IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received August 14, 2020, accepted August 25, 2020, date of publication August 27, 2020, date of current version September 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019956

Convolutional Neural Network Based Weakly Supervised Learning for Aircraft Detection From Remote Sensing Image

ZHI-ZE WU[1], **THOMAS WEISE**[1], **YAN WANG**[2], **AND YONGJUN WANG**[3]
[1]School of Artificial Intelligence and Big Data, Institute of Applied Optimization, Hefei University, Hefei 230601, China
[2]School of Art, Anhui Jianzhu University, Hefei 230601, China
[3]School of Design, Hefei University, Hefei 230601, China

Corresponding author: Yongjun Wang (wyj@hfuu.edu.cn)

This work was supported in part by the Anhui Provincial Project of Social Sciences Union Research under Grant 2019CX058, in part by the Anhui Provincial Education Department Project under Grant SK2015A484, in part by the Youth Project of the Provincial Natural Science Foundation of Anhui Province under Grant 1908085QF285, in part by the National Natural Science Foundation of China under Grant 61673359, in part by the Hefei Specially Recruited Foreign Expert Program; in part by the Scientific Research and Development Fund of Hefei University under Grant 19ZR05ZDA, and in part by the Talent Research Fund Project of Hefei University under Grant 18-19RC26.

ABSTRACT Object detection methods based on Convolutional Neural Networks (CNNs) require a large number of images with annotation information to train. In aircraft detection from remote sensing images (RSIs), aircraft targets are usually small and the cost of manual annotation is very high. In this article, we tackle the problem of weakly supervised aircraft detection from RSIs, which aims to learn detectors with only image-level annotations, i.e., without bounding-box labeled data during the training stage. Based on the fact that the feature maps learned from the CNN network are localizable, we propose a simple yet efficient aircraft detection algorithm called Weakly Supervised Learning in AlexNet (AlexNet-WSL). In AlexNet-WSL, we utilize the AlexNet CNN as backbone network, but replace the last two fully connected layers with a Global Average Pooling (GAP) and two convolutional layers. Based on the class activation maps, we generate heat maps via reverse weighting for locating the target object. Unlike object detection methods that require object location data for training, our proposal only needs image-level labelled data. We furthermore build a set of remote sensing aircraft images, the Weakly Supervised Aircraft Detection Dataset (WSADD) for algorithm benchmarking. The experimental results on the WSADD show that AlexNet-WSL effectively detects the aircraft and achieves a detection effect equivalent to the Faster R-CNN method and the YOLOv3 method, which both require bounding-box labelled data for training, with a lower false alarm rate and a shorter training time.

INDEX TERMS Aircraft detection, remote sensing image, weakly supervised learning, convolutional neural network.

I. INTRODUCTION

Remote sensing images (RSIs) are generated by acquiring target information from reflected, radiated, or scattered electromagnetic waves through sensors mounted on various remote platforms, which are far away from the target object. With the rapid development of remote sensing technology, various platforms with different imaging methods and different spatial resolutions have emerged and generate a large number of images. Nowadays, remote sensing images have become an indispensable and important resource that are widely used in civil and military applications [1]. As an important research direction of remote sensing image interpretation and analysis, object detection has attracted the interest of academia and industry. Aircraft are one of the most important targets in this field, which brings their recognition from remote sensing image into the focus of attention.

As the spatial resolution of remote sensing image becomes higher and higher, the information contained in the images becomes more and more abundant. With the increasing processing speed of computers, the image processing ability rises and researchers use computer vision to process and analyze

The associate editor coordinating the review of this manuscript and approving it for publication was Qiangqiang Yuan.

large numbers of RSIs. The research goal often is to get rid of the bottleneck of manual interpretation and processing and to achieve automated, intelligent interpretation of RSIs.

With the rapid improvement of Graphics Processing Unit (GPU) performance, deep learning-based methods have been used with overwhelming success in computer vision, data mining, and recommendation systems. Many fields such as smart medical care and unmanned driving have made remarkable achievements [2]. In the field of computer vision, Convolutional Neural Networks (CNNs) based deep learning methods have made breakthrough progress in image classification [3]–[5], image segmentation [6], [7], and object detection [8]–[12]. Compared with traditional methods, CNNs can extract richer semantic features and high-level representations, and also can better identify the discriminations between distinct objects. More importantly, the CNN is an end-to-end model structure. The original image is used as the input and the output of the network is the final result for the end user, which removes the need for previously required complex manual operations such as data preprocessing, feature extraction, and characterization.

Researchers have tried to introduce CNNs into aircraft recognition from RSIs, but this application field is still in its infancy and there are many difficulties and challenges: (1) Unlike natural scene images, target objects in RSIs occupy a relatively small part of the image. (2) Under the condition of high spatial resolutions, more complex backgrounds and interference factors that are difficult to distinguish from aircraft targets appear. (3) CNN methods require a large amount of image data with labeled information to learn, while labeling RSIs requires high labor costs.

Feature maps extracted by a CNN network are localizable representations of the image [13], [45]. With this in mind, we propose a simple yet efficient weakly supervised aircraft detection pipeline. Based on class activation maps, it locates the target object in RSIs. With only image-level annotations, i.e., without bounding-box labeled data, the CNN AlexNet [5] is used to implement the aircraft detection. We call our method Weakly Supervised Learning in AlexNet (AlexNet-WSL). Additionally, we build and provide a new remote sensing aircraft detection dataset using Google Earth satellite imagery. The experimental results on the newly-built dataset show that our proposed method effectively detects the aircraft. It achieves a detection effect equivalent to both Faster R-CNN [10] and YOLOv3 [11], which, however, need to use bounding-box labelled data. Additionally, our AlexNet-WSL has a lower false alarm rate and a shorter training time.

Our contributions are summed up as follows:

- We design a new aircraft detection model for RSIs, which integrates Weakly Supervised Learning into a CNN. The proposed method effectively extracts the semantic features of the data with only image-level annotations, thus significantly improving the aircraft detection process.
- We propose a heat map generation method during the feedforward step of the test phase. With the highlighted parts in the heat map, we can generate a binary image indicating the object location.
- We build a benchmark dataset for remote sensing aircraft detection, called the Weakly Supervised Aircraft Detection Dataset (WSADD). We make the WSADD available in an online repository.

The rest of this article is organized as follows. Section 2 provides a brief overview of the related work on aircraft detection and weekly supervised learning. In Section 3, we introduce our novel weekly supervised aircraft detection algorithm along with a model architecture analysis. Then, Section 4 details the experimental results and discussions on the WSADD dataset. Finally, we conclude the paper in Section 5.

The WSADD is downloadable at the website of: https://doi:10.5281/zenodo.3843229.

II. RELATED WORK

A. AIRCRAFT DETECTION FROM RSIs

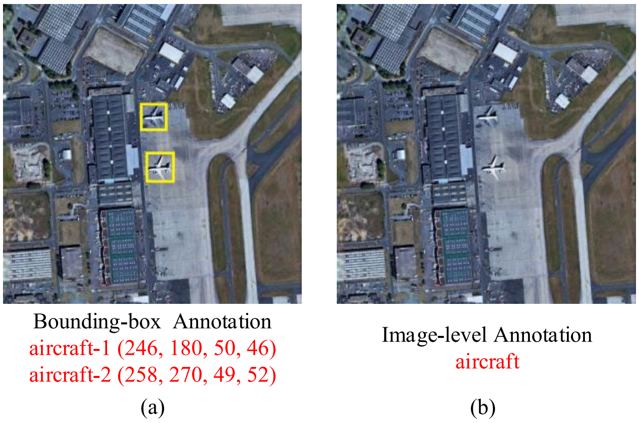
The process of aircraft detection from RSIs involves processing and analyzing images to estimate whether they contain aircraft targets and then to calibrate the positions of the detected aircraft. Similar to general object detection methods, aircraft target detection from RSIs can be divided into the following three steps: candidate region selection, candidate region feature extraction, and aircraft target recognition.

For candidate region selection, sliding window methods [14], [15] based on saliency [16] are commonly used. The typical approaches for candidate region feature extraction are based on general low-level features [17], [18], middle-level features [19], [20], and the specific design features of the aircraft targets [21], [22]. The methods for aircraft target recognition mainly employ template matching [23], [24] and model-based learning [25]–[27]. These methods have achieved some good results, but also have limitations of ow precision and long runtime [28]. There is a strong correlation among these three steps, as the results of each step will directly affect the next step. Furthermore, the serial connection between each step needs to be carried out manually, which complicates the whole detection process.

Deep neural networks have made rapid progress in the application of object detection in computer vision. Among them, the R-CNN [29] is considered as a milestone work. Based on the R-CNN framework, SPP-Net [8], Fast R-CNN [9], Faster R-CNN [10], YOLO [11], and SSD [12] were developed successively, continuously improving the efficiency of object detection.

As a result, the application of deep neural networks has led to breakthroughs in aircraft detection from RSIs [1]. The work [30] proposes a DBNs based pixel-wise learning method for aircraft detection from RSIs. In [31], an aircraft landmark points matching mechanism is integrated into a vanilla network for object detection in satellite images. Zuo *et al.* [32] perform a combination of aircraft detection and semantic segmentation using a CNN method. A conditional generative adversarial networks (GANs) based unsupervised representation learning is proposed for aircraft detection in [33]. Based on multi-class activation mapping, the work [34] constructs two subnetworks, i.e., the target network and the object network for aircraft detection from very high resolution (VHR) images.

Although promising results have been reported in the aforementioned methods, there is still a challenge problem for the aircraft detection in RSIs: It requires a high number of data with object location annotation information, such as the bounding-box annotation shown in Figure 1 (a). The obstacle of high labor cost for the annotation of the object location in the image is encountered. Because of the large sizes of RSIs and the small sizes of aircraft targets, it is difficult to observe the relevant target object with the human eye. Moreover, there is the problem of human subjectivity in labeling, which can easily affect the results of image annotation. Thus, the problem of artificial object annotation is particularly prominent for detecting an aircraft in a remote sensing image.



Bounding-box Annotation
aircraft-1 (246, 180, 50, 46)
aircraft-2 (258, 270, 49, 52)
(a)

Image-level Annotation
aircraft
(b)

FIGURE 1. Different annotations for aircraft detection in a remote sensing image.

B. WEAKLY SUPERVISED LEARNING

Weakly supervised learning (WSL) only requires image-level annotations that indicate whether or not an image contains the target objects [35]–[40], [46], [47], as shown in Figure 1 (b). The first attempt to utilize WSL for object detection from RSIs was proposed in [40]. Although this work solves the problem of aircraft detection in a RSI under the image-level annotation, the representation of the aircraft feature still utilizes the bag of visual words (BoVW) [41] method. In [35], Zhang *et al.* propose to heuristically mine the positive samples and to evaluate the learned detector based on the negative data. A similar idea is used in [36], Han *et al.* perform geospatial object detection by combining saliency, intra-class compactness, and inter-class separability using a deep Boltzmann machine. A negative bootstrapping idea is used in [37]. This work iteratively learns the detector by selecting the most informative negative samples. In [38], a novel weakly supervised, multi-instance learning algorithm is designed to learn instance-wise vehicle detectors from the region-level group annotation. Zhang *et al.* [39] build a coupled CNN for simultaneously generating the object proposals and locating the aircraft target from large-scale VHR images.

The latest works on WSL based object detection from RSIs are [46], [47]. In [46], Yao *et al.* propose a dynamic curriculum learning strategy to perform weakly supervised object detection from high-resolution RSIs. This work can progressively learn the object detectors by feeding training images with increasing difficulty that matches current detection ability. The work [47] designs a dual-contextual instance refinement strategy to divert the focus of detection network from local distinct part to the object and further to other potential instances by leveraging both local and global context information. With this, it can significantly boost object detection accuracy compared with the state of the arts.

Different from the above methods, we propose a simple yet efficient weakly supervised detection strategy for aircraft detection from RSI. We utilize Alex-Net [5] as backbone network, but replace last two fully connected layers with Global Average Pooling (GAP) [42] and two convolutional layers. Based on the class activation maps, we generate heat maps via reverse weighting for locating the target object.

III. AIRCRAFT DETECTION BASED ON WEAKLY SUPERVISED LEARNING

The framework for aircraft detection using our AlexNet-WSL is shown in Figure 2. It has two stages, namely the training and the testing step.

In the training stage, the labels of the training dataset indicate only whether the image contains an aircraft or not. No annotation information about its position, shape, or size are required. The training dataset is divided into the ‘‘positive sample set’’ with images containing aircraft and a ‘‘negative sample set’’ with those that do not. The AlexNet-WSL is used for a binary classification task, i.e., determining whether a remote sensing image contains aircraft or not.

When the training of AlexNet-WSL is complete, a test image is input into the trained AlexNet-WSL network model for forward propagation. Then, the weight corresponding to the image classification result of ‘‘with aircraft’’ in the fully-connected layer 9 is extracted, weighted, and averaged with the characteristic map output by convolutional layer 7 and then overlapped into a Heat Map. The highlighted area is the basis for the AlexNet-WSL network model to distinguish the test image using the ‘‘with aircraft’’ category. The biggest difference between the ‘‘with aircraft’’ images and the ‘‘without aircraft’’ images is where the former contain an aircraft target, so these areas correspond to the position of the aircraft target in the test image. Then, adaptive threshold based segmentation is used to obtain a binary graph. Finally, according to the minimum circumscribed rectangle corresponding to each connected region in the binary graph, the detection of the aircraft in the test image is completed. Next, we will describe the AlexNet-WSL network model, the pipeline of Heat Map generation, and how to locate aircraft in detail.

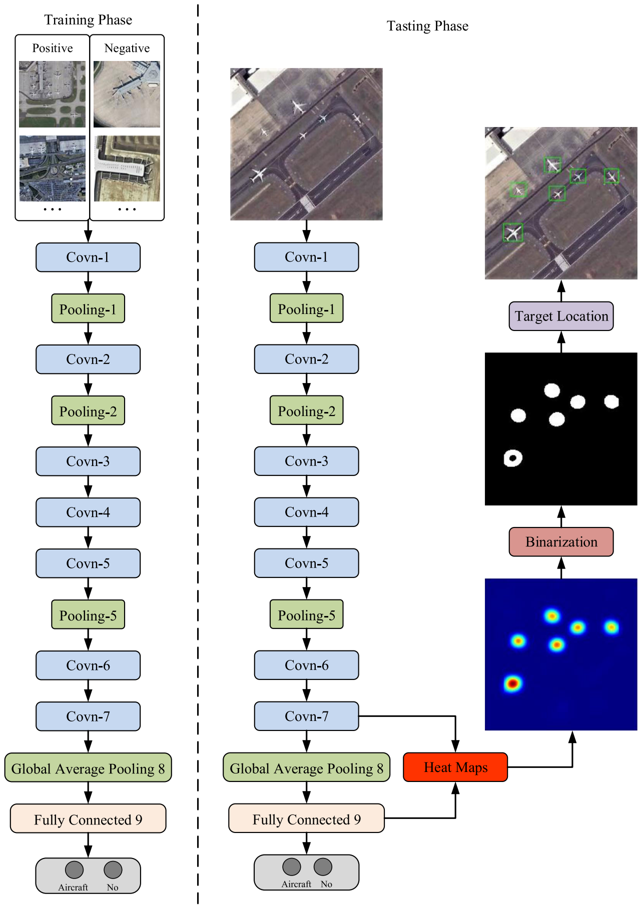


FIGURE 2. Framework for aircraft detection using AlexNet-WSL.

A. AlexNet-WSL NETWORK MODEL

We utilize Alex-Net [5] as our backbone network. It consists of 12 layers, including 7 convolutional layers, 3 pooling layers (pooling layer 1, 2, and 5), 1 GAP layer (layer 8), and 1 fully-connected layer (layer 9). Table 1 illustrates the structural parameters of the AlexNet-WSL network model.

Every convolutional unit in the CNN is essentially a detector that can locate the target in the image [10]. For example, if a target in the image is located in the upper left corner of the image, the upper left corner of the feature image after the convolutional layer will produce a greater response. If the target appears in the lower right corner, the region in the lower right corner of the feature map will have a larger response. The fully-connected layer "flattens" the output feature map of the convolutional layer into a one-dimensional characteristic vector, which loses the spatial position information of the image. Most of the parameters in the CNN are occupied by the fully-connected layer. Therefore, we replace the first two fully-connected layers by two convolutional layers and a GAP layer. The last fully-connected layer is used for classification.

This way, the image spatial position information is propagated to the last layer of the network, which provides the basis for the subsequent generation of a heat map.

TABLE 1. Specific structural parameters of AlexNet-WSL network.

Layer	Input	Kernel	Stride	Pad	Output
Conv-1	227×227×3	11×11	4	0	55×55×96
Pool-1	55×55×96	3×3	2	0	27×27×96
Conv-2	27×27×96	5×5	1	2	27×27×256
Pool-2	27×27×256	3×3	2	0	13×13×256
Conv-3	13×13×256	3×3	1	1	13×13×384
Conv-4	13×13×384	3×3	1	1	13×13×384
Conv-5	13×13×384	3×3	1	1	13×13×384
Pool-5	13×13×384	3×3	1	0	11×11×384
Conv- 6	11×11×384	3×3	1	1	11×11×512
Conv-7	11×11×512	3×3	1	1	11×11×512
GAP-8	11×11×512	11×11	11	0	1×1×512
FC-9	1×1×512	1×1	1	0	1×1×2

Another benefit is that the number of parameters in the network model is decreased, which reduces the chance of model overfitting.

After the last convolutional layer in the CNN, a pooling layer is added and an average pooling operation is applied. The size of the pooling window is set as the size of the output feature map of the convolutional layer. For example, in the AlexNet-WSL network model, the last convolutional layer is layer 7, whose output is 43 × 43×512, that is, 512 feature maps of size 43 × 43. We use GAP to calculate the average value of all pixels on each feature map as output. This way, 512 feature maps are mapped to 512 average values, which will form a one-dimensional feature vector as the output of the GAP layer.

B. HEAT MAP GENERATION

Based on the class activation mapping method proposed in [13], we generate a heat map of the interesting regions in the remote sensing image. We illustrate the specific process in Figure 3. For the structure and specific implementation of the AlexNet-WSL network model, the decision for the image classification is based on the feature vector input into the classifier (i.e., the GAP feature vector in Figure 3). However, since the feature vector is one-dimensional, it is impossible to examine the AlexNet-WSL network model to determine the image category according to specific regions of the image.

The above-mentioned one-dimensional feature vector is obtained via the GAP of the feature map output by the 7th convolutional layer. Thus, the feature vector corresponds to these feature maps aggregated one by one. Then, in the final classification step, if a certain value in the eigenvector contributes

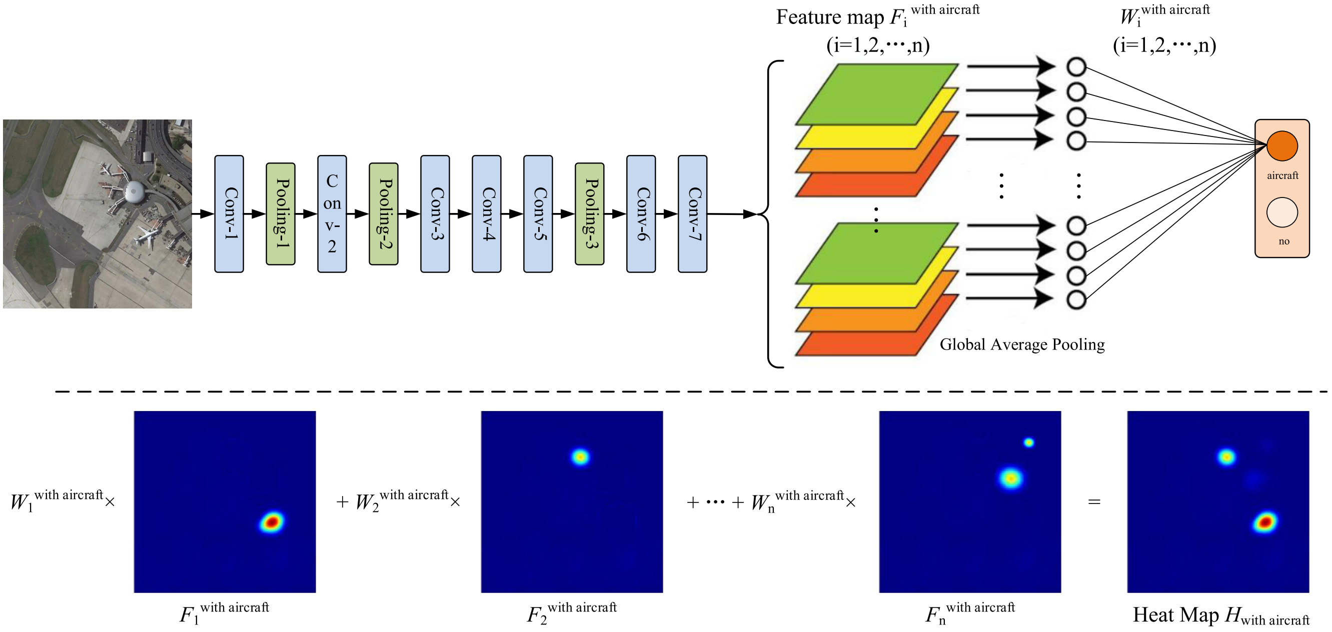

Conv-1
Pooling-1
Conv-2
Pooling-2
Conv-3
Conv-4
Conv-5
Pooling-3
Conv-6
Conv-7
Feature map $F_i^{\text{with aircraft}}$
(i=1,2,⋯,n)
$W_i^{\text{with aircraft}}$
(i=1,2,⋯,n)
aircraft
no
Global Average Pooling
$W_1^{\text{with aircraft}}\times$
$+ W_2^{\text{with aircraft}}\times$
$+ \cdots + W_n^{\text{with aircraft}}\times$
=
$F_1^{\text{with aircraft}}$
$F_2^{\text{with aircraft}}$
$F_n^{\text{with aircraft}}$
Heat Map $H_{\text{with aircraft}}$

FIGURE 3. Illustration of the generation process of a heat map.

to image discrimination with the result that "there is an aircraft", then the corresponding heat map of this value will reflect this contribution. The weight corresponding to "with aircraft" is extracted directly from the output characteristic map of the 7th convolutional layer and the classification result of the 9th fully-connected layer. A heat map is obtained in via reverse weighting using Formula (1).

In Formula (1), H_c represents the generated heat map for category c, F_i^c represents the output feature map of the 7th convolutional layer in the AlexNet-WSL network model, and W_i^c represents the connection weight between the output characteristic vector of the 8th GAP layer and the 9th fully-connected layer.

$$H_c = \sum\nolimits_{i=1}^{n} W_i^{c*} F_i^c \tag{1}$$

C. OBJECT LOCATION

To calibrate the aircraft position information in the remote sensing test images, we transform the heat map into a binary image. Then, the maximum inter-class variance is adaptively used to find the threshold values of the foreground and background in the heat map. By traversing different thresholds, our method calculates the difference between the background and foreground based on gray values. The larger the inter-class difference is, the larger is also the variance between the foreground and background. The threshold is thus the value at which the inter-class difference reaches its maximum.

For a remote sensing image with a size of M^*N, the foreground and background are segmented according to the gray-scale threshold, in which pixels larger than the threshold are marked as the foreground and the rest is the background. Let the number of pixels in the foreground be N_f, the average gray value be G_f, the number of pixels of the background be N_b, and the average gray value be G_b. The average gray value of the whole image is G, and the class variance between the background and foreground is σ. Then

$$N_f + N_b = M^*N \tag{2}$$

$$G = G_f^* \frac{N_f}{M^*N} + G_b^* \frac{N_b}{M^*N} \tag{3}$$

$$\sigma = \frac{N_f}{M^*N}\left(G - G_f\right)^2 + \frac{N_b}{M^*N}\left(G - G_b\right)^2 \tag{4}$$

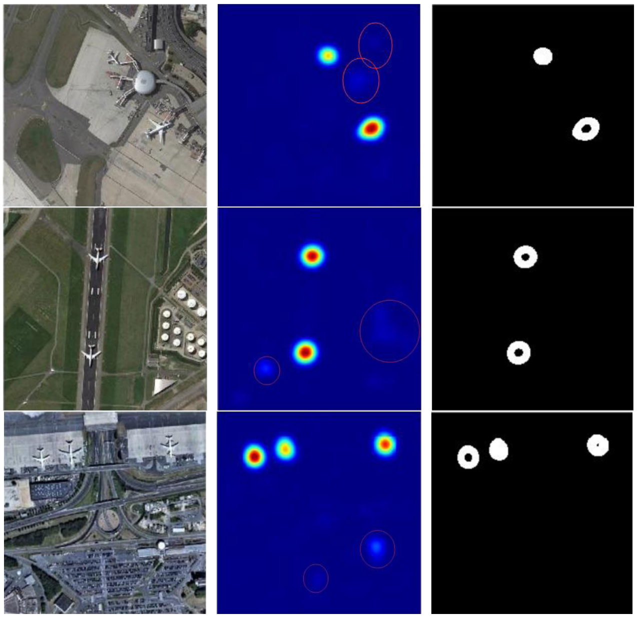
Substitute (3) into (4) to get the equivalent formula:

$$\sigma = \frac{N_f}{M^*N}{}^{*}\frac{N_b}{M^*N}\left(G_f - G_b\right)^2 \tag{5}$$

If $\sigma\ (T)$ is the variance between the background and foreground, then:

$$T_{opt} = \underset{T\in 0,1,\cdots 255}{\arg\max}\ (\sigma\ (T)) \tag{6}$$

where T_{opt} is the threshold value of the method of maximum variance of inter-class. Then, the foreground and background are segmented and binarized using the obtained threshold T_{opt}. Figure 4 shows that binarization can effectively suppress the response to the background area (the area surrounded by the red circles in the figure). After binarization, the minimum circumscribed rectangles of each connected region in the binary image are calculated. The coordinates of these rectangles are then mapped to the original image to detect aircraft targets.



(a) Input Image (b) Heat Map (c) Binary Image

FIGURE 4. Binary process of the heat map.

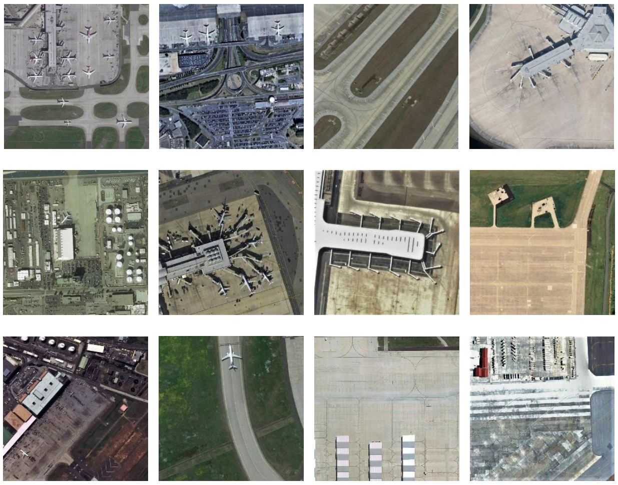
IV. EXPERIMENTAL STUDY

The environment for all experiments in this section is a Microsoft Window 10 operating system on an Intel Core i7-7700 processor with 16 GB memory and an Nvida GTX Titan X graphics card. We use the Caffe library [43] and MATLAB interfaces. We now first describe the data set and the criteria for experiment evaluation used. Next, the implementation details and the parameter analysis are provided. Then, we present the quantization and visualization evaluations.

A. DESCRIPTION OF THE DATA SET AND EXPERIMENTAL SETUP

1) WSADD

In order to evaluate the aircraft target detection algorithm based on weak supervised learning, we construct the WSADD data set and make it available in the online repository[1]. The images in this dataset include airports and nearby areas of different countries (mainly from China, the United States, the United Kingdom, France, Japan, and Singapore) taken from the Google Earth satellite. The dataset comprises 700 RSIs in total, of which 400 images contain an aircraft target (the ‘‘positive sample set’’) and the other 300 do not (the ‘‘negative sample set’’). In the process of dataset construction, the spatial resolution of the images was controlled between 0.3 m and 2 m, and the size was fixed to 768 × 768 pixels. We sought to collect images from different sensors during different daytimes, different seasons, and different light intensities to ensure that the dataset has a high diversity. Some image samples from the dataset are shown in Figure 5.



(a) Positive samples (b) Negative samples

FIGURE 5. Sample image of the WSADD dataset.

As demonstrate in Figure 5, the image scene in the positive sample set of WSADD is very complex. The proportion of aircraft targets inside the whole images is small. The images also contain a large number of background targets, such as oil tanks, hangars, and boarding buildings. The negative sample set mainly includes runway and apron images without aircraft.

2) DATA AUGMENTATION

A data augmentation strategy is adopted to generate new images from the original input images to expand the volume of the dataset and to reduce the chance of overfitting. As shown in Table 1, the input size of the convolutional layer 1 in the AlexNet-WSL network is 736*736. Therefore, the original images of 768*768 pixels can be randomly cropped with a window of 736*736 as well as inverted. Through this process, the data volume can be expanded by factor $(768-736)^2 * 2 = 2048$. During the network test stage, we crop the four corners and the middle of the test image and flip them. This way, ten images are fed into the network and the corresponding results are averaged as output.

3) CRITERIA FOR EXPERIMENT EVALUATION

Beside running time, following [39], we use three criteria for evaluating the aircraft detection performance: false positive rate (FPR), missing ratio (MR), accuracy (AC). Their calculation formulas are as follows:

$$FPR = \frac{\text{Number of falsely detected aircraft}^*}{\text{Number of detected aircraft}} 100\% \quad (7)$$

$$MR = \frac{\text{Number of undetected aircraft}^*}{\text{Number of aircraft}} 100\% \quad (8)$$

$$AC = \frac{\text{Number of detected aircraf}^*}{\text{Number of aircraft}} 100\% \quad (9)$$

When detected aircraft has a fractional intersection overlap with a test aircraft greater than 0.5, we take it as a true detected aircraft. On the other side, we take it as a missed aircraft.

[1] The WSADD is downloadable at https://doi:10.5281/zenodo.3843229

B. PARAMETER SETTING AND KEY PARAMETER ANALYSIS

1) PARAMETER SETTING

We initialize the AlexNet-WSL by pertaining it on the ImageNet [44]. During the fine-tuning training, following [5], we set the initial learning rate of each layer to 0.001 and reduce it to the previous 10% after 5'000 iterations. For the momentum and weight decay, we set to 0.9 and 0.005, respectively. The batch size of each iteration is set to 40, and the total number of iterations depends on the convergence of the net. During testing, we utilize our proposed Heat Map based object location for calibrate the aircraft position information in the test remote sensing image.

2) KEY PARAMETER ANALYSIS

In order to further study the inner characteristics of deep neural networks for processing the RSIs, like our other work [1], we now analyze the key parameter of the learning rate by investigating the different layers of the AlexNet-WSL using different learning rates.

After the initialization of the AlexNet-WSL, we study the impact of the parameter transfer using different initial learning rate settings for the layers in the AlexNet-WSL network. For the 7 convolutional layers and 1 fully-connected layer in the AlexNet-WSL, we orderly increase a "0" learning rate. By this, we can totally get 8 different settings. As shown in Table 2, each setting can be regarded as a model. In Table 2, M indicates "Model", the learning rate "0" indicates that the layer parameters are transferred from the pre-training stage without fine-tuning on WSADD, and "1" indicates the initial learning rate is 0.001.

TABLE 2. Different learning rate settings the and corresponding model.

M	C-1	C-2	C-3	C-4	C-5	C-6	C-7	FC-9
M-1	1	1	1	1	1	1	1	1
M-2	0	1	1	1	1	1	1	1
M-3	0	0	1	1	1	1	1	1
M-4	0	0	0	1	1	1	1	1
M-5	0	0	0	0	1	1	1	1
M-6	0	0	0	0	0	1	1	1
M-7	0	0	0	0	0	0	1	1
M-7	0	0	0	0	0	0	0	1

We illustrate the detected results in terms of the AC for different models in Figure 6. As can be seen from Figure 2, when using only low-level parameters of the pre-trained AlexNet-WSL without fine-tune in target domain, the models can achieve an AC equivalent to Model-1. It shows that low-level features of CNN have similar semantics. When high-level parameters is not performed fine-tuning, the experimental results will be affected. We can find such phenomenon for Model-4 to Model-8. Based on the observations, we argue that the low-level parameters of the pre-trained CNNs can be directly used for the target task, and it is necessary for high-level parameter fine-tuning.

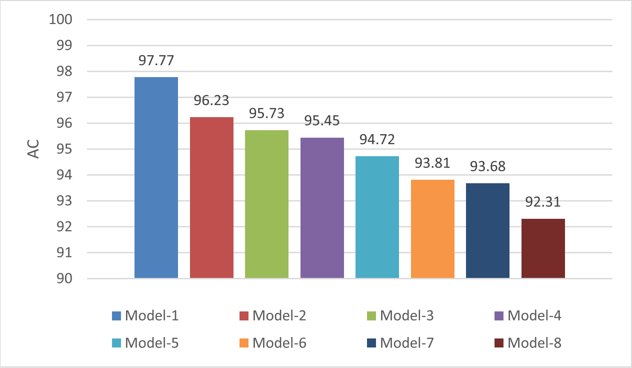


FIGURE 6. Corresponding detected result of different models.

C. EVALUATION OF THE AIRCRAFT DETECTION

In order to evaluate the effectiveness of the proposed algorithm, we compare it with Faster R-CNN [23] and YOLOv3 [24], which have achieved excellent results in natural scene object detection. To train the AlexNet-WSL network model, first, 300 images are randomly selected from the positive sample set of the WSADD dataset and labeled as "with aircraft". The remaining 100 images are used as test data. Then, all 300 images in the negative sample set are selected and labeled as "without aircraft." The 6th fully-connected layer and 7th layer in the original Alexnet network model were removed in the design of the AlexNet-WSL network model, which reduces the number of parameters of the network model. However, using training data with only 600 images still makes it difficult to effectively train the AlexNet-WSL network model and overfitting might still occur.

Therefore, a strategy of transfer learning is adopted, as discussed in the previous section. The selected 600 images are only used to fine-tune and train the AlexNet-WSL network model. For Faster R-CNN and YOLOv3, pretraining is also carried out on the ImageNet dataset. Fine-tuning training uses the same dataset also used for fine-tuning the AlexNet-WSL, but the bounding box annotation information of the aircraft position is given for the 300 positive samples. Next, the remaining 100 images with aircraft, including 358 aircraft targets in total, are used for testing. The experimental results are listed in Table 3.

In this study, several commonly used performance indicators for the detection of aircraft targets are compared, including the detection rate, missed alarm rate, and false alarm rate.

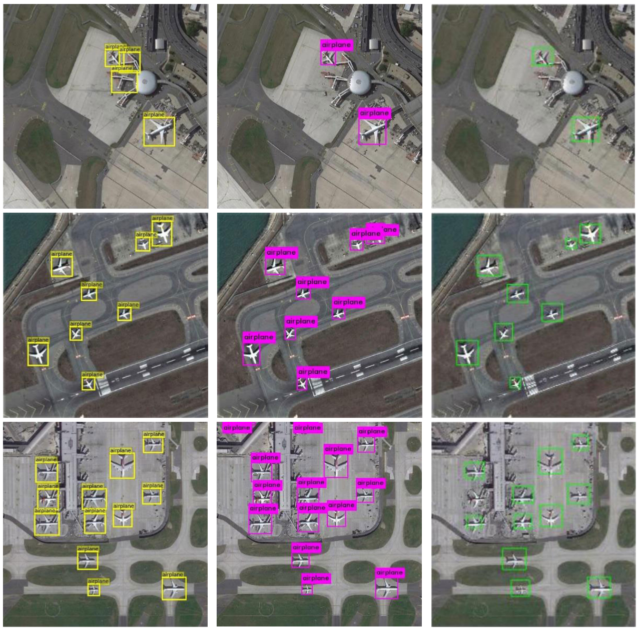
The higher the detection rate of the aircraft target, the better is the performance of the algorithm. In practice, the targets detected by the algorithm may not all be aircraft targets, but also include some interference targets that are mistaken as aircraft targets. The false alarm rate measures this phenomenon. The lower the false alarm rate, the better the ability of the algorithm to distinguish aircraft from the jamming targets.

TABLE 3. Experimental results of the different target detection algorithms on the WSADD dataset.

Method	Faster R-CNN	YOLOv3	Proposed method
Bounding-box annotation	Yes	yes	no
Convergence time	3 hours	10 hours	10 minutes
AC	98.88% (354/358)	**99.16% (355/358)**	97.77% (350/358)
MR	1.12% (4/358)	**0.84%(3/358)**	2.23% (8/358)
FAR	1.39% (5/359)	0.84% (3/358)	**0.57% (2/352)**

Table 2 shows that the detection rate of the aircraft target in the images based on the weakly supervised learning proposed in this article is 97.77%. This is remarkable, as the algorithm does not use any target location annotation information, but performs equivalent to the Faster R-CNN and YOLOv3 methods, which require target location annotation information. Moreover, the false alarm rate of our method is low and it only needs to train a CNN for classification. After adopting the migration learning strategy, we only needed to train on the WSADD dataset for 10 minutes, after which the network model had converged. For Faster R-CNN and YOLOv3, although transfer learning is also used, the network model converged only after training for 3 hours and 10 hours, respectively.

Figure 7 shows some resulting images. The false alarms by Faster R-CNN are mainly due to the recognition of other backgrounds (such as the boarding building) as aircraft, while the false alarms by YOLOv3 appear at the edges of the images. These problems, however, do not affect our method.



(a) Faster R-CNN (b) YOLOv3 (c) Proposed method

FIGURE 7. Selected results for RS aircraft target detection.

From the detected results of the third test image in Figure 7, we notice that our method also performs well for the dense aircraft detection. For this effect, we argue that there are four main reasons: (1) AlexNet is a very successful deep network architecture. (2) We have introduced a transfer learning strategy in the experiment. The network model used is first pre-trained on the large-scale data set ImageNet and is further fine-tuned on the domain data set. (3) The feature maps extracted by a CNN classifier are localizable representation of the image. (4) The weakly supervised learning mechanism based on the heat map proposed in this work is feasible.

Our method also has some limitations. For example, the discovered bounding boxes of the aircraft may be slightly too large or slightly too small in some cases. Since the training data used for our method are not labeled with the target position, our method cannot use this information to modify the target position information output.

V. CONCLUSION

At present, object detection methods based on CNNs require a large amount of data with target location annotation information to be trained. However, in the field of remote sensing, the labor cost of image annotation is high. Therefore, we introduced the idea of weakly supervised learning and combined it with a CNN to develop an aircraft target detection algorithm for RSIs. In the experiments given in Section 4, the proposed AlexNet-WSL algorithm achieves similar detection results as the Faster R-CNN and YOLOv3. These two methods require target location annotation information for their training. The FAR of the proposed method is slightly lower than that of Faster R-CNN and YOLOv3, so the effectiveness of AlexNet-WSL is verified. Furthermore, training AlexNet-WSL is about 18 times faster than Faster R-CNN and about 60 times faster than YOLOv3.

Our AlexNet-WSL does not use boundary regression to modify and optimize the position information of the detected aircraft and thus does not need target position annotation information in the training data. Therefore, the coverage of the target position output is naturally a bit lower than what could potentially be achieved in a fully-supervised learning scenario. This drawback is very small compared to the gained ability to train without needing target position annotations and the much higher training speed.

DISCLOSURE

All authors declare that there are no conflicts of interests. All authors declare that they have no significant competing financial, professional or personal interests that might have influenced the performance or presentation of the work described in this manuscript. Declarations of interest: none. All authors approve the final article.

REFERENCES

[1] Z.-Z. Wu, S.-H. Wan, X.-F. Wang, M. Tan, L. Zou, X.-L. Li, and Y. Chen, "A benchmark data set for aircraft type recognition from remote sensing images," *Appl. Soft Comput.*, vol. 89, Apr. 2020, Art. no. 106132, doi: 10.1016/j.asoc.2020.106132.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[4] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, and Y.-Z. Song, "The devil is in the channels: Mutual-channel loss for fine-grained image classification," *IEEE Trans. Image Process.*, vol. 29, pp. 4683–4695, 2020, doi: 10.1109/TIP.2020.2973812.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105, doi: 10.1145/3065386.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[7] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei, "Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 82–92, doi: 10.1109/CVPR.2019.00017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015, doi: 10.1109/TPAMI.2015.2389824.

[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSS: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[13] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2921–2929, doi: 10.1109/CVPR.2016.319.

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010, doi: 10.1109/TPAMI.2009.167.

[15] W. Zhang, X. Sun, K. Fu, C. Wang, and H. Wang, "Object detection in high-resolution remote sensing images using rotation invariant parts based model," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 74–78, Jan. 2014, doi: 10.1109/LGRS.2013.2246538.

[16] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998, doi: 10.1109/34.730558.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[19] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Stat. Learn. Comput. Vis. (ECCV)*, 2004, pp. 1–22.

[20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2169–2178, doi: 10.1109/CVPR.2006.68.

[21] A. Wahi, C. Palamsamy, and S. Sundaramurthy, "Rotated object recognition-based on hu moment invariants using artificial neural system," in *Proc. World Congr. Inf. Commun. Technol.*, Oct. 2012, pp. 45–49, doi: 10.1109/WICT.2012.6409048.

[22] F. Zhang, S.-Q. Liu, D.-B. Wang, and W. Guan, "Aircraft recognition in infrared image using wavelet moment invariants," *Image Vis. Comput.*, vol. 27, no. 4, pp. 313–318, Mar. 2009, doi: 10.1016/j.imavis.2008.08.007.

[23] U. Erkan and D. N. H. Thanh, "Autism spectrum disorder detection with machine learning methods," *Current Psychiatry Res. Rev.*, vol. 15, no. 4, pp. 297–308, Jan. 2020, doi: 10.2174/2666082215666191111121115.

[24] G. Liu, X. Sun, K. Fu, and H. Wang, "Aircraft recognition in high-resolution satellite images using coarse-to-fine shape prior," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 573–577, May 2013, doi: 10.1109/LGRS.2012.2214022.

[25] G. Cheng, P. Zhou, J. Han, J. Han, and L. Guo, "Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images," *IET Comput. Vis.*, vol. 9, no. 5, pp. 639–647, Oct. 2015, doi: 10.1049/iet-cvi.2014.0270.

[26] L. Zhang, L. Zhang, D. Tao, and X. Huang, "Sparse transfer manifold embedding for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1030–1043, Feb. 2014, doi: 10.1109/TGRS.2013.2246837.

[27] X. Yao, J. Han, L. Guo, S. Bu, and Z. Liu, "A coarse-to-fine model for airport detection from remote sensing images using target-oriented visual saliency and CRF," *Neurocomputing*, vol. 164, pp. 162–172, Sep. 2015, doi: 10.1016/j.neucom.2015.02.073.

[28] F. Zeng, L. Cheng, N. Li, N. Xia, L. Ma, X. Zhou, and M. Li, "A hierarchical airport detection method using spatial analysis and deep learning," *Remote Sens.*, vol. 11, no. 19, p. 2204, Sep. 2019, doi: 10.3390/rs11192204.

[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[30] W. Diao, X. Sun, F. Dou, M. Yan, H. Wang, and K. Fu, "Object recognition in remote sensing images using sparse deep belief networks," *Remote Sens. Lett.*, vol. 6, no. 10, pp. 745–754, Oct. 2015.

[31] A. Zhao, K. Fu, S. Wang, J. Zuo, Y. Zhang, Y. Hu, and H. Wang, "Aircraft recognition based on landmark detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1413–1417, Aug. 2017, doi: 10.1109/LGRS.2017.2715858.

[32] J. Zuo, G. Xu, K. Fu, X. Sun, and H. Sun, "Aircraft type recognition based on segmentation with deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 282–286, Feb. 2018, doi: 10.1109/LGRS.2017.2786232.

[33] Y. Zhang, H. Sun, J. Zuo, H. Wang, G. Xu, and X. Sun, "Aircraft type recognition in remote sensing images based on feature learning with conditional generative adversarial networks," *Remote Sens.*, vol. 10, no. 7, p. 1123, Jul. 2018, doi: 10.3390/rs10071123.

[34] K. Fu, W. Dai, Y. Zhang, Z. Wang, M. Yan, and X. Sun, "MultiCAM: Multiple class activation mapping for aircraft recognition in remote sensing images," *Remote Sens.*, vol. 11, no. 5, p. 544, Mar. 2019, doi: 10.3390/rs11050544.

[35] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu, and L. Guo, "Weakly supervised learning for target detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 701–705, Apr. 2015.

[36] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[37] P. Zhou, G. Cheng, Z. Liu, S. Bu, and X. Hu, "Weakly supervised target detection in remote sensing images based on transferred deep features and negative bootstrapping," *Multidimensional Syst. Signal Process.*, vol. 27, no. 4, pp. 925–944, Oct. 2016.

[38] L. Cao, F. Luo, L. Chen, Y. Sheng, H. Wang, C. Wang, and R. Ji, "Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning," *Pattern Recognit.*, vol. 64, pp. 417–424, Apr. 2017.

[39] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.

[40] D. Zhang, J. Han, D. Yu, and J. Han, "Weakly supervised learning for airplane detection in remote sensing images," in *Proc. 2nd Int. Conf. Commun., Signal Process., Syst.* Cham, Switzerland: Springer, 2014, pp. 155–163, doi: 10.1007/978-3-319-00536-2_18.

[41] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Stat. Learn. Comput. Vis. (WSLCV)*, 2004, pp. 1–2.

[42] Y. Pang, M. Sun, X. Jiang, and X. Li, "Convolution in convolution for network in network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1587–1597, May 2018, doi: 10.1109/TNNLS.2017.2676130.

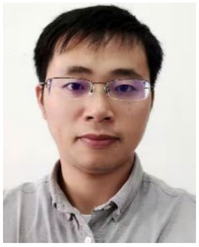
[43] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 675–678, doi: 10.1145/2647868.2654889.

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[45] R. Qiao, A. Ghodsi, H. Wu, Y. Chang, and C. Wang, "Simple weakly supervised deep learning pipeline for detecting individual red-attacked trees in VHR remote sensing images," *Remote Sens. Lett.*, vol. 11, no. 7, pp. 650–658, Jul. 2020, doi: 10.1080/2150704X.2020.1752410.

[46] X. Feng, J. Han, X. Yao, and G. Cheng, "Progressive contextual instance refinement for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Apr. 27, 2020, doi: 10.1109/TGRS.2020.2985989.

[47] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning," *IEEE Trans. Geosci. Remote Sens.*, early access, May 18, 2020, doi: 10.1109/TGRS.2020.2991407.



ZHI-ZE WU received the Ph.D. degree from the School of Computer Science and Technology (SCST), University of Science and Technology of China (USTC), in 2017. He is currently a Researcher with the School of Artificial Intelligence and Big Data, Institute of Applied Optimization, Hefei University. His research interests include image processing, neural networks, and machine learning.

THOMAS WEISE received the Diplom Informatiker (equivalent to M.Sc. degree in computer science) from the Chemnitz University of Technology, in 2005, and the Ph.D. degree in computer science from the University of Kassel, in 2009. He then joined the University of Science and Technology of China (USTC), as a Postdoctoral Researcher, and subsequently became an Associate Professor with the USTC-Birmingham Joint Research Institute in Intelligent Computation and Its Applications (UBRI), USTC. In 2016, he joined Hefei University, as a Full Professor, to found the Faculty of Computer Science and Technology, Institute of Applied Optimization (IAO). He has more than 80 scientific publications in international peer reviewed journals and conferences. His book *Global Optimization Algorithms–Theory and Application* has been cited 840 times. He has acted as a Reviewer, an Editor, or a Programme Committee Member at 70 different venues, and is a member of the editorial board of the *Applied Soft Computing* journal.

YAN WANG received the master's degree from the School of Fine Arts, Central South University for Nationalities, in 2019. She is currently a Lecturer with the School of Art, Anhui Jianzhu University. Her research interests include animation and visual communication in the field of art.

YONGJUN WANG received the bachelor's degree in educational technology from Anhui Normal University, in 1999, and the master's degree in educational technology from East China Normal University, in 2007. From 2010 to 2020, he has served as an Associate Professor with Hefei University. His research interests include computer-aided design and animation design.