.14	0.67
CNNSS	70.62	67.28	68.91	0.68
MSCNNA	82.10	73.92	77.79	0.61

As can be seen from Table 3, the detection accuracy, recall rate and F values of MSCNNA are the highest, which are 80.10%, 73.92% and 77.79%, respectively. Although M-FCN, CNNWSL and MSCNNA are based on the same network architecture, MSCNNA has high performance evaluation indexes, mainly because MSCNNA utilizes multi-scale and global attention mechanism (GAM) to obtain multi-scale significant classification characteristics. M-FCN uses deconvolution process to restore image resolution and optimize segmentation results. However, the down-sampling operation of this method weakens the feature extraction ability of the model, leading to its poor segmentation ability of aircraft region, and therefore its detection performance index is poor. Compared with FCN, M-FCN is far superior to FCN in segmentation performance, mainly because FCN is used to segment images by setting fixed thresholds, so the segmentation results are poor. FCN is trained to be sensitive to aircrafts, while M-FCN replaces the last two fully-connected layers with a global average pooling (GAP) and two convolutional layers. Unlike DCNN and CNNSS that require object location data for training, M-FCN only needs image-level labeled data. Both DCNN and SCNN extract deep-level classification features by adding convolution layers, but due to aircrafts in RSIs occupy a relatively small part, the deep-level features may loss the target features. Therefore, these two methods are weaker than MSCNNA in performance evaluation indexes. According to the above results, it can be judged that MSCNNA has a good detection effect, and the detection result of aircraft region is closest to the marking value, which can meet the high precision requirements of aircraft detection in natural state. However, due to the small area of the aircraft image and the high similarity between some aircraft areas and background areas, it is difficult to distinguish the boundary of the aircraft area from the background area. Therefore, it can be seen from Table 3 that the aircraft detection performance evaluation index value of the algorithm is still low.

In terms of single image detection time, the detection time of MSCNNA is smaller than that of other detection algorithms, mainly because MSCNNA adopts shallow cascade mode for training and uses GAM to further select the obtained multi-scale features, which not only reduces the hardware requirements of the network model, but also reduces the model training time. Because DCNN and CNNSS adopt deep convolutional layers, so they take a long time to detect a single image. In FCN, deconvolution operation is required to restore the resolution of the image, thus increasing the detection time of a single image. Based on the class activation maps, M-FCN generates heat maps via reverse weighting for locating the aircraft object, reducing the model training time.

From the above analysis, the comparative experiments of Figs.5 to 8 and Tables 2 and 3 reveal that MSCNNA has a great advantage in small aircraft detection. From Fig.8 and Table 3, it is also seen that MSCNNA is capable of detecting aircraft targets of different resolutions and shows strong feature representation ability in the detection of small aircrafts.

5. Conclusion

Accurate aircraft detection is always an important but challenging task for the various aircraft poses, sizes and complex backgrounds, and in particular, the shapes of the same aircraft RSI are multi-size in different resolution images and their sizes are relatively small and vary greatly in the same resolution image. Although the standard DCNN and FCN can extract rich semantic features, they may loss the bottom-level location information. The features of small aircrafts may also be submerged by redundant top-level features, resulting in poor detection. To solve the problems, a novel network namely multi-scale convolution neural network with attention (MSCNNA), is constructed for aircraft detection in RSIs. Multi-size images and global attention mechanism (GAM) are used to extract the

multi-scale significant classification characteristics to improve the performance of detecting multi-scale aircrafts in RSIs. GAM gradually suppresses irrelevant regions in the input feature maps. Comparison results show that MSCNNA is effective and feasible for aircraft detection in RSIs. Still, there is room for improvement to detect small aircrafts precisely in RSIs, especially in some complicated conditions. In the future work, MSCNNA will be applied to the Weakly Supervised Aircraft Detection Dataset (WSADD) for algorithm benchmarking, and does not need aircraft annotation information in the training data.

References

- [1] Sun Hao, Sun Xian, Wang Hongqi, et al. Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. *IEEE Geoscience and Remote Sensing Letters*, 2012, 9(1): 109-113
- [2] X. Zhu, Z. Zhu, Y. Xiong. Aircraft Recognition Based on Feature Fusion and Feature Selection. *IEEE 5th International Conference on Computer and Communications (ICCC)*, 2019:1676-167
- [3] Zhao A, Fu K, Sun H, et al. An Effective Method Based on ACF for Aircraft Detection in Remote Sensing Images. *IEEE Geoscience & Remote Sensing Letters*, 2017, 14(5):744-748.
- [4] Yan H. Aircraft detection in remote sensing images using centre-based proposal regions and invariant features. *Remote Sensing Letters*, 2020, 11(8):787-796.
- [5] Bin P, Jianhao T, Qi Z, et al. Cascade Convolutional Neural Network Based on Transfer-Learning for Aircraft Detection on High-Resolution Remote Sensing Images. *Journal of Sensors*, 2017, 2017:1-14.
- [6] Guo Z, Song P, Zhang Y, et al. Aircraft Detection Method Based on Deep Convolutional Neural Network for Remote Sensing Images. *Journal of Electronics & Information Technology*, 2018.
- [7] Hu G, Yang Z, Han J, et al. Aircraft detection in remote sensing images based on saliency and convolution neural network. *Eurasip Journal on Wireless Communications & Networking*, 2018, 2018(1):26.
- [8] Yang Y, Yin Z, Bi F, et al. M-FCN: Effective Fully Convolutional Network-Based Airplane Detection Framework. *IEEE Geoscience & Remote Sensing Letters*, 2017, 14(8):1293-1297.
- [9] Zhang Y, Fu K, Sun H, et al. A multi-model ensemble method based on convolutional neural networks for aircraft detection in large remote sensing images. *Remote Sensing Letters*, 2018, 9(1):11-20.
- [10] Zhong C, Ting Z, Chao O. End-to-End Airplane Detection Using Transfer Learning in Remote Sensing Images. *Remote Sensing*, 2018, 10(1):139.
- [11] Li Y, Zhang S, Zhao J, et al. Aircraft Detection in Remote Sensing Images Based On Deep Convolutional Neural Network. *IOP Conference Series: Earth and Environmental Science*, 2019, 252(5):052122.
- [12] Wang T, Cao C, Zeng X, et al. An Aircraft Object Detection Algorithm Based on Small Samples in Optical Remote Sensing Image. *Applied Sciences*, 2020, 10(17):5778.
- [13] Fu K, Chang Z, Zhang Y, et al. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020, 161:294-308.
- [14] Liu Q, Xiang X, Wang Y, et al. Aircraft detection in remote sensing image based on corner clustering and deep learning. *Engineering Applications of Artificial Intelligence*, 2020, 87(Jan.):103333.

- [15] Mustafa H T, Yang J, Zareapoor M. Multi-scale convolutional neural network for multi-focus image fusion. *Image and Vision Computing*, 2019, 85(MAY):26-35.
- [16] Zhang X, Wang T, Yang Y. Hyperspectral image classification based on multi-scale residual network with attention mechanism. *Geosciences Journal*, 2020.
- [17] Ghorbani Z, Behzadan A H. Monitoring offshore oil pollution using multi-class convolutional neural networks. *Environmental Pollution*, 2021(2):117884.
- [18] Lim L A, Keles H Y. Learning multi-scale features for foreground segmentation. *Pattern Analysis and Applications*, 2020, 23(3):1369-1380.
- [19] Zhao L, Jia K. Multiscale CNNs for Brain Tumor Segmentation and Diagnosis. *Comput. Math Methods Med*, 2016, 2016:1-7.
- [20] Zhipeng Deng, Hao Sun, Shilin Zhou, et al. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018, 145:3-22.
- [21] Wang Y, Li H, Jia P, et al. Multi-Scale DenseNets-Based Aircraft Detection from Remote Sensing Images. *Sensors*, 2019, 19(23):5270.
- [22] J. Fu, J. Liu, H. Tian. Dual Attention Network for Scene Segmentation. *IEEE Conference on Computer Vision and Pattern (CVPR)*, 2019:3141-3149.
- [23] Zhang Q, Cong R, Li C, et al. Dense Attention Fluid Network for Salient Object Detection in Optical Remote Sensing Images. *IEEE Transactions on Image Processing*, 2021,30: 11305-1317.
- [24] Tan T, Wang Z, Du H, et al. Lightweight pyramid network with spatial attention mechanism for accurate retinal vessel segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 2021, 16(1).
- [25] Dayananda C, Choi J Y, Lee B. Multi-Scale Squeeze U-SegNet with Multi Global Attention for Brain MRI Segmentation. *Sensors*, 2021, 21(10):3363.
- [26] Weng Y T, Chan H W, Huang T Y. Automatic Segmentation of Brain Tumor from 3D MR Images Using SegNet, U-Net, and PSP-Net. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 2019:226-233.
- [27] Wang Z, Zhang S. Sonar Image detection based on Multi-Scale Multi-Column Convolution Neural Networks. *IEEE Access*, 2019, PP(99):1-1.
- [28] Ju M, Luo J, Wang Z, et al. Adaptive feature fusion with attention mechanism for multi-scale target detection. *Neural Computing and Applications*, 2021, 33(7): 2769-2781.
- [29] Keys R G. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 2003, 29:1153-1160.
- [30] Zhang X, Gao Y, Wang H, et al. Improve YOLOv3 using dilated spatial pyramid module for multi-scale object detection. *International Journal of Advanced Robotic Systems*, 2020, 17(4):172988142093606.
- [31] Zhize Wu, Thomas Weise, Yan Wang, et al. Convolutional Neural Network Based Weakly Supervised Learning for Aircraft Detection From Remote Sensing Image. *IEEE ACCESS*, 2020. <https://creativecommons.org/licenses/by/4.0/>
- [32] Hwang S, Cho S, Shim H. Aircraft Detection using Deep Convolutional Neural Network based Semantic Segmentation. *Journal of Institute of Control Robotics & Systems*, 2017,23(8):625-634.
- [33] GS Xia, J Hu, F Hu, et al. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2017, 55(7):3965-3981.

- [34] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015,115: 211–252.
- [35] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2014, 39(4):640-651.
- [36] Badrinarayanan V, Kendall a, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for scene segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, 39(12):2481-2495.
- [37] Xiang, D, Tang Tao, Hu C, et al. A Kernel Clustering Algorithm with Fuzzy Factor: Application to SAR Image Segmentation. *Geoscience and Remote Sensing Letters*, 2014, 11(7):1290-1294.
- [38] Chen Y, Zuo J, Gui Y, et al. The Combination with Global Search Algorithm and Fuzzy C-means Clustering Method in Image Segmentation. *Journal of Computational Information Systems*, 2014, 10(22):9547-9554.
- [39] Yin S, Zhang Y, Karim S. Large Scale Remote Sensing Image Segmentation Based on Fuzzy Region Competition and Gaussian Mixture Model. *IEEE Access*, 2018:1-1.
- [40] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2018, 40(4):834-848.