IEEE Access
Multidisciplinary Rapid Review Open Access Journal

Received November 6, 2019, accepted December 9, 2019, date of publication December 19, 2019, date of current version January 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2960931

Rotation-Invariant Feature Learning for Object Detection in VHR Optical Remote Sensing Images by Double-Net

ZHI ZHANG[1,3], RUOQIAO JIANG[2], SHAOHUI MEI[2], SHUN ZHANG[2], AND YIFAN ZHANG[2]

[1] State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100101, China
[2] School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China
[3] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China

Corresponding author: Shaohui Mei (meish@nwpu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671383 and Grant 61301235, in part by the Fundamental Research Funds for the Central Universities under Grant 3102018AX001, in part by the Natural Science Foundation of Shaanxi Province under Grant 2018JM6005, and in part by the China Postdoctoral Science Foundation under Grant 2014M550872.

ABSTRACT Rotation-invariant feature extraction is crucial for object detection in Very High Resolution (VHR) optical remote sensing images. Although convolutional neural networks (CNNs) are good at extracting the translation-invariant features and have been widely applied in computer vision, it is still a challenging problem for CNNs to extract rotation-invariant features in VHR optical remote sensing images. In this paper we present a novel Double-Net with sample pairs from the same class as inputs to improve the performance of object detection and classification in VHR optical remote sensing images. Specifically, the proposed Double-Net contains multiple channels of CNNs in which each channel refers to a specific rotation direction and all CNNs share identical weights. Based on the output features of all channels, multiple instance learning algorithm is employed to extract the final rotation-invariant features. Experimental results on two publicly available benchmark datasets, namely Mnist-rot-12K and NWPU VHR-10, demonstrate that the presented Double-Net outperforms existing approaches on the performance of rotation-invariant feature extraction, which is especially effective under the situation of small training samples.

INDEX TERMS Convolution neural network (CNN), feature learning, object detection, rotation-invariant, very high resolution (VHR).

I. INTRODUCTION

Extracting rotation-invariant features is crucial for object detection in very high resolution (VHR) optical remote sensing images. VHR optical remote sensing images are captured at high altitude, and they usually contain sophisticated challenges such as rotation variations, geometric transformation changes, scale changes, illumination changes, etc. Consequently, object detection and classification in VHR optical remote sensing images is significantly different from that of nature image dataset (like ImageNet [1], CIFAR-10) which rarely considers rotation variations.

To address the rotation-invariant problem, most of conventional machine learning approaches employ handcrafted or shallow-learning-based features [2], [3]. One of the most famous handcrafted feature is Scale-invariant feature transform (SIFT) [4], [5] which has been widely applied in image matching. SIFT is able to extract rotation-invariant features and transformation-invariant features. However, such a local feature descriptor fails to represent the whole object in object detection and classification tasks. Histogram of oriented gradient (HOG) feature [6] has also widely used in remote sensing image classification and object detection. The core idea of HOG feature is to capture the edge or local shape information of the objects by using edge direction histogram of cell units of small connected regions in an image. The combination of HOG feature and SVM classifier has shown great success in object detection task. HOG has also been extended to rotation-invariant version and applied to object detection in VHR optical remote sensing images [7]. These handcrafted or shallow-learning-based features, such as SIFT, HOG, and bag-of-words (BoW) [8], have gained

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

satisfying performance in remote sensing image classification and object detection due to their simplicity, efficiency, and invariance under viewpoint changes and background clutter. However, they expose deficiencies in description capability and generalization capability when the application scene is becoming more complex.

The research of feature invariance by neural network (NN) can date back to 1990s. Yoshitaka *et al.* [9] used image preprocessing methods to obtain rotation-invariant and transformation-invariant features. They adopted a three layer feed-forward network with back-propagation for learning and recognition, in which data augmentation is used to obtain more training samples. This idea motivates recent studies on invariant feature learning by deep NN (DNN) a lot. Recently, after the AlexNet model [10] showing powerful feature representation capability and generalization capability in nature scene images, many researches have been carried out based on Convolutional Neural Networks (CNNs) to extract both rotation-invariant and transformation-invariant features. For example, Cheng *et al.* [11] modified the architecture of AlexNet [10] by replacing 'FC8' layer with a new rotation-invariant layer, which averages the features extracted from different angles of the same training sample to obtain a rotation-invariant feature. The TI-Pooling (transformation-invariant pooling operator) takes advantage of the tiny rotation invariance of max-pooling, which uses many angles of input sample to expand the rotation invariance of max-pooling [12]. Dielemanet *et al.* [13] presented a deep neural network for galaxy morphology classification by exploiting translational and rotational symmetry. The Deep-HiTS [14] builded a rotation-invariant convolutional neural network for classifying images of transient candidates into artifacts or real sources for the High Cadence Transient Survey (HiTS).

Though CNN-based approaches have achieved success in rotation-invariant and transformation-invariant feature extraction, the size of objects in VHR remote sensing images is much smaller and their background is more complex than nature images. Moreover, the labeled samples are often rare for VHR remote sensing images, especially for that acquired over aviation or space platforms. Therefore, more powerful feature extraction algorithms are required for object detection in VHR remote sensing images. In this paper, a novel Double-Net with sample pairs from the same class as inputs is proposed to improve the performance of object detection and classification in VHR optical remote sensing images. Specifically, the proposed Double-Net contains multiple channels of CNNs in which each channel refers to a specific rotation direction and all CNNs share identical weights. Based on the output features of all channels, multiple instance learning algorithm is employed to extract the final rotation-invariant features. Moreover, a new objective function is designed to train the proposed Double-Net with the within-class distance and softmax loss. We evaluate the proposed method on two publicly available benchmark datasets: Mnist-rot-12K and NWPU VHR-10, to validate rotation-invariant performance by comparing it with state-of-the-art rotation-invariant algorithms, including RICNN [11], TI-Pooling [12] and traditional CNN features. This work is an extension of our preliminary work published in [15]. Here, we improve the rotation-invariant feature extraction algorithm by designing new loss function, and conduct more extensive and comprehensive experiments including verification analysis and image segmentation to verify the performance of object detection in VHR optical remote sensing images.

The rest of this paper is organized as follows. Section II briefly introduces the preliminary knowledge including multiple instance learning and metric learning involved in this paper. We describe details of our Double-Net and new objective function in Section III. Section IV introduces the object detection framework with Double-Net in VHR optical remote sensing images. Section V shows comparative experimental results on two datasets. Finally, conclusions are drawn in Section VI.

II. PRELIMINARY KNOWLEDGE

The presented Double-Net model uses two different training samples from the same class as inputs to extract rotation-invariance features in images. Our approach involves two aspects of preliminary knowledge: multiple instance learning and metric learning. We briefly introduce them in this section.

A. MULTIPLE INSTANCE LEARNING

Multiple instance learning (MIL) [16]–[18] is a method evolved from the supervised learning algorithm. Instead of giving individual label for each instance, MIL just receives a set of 'bags' with the bag labels. Each bag contains many individual instances and an instance can be a single image sample or feature. A 'bag' X_{bag} is defined as:

$$X_{bag} = \{X_1, X_2, \ldots, X_n\}, \tag{1}$$

where X_i is the i^{th} instance, and n is the number of instance in the bag. For the binary classification task, the bag is considered positive if it contains at least one positive instance. In other words, if all instances of bag are negative, the bag will be labeled as negative. Obviously, relativity of sample from the same class is considered in multiple instance learning (MIL). In this paper, we adopt the multiple instance learning algorithm to solve within-class diversity with metric learning during training our Double-Net.

B. METRIC LEARNING

Metric Learning [19], [20] is also known as similarity learning, which aims to maximize the inter-class distances and minimize the intra-class distances. Euclidean distance and Mahalanobis distance are usually used to measure the distance between different features. And the Euclidean distance D_e and Mahalanobis distance D_m are computed by (2) and (3), respectively:

$$D_e^2(x_i, x_j) = \|x_i - x_j\|_2^2, \tag{2}$$

$$D_m(x_i, x_j) = \sqrt{(x_i - x_j)^T \Sigma^{-1}(x_i - x_j)}, \tag{3}$$

where x_i and x_j are d-dim features, and Σ is the covariance matrix.

The existing metric learning works based on deep CNN can be roughly divided into two classes: Siamese network and triplet network.

1) SIAMESE NETWORK

Siamese network [21] is trained with training pair (x_i, x_j) and it employs contrastive loss function:

$$L_{con} = \frac{1}{2}y_{i,j}D_e^2(x_i, x_j) + \frac{1}{2}(1 - y_{i,j})h(\alpha - D_e^2(x_i, x_j)), \quad (4)$$

where $y_{i,j}$ describes the sample pair (x_i,x_j) has the same identity or different identity. For example, if x_i and x_j are from the same class, $y_{i,j} = 1$; otherwise $y_{i,j} = 0$. $h(\alpha - D_e^2(x_i, x_j))$ is a hinge loss function:

$$h(\alpha - D_e^2(x_i, x_j)) = \max(\alpha - D_e^2(x_i, x_j), 0), \quad (5)$$

where α is a predefined margin.

Contrastive loss function can effectively handle with the relationship of a sample pair (x_i, x_j) in neural networks. If the sample pair (x_i, x_j) is from the same class, the loss is $L_{con} = \frac{1}{2}D_e^2(x_i, x_j)$, and we aim to decrease $D_e^2(x_i, x_j)$. Otherwise, the loss is $\frac{1}{2}\max(\alpha - D_e^2(x_i, x_j), 0)$ and we increase $D_e^2(x_i, x_j)$ until it is larger than the margin α.

2) TRIPLET NETWORK

Unlike the Siamese network takes a sample pair as input, the input of the Triplet network [22] is a triplet, (x_a, x_p, x_n), where x_a is called as anchor, (x_a, x_p) from same class, and x_n from different classes. It is trained with the triplet loss function:

$$L_{trip} = max(D_e^2(x_a, x_p) - D_e^2(x_a, x_n) + \alpha, 0)^2. \quad (6)$$

Intuitively, we hope that the negative pair distance is larger than the positive pair distance with a margin. To achieve this, we aim to penalize the negative pair distances for being smaller than positive pair distances plus a pre-defined margin. In this paper, we adopt the contrastive loss to train our Double-Net. However, different from most of existing metric learning works, the metric learning regularization is not only used to learn discriminative feature representations but also explored to train an effective classifier simultaneously.

III. PROPOSED METHOD

The Figure 1 illustrates the overall architecture of our proposed Double-Net for feature learning and feature extraction. In summary, it consists of the following three steps:

1) Image preprocessing: The goal of our proposed method is to address the challenges of within-class diversity in VHR optical remote sensing image. Thus, the input of our proposed Double-Net is sample pair (X_i, X_j), where X_i and X_j are from the same class. In image preprocessing, we select arbitrary two samples from the same class as training sample pairs.
2) Feature Learning: the proposed Double-Net is trained with sample pairs from the same class as input to feature learning. The two samples are input to two blocks of Double-Net respectively, for rotation-invariant feature learning. Their rotation-invariant features are combined into a new feature. Such input of sample pairs can also be viewed as a training augmentation that the number of training is increased to be much larger than the number of available training samples. And it also has another advantage of changing the distribution of within-class feature. In the classification part of the network, except for using the cross entropy loss, we also use the metric learning regularization term on learning the rotation-invariant features to enforce the Double-Net models to be more discriminative.
3) Feature Extraction: The proposed Double-Net takes only one sample as input to extract rotation-invariant feature.

A. FEATURE LEARNING BY DOUBLE-NET

The total architecture of the proposed Double-Net is shown in the Figure 1. As shown in Figure 1, the Double-Net model contains two parts: the upper half is feature learning model for training, and the lower half is feature extraction model for testing.

The proposed Double-Net is divided into two identical blocks. Sample pairs from the same class are fed to these block such that two blocks of Double-Net receive two different samples from the same class. As shown in Figure 1, each block contains multiple CNNs to handle a sample with different rotations to extract rotation-invariant features. The architecture of the CNN in each channel of the block is shown in Figure 2. To facilitate training and, particularly, to avoid too large number of model parameters, the parameters (weights and biases) of layers $C1$, $C2$, $C3$, $FC4$ of CNN in each channel of the block, denoted by $\{W1, W2, W3, W4\}$ and $\{B1, B2, B3, B4\}$, are shared. The summation of gradient in all channels is used to update the sharing weights.

We define L rotation angles $\theta = \{\theta_1, \theta_2, \ldots, \theta_L\}$, and apply rotation transformation in the input image, so that we have L samples wither different rotation transformation: $X = \{X_{\theta_1}, X_{\theta_2}, \ldots, X_{\theta_L}\}$. As Double-Net has two blocks, we have $(X_i, X_j) = (\{X_{i\theta_1}, X_{i\theta_2}, \ldots, X_{i\theta_L}\}, \{X_{j\theta_1}, X_{j\theta_2}, \ldots, X_{j\theta_L}\})$, which will be fed into the two blocks to train the Double-Net.

We define $O_4(X_{i\theta_1})$ as outputs of FC4 layer when the sample $X_{i\theta_1}$ is trained by the Double-Net. Thus the new rotation-variant feature g_r is computed by

$$g_r = max\{O_4(X_{i\theta}), O_4(X_{j\theta})\}, \quad (7)$$

where $O_4(X_{i\theta})$ is defined as a feature set $\{O_4(X_{i\theta_1}), O_4(X_{i\theta_2}), \ldots, O_4(X_{i\theta_L})\}$ and $O_4(X_{j\theta})$ is defined as a set $\{O_4(X_{j\theta_1}), O_4(X_{j\theta_2}), \ldots, O_4(X_{j\theta_L})\}$.

Motivated by the MIL algorithm, we want to construct the new feature g_r with MIL. Instead of the maximum operation in (7), many other operators can be used. As many previous work have proved that the maximum operator is the most effective, we adopt the maximum operator in this paper. This (7) satisfies the axioms of closure, associativity, invertibility and identity when the sample pair $(O_4(X_{i\theta_L}),$

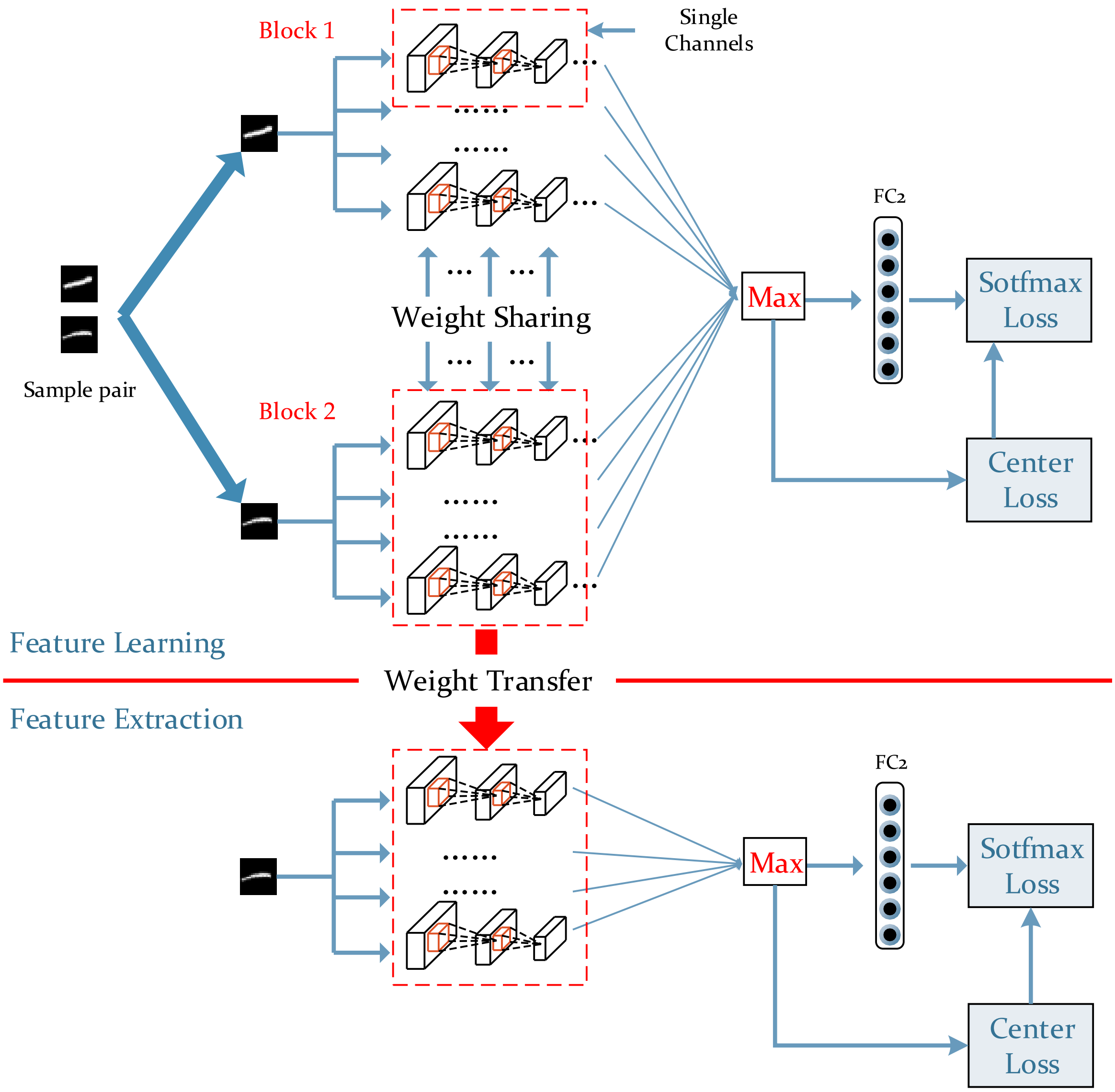


FIGURE 1. Process overview of feature learning and feature extraction.

$O_4(X_{j\theta_L})$) as input data, for the detailed proof please refer to [17].

In image classification tasks, generalization of recognition algorithms is eternal pursuit of researcher. The TI-Pooling solves a problem of view invariance, its group consists of a series of transformations of one sample. Different from TI-Pooling [12], we want that this new features g_r can possess rotation invariance and algorithm model can possess better generalization of recognition algorithms. Considering within-class variance in the new rotation-invariant feature we design, $(O_4(X_{i\theta_L}), O_4(X_{j\theta_L}))$ from the same class, and new features g_r can obtain the better generalization.

Thus the X_i and X_j in sample pair (X_i, X_j) can be combined arbitrarily, which result in number of training data set become larger. In order to prevent overfitting, we introduce the $L2$ regularization term in object function:

$$R_{L2} = \frac{\lambda_1}{n}(\|W\|_2^2), \tag{8}$$

where $W = \{W_1, W_2, W_3, W_4\}$, and λ_1 is tradeoff parameters that control the relative importance of the regularization term in object function.

By constructing a new invariant feature g_r, within-class diversity is considered by simple pair matching strategy and multiple instance learning (MIL), the generalization performance of algorithm model is improved. But beyond that, we also add metric learning regularization term on new invariant feature g_r to enforce the within-class feature distribution of algorithm model to be closer. The metric learning regularization term is defined as:

$$\begin{aligned} J_C &= \frac{\lambda_2}{2}\sum_{i=1}^{N}\|g_r - c_{y_i}\|_2^2 \\ &= \frac{\lambda_2}{2}\sum_{i=1}^{N}\|max\{O_4(X_{i\theta}), O_4(X_{j\theta})\} - c_{y_i}\|_2^2, \end{aligned} \tag{9}$$

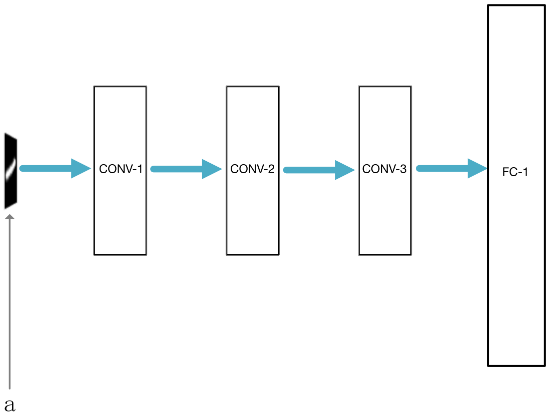


FIGURE 2. The architecture of CNN in the channels of the proposed Double-Net, in which 'a' is rotated image sample.

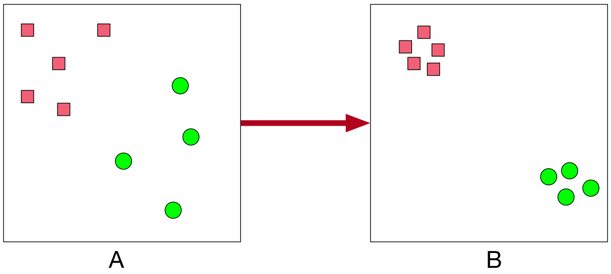


FIGURE 3. The feature space contrastive illustration, the Figure A denote original CNN feature space and Figure B denote the proposed Double-Net feature space, where red dot and blue dot denote the different feature.

where coefficient weight λ_2 is similar to λ_1 to control the relative importance of the regularization term in object function, total number of sample pair is N and c_{y_i} is defined as the y_i class center of deep features.

By using this regularization term, deep invariant feature g_r is forced to approach the its class center of deep features. How to extract center of deep feature will not be introduced in this paper, for detailed inference please refer to [21]. By applying this constraint item, the proposed Double-Net feature space is changed, which is shown in Figure 3.

In the end, the Double-Net algorithm model object function that we propose, which consists of three terms:

$$J = min(J_S + \lambda_1 J_C + \lambda_2 R_{L2}) \tag{10}$$

where J_S is a cross entropy loss term, which is defined as:

$$\begin{aligned} J_S &= -\frac{1}{N}\sum_{N}^{i=1}\langle y_i, log(g_{ri})\rangle \\ &= -\frac{1}{N}\sum_{N}^{i=1}\sum_{M}^{k=1} y_{x_i}[k]\cdot log(g_{ri}[k]). \end{aligned} \tag{11}$$

B. FEATURE EXTRACTION BY DOUBLE-NET

In the feature extraction process, the feature extraction network of Double-Net is different from the feature learning network of Double-Net. At first, we take one single sample as input for feature extraction with Double-Net, and it is just like feature extraction with traditional CNNs. Then, as shown in lower half of Figure 1, the testing sample is fed to a CNN consisting three convolutional layers and a fully connected layer. The weight of this feature extraction CNNs channels is transferred from the Double-Net, feature extraction of Double-Net only change input data flow. For example, if Double-Net have N channels, which means that sample pair in feature learning of Double-Net will be rotated N angles randomly or one sample in sample pair will produce $\frac{N}{2}$ rotated sample, and one rotated sample will be fed to one channel. So one testing sample will be rotated N angles randomly to fed the channels in feature extraction of Double-Net, they will reuse weight of channels after training. Thus, the deep rotation invariant feature is different from the (11), it is computed by:

$$g_r = max\{O_4(X_{i\theta})\}, \tag{12}$$

where $O_4(X_{i\theta})$ is defined as a set $\{O_4(X_{i\theta_1}), O_4(X_{i\theta_2}), \ldots, O_4(X_{i\theta_{2L}})\}$, and the L also denote the number of channels in each Block.

IV. OBJECT DETECTION WITH DOUBLE-NET

As shown in Figure 4, the object detection system that we propose has two stages: region proposal and image classification. In the region proposal stage, its main feature is to search the possible regions of target in images. In theory, one image can contain unlimited region proposals. To avoid too large number of region proposals, we adopt the selective search strategy to solve this problem. The image classification stage takes the inputs from previous stage, and it aims to find the object with high probability from region proposed.

In this section, we will introduce how to extract rotation-invariant for object detection system by Double-Net in details.

A. REGION PROPOSALS

Region proposal is an important stage in object detection task. Earlier image segmentation algorithm is based on sliding-window search, and it will scan each image at all positions with different scales. With the development of computer vision technology, more and more efficient region proposal algorithms are applied in object detection. Due to our main goal is to validate the Double-Net model and new rotation-invariant feature rather than develop an new object proposal algorithm, we adopt existing region proposal algorithm named selective search [23] to produce the object hypotheses. In addition, compared with the other region proposal algorithm, selective search can generate a comparable number of proposals that have a higher overlap with the objects.

In application, selective search algorithm could produce more than 20 thousand region proposals from one image, which makes the speed of object detection slow. To solve this problem, we make a rule to limit the number of region proposals: we only use segmented image that length-width radio equal approximately 1 as possible region of object.

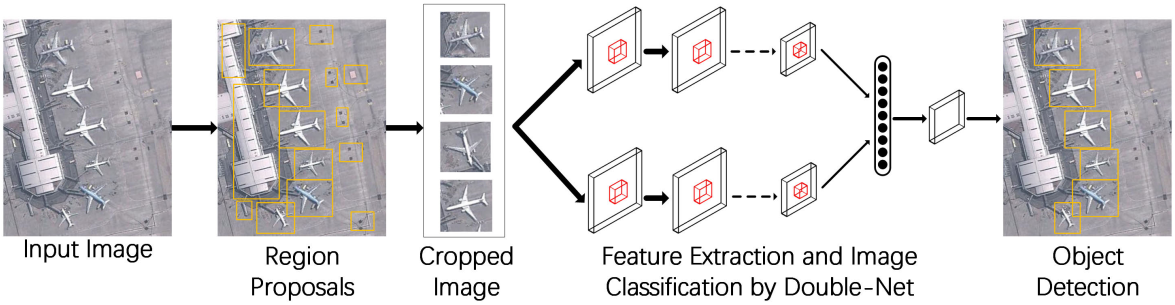


FIGURE 4. Process overview of proposed object detection system.

B. IMAGE CLASSIFICATION

Using the trained Double-Net model, we can extract our new rotation-invariant features from each region proposal. As each size of region proposal is not completely compatible to the input size of the trained Double-Net model, we need transform region proposals to the fixed size of 32×32 pixels. In other hand, we consider image contexts around the original region of object proposal in our algorithm.

In application, as candidate bounding boxes generated by region proposals and ground truth bounding box is not completely overlapping, we use IoU (Intersection-over-Union) metric to determine whether the classification is correct, the IoU metric is computed by

$$a_o = \frac{B_{op} \bigcap B_{gt}}{B_{op} \bigcup B_{gt}}, \tag{13}$$

where B_{gt} denotes the ground truth bounding box and B_{op} denotes the object proposal bounding box.

To train the object detection system with Double-Net, the training dataset is divided into two parts: positive samples and negative samples. The positive samples contain all ground truth bounding boxes and their labels, and the negative samples are obtained by a hard-negative mining technique from region proposals with $a_o < 0.2$. To validate our performance of rotation-invariant feature and train multi-class object detection system with Double-Net, the region proposals with $a_o \geq 0.5$ are defined as correct classification results.

V. EXPERIMENTS

As our paper mainly presents a method of extracting rotation-invariant feature rather than an object detection system, the first experiment is to validate capacity of rotation-invariant feature on Mnist-rot-12K dataset. In order to show the application of proposed new rotation-invariant in VHR optical remote sensing images object detection, the second experiment is validated on the NWPU VHR-10 dataset. To quantitatively evaluate the proposed Double-Net method, we compare our method with two state-of-the-art convolution neural networks about rotation-invariant or transformation-invariant feature extraction in this section.

A. EXPERIMENT ON MNIST-ROT-12K DATASET

Mnist-rot-12k dataset is commonly used to evaluate the performance of rotation-invariant feature extraction. This dataset is based on the original MNIST dataset [24], [25], which is designed to test artificially-introduced variations. It contains some images from original MNIST dataset rotated by a random angle from 0 to 360 degree (full circle). And this dataset has 12,000 training images and 50,000 testing sample. Figure 5 shows some samples from two different categories.

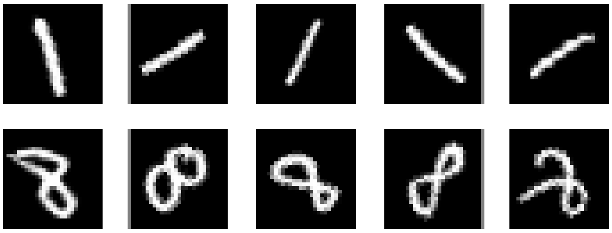


FIGURE 5. Some sample in the Mnist-rot-12k data set.

In order to evaluate the rotation-invariant performance of the proposed Double-Net method, we select 10 to 20 training samples for each class to validate performance in small amount of training samples. To validate performance in a large amount of training samples, we use 10,000 training samples.

1) SMALL TRAINING DATASET

First, in order to evaluate the influence of parameter λ_1 and λ_2 in the object function defined by (10), λ_1 is set from $\{0.025, 0.005, 0.0005\}$ and λ_2 is set from $\{0.0005, 0.00005\}$. In this experiment, 10 training samples per class are randomly selected from Mnist-rot-12k training dataset, while the rest is used for testing. The number of channels is set as 4. To train network model, we set the learning rate to 1×10^{-4}. In addition, we train every experiment for 2,000 epochs, and use its highest accuracy as the final result. The experimental results of the proposed Double-Net with varied λ_1 and λ_2 are shown in Fig.6. It is observed that the best accuracy is obtained when $\lambda_1 = 0.025$ and $\lambda_2 = 0.0005$.

In addition, since testing accuracy is affected by the number of channels, we test the experiments with different number of channels. The experimental results with varied number of channels is shown in Figure 7, in which the number of channels is varied from 2 to 8. Other parameters are set as the same as that in previous experiment. As can be seen from the Figure 7, the classification accuracy increase continuously when more channels are selected and achieves its best value when the number of channels is set as 8.

TABLE 1. Small data set image classification accuracy contrast table the bold numbers denote the highest values in each column.

Method \ Number of Class	10	11	12	13	14	15	16	17	18	19	20
Double-Net	**86.11%**	**88.12%**	**88.34%**	**89.34%**	**89.56%**	**89.06%**	**89.78%**	**90.22%**	**90.24%**	**90.92%**	**91.32%**
Ti-Pooling	78.23%	82.01%	82.38%	83.72%	83.63%	86.33%	86.33%	86.23%	86.33%	86.23%	86.85%
RICNN	80.66%	84.61%	86.01%	85.15%	86.39%	85.45%	85.64%	86.48%	87.74%	87.78%	88.65%

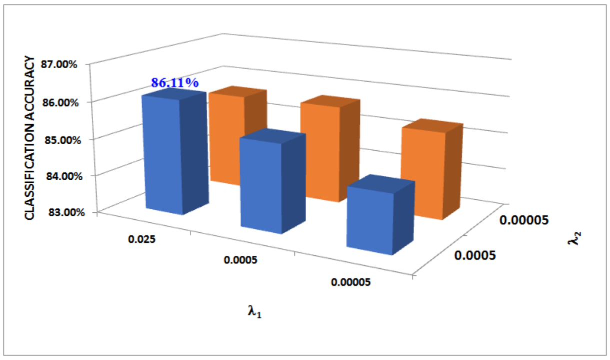


FIGURE 6. Image classification results by the proposed Double-Net with varied λ_1 and λ_2. The best result is highlighted with its accuracy over its corresponding bar.

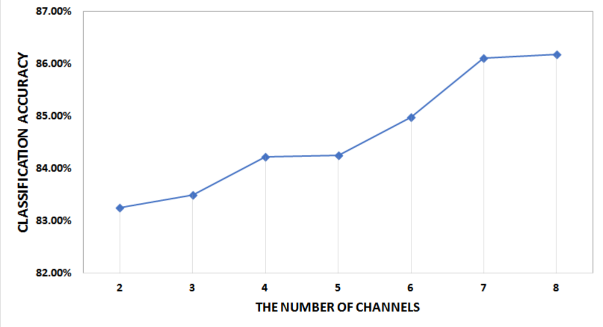


FIGURE 7. Image classification results of the proposed Double-Net with different the number of channels.

In order to further verify the performance of proposed method in small data set, different number of training samples per class are considered by varying it from 10 to 20. Two state-of-the-art CNN methods, namely RICNN [11] and Ti-Pooling [12], are selected for comparison. The experimental results with different number of training samples are listed in Table 1. As can be seen from the Table 1, the proposed Double-Net outperforms the RICNN [11] and Ti-Pooling [12] in all small training dataset experiment, indicating that the proposed Double-Net method has higher generalization ability and robustness.

2) BIG TRAINING DATA SET

Since training input sample of Double-Net method is sample pair matching, which will generate the more than 10 times number of the original training data set, and how to avoid overfitting in training big data set is necessary to be considered in proposed Double-Net method. Therefore this experiment has two purposes: on the one hand it is to validate performance of the proposed Double-Net in the big training data set; on the other hand, it can help figure out how the number of sample pairs affects experiment accuracy.

TABLE 2. Image classification accuracy in Big dataset.

Method	Testing Accuracy
Double-Net	**98.444%**
Ti-Pooling [12]	97.945%
RICNN [11]	98.193%
P4CNN [26]	97.72%
H-Net [27]	98.242%
OR-Ti-Pooling [28]	98.312%

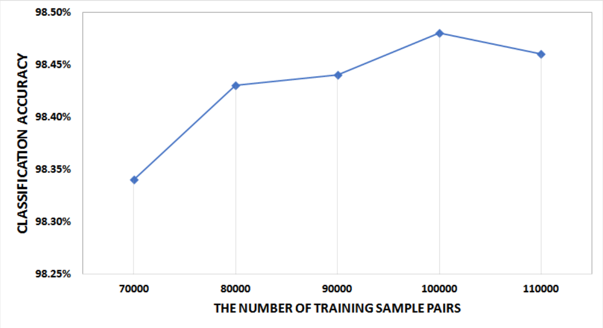


FIGURE 8. Image classification results of the proposed Double-Net by adopting different number of training sample pairs.

In this experiment, the training dataset has 10,000 samples for training, and per class have 10,000 training samples. In addition, the training dataset is obtained by sample matching randomly, the number of sample pairs is 160,000. And the number of channels of each block in the Double-Net is set to 4 in this experiment. We train this Double-Net method on GPU for 4,000 epochs repeatedly, and get the highest result.

The comparative results with TI-Pooling and RICNN are shown in Table 2. It shows that our proposed Double-Net method consistently outperforms TI-Pooling and RICNN, generally 0.3% higher in classification accuracy.

Due to the input of the proposed Double-Net is sample pairs, so how many samples are generated can obtain the best testing accuracy, and avoid overfitting because the number of training sample pair is too large. A set of experiment is set up, their training sample pairs are generated into 5 different sets of number by 10000 images. The result is shown in Figure 8, and the numbers of training sample pairs of 5 experiments are {70000, 80000, 90000, 100000, 110000}. From the result comparison chart, the number of training sample pairs is not important to classification results, which only cause that results fluctuation is very small.

B. EXPERIMENT ON NWPU VHR-10 DATA SET

NWPU VHR-10 is a challenging ten-class VHR optical remote sensing image object detection dataset [29], [30], which is used for multi-class object detection. It includes the following classes: airplane, storage tank, tennis court, basketball court, harbor, bridge, ship, ground track field

TABLE 3. Performance comparisons of two state-of-the-art methods in terms of AP values. The bold numbers denote the highest values in each row.

	Ti-Pooling [12]	RICNN [11]	Double-Net
Airplane	**86.31**%	81.58%	82.26%
Ship	54.90%	55.74%	**60.19**%
Storage tank	**97.13**%	92.40%	93.30%
Baseball diamond	**91.04**%	74.64%	70.39%
Tennis court	77.05%	78.05%	**78.73**%
Basketball court	11.43%	15.65%	**15.74**%
Ground track field	21.98%	31.64%	**32.10**%
Harbor	48.67%	**49.62**%	45.56%
Bridge	9.7%	14.63%	**15.30**%
Vehicle	69.42%	81.36%	**82.79**%
Mean Average Precision	70.28%	71.33%	**72.20**%

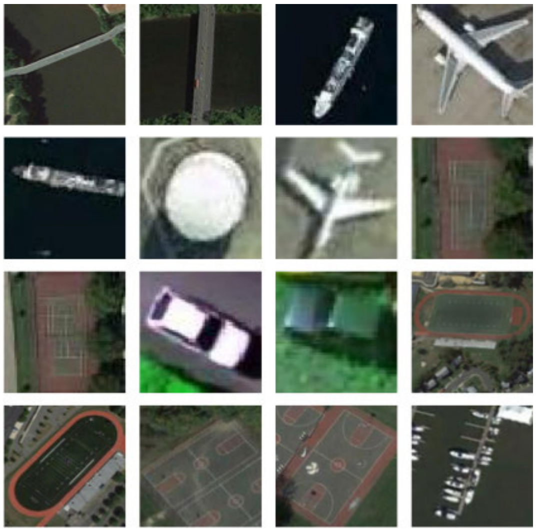


FIGURE 9. Some sample in the NWPU VHR-10 data set.

and vehicle. Images in this dataset are obtained from Google Earth with the spatial resolution ranging from 0.5m to 2m and Vaihingen data with a spatial resolution of 0.08m. NWPU VHR-10 dataset is divided into two image sets: a positive dataset and a negative dataset. The positive dataset contains 757 airplane objects, 302 ship objects, 390 baseball diamonds objects, 655 storage tank objects, 524 tennis courts objects, 159 basketball court objects, 124 bridge objects, 244 harbor objects, 163 ground track field objects, and 477 vehicle objects. The negative dataset includes 150 samples and does not contain any objects that need to detect. Some sample of NWPU VHR-10 is shown in Figure 9,

1) EVALUATION METRICS

To evaluate the performance, we adopt precision recall curve (PRC) to evaluate the performance of feature extraction CNNs for object detection in VHR optical remote sensing images.

To evaluate the result, we adopt average precision (AP) to evaluate the performance of feature extraction CNNs for object detection in VHR optical remote sensing images. The AP computes the average value of precision over the interval from 0 to 1. The higher AP value represents the better performance, and vice versa.

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

TABLE 4. The parameters involved in the proposed Double-Net used in the experiments.

Layer	Parameters
input	size: 32×32
Covn-1	kernel: 3×3
max pooling	kernel: 2×2, stride: 2
Covn-2	kernel: 3×3
max pooling	kernel: 2×2, stride: 2
Covn-3	kernel: 3×3
dropout	rate: 0.5

where *tp* denotes the number of true positives and *fn* denotes the number of false negatives.

2) IMPLEMENTATION DETAILS

In this experiment, we use 30% of the NWPU VHR-10 dataset for training and 70% for testing. Table 3 shows the quantitative comparison results of two different CNNs method, which is measured by AP. Due to this different rotation-invariant feature extraction method need set the number of rotating image, so the RICNN and Ti-Pooling is set as 8 rotation channels, and the proposed Double-Net is set as 4 channels for every block. The parameter setting of our proposed Double-Net model is list in Table 4. To train network model for image classification in the proposed object detection task, the learning rate is set to 0.00001. In addition, λ_1 and λ_2 is set as 0.0025 and 0.00005, respectively, and the batch size is set as 64. We train the model for 3,000 epochs to obtain the best accuracy.

3) EXPERIMENTAL RESULT AND ANALYSIS

Figure 10 shows an example of object detection with the proposed Double-Net. Table 3 further shows the comparison results of two state-of-the-art method [11], [12]. As can be seen from the Table 3, the proposed method outperform all comparison method for mean average precision, compared with TI-Pooling [12] and RICNN [11], the proposed method obtained 1.92%, 0.87% performance gains.

In summary, compared with other state-of-the-art methods, the proposed Double-Net method achieves superior performance in rotation-invariant feature extraction of VHR optical remote sensing images:

1) Since the simple pairs consist of samples from same class and they are combined arbitrarily, the proposed Double-Net can reduce within-class variance in

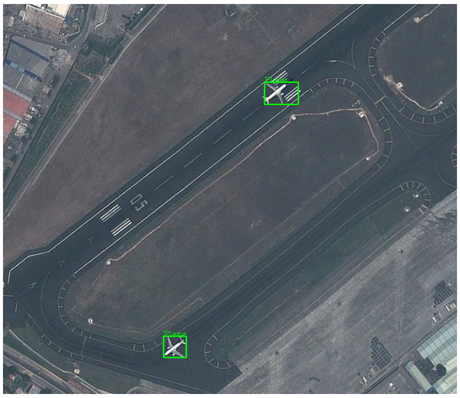


FIGURE 10. Object detection result by the proposed Double-Net. The true positives are indicated by green rectangles.

feature space, which has similar function as bagging algorithm in ensemble learning.

2) Metric learning is used in the proposed Double-Net, leading to a more discriminative feature space.

VI. CONCLUSION

In this paper, a novel Double-Net method is constructed to extract invariant for image classification for VHR optical remote sensing image object detection. For this Double-Net method, we have proposed a novel and effective approach to learn Double-Net model by optimizing a new object function, which use the multiple instance learning (MIL) to construct the new rotation-invariant feature and enforce they gather toward the center of class. In quantitative comparison experiments on Mnist-rot-12k dataset and NWPU VHR-10 dataset, compared with state-of-the-art methods, the proposed Double-Net achieves good performance, especially in small training dataset.

However, as far as we know, metric learning in CNNs will cause the slow converging speed, there are rarely any exceptions, even for the proposed Double-Net. Therefore, in our future work, we will focus on how to speed up convergence for the proposed Double-Net. Meanwhile, MIL will become our new research direction.

ACKNOWLEDGMENT

(Zhi Zhang and Ruoqiao Jiang contributed equally to this work.)

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. 2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[2] M. Ma, S. Mei, S. Wan, J. Hou, Z. Wang, and D. D. Feng, "Video summarization via block sparse dictionary selection," *Neurocomputing*, vol. 378, pp. 197–209, Feb. 2019, doi: 10.1016/j.neucom.2019.07.108.

[3] G. Tao, Z. Liu, J. Cao, and S. Liang, "Local difference ternary sequences descriptor based on unsupervised min redundancy mutual information feature selection," *Multidimensional Syst. Signal Process.*, pp. 1–21, 2018, doi: 10.1007/s11045-018-0595-z.

[4] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

[5] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2005, pp. 886–893.

[7] 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.

[8] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2005, pp. 524–531.

[9] Y. Onodera, H. Watanabe, A. Taguchi, N. Iijima, M. Sone, and H. Mitsui, "Translation and rotation-invariant pattern recognition method using neural network with back-propagation," in *Proc. ICCS/ISITA*, Jan. 2003, pp. 548–552.

[10] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[11] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[12] D. Laptev, N. Savinov, J. M. Buhmann, and M. Pollefeys, "TI-POOLING: Transformation-invariant pooling for feature learning in convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 289–297.

[13] S. Dieleman, K. W. Willett, and J. Dambre, "Rotation-invariant convolutional neural networks for galaxy morphology prediction," *Monthly Notices Roy. Astron. Soc.*, vol. 450, no. 2, pp. 1441–1459, Jun. 2015.

[14] G. Cabrera-Vives, I. Reyes, F. Förster, P. A. Estevez, and J.-C. Maureira, "Deep-HiTS: Rotation invariant convolutional neural network for transient detection," *Astrophys. J.*, vol. 836, no. 1, p. 97, Feb. 2017.

[15] S. Mei, R. Jiang, J. Ji, J. Sun, Y. Peng, and Y. Zhang, "Invariant feature extraction for image classification via multi-channel convolutional neural network," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Nov. 2017, pp. 491–495.

[16] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3460–3469.

[17] D. Laptev and J. M. Buhmann, "Transformation-invariant convolutional jungles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3043–3051.

[18] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3642–3649.

[19] E. P. Xing, A. Y. Ng, M. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2002, pp. 521–528.

[20] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, no. 1, pp. 207–244, 2009.

[21] Y. Wen, K. Zhang, and Z. Li, "A discriminative feature learning approach for deep face recognition," in *Proc. Comput. Vis. (ECCV)*, vol. 2016, pp. 499–515.

[22] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[23] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective Search for Object Recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[24] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 473–480.

[25] C.-L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: Benchmarking of state-of-the-art techniques," *Pattern Recognit.*, vol. 36, no. 10, pp. 2271–2285, Oct. 2003.

[26] T. S. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2990–2999.

[27] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Harmonic networks: Deep translation and rotation equivariance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5028–5037.

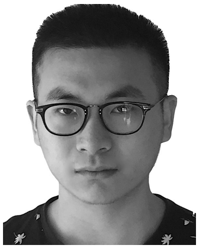
[28] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Oriented response networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 519–528.

[29] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.

[30] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2019.

ZHI ZHANG received the B.S. and master's degrees in applied mathematics from Air Force Engineering University, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from the National University of Defense Technology, Changsha, China, in 2010.

His research interests include neural networks and pattern recognition.

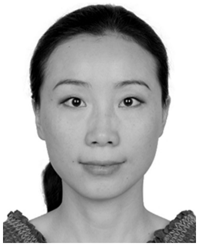


RUOQIAO JIANG received the B.S. degree in electronics and information engineering from Hohai University, Nanjing, China, in 2017. He is currently pursuing the master's degree in signal and information processing with Northwestern Polytechnical University.

His research interests include neural networks and pattern recognition.

SHAOHUI MEI received the B.S. degree in electronics and information engineering and the Ph.D. degree in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2005 and 2011, respectively. He was a Visiting Student with the University of Sydney, from October 2007 to October 2008. He is currently an Associated Professor with the School of Electronics and Information, Northwestern Polytechnical University. His research interests include hyperspectral remote sensing image processing, deep learning, video processing, and pattern recognition.

SHUN ZHANG received the B.S. and Ph.D. degrees in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 2009 and 2016, respectively. He is currently an Assistant Professor with the School of Electronic and Information, Northwestern Polytechnical University, Xi'an. His research interests include machine learning, computer vision, and human–computer interaction, with a focus on visual tracking, object detection, image classification, feature extraction, and sparse representation.



YIFAN ZHANG received the B.S. degree in electronics and information technology, and the M.S. and Ph.D. degrees in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2001, 2004, and 2007, respectively. From 2007 to 2010, she worked as a Postdoctoral Researcher at the Vision Laboratory, Department of Physics, University of Antwerp, Antwerp, Belgium. She is currently an Associate Professor with the School of Electronics and Information, Northwestern Polytechnical University. Her research interests include hyperspectral image analysis, image fusion, and image restoration.