Weakly Supervised Learning Based on Coupled Convolutional Neural Networks for Aircraft Detection

Fan Zhang, Bo Du, *Senior Member, IEEE*, Liangpei Zhang, *Senior Member, IEEE*, and Miaozhong Xu

***Abstract*—Aircraft detection from very high resolution (VHR) remote sensing images has been drawing increasing interest in recent years due to the successful civil and military applications. However, several challenges still exist: 1) extracting the high-level features and the hierarchical feature representations of the objects is difficult; 2) manual annotation of the objects in large image sets is generally expensive and sometimes unreliable; and 3) locating objects within such a large image is difficult and time consuming. In this paper, we propose a weakly supervised learning framework based on coupled convolutional neural networks (CNNs) for aircraft detection, which can simultaneously solve these problems. We first develop a CNN-based method to extract the high-level features and the hierarchical feature representations of the objects. We then employ an iterative weakly supervised learning framework to automatically mine and augment the training data set from the original image. We propose a coupled CNN method, which combines a candidate region proposal network and a localization network to extract the proposals and simultaneously locate the aircraft, which is more efficient and accurate, even in large-scale VHR images. In the experiments, the proposed method was applied to three challenging high-resolution data sets: the Sydney International Airport data set, the Tokyo Haneda Airport data set, and the Berlin Tegel Airport data set. The extensive experimental results confirm that the proposed method can achieve a higher detection accuracy than the other methods.**

***Index Terms*—Aircraft detection, convolutional neural networks (CNNs), weakly supervised learning.**

I. Introduction

THE past decade has witnessed a rapid development in modern remote sensing technologies, with optical imageries with very high spatial resolutions now being available, which facilitates a wide range of applications such as disaster control, land planning, urban monitoring, and traffic planning [1]–[3]. In these applications, object detection from very high resolution (VHR) remote sensing images has gained increasing interest in recent years, particularly aircraft detection, due to the successful civil and military applications [4], [5]. However, aircraft detection is still a challenging problem because of aircraft appearance variation, the complex and cluttered background, and the multiple resolutions of satellite images [6].

Various aircraft detection methods have been proposed for automatically detecting aircraft in remote sensing images. The most common method is to extract a low-level feature such as the shape feature, context feature, or local image feature [e.g., scale-invariant feature transform (SIFT) and histogram of oriented gradients (HOG)]; construct a bag-of-visual-words representation or dictionary; and run a statistical classifier [7]– [10]. For example, Liu and Shi [11] developed an algorithm which focuses on the airplane feature being rotation invariant and combines sparse coding with radial gradient transform, with a support vector machine (SVM) as the classifier. Yao *et al.* [12] proposed to detect aircraft based on a target-oriented saliency model and discriminative learning of sparse coding. Cheng *et al.* [13] used the HOG feature and a latent SVM to train deformable part-based mixture models for each object category. By heavily relying on the human-labeled training examples and manually designed feature descriptors, the aforementioned supervised learning methods can achieve promising performances, when there is a large amount of training data and the feature representation is efficient [14].

However, with the rapid development of remote sensing technology, a large amount of VHR remote sensing images are now being obtained each day. This brings about several obstacles for the aircraft detection task in remote sensing images. The very high resolution gives more information and details, whereas the traditional methods can only extract the low-level feature representations from the image. The fact that the high-level features and the hierarchical feature representations of the image are ignored leads to a tremendous semantic gap for the object detection task. A lot of work has been focused on the learning of hierarchical internal feature representations from image data sets [15]–[17]. These methods are called "deep learning" methods. Good internal feature representations are hierarchical. In an image, pixels are assembled into edgelets; edgelets are assembled into motifs; motifs are assembled into parts; parts are assembled into objects; and finally, objects are assembled into scenes [18]. Convolutional neural networks (CNNs) are among the most prevalent deep learning methods [19]. Due to the

Manuscript received December 23, 2015; revised March 23, 2016 and April 29, 2016; accepted May 8, 2016. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2012CB719905, by the National Natural Science Foundation of China under Grant 41431175 and Grant 61471274, by the Natural Science Foundation of Hubei Province under Grants 2014CFB193, and by the Fundamental Research Funds for the Central Universities. *(Corresponding author: Miaozhong Xu.)*

F. Zhang, L. Zhang, and M. Xu are with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: rszhang@whu.edu.cn; zlp62@whu.edu.cn; mzxu6319@whu.edu.cn).

B. Du is with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: remoteking@whu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2016.2569141

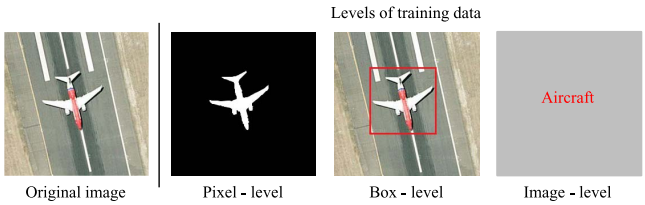


Fig. 1. Different levels of training data for aircraft detection.

recent development of large public image repositories such as ImageNet [20], and high-performance computing systems such as graphics processing units (GPUs) and large-scale distributed clusters [21], CNNs have enjoyed great success in large-scale visual recognition. Extracting the high-level features and the hierarchical feature representations of the objects is, therefore, a promising approach for the object detection task.

On the other hand, supervised-learning-based aircraft detection approaches often require a large amount of training data with manual annotation, such as labeling a bounding box around each object to be detected, or even every pixel, as shown in Fig. 1. However, the manual annotation of objects in large image sets is generally expensive and sometimes unreliable. For example, for objects such as aircraft, the coverage of the target objects is very small, there can be significant appearance variations, the background is often complex and cluttered, and the images can be at multiple scales. As a result, it is difficult to achieve accurate annotation with a small amount of training data. However, training the object detectors with weakly supervised learning for VHR remote sensing images can alleviate the need for human annotation [22]. Han *et al.* [22] proposed an iterative detector training method, in which the detector is iteratively trained using refined annotations until the model converges. Weakly supervised learning means that we only require weak labels, such as image-level labels, for the training images, to indicate whether an image contains the object of interest or not and automatically mine the related data. As a result, weakly supervised learning methods require significantly less annotation effort during training.

Most approaches to object detection build a sliding-window detector, which slides across the whole image [23]. However, the increased size and spatial resolution of remote sensing images makes precise location within the sliding-window paradigm an open problem. As shown in Fig. 2, the traditional method only focuses on a small-scale VHR image data set, using the sliding-window detector to perform the detection task, and ignores the entire airport area image, which may have a dimension of about $10\,000 * 10\,000$ and covers an area of 30–50 km^2. These small-area remote sensing images are often sampled from a large-scale VHR image, and they only have a dimension of about $1000 * 1000$ and cover an area of 1–2 km^2. How to directly detect the aircraft in the large-scale VHR image covering the whole airport area is much more practical. Here, we focus on large-scale VHR image data sets and solve the detection problem by operating a "region proposal method" [24], which has been a successful approach in both natural object detection and semantic segmentation [25]. Region proposal methods first extract some potential regions which may contain the object and then train a classifier to detect the objects in these

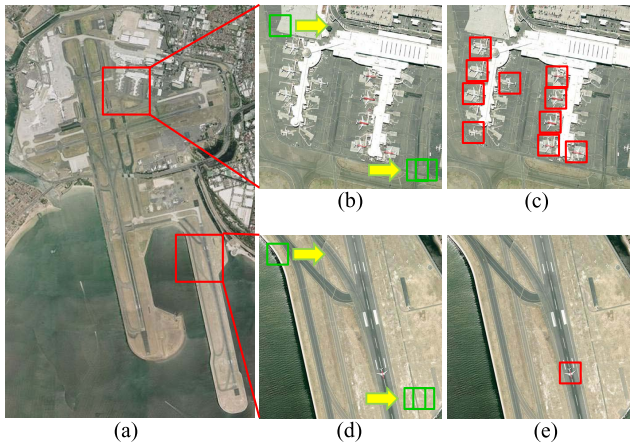


Fig. 2. Traditional method for sliding-window-based aircraft detection. (a) Large-scale VHR image. (b)–(d) Small-scale VHR image detection using the sliding-window detector. (c)–(e) Detection maps.

regions, which dramatically reduces the search area. Selective search (SS) [25], which is one of the most popular methods in natural scene image analysis, greedily merges superpixels based on engineered low-level features. Nevertheless, the SS method cannot be directly used on large-scale VHR images when the image size is very large, such as $10\,000 * 10\,000$ pixels. In [26], the authors proposed a circle-frequency filter (CFF)-based method to extract the airplane candidate regions, which considers the special characteristics of aircraft. Although this method is very fast, the results are noisy and the filter size significantly influences the result with different-size aircraft. How to extract good potential regions is a critical task for accurate aircraft detection in large-scale VHR images.

In this paper, we simultaneously tackle these problems by proposing a weakly supervised learning framework using coupled CNNs for large-scale VHR image detection. To the best of our knowledge, this is the first time that coupled CNNs have been effectively constructed in a weakly supervised learning manner, which is aimed at improving the aircraft detection accuracy. First, we develop a CNN-based method to extract the high-level features and the hierarchical feature representations of the objects. We then employ an iterative weakly supervised learning framework to automatically mine and augment the training data set from the original image, which can significantly reduce the human labor cost. Based on the recent success of pretraining technology in deep learning, unsupervised pretraining can help to prevent overfitting, leading to significantly better generalization when the number of labeled examples is small, or in a transfer setting where we have lots of examples for some "source" tasks, but very few for some "target" tasks [27]. We also explore the weakly supervised learning framework to learn a cross-domain and discriminative feature representation from related tasks such as scene classification, which further reduces the human labor cost, even without any labeled examples. Finally, we propose a coupled CNN model which combines a candidate region proposal network (CRPNet) and a localization network (LOCNet) to extract the proposals and simultaneously locate the aircraft, which is suitable for large-scale VHR images, and only requires an image-level training data set.

The major contributions of this paper are presented as follows.

1) We develop a CNN-based method to extract the high-level features and the hierarchical feature representations of the objects.
2) We propose a weakly supervised learning framework which only needs an image-level training data set and automatically mines and augments the training data set from the original image, which can significantly reduce the human labor cost.
3) We explore the weakly supervised learning framework to learn cross-domain and discriminative feature representations from the related tasks, and we do not require any labeled examples in the detection task.
4) We develop a novel coupled CNN model, which for the first time combines a CRPNet and a LOCNet, to extract the proposals and simultaneously locate the aircraft.

The remainder of this paper is organized as follows. In Section II, we briefly introduce weakly supervised learning and the deep learning method. In Section III, we describe the proposed method in detail. The details of our experiments and the results are presented in Section IV. Finally, Section V concludes this paper with a discussion of the results.

II. Related Work

Here, we briefly introduce weakly supervised learning and the classical deep learning method.

A number of weakly supervised learning approaches have been applied to natural scene image analysis [28]–[30], but these existing methods cannot be directly used in VHR image analysis because they are not able to handle the large-scale complex background and large image size. However, a few efforts [22], [31] have been made to adopt weakly supervised learning for aircraft detection in VHR remote sensing images. In [31], weakly supervised learning was adopted and heuristically combined with saliency-based self-adaptive segmentation, a negative mining algorithm, and a negative evaluation mechanism for target detection in remote sensing images. Han *et al.* [22] proposed a weakly supervised learning framework based on Bayesian principles, for detecting objects from optical remote sensing images, and unsupervised feature learning via deep Boltzmann machines (DBMs), to build a high-level feature representation for various geospatial objects. Although this method generates encouraging results, the feature learning framework ignores the relationship between the different-level features. The low-level features are first extracted using the SIFT descriptor, and the locality-constrained linear coding (LLC) model [32] is then applied to encode the local low-level features into a mid-level feature representation. Finally, a three-layer DBM is adopted to learn the high-level representation. However, all of these methods only concentrate on small-area remote sensing image data sets with a maximum dimension of about $1000 * 1000$, and they cannot directly handle large-area VHR images.

Compared with the method proposed in [22], the deep learning method of CNNs [15], [16] is a trainable multilayer architecture composed of multiple feature extraction stages. Each stage consists of three layers: 1) a convolutional layer; 2) a nonlinear layer; and 3) a pooling layer. The architecture of a CNN is designed to take advantage of the 2-D structure of the input image. A typical CNN is composed of one, two, or three such feature extraction stages, followed by one or more traditional fully connected layers, and a final classifier layer, as shown in Fig. 3. Each layer type is described as follows.

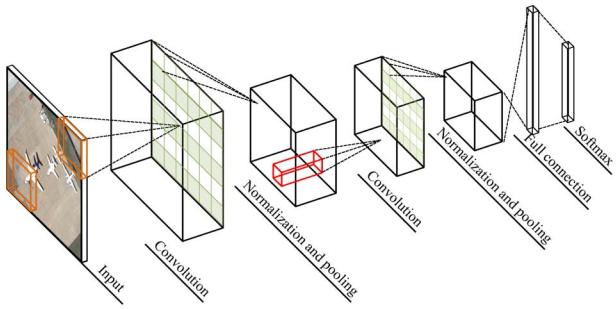


Fig. 3. Typical CNN architecture with two feature extraction stages.

Convolutional Layer: The input to the convolutional layer is a 3-D array with r 2-D feature maps of size $m \times n$. Each component is denoted by $x^i_{m,n}$, and each feature map is denoted by x^i. The output is also a 3-D array $m_1 \times n_1 \times k$, which is composed of k feature maps of size $m_1 \times n_1$. The convolutional layer has k trainable filters of size $c \times c \times q$, which is also called the filter bank W, where c is smaller than the dimension of the image, and q is equal to the number of channels r. The convolutional layer computes the output feature map $z^s = \sum_i^q W_i^s * x^i + b_s$, where $*$ is the 2-D discrete convolution operator, b is a trainable bias parameter, and s indexes the filter number.

Nonlinear Layer: In a traditional CNN, this layer simply consists of a pointwise nonlinear function applied to each component in the feature map. The nonlinear layer computes the output feature map $a^s = f(z)^s$, $f(\cdot)$ is commonly chosen to be a rectified linear unit (ReLU), and $f(x) = \max(0, x)$.

Pooling Layer: The pooling layer involves executing a max operation over the activations within a small spatial region G of each feature map: $p^s_G = \max_{i \in G} a^s_i$. To be more precise, a pooling layer can be thought of as consisting of a grid of pooling units spaced s pixels apart, each summarizing a small spatial region of size $p * p$ centered at the location of the pooling unit.

After the multiple feature extraction stages, the entire network is trained with the backpropagation [20] of a supervised loss function such as the classic cross entropy of a softmax classifier output $\hat{y}_i = \text{soft}\max(a_i) = e^{a_i} / \sum_j^{N_c} e^{a_j}$; $\hat{y}_i$ is the activations of the previous layer node i pushed through a softmax function. The target output y is represented as a 1-of-K vector, where K is the number of outputs, and L is the number of layers, as follows:

$$J(W, b) = -\sum_i^{N_c} y_i \log(\hat{y}_i) + \lambda \sum_l^L \text{sum}\left(\left\|W^{(l)}\right\|^2\right) \quad (1)$$

where $\hat{y}_i$ is the activations of the previous layer, λ is a regularization term (also called a weight decay term), and l indexes

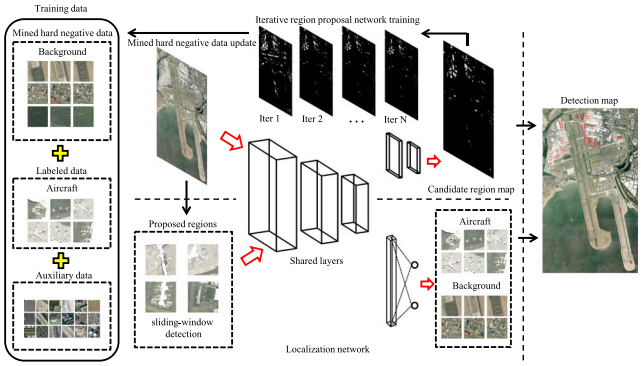


Fig. 4. Overall network architecture of the coupled CNN model.

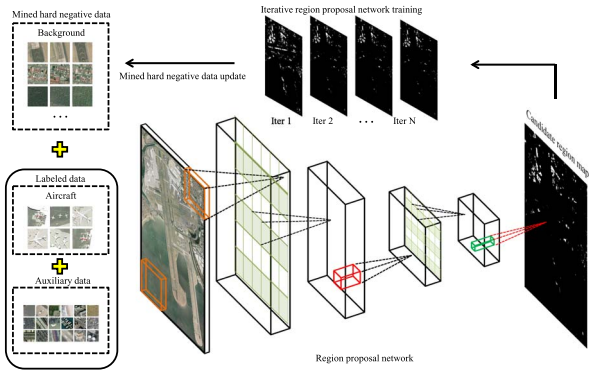


Fig. 5. Overall network architecture of the CRPNet.

the layer number. Our goal is to minimize $J(W, b)$ as a function of W and b. To train the CNN, we apply stochastic gradient descent, with backpropagation to optimize the function [21].

III. Overview of the Proposed Method

The proposed object detection method consists of two major components: the coupled CNN model and the weakly supervised learning framework. The flowchart of the object detection framework is shown in Fig. 4. The coupled CNN model consists of a CRPNet and a LOCNet, which share the same feature extraction stage. As mentioned earlier, due to the huge size of satellite images, we use the CRPNet to extract the candidate regions from the image, which can dramatically reduce the search area. The LOCNet is then used to extract the high-level features and locate the final aircraft positions.

Each component is described as follows.

Based on the coupled CNN model, the training of the weakly supervised learning framework consists of two steps: training data initialization and iterative network training. The training data consist of three parts: mined hard negative data, labeled data, and auxiliary data. As mentioned earlier, due to the huge size and complex background, it is difficult to manually label training data in a VHR image. Since we only use an image-level label, collecting the different scene classes from correlated tasks such as scene classification is very efficient [33], [34]. We use these correlated data as auxiliary data with image-level labels to pretrain our networks and to add to the training data set, which helps to learn a cross-domain and discriminative feature representation covering the different background classes.

Given the auxiliary data, we only need to label the aircraft images to construct our positive training data. We then use the auxiliary data as negative training data and the labeled positive training data, to train our coupled CNN model iteratively. During each iteration, the CRPNet generates a candidate region map and automatically mines the negative training data from the original image to construct the background data set and update the training data set, to retrain the networks, which can effectively suppress the background classes.

A. CRPNet

A CRPNet takes an image (of any size) as input and outputs a map of "objectness probability," where each pixel denotes a rectangular region in the original image. We model this process as a fully convolutional neural network [35], which is described as follows.

1) Overall Architecture: Typical CNNs, including LeNet [15], AlexNet [20], and its deeper successors [21], [24], ostensibly take fixed-size inputs and produce nonspatial outputs. The fully connected layers of these networks have fixed dimensions and discard the spatial coordinates. However, for the detection task, the size of the input image is not fixed. In order to handle the different sizes of image, these fully connected layers can be viewed as convolutions with kernels that cover their entire input regions and changed into convolutional layers. Doing so casts them as a fully convolutional neural network that takes an input of any size and outputs a classification map. As illustrated in Fig. 5, the CRPNet is a fully convolutional neural network which only contains convolutional layers.

Although the CRPNet only contains convolutional layers, the training procedure is generally the same as for a typical CNN. To train the CRPNet, we assign a binary class label to each item of the training data. During each iteration, we feed a batch of training data and their labels to the network, to update the parameters. At testing time, given the trained CRPNet and an input large-scale satellite image, we generate the candidate region map, as shown in Fig. 5.

2) Pretraining: For the detection task, labeled data are often very difficult and expensive to obtain. Using an auxiliary data set to "pre-train" the network has recently been attracting a lot of attention [27]. For smaller data sets, pretraining technology in deep learning helps to prevent overfitting, leading to significantly better generalization when the number of labeled examples is small, and helps us to learn a cross-domain and discriminative feature representation covering the different background classes.

In this paper, we used the University of California (UC) Merced data set [9] as auxiliary data to pretrain the network. Fig. 6 shows a few example images representing various aerial scenes that are included in this data set. The images have a resolution of one foot per pixel and are of 256 $\times$ 256 pixels. For the proposed method, we resized the images to 128 $\times$ 128 pixels. The data set contains 21 challenging scene categories, with 100 image samples per class. The data set represents various classes such as agricultural, buildings, and overpass, which cover most of the typical background classes for aircraft detection.

Fig. 6. Example images associated with the 21 land-use categories in the UC Merced data set: (1) agricultural; (2) airplane; (3) baseball diamond; (4) beach; (5) buildings; (6) chaparral; (7) dense residential; (8) forest; (9) freeway; (10) golf course; (11) harbor; (12) intersection; (13) medium residential; (14) mobile home park; (15) overpass; (16) parking lot; (17) river; (18) runway; (19) sparse residential; (20) storage tanks; and (21) tennis court.

3) Hard Negative Mining: Using auxiliary data is an efficient and easy way to collect the background classes for aircraft detection. Due to the complex composition and cluttered background in VHR images, it is also important to collect negative examples which cover the different background classes in the original image. When training a model for object detection, we usually employ an iterative process by which hard negatives are automatically collected for retraining the model.

In this paper, we develop a data-mining process motivated by the idea of iteratively training the CRPNet. As shown in Fig. 5, the proposed method first trains an initial CRPNet using the negative training data and the labeled positive training data. We then use the output map of the CRPNet to collect the negative data which have a high score in the background classes. In each iteration, we add new hard examples to the negative training data which have high scores. The process eventually converges when the number of negative samples generated is below a certain threshold.

4) Candidate Region Generation: When the training of the CRPNet is finished, we use the candidate region map to generate candidate regions for the aircraft detection. A candidate region map is a map of objectness probability, and each pixel denotes a rectangular region in the original image. We denote a probability of > 0.5 as a candidate region and a probability of < 0.5 as a background one. We can also view this as a binary classification problem. If the pixel has a probability of > 0.5, then we extract the region centered at this pixel with a size of $196 * 196$, which is slightly larger than the training image size. Given all the candidate regions, we apply greedy nonmaximum suppression [21], which will reject a region if it has an intersection-over-union overlap with a higher scoring selected region larger than 0.75.

B. LOCNet

A LOCNet takes a candidate region as the input and uses greedy sliding-window searching to locate the position of the aircraft and output the probability. We model this process with a traditional CNN.

1) Multiscale Training: In general, the actual object size in an image will be unknown, and objects appear at a continuous range of scales. This constitutes a significant mode of intraclass variability in detection problems [36]. The dominant perspective suggests that we should learn scale-invariant representations from the data set. One possible solution would be to run multiple parallel networks for different image scales, and to average their probability. However, training a network is difficult and time consuming. Here, we opt for a different less memory-intensive solution. Instead, we train a single LOCNet by rescaling the input image to multiple different sizes, and we extract image patches with different spatial resolutions.

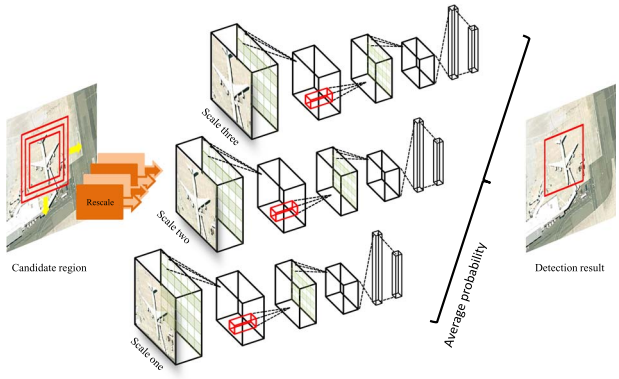


Fig. 7. Overall network architecture of the LOCNet.

In detail, all the images are first resized by a scale uniformly sampled between 0.75 and 1.25. Each training image is then randomly sampled from the resized images with a fixed size to construct the training minibatch. This allows the network to see the objects in the image at various scales. In addition, this type of multiscale training also induces some scale invariance in the network.

2) Sharing Convolutional Features for the CRPNet and LOCNet: Both the CRPNet and the LOCNet can be trained independently. Note that, rather than learning two separate networks, sharing convolutional layers between the two networks can reduce the training time and speed up the convergence. We use alternating optimization to learn shared features between the CRPNet and the LOCNet.

In the first step, we train the CRPNet as described earlier. This network is pretrained with the UC Merced data set and fine-tuned using the labeled data and auxiliary data. In the second step, we train a separate LOCNet using the updated data set generated by the CRPNet. The convolutional layers of the LOCNet are initialized by the previous CRPNet. In the third step, we use the LOCNet to initialize the convolutional layers of the CRPNet, and we train it with the updated data set. As such, both networks share the same convolutional layers and form the coupled CNN model.

3) Detection: At the localization step, we apply the sliding-window procedure at multiple finely sampled scales, as shown in Fig. 7. In detail, we extract subimages with different window sizes [such as $128*128$, $(128 * 0.75) * (128 * 0.75)$, and $(128 * 1.25) * (128 * 1.25)$] from the candidate region. We then rescale the subimages to a fixed image size and input them into the three LOCNets, which have the same parameters. Rescaling the image at large scales allows the networks to find even very small objects. For each scale, the aircraft and

TABLE I
INFORMATION ABOUT THE THREE DATA SETS

Data set	Image size (pixel)	Spatial resolution (m)	Aircraft size (pixels)	Aircraft number	Cover area (km^2)
Sydney International Airport	4992*8256	1.0	300~3500	46	41.2
Tokyo Haneda Airport	6528*7488	1.0	220~3800	65	48.75
Berlin Tegel Airport	8160*3456	1.0	200~3000	31	28.5

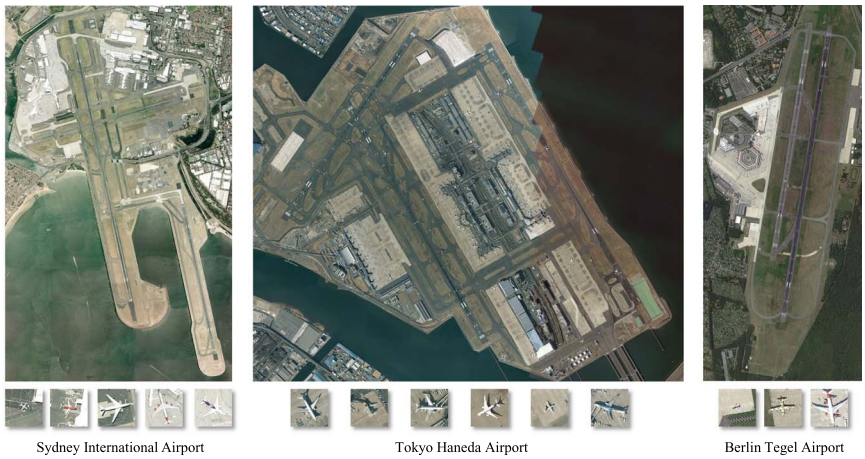


Fig. 8. Example images associated with the Sydney International Airport, Tokyo Haneda Airport, and Berlin Tegel Airport data sets.

TABLE II
DIFFERENT NET ARCHITECTURES

Net configurations			
Net-A	Net-B	Net-C	Net-D
Input image			
Conv3*3-96	Conv3*3-96	Conv3*3-128	Conv3*3-128
Maxpool 9*9-5	Maxpool 5*5-3	Maxpool 5*5-3	Maxpool 5*5-3
Conv3*3-128	Conv3*3-128	Conv3*3-192	Conv3*3-192
Maxpool 3*3-2	Maxpool 3*3-2	Maxpool 3*3-2	Maxpool 3*3-2
	Conv3*3-128	Conv3*3-192	Conv3*3-192 Conv3*3-192
	Maxpool 2*2-2	Maxpool 2*2-2	Maxpool 2*2-2
CRPNet : Conv8*8-420 LOCNet: Full-connect-420			
CRPNet : Conv1*1-10 LOCNet: Full-connect-10			
Softmax			

background scores are computed and averaged into single scores. The aircraft can then be identified by choosing a probability threshold τ. If the probability is > 0.5, then we consider the object to be a detected aircraft.

IV. EXPERIMENTS AND ANALYSIS

Here, we first describe the data sets used for the experiments and then the parameter settings of the proposed method. The results obtained for the aircraft detection of three benchmark high-resolution satellite images are then discussed.

A. Description of the Data Sets and Parameter Settings

Three data sets were used in the experiments. All of the data sets were constructed from large-scale satellite images which were acquired from Google Earth. The details of these data sets are shown in Table I. Fig. 8 shows the three data sets and some image samples. As can be observed, the targets in the different data sets have various sizes, orientations, and colors. The three different airport areas also have a complex composition and cluttered background classes.

For these three data sets, we trained the proposed method using stochastic gradient descent, with a batch size of 12, a momentum of 0.9, a weight decay of 0.0005, and a learning rate of 0.001. For the hard negative mining procedure, we set a negative score of > 0.9 as a negative sample. At each iteration, when the new generated negative samples were less than 5% of the whole training data set, we considered the process to have converged.

For the Sydney International Airport and Tokyo Haneda Airport data sets, we randomly selected ten aircraft as the labeled training data, and we set the remaining aircraft for testing. In order to illustrate the generalization ability of the proposed method, for the Berlin Tegel Airport data set, we did not select any labeled training data, and we directly used the previous trained coupled CNNs. We can view this as an offline detection experiment, which did not need any labeled training data.

TABLE III
COMPARISON OF THE DIFFERENT CNN ARCHITECTURES

Dataset	Method	CRPNet-A	CRPNet-B	CRPNet-C	CRPNet-D
Sydney Airport	FPR	70.30%(277/394)	71.31%(261/366)	**60.00%(153/255)**	63.55%(190/299)
	MR	4.35%(2/46)	**2.17%(1/46)**	**2.17%(1/46)**	**2.17%(1/46)**
Tokyo Airport	FPR	48.61%(157/323)	36.23%(96/265)	**31.98%(79/247)**	32.30%(83/257)
	MR	**0.00%(0/65)**	**0.00%(0/65)**	**0.00%(0/65)**	**0.00%(0/65)**
Berlin Airport (off-line)	FPR	40.55%(37/91)	52.05%(89/171)	**18.07%(15/83)**	26.17%(28/107)
	MR	16.13%(5/31)	**3.23%(1/31)**	**3.23%(1/31)**	**3.23%(1/31)**

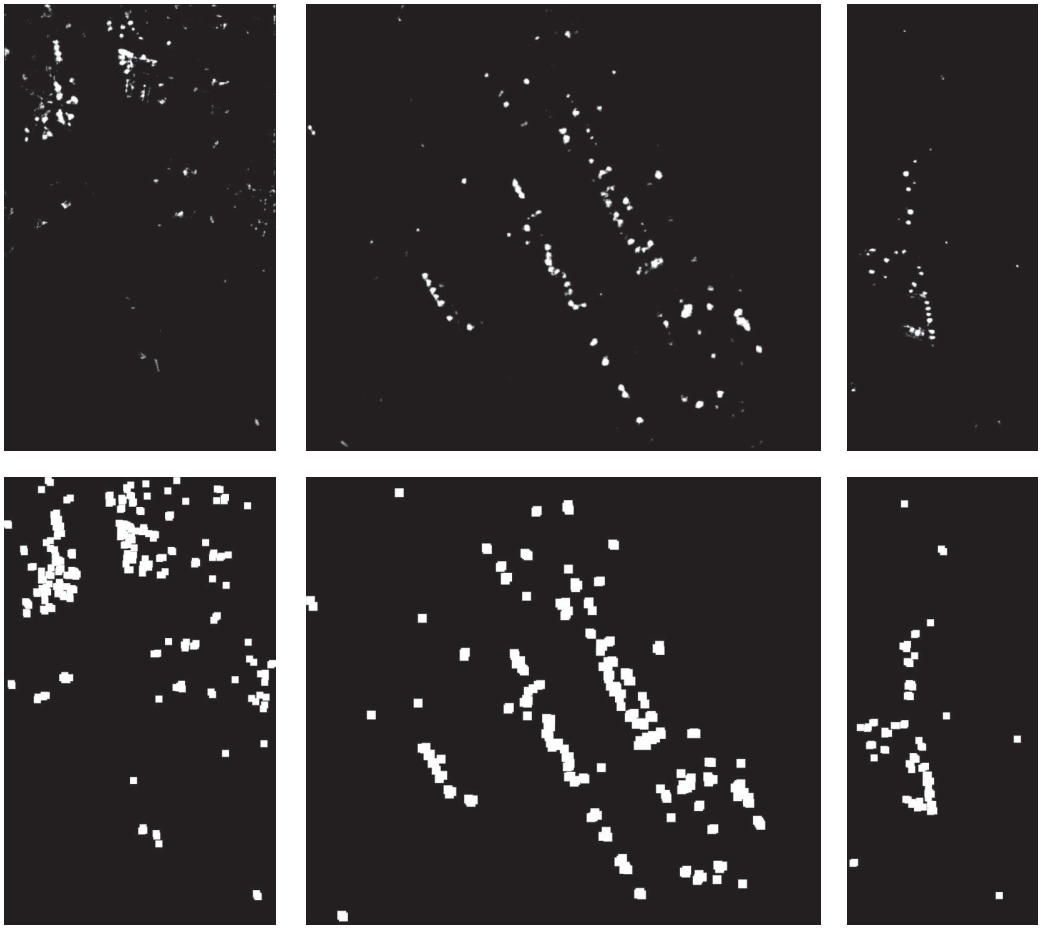


Fig. 9. (Top) Candidate region map for the three data sets. (Bottom) Generated candidate regions for the three data sets.

The experiments were run on a PC with a single Intel core i7 CPU, an NVIDIA Titan GPU, and 6 GB of memory. The operating system was Windows 7, and the implementation environment was under MATLAB 2014a with CUDA kernels.

B. Evaluation of the Region Proposal Performance

To measure the region proposal performance, we first compared the proposal accuracy with CRPNets of different architectures. The different network architectures are outlined in Table II, one per column. The convolutional layer parameters are denoted as "Conv(receptive field size)-(number of features)." The pooling layer parameters are denoted as "Maxpool(pooling region size)-(pooling stride)."

For the region proposal performance, two commonly used criteria were computed, i.e., the false positive rate (FPR) and the missing ratio (MR), which are defined as follows:

$$\text{FPR} = \frac{\text{Number of falsely proposed regions}}{\text{Number of proposed regions}} * 100\%$$

$$\text{MR} = \frac{\text{Number of missing aircraft}}{\text{Number of aircraft}} * 100\%.$$

TABLE IV
COMPARISON OF THE DIFFERENT METHODS

Dataset	Method	CRPNet-C	CFF
Sydney Airport	FPR	**60.00%(153/255)**	96.83%(2537/2620)
	MR	**2.17%(1/46)**	10.87%(5/46)
Tokyo Airport	FPR	**31.98%(79/247)**	91.92%(1629/1786)
	MR	**0.00%(0/65)**	0.00%(0/65)
Berlin Airport (off-line)	FPR	**18.07%(15/83)**	95.94%(1820/1897)
	MR	3.23%(1/31)	**0.00%(1/31)**

TABLE V
COMPARISON OF THE DIFFERENT CNN ARCHITECTURES

Dataset	Method	LOCNet-A	LOCNet-B	LOCNet-C	LOCNet-D
Sydney Airport	FPR	55.63%(173/312)	48.47%(111/230)	48.54%(100/207)	**32.98%(62/189)**
	MR	**10.87%(5/46)**	**10.87%(5/46)**	**10.87%(5/46)**	**10.87%(5/46)**
	AC	**89.13%(41/46)**	**89.13%(41/46)**	**89.13%(41/46)**	**89.13%(41/46)**
	ER	66.50%	59.34%	59.41%	**43.85%**
Tokyo Airport	FPR	27.61%(74/269)	25.21%(61/243)	20.80%(47/227)	**17.26%(39/227)**
	MR	**1.54%(1/65)**	**1.54%(1/65)**	**1.54%(1/65)**	3.08%(2/65)
	AC	**98.46%(64/65)**	**98.46%(64/65)**	**98.46%(64/65)**	96.92%(63/65)
	ER	29.15%	26.75%	22.33%	**20.33%**
Berlin Airport (off-line)	FPR	20.00%(13/66)	37.89%(61/161)	7.69%(6/79)	**4.60%(4/88)**
	MR	29.03%(9/31)	3.23%(1/31)	**3.23%(1/31)**	16.13%(5/31)
	AC	70.97%(22/31)	96.77%(30/31)	**96.77%(30/31)**	83.87%(26/31)
	ER	49.03%	41.12%	**10.92%**	20.27%

TABLE VI
COMPARISON OF THE DIFFERENT METHODS

Dataset	Method	LOCNet-C	SPMK	UFL
Sydney Airport	FPR	48.54%(100/207)	**34.33%(23/67)**	45.24%(19/43)
	MR	**10.87%(5/46)**	43.48%(20/46)	65.22%(30/46)
	AC	**89.13%(41/46)**	56.52%(26/46)	34.78%(16/46)
	ER	**59.41%**	77.81%	110.46%
Tokyo Airport	FPR	20.80%(47/227)	**12.38%(13/106)**	16.92%(11/65)
	MR	**1.54%(1/65)**	20.00%(13/65)	29.23%(19/65)
	AC	**98.46%(64/65)**	80.00%(52/65)	70.77%(46/65)
	ER	**22.33%**	32.38%	46.15%
Berlin Airport (off-line)	FPR	**7.69%(6/79)**	17.31%(9/53)	9.09%(3/33)
	MR	**3.23%(1/31)**	22.58(7/31)	22.58(7/31)
	AC	**96.77%(30/31)**	77.42%(24/31)	77.42%(24/31)
	ER	**10.92%**	39.89%	31.67%

For every candidate region, if it had an intersection overlap with a test aircraft of greater than 0.5, we considered it to be a true region. For every candidate region, if a test aircraft was not found in the candidate region, we considered it to be a missing aircraft.

Table III shows the performance with the different sizes of feature extraction architecture. In Table III, we can observe that CRPNet-C using three feature extraction stages achieves the best FPR and MR accuracy values. The CRPNet can therefore benefit from the increased depth of network. However, with the increasing depth of network to the four-layer CRPNet-D, the FPR and MR values decrease slightly where the deep architecture has more parameters to train, and the limitation of the training samples restricts the performance of the deep architecture.

To visualize the extracted candidate regions, we show the candidate map generated from CRPNet-C in Fig. 9. Here, it is shown that the proposed method dramatically reduces the number of search regions and maintains a very high proposal accuracy.

Finally, we also compared the region proposal performance of the CRPNet with the CFF method [26]. A comparison of the results of the CRPNet and CFF methods is shown in Table IV, where it is shown that the CRPNet obtains lower FPR and MR values.

C. Evaluation of the Aircraft Detection

To measure the detection performance of the proposed LOCNet, we compared the detection accuracy with different architectures, as shown in Table II, and with the two supervised methods of spatial pyramid matching kernel (SPMK) [8] and the unsupervised feature learning method (UFL) [23], [37]. For a fair comparison, we used the auxiliary data, labeled data, and mined hard negative data that were generated by the CRPNet to train the SPMK and UFL methods. All these methods were evaluated in the candidate regions proposed by the CRPNet.

For the aircraft detection performance, four commonly used criteria were computed: FPR, MR, accuracy (AC), and error ratio (ER). These criteria are defined as follows:

$$\text{FPR} = \frac{\text{Number of falsely detected aircraft}}{\text{Number of detected aircraft}} * 100\%$$

$$\text{MR} = \frac{\text{Number of missing aircraft}}{\text{Number of aircraft}} * 100\%$$

$$\text{AC} = \frac{\text{Number of detected aircraft}}{\text{Number of aircraft}} * 100\%$$

$$\text{ER} = \text{FPR} + \text{MR}.$$

For every detected aircraft, if it had an intersection overlap with a test aircraft of greater than 0.5, we considered it to be a true detected aircraft. For every detected aircraft, if a test aircraft was not found within the detected aircraft with an intersection overlap of greater than 0.5, we considered it to be a missing aircraft.

We first compared the detection accuracy with LOCNets of different architectures. Table V shows the detection performance of the LOCNet method with different architectures. A similar conclusion can be made for these three data sets, in that

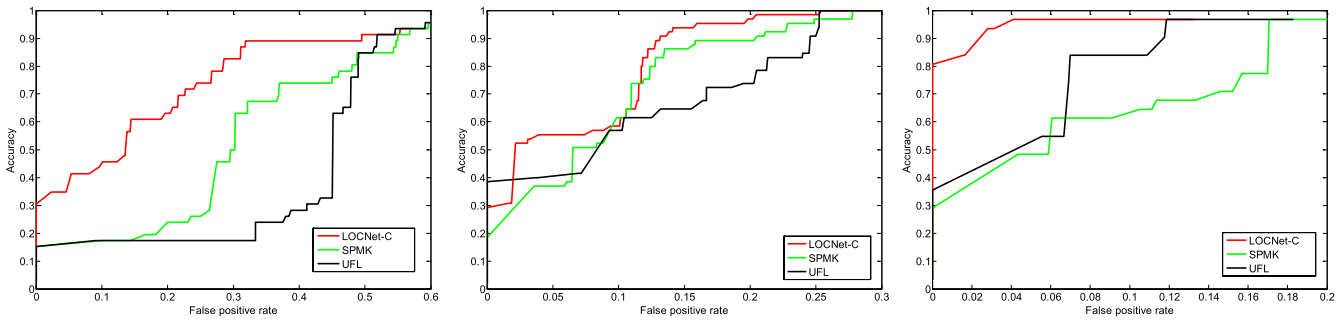


Fig. 10. AC/FPR curves for the comparison of the three methods. (a) Sydney Airport. (b) Tokyo Airport. (c) Berlin Airport.

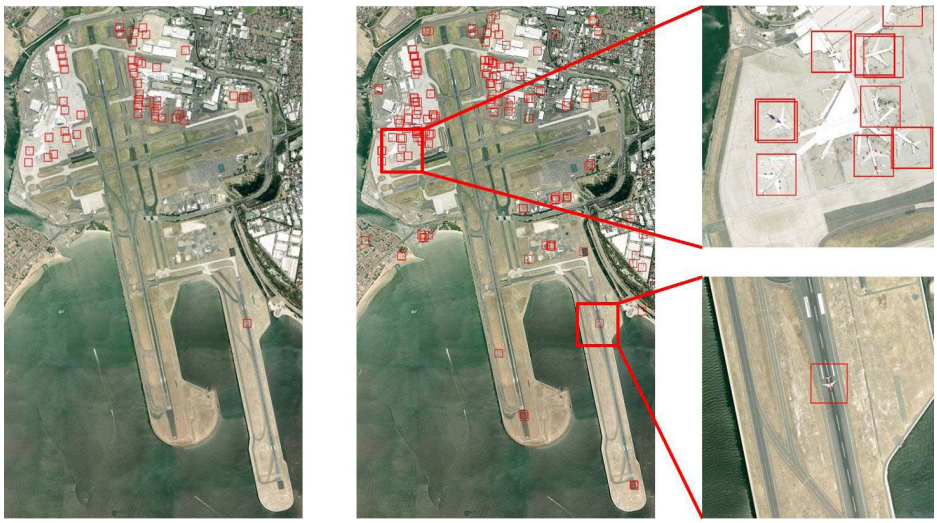


Fig. 11. (Left) Aircraft ground truth. (Right) Detection map from LOCNet-C for Sydney Airport.

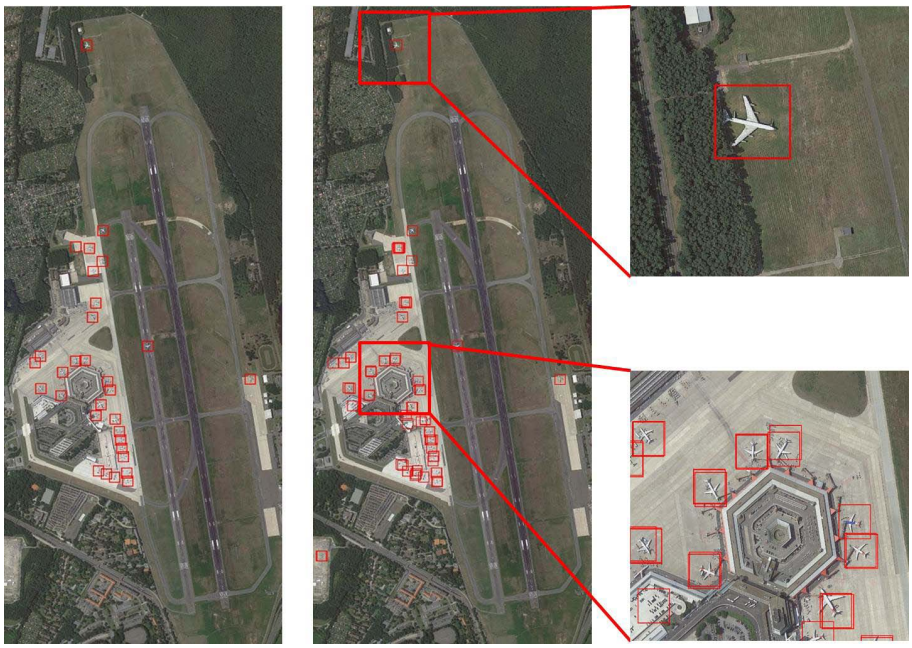
Fig. 12. (Left) Aircraft ground truth. (Right) Detection map from LOCNet-C for Tokyo Airport.

the LOCNet benefits from an increased depth of network, but increasing the depth of the network to the four-layer LOCNet-D results in a slightly decreased accuracy.

We then compared the detection performance of LOCNet-C with SPMK [8] and the UFL method in [23]. We compared the reported detection performances for the three airport data sets. Of the three strategies that we compared, LOCNet-C produces the best performance, as shown in Table VI. For the quantitative evaluation, we also plotted the AC/FPR curves of the object detection results, as shown in Fig. 10. Specifically, the AC/FPR curves were plotted based on the AC and FPR values under different values of the probability threshold τ. As shown in Fig. 10, the proposed LOCNet-C always outperforms the other methods, which confirms that extracting high-level features and the hierarchical feature representations of the objects is useful and efficient. To visualize the detection performance, we show the detection maps from LOCNet-C in Figs. 11–13.

V. Conclusion

In this paper, we have proposed a weakly supervised learning framework based on coupled CNNs, which combines a CRPNet



Aircraft ground truth Detection map

Fig. 13. (Left) Aircraft ground truth. (Right) Detection map from LOCNet-C for Berlin Airport (offline).

and a LOCNet to extract the proposals and to simultaneously locate the aircraft, using only an image-level training data set. The proposed framework presented a convincing performance with three challenging high-resolution data sets. The experiments showed that: 1) the weakly supervised learning framework is a promising and efficient way to alleviate the human labor cost of annotation and automatically collect training data; 2) the proposed coupled CNNs can generate more accurate results than the traditional methods.

REFERENCES

[1] M. A. Hossain, X. Jia, and M. Pickering, "Subspace detection using a mutual information measure for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 2, pp. 424–428, Feb. 2014.

[2] G. Zhang, X. Jia, and J. Hu, "Superpixel-based graphical model for remote sensing image mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 5861–5871, Nov. 2015.

[3] S. Aksoy, I. Z. Yalniz, and K. Tasdemir, "Automatic detection and segmentation of orchards using very high resolution imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 8, pp. 3117–3131, Aug. 2012.

[4] X. Bai, H. Zhang, and J. Zhou, "VHR object detection based on structural feature extraction and query expansion," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6508–6520, Oct. 2014.

[5] Y. Yu, H. Guan, and Z. Ji, "Rotation-invariant object detection in high-resolution satellite imagery using superpixel-based deep Hough forests," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2183–2187, Nov. 2015.

[6] Z. Lei, T. Fang, H. Huo, and D. Li, "Rotation-invariant object detection of remotely sensed images based on Texton forest and Hough voting," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1206–1217, Apr. 2012.

[7] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 13–16, 2003, vol. 2, pp. 1470–1477.

[8] 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.*, 2006, vol. 2, pp. 2169–2178.

[9] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM GIS*, 2010, pp. 270–279.

[10] F. Hu *et al.*, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2015–2030, May 2015.

[11] L. Liu and Z. Shi, "Airplane detection based on rotation invariant and sparse coding in remote sensing images," *Optik*, vol. 125, no. 18, pp. 5327–5333, 2014.

[12] X. Yao, J. Han, and L. Guo, "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.

[13] G. Cheng *et al.*, "Object detection in remote sensing imagery using a discriminatively trained mixture model," *ISPRS J. Photogramm. Remote Sens.*, vol. 85, pp. 32–43, Nov. 2013.

[14] B. Zhao, Y. Zhong, G. S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.

[15] Y. Le Cun *et al.*, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information*. San Francisco, CA, USA: Morgan Kaufmann, 1990.

[16] Y. Le Cun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[17] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.

[18] Y. Le Cun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. IEEE ISCAS*, Jun. 2010, pp. 253–256.

[19] J. Dean *et al.*, "Large scale distributed deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1232–1240.

[20] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25, pp. 1106–1114.

[21] P. Sermanet *et al.*, "OverFeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. ICLR*, 2014, pp. 1–15.

[22] J. Han *et al.*, "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.

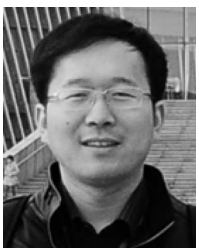
[23] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.

[24] S. Ren *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[25] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.
[26] Z. An, Z. Shi, and X. Teng, "An automated airplane detection system for large panchromatic image with high spatial resolution," *Optik*, vol. 125, no. 12, pp. 2768–2775, Jun. 2014.
[27] Y. Le Cun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, pp 436–444, 2015.
[28] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.
[29] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *Int. J. Comput. Vis.*, vol. 100, no. 3, pp. 275–293, Dec. 2012.
[30] P. Siva, C. Russell, and T. Xiang, "In defense of negative mining for annotating weakly labeled data," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 594–608.
[31] D. Zhang *et al.*, "Weakly supervised learning for target detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 701–705, Apr. 2015.
[32] J. Wang *et al.*, "Locality-constrained linear coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3360–3367.
[33] R. Girshick, "*Fast R-CNN*," in *Proc. ICCV*, 2015, pp. 1–9.
[34] N. Wang, S. Li, and A. Gupta, "Transferring rich feature hierarchies for robust visual tracking," unpublished paper. [Online]. Available: http://arxiv.org/abs/1501.04587
[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3431–3440.
[36] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 241–254.
[37] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

Fan Zhang received the B.S. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2012. He is currently working toward the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University.

His research interests include high-resolution image and hyperspectral image classification, machine learning, and computation vision in remote sensing applications.



Bo Du (M'10–SM'15) received the B.S. degree and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2005 and 2010, respectively.

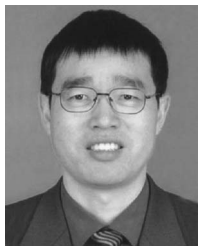
He is currently a Professor with the School of Computer Science, Wuhan University. He has more than 40 research papers published in the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE JOURNAL OF SELECTED TOPICS IN EARTH OBSERVATIONS AND APPLIED REMOTE SENSING (JSTARS), the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL), etc. His main research interests include pattern recognition, hyperspectral image processing, and signal processing.

Dr. Du was a recipient of the Best Reviewer Awards from the *IEEE Geoscience and Remote Sensing Society* for his service to the IEEE JSTARS in 2011 and the ACM Rising Star Awards for his academic progress in 2015. He was the Session Chair for the Fourth IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing. He also serves as a reviewer of 20 Science Citation Index magazines, including IEEE TGRS, TIP, JSTARS, and GRSL.

Liangpei Zhang (M'06–SM'08) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He is currently the Head of the Remote Sensing Division, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. He is also a Chang-Jiang Scholar Chair Professor appointed by the Ministry of Education of China. He is currently a Principal Scientist for the China State Key Basic Research Project (2011–2016) appointed by the Ministry of National Science and Technology of China to lead the remote sensing program in China. He has more than 410 research papers. He is the holder of 15 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a Fellow of the Institution of Engineering and Technology, an Executive Member (Board of Governors) of the China National Committee of the International Geosphere-Biosphere Programme, and an Executive Member of the China Society of Image and Graphics. He was a recipient of the 2010 Best Paper Boeing Award and the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing. He regularly serves as a Cochair of the series of SPIE Conferences on Multispectral Image Processing and Pattern Recognition, the Conference on Asia Remote Sensing, and many other conferences. He edits several conference proceedings, issues, and geoinformatics symposiums. He also serves as an Associate Editor for the *International Journal of Ambient Computing and Intelligence*, the *International Journal of Image and Graphics*, the *International Journal of Digital Multimedia Broadcasting*, the *Journal of Geo-Spatial Information Science*, the *Journal of Remote Sensing*, and the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.



Miaozhong Xu received the B.S. degree from the Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1989 and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2004.

He is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. His main research interests include high-resolution image processing, photogrammetry, and cartography.