

Received June 22, 2021, accepted June 30, 2021, date of publication July 5, 2021, date of current version July 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3094825

# The Recognition Framework of Deep Kernel Learning for Enclosed Remote Sensing Objects

LONG SUN<sup>1,2,3</sup>, JIE CHEN<sup>3,4</sup> (Member, IEEE), DAZHENG FENG<sup>1,2</sup> (Member, IEEE),  
AND MENGDAO XING<sup>1,2</sup> (Fellow, IEEE)

<sup>1</sup>National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China

<sup>2</sup>Collaborative Innovation Center of Information Sensing and Understanding, Xidian University, Xi'an 710071, China

<sup>3</sup>38th Research Institute, China Electronics Technology Group Corporation, Hefei 230088, China

<sup>4</sup>Key Laboratory of Intelligent Computing and Signal Processing of the Ministry of Education, School of Electronics and Information Engineering, Anhui University, Hefei 230039, China

Corresponding author: MengDao Xing (xmd@xidian.edu.cn)

This work was supported in part by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61621005, in part by the National Natural Science Foundation of China under Grant 62001003, in part by the Natural Science Foundation of Anhui Province under Grant 2008085QF284, and in part by the China Postdoctoral Science Foundation under Grant 2020M671851.


**ABSTRACT** Remote sensing image target recognition is used in various fields, such as ships, tanks, airplanes, and vehicles, which are closed targets. The features of these targets include target outlines that are obvious and target discriminant features that are significantly different from the surrounding environment, and the targets are characterized as small and dense. Therefore, the recognition of these types of targets is a popular topic. We proposed a recognition framework consisting of a remote sensing image target recognition method based on deep saliency kernel learning analysis, which uses a target region extraction method based on the visual saliency mechanism and implements a nonlinear deep kernel learning saliency feature analysis method to realize target extraction and recognition. Experimental results show that a 95.9% recognition rate is achieved for SAR remote sensing target recognition on the public MSTAR data set, a 96% recognition rate on the UC Merced Land Use data set, and an 85% recognition rate on a self-built visible light remote sensing image data set. The recognition framework can be used for video recognition.

**INDEX TERMS** Saliency analysis, deep kernel learning, remote sensing target recognition.

## I. INTRODUCTION

In recent years, remote sensing data have been widely used in various fields. Remote sensing images play an important role in many aspects, such as the exploration of geographic information resources [1], important ground information observations, geographic mapping, meteorology, civilian-military communications [2], the detection of military information [3], the capture of sensitive information, and battlefield situational awareness [4]. In military fields, the automatic identification of sensitive targets is a very important research direction in military reconnaissance and military early warning systems. Integrating automatic target recognition technology into systems characterized by high practicality and high robustness also appears to be particularly important. At present, the realization of obtaining information from domestic remote sensing images is at the stage of

transformation from traditional manual determination methods to intelligent automatic identification methods. Many units at home and abroad are gradually carrying out platform and systematization work on the technology to extract target information from remote sensing images. Many universities and research institutions such as the University of Trento, the School of Computer Science at Carnegie Mellon University, and the Department of Geographic Information at the University of Maryland have conducted in-depth research on this topic. At present, the research on processing and recognition systems for remote sensing images at home and abroad is based on custom target recognition systems for specific targets of interest, such as the ship target recognition system mentioned by Kodors *et al.* [5]. An airport identification system was proposed by Liu *et al.* [6], and a building group identification system was proposed by Li *et al.* [7]. Most of these studies are highly targeted, aiming at specific targets in specific scenes and achieving good processing effects; however, the generalization of the system is relatively weak.

The associate editor coordinating the review of this manuscript and approving it for publication was Juan A. Lara .

On the other hand, universal remote sensing image automatic processing systems have been proposed with good universality, such as SAHARA, a semiautomated image scene understanding system based on multisource remote sensing images developed by Druyts P *et al.* of the Western European Joint Satellite Center [8], and the high-resolution remote sensing image processing system SCORPIUS developed by B Guindon and other research institutions in the United States [9] that provides classification recognition and tracking functions for certain specific military targets. However, much work remains to be performed for highly customized practical applications.

From the perspective of imaging, remote sensing images have rich imaging details and a single imaging angle, but their imaging scale and illumination conditions change greatly, and they are affected by weather conditions. There is also considerable clutter-induced noise in these images and a large amount of background information. The basis of remote sensing target recognition is an accurate description of the visual characteristics of the target in the studied remote sensing image and the construction and expression of the prior knowledge of the image target. In recent years, the research on target recognition algorithms for remote sensing images has mainly been aimed at roads [10], building clusters [11], aircraft [12], large bridges [13], highways [14], oil tanks [15] and other targets that are closely related to human or military activities. Synthesizing various processing methods developed for the recognition of targets of interest in traditional remote sensing images in recent years [32]–[34], we find that they can be roughly divided into two cases: feature-based and model-based methods.

The basis of model-based remote sensing image recognition methods is the construction of the target model of the researched remote sensing image. The construction of the model often depends on the accumulation of prior knowledge of the target and the background. Model-based methods focus on the salient features or combinations of particular structural primitives in the target, such as long and straight runway structures in airport detection, large parallel linear structures in the detection of bridges and highways, and dense, short lines in the detection of building clusters. Abstract modeling of the research object is accomplished by constructing special functions and vectors, such as road recognition methods based on geographic information models, as proposed by Barzohar and Cooper [16]. In addition, some model-based remote sensing images directly detect research targets by means of an energy function, such as a conditional random field [17].

Feature-based remote sensing image target recognition technology is a widely studied target recognition technology, and its research basis is the construction and description of target features in images. The features of an image include many aspects. One aspect refers to recognition based on statistical information such as the grayscale, texture and color of the image. Image statistical information is an important part of image features. Researchers at the German

Aerospace Center have proposed an automatic highway extraction method based on the statistical information of the target color in remote sensing images [18]. The second aspect refers to target recognition methods based on the corner features, geometric features, linear features, edge features, and area features of the remote sensing image [19], such as the ship detection method proposed by Jin *et al.* [20]; combining a Harris corner detector and an image local significance value calculation, a ship's features can be precisely constructed and described. Xin *et al.* [21] designed a feature structure based on the combination of image linear feature detection with the SIFT operator [22]. This method, supplemented by a tree classifier as a decision strategy, achieved the detection of airport targets. In these detection methods, many scholars use one or more characteristics to express and describe the image and target and use one or more methods to express each characteristic. For example, the gray feature co-occurrence matrix and the LBP local binary mode are used to describe the texture features [31]. SVMs (support vector machines), decision trees, AdaBoost cascading classifiers and other machine learning methods have been introduced to assist decision making and have achieved good recognition results.

At present, most of the existing traditional remote sensing target recognition and detection algorithms use the strategy of coarse detection and fine detection. First, the candidate target regions with possible targets are extracted from the input image; that is, ROIs are extracted. Then, based on the candidate area, more accurate target recognition and confirmation are performed, and the false detection area is removed to find and confirm the final target. For the extraction process of candidate areas in original images, scholars have proposed many ideas. Among them, Zhu CR *et al.* proposed a method of threshold segmentation combining gray information and edge information in the image [23]. Li ZM, Yang DQ *et al.* proposed a multilayer sparse coding method to obtain the sparse description features of the image [24], calculated its saliency value based on the sparse feature of the image, and segmented the image according to the saliency value to obtain the candidate target area of the image. For the process of target recognition and confirmation after obtaining each candidate region, most researchers adopt the method of using various feature descriptors to describe the detected target and then perform further confirmation by means of a support vector machine or an AdaBoost classifier. For example, in reference [25], the author extracted the texture features of the image and combined them with the shape features of the target to obtain a high-dimensional feature descriptor to describe a ship target. Similarly, Yang F *et al.* [26] calculated a local image binary model as the image target feature; a deformable component model was also proposed in reference [27] to describe the target in an image. References [28] and [29] proposed target recognition and detection technology using feature analysis and the mean shift algorithm. These algorithms were aimed at some specific close-up targets but are not related to the target identification of remote sensing targets such as aircraft, oil tanks and ships. Therefore, it is necessary

to conduct a customized analysis of the characteristics of these targets.

In summary, this paper proposes a remote sensing image target recognition method based on deep saliency kernel learning analysis that uses target region extraction based on the visual saliency mechanism and uses a nonlinear deep kernel learning saliency feature analysis method to realize target extraction and recognition. Given the problem that single kernel model learning with fixed parameters is no longer suitable for remote sensing image classification, a new deep kernel mapping architecture for remote sensing image classification is proposed by combining kernel mapping and deep learning. This architecture solves the problems of the similarity degree of the input vector of the kernel mapping function, the structure of the kernel mapping function and the structure of the kernel mapping learning network. The proposed approach can effectively improve the adaptability of the learning model to target extraction, more accurately describe the data from the input space in terms of the nonlinear mapping relationship, and enable the data belonging to different classes to achieve better discrimination in the nonlinear mapping space. Compared with the existing algorithms, the proposed method can extract features with higher resolution and improve the target classification accuracy.

## II. METHODS

### A. ALGORITHM FRAMEWORK

As shown in Fig. 1, we proposed a recognition framework for the remote sensing image target recognition method based on deep saliency kernel learning analysis. The framework extracts the features of targets, including target outlines that are obvious, target discriminant features that are significantly different from the surrounding environment, and targets that are small and dense. The detection process can be explained in two parts. The first part is the rapid acquisition of candidate regions. The second part is the description and recognition of the target. The main steps are to calculate the saliency map of the input image by using the improved FT saliency algorithm and to use the mean shift algorithm to segment the original image and merge small fragments in the original image. Then, the region result obtained by segmentation and the basic shape characteristics of the research object are synthesized to conduct the subsequent screening of various ROIs. The specific region extraction method will be described in detail later. For the collected target sample files, a sample data set is constructed with a 1:3 ratio of positive and negative samples, and a sample description file is generated. Then, for the sample description file, appropriate features are designed, and deep kernel learning is used to extract and describe the target features. The feature descriptors are sent to the classifier for learning classification. Finally, the features of each ROI are extracted and described and then implemented in a previously trained deep kernel learning detector, which outputs a final recognition result.

### B. STEP 1: OBJECT EXTRACTION BASED ON THE VISUAL SALIENCY MECHANISM

The size of the target is smaller than the entire input image, and more of the target's background area is in the image. Therefore, quickly locating the area where the target may appear, removing redundant information, and extracting areas with a high degree of similarity to the target is necessary. One of the commonly used research ideas is to extract regions with the help of visual saliency. The ROI extraction algorithm based on the visual saliency mechanism draws on the human visual selection attention mechanism, and the pixels with significant local optical features in the input image are aggregated into the region of interest. The saliency detection method can be divided into two types based on this idea; one is a bottom-up data-driven saliency detection method, which usually uses the local features, spectrum, and local contrast and other information to measure the saliency of the image; the second can be considered a top-down visual saliency model. The generation of saliency maps under this idea is mostly achieved by combining the bottom-up detection results with scale, position, size, contour, and other features according to the specific scene requirements. The calculation of the saliency model under this idea is mostly more complicated than the first idea. A saliency detection algorithm based on the color characteristics of the image is used to calculate the saliency map of the input image. Based on the visual saliency map, to extract the ROI in high-resolution remote sensing target recognition in this paper, the implementation process is shown in Fig. 2.

For the process of ROI extraction, the main steps are as follows:

Step (1) For the input image, use the visual saliency detection algorithm to calculate the saliency value  $S(x,y)$  of each pixel and generate a saliency image;

Step (2) Calculate the average significant value  $S_{\text{mean}}$  of the image;

Step (3) For the input image, use the mean shift algorithm (shift of the mean value) to segment the Gaussian-filtered image, merge small areas in the image, and merge as many similar parts in the background as possible;

Step (4) Use the UNICOM-detected area, combine the geometric features of the target in the image to eliminate part of the background area, and then initially obtain the area of interest before finally calculating the average significant value  $S(k)$  of each area ( $k$ );

Step (5) Compare  $S(k)$  with the aforementioned  $S_{\text{mean}}$ , and if  $S(k) > 2 * S_{\text{mean}}$ , keep the area;

Step (6) Use the target shape features to filter the regions, and then merge adjacent regions;

Step (7) Combine the original image to output the final candidate region.

The process of saliency detection is shown in Fig. 3. This method mainly uses the color and brightness information of the image in Lab space. In this method, for an input image, it is first filtered using a Gaussian filter kernel to remove some

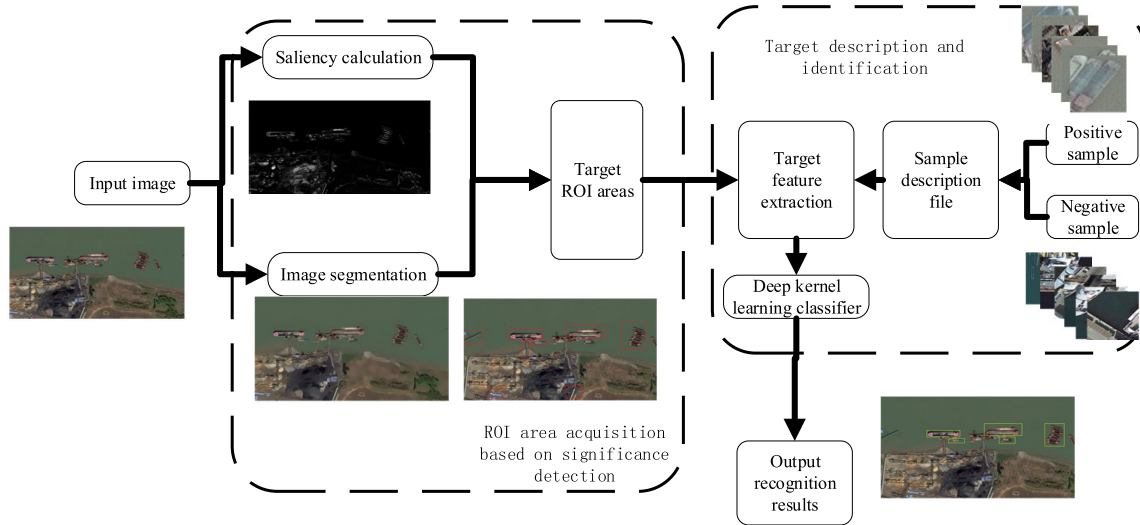


FIGURE 1. Closed remote sensing target recognition method based on deep saliency kernel learning analysis.

noise, and the significant value calculated by this method is shown in the following formula:

$$S(x, y) = \|I_\mu - I_{\omega hc}(x, y)\| \quad (1)$$

where  $S(x, y)$  is the significant value of the pixel in the image with coordinate  $(x, y)$ .  $I_\mu$  is the mean value of each channel of the image mapped to Lab space after Gaussian filtering, and its expression is:

$$I_\mu = \begin{bmatrix} L_\mu \\ a_\mu \\ b_\mu \end{bmatrix} \quad (2)$$

where  $L_\mu, a_\mu, b_\mu$  are average color values of the three channels L, a and b, respectively.  $I_{\omega hc}(x, y)$  is the description vector after the point  $(x, y)$  is mapped to Lab space, as shown in the formula:

$$I_{\omega hc}(x, y) = \begin{bmatrix} L_{\omega hc} \\ a_{\omega hc} \\ b_{\omega hc} \end{bmatrix} \quad (3)$$

The mean shift (shift of the mean value) that appears in the flow of Fig. 4 is essentially a clustering algorithm. It was first proposed by Fukunage in 1975. The mean shift algorithm used is now mostly an improved version of the method of Yizong Cheng *et al.* in terms of the kernel function and weight coefficient [1], [2]. The mean shift algorithm is a parameterless clustering algorithm for the feature space. Its calculation method essentially depends on a probability density estimation. This kind of clustering does not need the user to input the number of clusters, and there is no requirement on the shape of the cluster. Its feature space can be regarded as a type of its posterior probability density function. Within its unknown probability density, the mode corresponds to the higher density part, and the same cluster is composed of all data with the same mode. If there is some d-dimensional space Rd, given n sample points  $x_i(i = 0, 1, 2, 3, \dots, n)$ ,

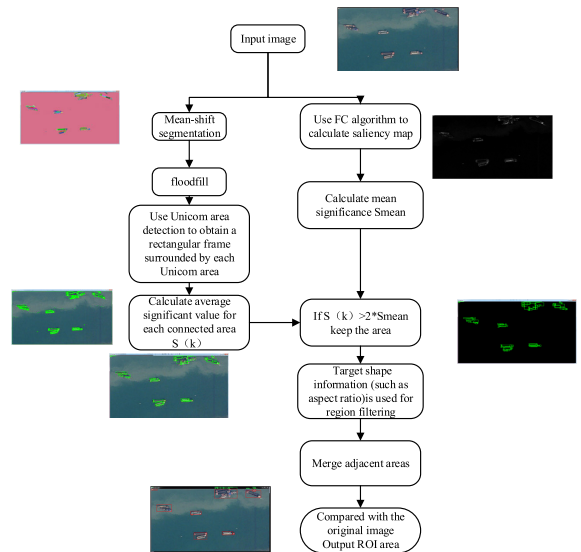


FIGURE 2. ROI extraction process.

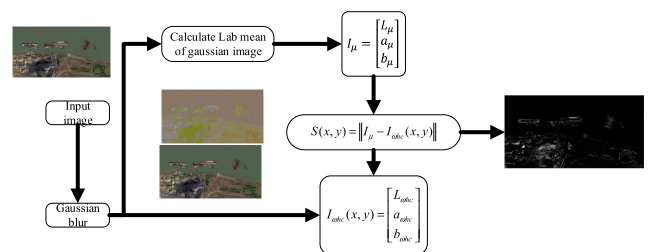


FIGURE 3. Acquisition of the saliency map.

then for the point x, its corresponding mean shift vector can be described as the form of equation 3-12:

$$M_h(x) = \frac{1}{k} \sum_{x_i \in S_h} (x_i - x) \quad (4)$$

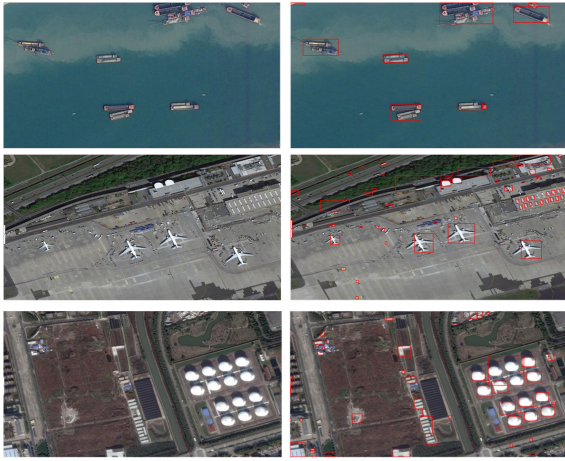


FIGURE 4. Schematic diagram of the ROI extraction results.

$S_h$  refers to a high-dimensional spherical area with a radius  $h$  and is defined as:

$$S_h(x) = \{y | (y - x)(y - x)^T \leq h^2\} \quad (5)$$

In actual situations, in the region of  $S_h$ , each point contributes differently to  $x$ . Therefore, to better describe this situation, the two concepts of the kernel function and sample weight are introduced into the mean shift vector. With the introduction of the kernel function, as the sample distance changes with the offset point, the influence of the offset on the mean shift vector also changes with the distance. The mean shift vector after introducing the kernel function concept and the sample weight parameter can be expressed as:

$$M_h(x) = \frac{\sum_{i=1}^n G_H(x_i - x)\omega(x_i)(x_i - x)}{\sum_{i=1}^n G_H(x_i - x)\omega(x_i)} \quad (6)$$

where  $G(x)$  is the introduced kernel function and  $\omega(x_i)$  is the introduced sample weight parameter.  $H$  is a positive definite bandwidth matrix, and the form of  $H$  is shown in equation 3-15:

$$H = \begin{pmatrix} h_1^2 & 0 & \dots & 0 \\ 0 & h_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & h_d^2 \end{pmatrix}_{d \times d} \quad (7)$$

Then, the mean shift vector introduced with the kernel function and weights described in equation 3-14 can be rewritten in the form of equation 3-16.

$$M_h(x) = \frac{\sum_{i=1}^n G(\frac{x_i - x}{h_i})\omega(x_i)(x_i - x)}{\sum_{i=1}^n G(\frac{x_i - x}{h_i})\omega(x_i)} \quad (8)$$

The mean shift vector can be essentially regarded as a regularized probability density gradient, and its gradient direction is the weighted average of the direction vectors of each data point. The mean shift algorithm uses the probability density gradient to obtain the local optimal solution in the sample points. Using the mean shift algorithm to solve the problem means that the problem must be converted into a density

estimation problem. For image processing, the image mainly contains two types of information: color and coordinates. The segmentation problem can be regarded as finding the class center for each pixel in the input image, and all points with the same class center can be regarded as a set of clusters. When segmenting with the mean shift, each point on the input image  $(x, y)$  can be regarded as a set of multidimensional data consisting of coordinates and RGB color values  $(x, y, r, g, b)$ , and the mean shift algorithm uses the window scan space to find the region with the highest density in the input multidimensional data. For one image, the spatial position of each point, that is, the range of the corresponding coordinates  $(x, y)$  in the image, is significantly different from the range of the RGB color values, so for these two dimensions, two different size windows can be used for clustering in the algorithm. When the mean shift window moves, all the points that converge to the same peak after the window transformation will form the same cluster and form the cluster under this peak. This relationship acts on the image to form the region segmentation effect of the image.

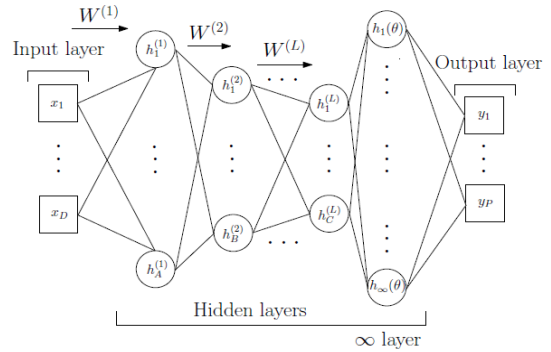


FIGURE 5. Network structure combining deep structure and kernel mapping [30].

C. STEP 2: TARGET RECOGNITION BASED ON THE DEEP KERNEL CLASSIFIER

Kernel regularization technology is used to achieve fast visual recognition based on a neural network, and the use of stochastic gradient descent effectively improves the accuracy and speed of visual recognition and provides a research framework for subsequent deep kernel learning. Therefore, before selecting a suitable deep kernel mapping structure, we investigated the corresponding theories of deep multicore mapping. During the accumulation and investigation of deep multicore mapping theory, after collating the data and literature, we mainly focused on three common methods from shallow core mapping to deep core mapping at this stage:

$$k(x_i, x_j | \theta) \rightarrow k(g(x_i, w), g(x_j, w) | \theta, w) \quad (9)$$

where  $g(x_i, w)$  is a nonlinear feature map obtained from the deep structure. In this way, the results of feature extraction from the deep structure are further processed by kernel mapping, and the performance of feature extraction is improved by combining the two methods. The network structure is shown Fig. 5 [30].

For the application of core mapping in support vector machines, there are also corresponding structural extensions of multilayer and multicore SVMs. The network structure extension from the multicore SVM to the multilayer multicore SVM is shown in Fig. 6.

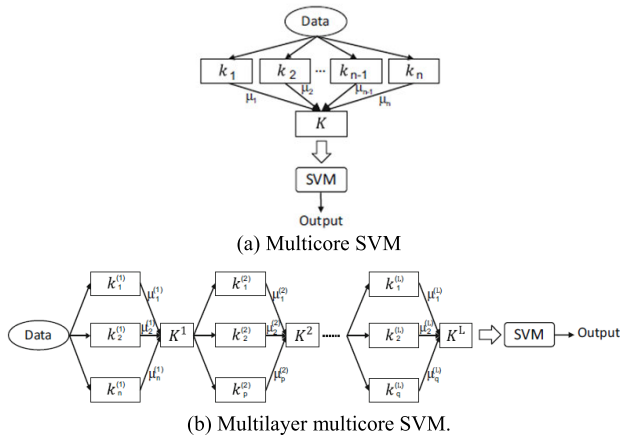


FIGURE 6. Schematic diagram of a Sallen-Key filter circuit.

First, the single-core mapping is extended to a linearly connected multicore structure, and then each single-core mapping is extended to a multicore mapping structure using the multilayer multicore structure to improve the feature extraction capability, thereby improving the classification result of the SVM. Although the multilayer multicore structure is also a deep structure, its composition is still linearly connected. The feature extraction structure of deep multicore mapping based on the network structure is shown in Fig. 7.

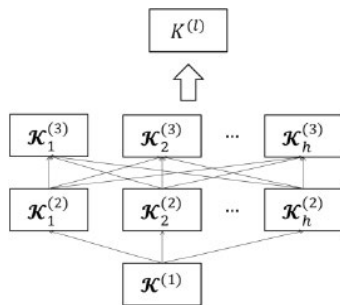


FIGURE 7. Basic deep multicore mapping structure.

In the deep multicore mapping structure above, the fully connected structure of the deep belief network is used as a reference. Each unit is a basic core. The basic cores in the layer are independent of each other. The basic cores between the layers are fully connected structures. Therefore, the weight coefficients of the basic kernel are also optimized in layer-by-layer optimization during training. The entire depth structure is functionally equivalent to an extension of kernel mapping, and the feature extraction performance has been improved.

To propose a deep multicore mapping structure suitable for the target feature extraction of remote sensing images, in the previous work, the basic deep multicore structure was introduced into the target feature extraction and recognition of remote sensing images. While verifying its adaptability, it was compared with the traditional single-core single-layer mapping with other deep learning structures. For single-layer single-core mapping feature extraction, the general method can use a combination of multiple basic cores to construct a single-layer multicore mapping to improve the feature extraction capabilities. The deep multicore mapping method can be regarded as the nesting of kernel functions to achieve multi-level expansion, and the kernel mapping expression formula for layer l is:

$$K^{(l)}(x, y) = \Phi^{(l)}(\dots \Phi^{(1)}(x)) \cdot \Phi^{(l)}(\dots \Phi^{(1)}(y)) \quad (10)$$

When the core mapping of each layer is expanded from a single core to a combination of multiple basic cores, a multilayer multicore mapping network is realized, and its expression formula is:

$$K^{(l)}(x, y) = \left\{ \theta_{1,1}^{(l)} K_{1,1}^{(l)} \left( \theta_{1,1}^{(l-1)} K_{1,1}^{(l-1)} + \dots \right) + \dots + \theta_{h,m}^{(l)} K_{h,m}^{(l)}(\dots) \right\} \quad (11)$$

$K_{h,m}^{(l)}$  represents the m-th basic kernel in the l-th row and h-th column, and  $\theta_{h,m}^{(l)}$  is the weight value corresponding to the basic core. The basic deep multicore structure used in this paper is a deep structure that adjusts the number of layers and the number of units. The basic deep multicore structure is similar to the deep confidence network in the connection method and uses full connections between units at different layers and no connection structure between the units in the same layer. The difference is that the deep belief network optimizes the network by optimizing the probability between the connected units, while each unit of the deep multicore structure consists of a basic core. Each basic kernel selects its type and internal parameters before training, and each iteration during the training process uses a layer-by-layer optimization strategy to obtain the optimal solution of the weights of each basic kernel. The weight update formula is as follows:

$$\theta_k^{t+1} \leftarrow \theta_k^t - \gamma_k \frac{\partial T_{Span}}{\partial \theta_k} \quad (12)$$

where  $\gamma$  is the learning rate,  $T_{Span}$  comes from the loss function, and the leave-one-out method is used to solve the loss function. The loss function formula is shown in equation 3-x:

$$L((x_1, y_1), \dots, (x_l, y_l)) \leq \sum_{p=1}^l \phi(\alpha_p^0 S_p^2 - 1) =: T_{Span} \quad (13)$$

Data  $x_p$  are mapped to a high-dimensional space through kernel mapping to obtain  $\Phi_{K_p}(x_p)$ , and  $S_p$

is the distance from point  $\Phi_{K_p}(x_p)$  to  $\Gamma_p$ ,  $\Gamma_p = \left\{ \sum_{i \neq p, \alpha_i^0 > 0} \lambda_i \Phi_{K_\theta}(x_i) \mid \sum_{i \neq p} \lambda_i = 1 \right\}$ .

The training process of deep multicore mapping feature extraction is as follows:

---

**Algorithm** Deep Multicore Mapping Feature Extraction Network Training Algorithm Process

---

1. Input: training data and label, set initial value of the learning

rate  $\gamma$  and initial value of the weight  $\theta_i^l = \frac{1}{m}$ , for  $i = 1, \dots, m$

2. for  $t = 1, 2, \dots$  execute

Use  $K^{(l)}(\theta^t)$  to optimize the deep kernel mapping structure

3. for  $k = 1, 2, \dots$  execute  $\theta_k^{t+1} \leftarrow \theta_k^t - \gamma_k \frac{\partial T_{Span}}{\partial \theta_k}$

4. When the cycle number or stop condition is reached, exit the cycle and complete the network training

---

### III. EXPERIMENTAL RESULTS

#### A. PUBLIC DATA SET VALIDATION

##### 1) EXPERIMENT 1: UC MERCED LAND USE DATA SET VALIDATION

The UC Merced Land Use data set is a 21-level remote sensing data set of land use images for research, with a total of 100 types of images extracted from the USGS National Map urban area image series and used in urban areas all over the country. The pixel resolution of the public domain image in this data set is 1 foot, and the pixel size of the image is  $256 * 256$ . There are 2100 scene images in 21 categories, including 100 in each category.

In the experiment, 80% of the UC Merced Land Use data set was used as the training set, and 20% was used as the test set. In the first experiment, no transfer learning method was used, and the remote sensing target recognition network was directly trained with a small data set. As seen from the figure, the accuracy of the test set is only 59%. For the direct training and learning of small data sets, the performance results of the convolutional neural network model are not ideal. The main reason is that the quantity of remote sensing target data sets is too small to meet the demand of the deep network. To solve the problem of small samples in deep learning networks, transfer learning is introduced in deep learning networks to improve the experimental results.

First, the ImageNet data set is used for pretraining. The ImageNet data set is quite large, and the targets in its images and the targets in remote sensing images have certain similarities in their nature, so the method of transfer learning can be used to optimize the deep convolutional network. The ImageNet data set is used to pretrain the original model of remote sensing target recognition, initialize its interlayer parameters, and retain the convolutional layer weights. Then, its loss layer and softmax layer are modified to suit our classification task, and then the UC Merced Land Use data set is used to fine-tune the loss layer and softmax layer. For the problem of insufficient data sets in the training process,

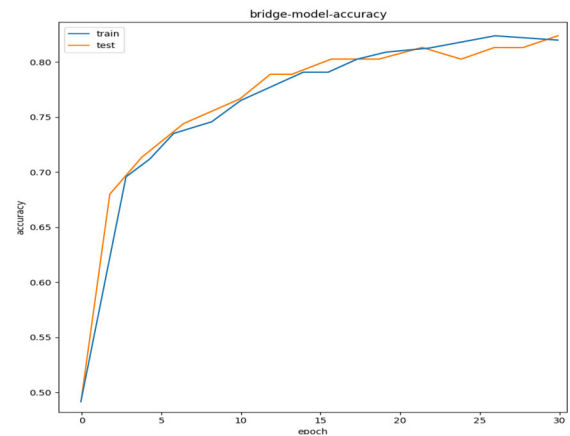


FIGURE 8. Accuracy of the data set before pretraining.

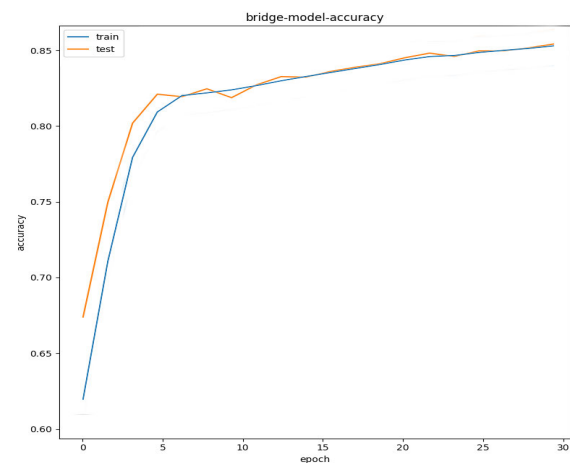


FIGURE 9. Accuracy of the data set after pretraining.

this approach expands the data set so that the number of labeled data is multiplied. The expanded data set is trained. Under the condition of other training parameters and the same environment, through many experiments on the target data set, the accuracy of target recognition is improved by 3 ~ 5 percentage points by expanding the data.

##### 2) EXPERIMENT 2: MSTAR DATA SET VALIDATION

This part of the experiment uses MSTAR as the experimental data. The experimental data use the measured SAR ground stationary target data published by the MSTAR program supported by the United States Defense Advanced Research Projects Agency (DARPA). The sensor collecting this data set is a high-resolution spotlight synthetic aperture radar with a resolution of  $0.3 \text{ m} \times 0.3 \text{ m}$ . It works in the X-band, preprocessing is performed on the collected data to extract sliced images with a pixel size of  $128 \times 128$  containing various types of targets. Most of these data are SAR slice images of stationary vehicles, including a variety of vehicle target images obtained at various azimuth angles. Some target images are shown in Fig. 10.

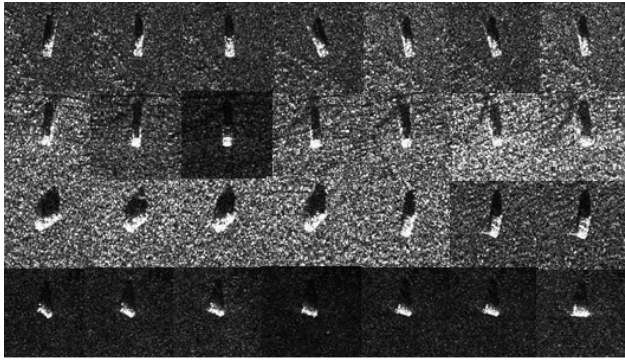


FIGURE 10. Partial SAR target data set image.

The deep belief network is used for feature extraction, and then the classifier is trained to obtain a classification model. The classification effect is tested using test data. The process is as follows:

- 1) Data preprocessing: The image data are converted into the corresponding input data form of a deep belief network.
- 2) Data reading: The input data and the labels for the data are read.
- 3) Model training: Network initialization and signal forward propagation is performed, and then the reverse feedback algorithm is used to train through iteration to obtain the training model.
- 4) The application effect of the extracted features in classification is verified on the test data.

In the SAR image classification experiment based on deep belief networks, the performance of the network extraction features is tested from three aspects and evaluated using two evaluation indicators: the operation time and classification error rate. The feature extraction effects of different layers of DBNs and different numbers of units in DBNs are shown in Table 1 and Table 2.

TABLE 1. Experimental results of SAR images using DBNs with different layers.

DBN layers	Number of units in each layer	Computing Time (h)	Error rate
2	500-2000	0.685	14%
3	500-1000-2000	1.069	6.5%
4	500-1000-1000-2000	1.011	7.5%
5	500-1000-1000-1000-2000	1.358	12.8%

It can be seen from the experimental results that the actual effects of deep belief networks are impacted by many factors, and the number of different hidden layers and the number of units affect the computing time and error rate. As the number of layers and the number of units in a layer increase, the computing time increases, and different network structures have

TABLE 2. Experimental results of SAR images using DBNs with different numbers of units.

Number of units in each layer	Computing time (h)	Error rate
500-1000-1000-2000	1.011	7.5%
500-500-500-2000	0.505	11.8%
1000-2000-2000-4000	2.723	15%
500-500-500-500	0.391	15.8%

TABLE 3. Comparison of the DBN with the single-kernel function extraction method.

Experimental method	Recognition error rate
4-layer DBN	7.5%
Polynomial-SVM	55.5%
RBF-SVM	16.5%
Sigmoid-SVM	26%

different classification error rates, but the classification error does not simply decrease with an increasing number of layers. Therefore, when the depth structure is introduced into the remote sensing image target application, the optimal structure and the corresponding optimization algorithm need to be analyzed. Compared with the shallow SVM based on a single kernel function, deep belief networks still present great improvements, which proves that the deep structure improves the performance of shallow kernel mapping in feature extraction. The results are shown in Table 3.

Through the collation and comparison of the experimental results, it can be shown that the deep structure greatly improves the feature extraction performance of the shallow mapping structure. Based on this result, the following work can improve the feature extraction performance of kernel mapping in remote sensing image targets through the deep structure.

In the experiment, the same SAR image target as the deep belief network experiment is used as the verification object. Ten types of tank targets at different angles are selected as the data set. The original pictures in the MSTR data set are preprocessed, and they are cropped to a consistent size. Then, 1000 images are selected as the training set, and 1000 images are selected as the test set. In the subsequent experiments, to test the effect of the number of samples in the data set on the feature extraction results, the other conditions remain unchanged, and a training set with 400 images and a test set with 400 images are used to perform a comparison test for the data objects.

Different data sets are constructed to verify the feature extraction effects in two modes and compare the target classification results after deep kernel mapping feature extraction with other common methods. First, the feature extraction effect is verified through classification tasks. The ordinary SVM only supports two classification methods, so different categories of targets are combined in pairs, and then a data



**TABLE 4.** Target classification results of deep kernel mapping.

Target category	Amount of training data	Amount of test data	Two-level classification results	Three-level classification results
2S1&BRDM_2	1000	1000	89%	92.9%
BRDM_2&ZSU234	1000	1000	89.7%	90.4%
2S1&BRDM_2	400	400	91.7%	92%
D7&T62	400	400	100%	100%
T62&2S1	400	400	85%	85.5%

**TABLE 5.** Target detection results of depth kernel mapping.

Detection target	Amount of training data	Amount of test data	Two-level classification results	Three-level classification results
2S1	1000	1000	80.5%	83.2%
BRDM_2	1000	1000	85.6%	85.2%
ZSU234	1000	1000	85.0%	89.2%
D7	1000	1000	65.3%	67%
ZIL131	1000	1000	68.5%	72%

set is constructed according to the combined structure. The effect of deep kernel mapping structure feature extraction on classification is verified under two results, and it can be shown that the deep structure greatly improves the feature extraction performance of the shallow mapping structure. Based on this result, the following work can improve the feature extraction performance of kernel mapping in remote sensing image targets through the deep structure.

In the experiment, the same SAR image target as the deep belief network experiment is used as the verification object. Ten types of tank targets at different angles are selected as the data set. The original pictures in the MSTR data set are preprocessed, and they are cropped to a consistent size. Then, 1000 images are selected as the training set and 1000 images as the test set. In the subsequent experiments, to test the effect of the number of samples in the data set on the feature extraction results, other conditions remain unchanged, and a training set with 400 images and a test set with 400 images are used to perform a comparison test for the data objects.

Different data sets are constructed to verify the feature extraction effect in two modes and compare the target classification results after deep kernel mapping feature extraction with other common methods. First, the feature extraction effect is verified through classification tasks. The ordinary SVM only supports two classification methods, so different categories of targets are combined in pairs and then a data set is constructed according to the combined structure. The effect of deep kernel mapping structure feature extraction on classification is verified under two different data set sizes. The results are shown in Table 4.

Through five groups of dichotomous tasks, the feature extraction results of deep kernel mapping are verified for different objects and different size data sets. The same set of network parameters performs slightly differently in the classification tasks of different targets, but the overall accuracy

**TABLE 6.** Comparison of classification results of deep kernel mapping and other algorithms.

RBF core SVM	4-level DBN structure	CNN structure	Deep multicore mapping
83.5%	88.2%	92.3%	95.9%

is improved with the deepening of the structure. An increase in the data volume of the data set improves the classification accuracy, but the computing time also increases. Therefore, in subsequent research, it will be necessary to optimize the data and network structure according to the operation speed and algorithm performance in practical applications. After verifying the results of the classification problem, the performance of the deep kernel mapping in monitoring is shown in Table 5.

There are differences in the detection accuracy among the five categories of targets, indicating that the differences between each category and other categories are dissimilar, and there are also differences in the performance of feature extraction for different targets. However, with an increase in the structure depth, the detection performance is improved. These findings show that the depth structure can improve the ability of kernel mapping in remote sensing image feature extraction, but for various specific targets and detection tasks, it is necessary to optimize the structural parameters to improve its performance.

Under the same computing resources and data objects, we compare the classification performance of the common methods of feature extraction in the target classification task. The support vector machine using the RBF core is selected in the single-core mapping, and a deep confidence network with a four-layer network structure is chosen. The convolutional network uses a common AlexNet model. The classification results of each method are shown in Table 6.



FIGURE 11. Some examples of experimental results.

By comparing the classification results of feature extraction based on deep kernel mapping with the classification results of other methods, it can be found that the deep multi-core mapping feature extraction has the best effect. Compared with ordinary single-core mapping algorithms, the accuracy rate is greatly improved, indicating that the deep structure can improve the feature extraction performance of the kernel mapping algorithm. This approach provides a basis for subsequent deep kernel mapping feature extraction research in this paper and optimizes the algorithm for the mapping structure and parameters based on the basic depth structure, the existing loss functions, and the number and types of kernel functions.

### B. VERIFICATION ON A SELF-BUILT DATA SET

To verify the performance of the algorithm on full-color (black and white), multispectral (color), and infrared and SAR satellite remote sensing images, a full-color, multi-spectral remote sensing image data set is employed. The statistics of the remote sensing images currently available are shown in Table 7, with a total of 760 categories and a resolution of 0.5 m. Collected port source data have a total of 1121 images containing targets of the type <Port> with a resolution of 0.5 m. The collected <Oil tank> target images are high-resolution images with a total of 900 images with a resolution of 0.5 m. The collected <Ship> target images are high-resolution images with a total of 533 images with

TABLE 7. Visible light remote sensing data set.

Resolution type	Target	Quantity (images)	Resolution
High resolution	Port	1121	0.5 m
	Oil tank	500	0.5 m
	Ship	780	0.5 m
	Aircraft	760	0.5 m
Low resolution	Airport	500	6 m or more
	Bridge	828	6 m or more

a resolution of 0.5 m. For the target type of <Airport>, a total of 500 instances of airport data with a resolution of 6 m or more were collected. For the target of <Bridge>, a total of 558 instances of source data were collected with a resolution of 6 m or more.

In supervised learning, the quality of the data labels is very important. High-quality labeled images can make the model achieve better detection results. For the aforementioned remote sensing image data, it is necessary to assign labels, generate a training set and send it to the network for training. The image annotation tool is used to circle the target position in the image, and a corresponding XML file for each image is generated. The XML file contains the coordinates of the target position, the label of the target, and other information.

Although we have obtained as much remote sensing data as possible, the amount of data is still slightly insufficient for

TABLE 8. Visible light remote sensing data set.

Target	Quantity (images)	Resolution	Training samples	Test samples	Recognition rate
Port	1121	0.5 m	800	200	85.5%
Oil tank	500	0.5 m	300	200	86.0%
Ship	780	0.5 m	500	200	82.5%
Aircraft	760	0.5 m	500	200	82.5%

deep learning-based training networks. To better enhance the training data, the data set is expanded in the following two ways. The lighting of the source data is changed, the data are rotated at different angles, and the label file is modified accordingly to accommodate the rotation.

The experimental results are shown in Fig. 11, and the statistical experimental results are shown in Table 8. The average recognition rate is 84.125%.

### C. DISCUSSION

After a comprehensive verification on public data sets and self-built data sets, the method proposed in this paper obtains the best recognition efficiency. Compared with the traditional RBF core SVM, DBN structure and CNN structure, the proposed method based on deep multicore mapping achieves the best performance under the same computing resource conditions. Therefore, it shows that this method has an advantage in achieving closed target recognition. At the same time, this method has no strict restrictions on the imaging methods of remote sensing images and has good performance for SAR and visible light imaging.

### IV. CONCLUSION

This paper proposes a remote sensing image target recognition method based on deep saliency kernel learning analysis that uses target region extraction based on the visual saliency mechanism and implements a nonlinear deep kernel learning saliency feature analysis method to realize target extraction and recognition. A deep kernel mapping architecture for remote sensing image classification is proposed by combining kernel mapping and deep learning, which solves the problems of the similarity degree of the input vector of the kernel mapping function, the structure of the kernel mapping function and the structure of the kernel mapping learning network. The proposed architecture can effectively improve the adaptability of the learning model to the target extraction, more accurately describe the data from the input space in terms of the nonlinear mapping relationship, and enable the data belonging to different classes to achieve better discrimination in the nonlinear mapping space. The experiments show that the features with higher resolution can be used to improve the target classification accuracy.

### REFERENCES

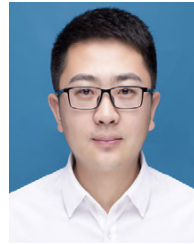
- [1] J. T. Al-Bakri and Y. Y. Al-Jahmany, "Application of GIS and remote sensing to groundwater exploration in Al-Wala basin in Jordan," *J. Water Resource Protection*, vol. 05, no. 10, pp. 962–971, 2013.
- [2] J. N. Sweet, "The spectral similarity scale and its application to the classification of hyperspectral remote sensing data," in *Proc. IEEE Workshop Adv. Techn. Anal. Remotely Sensed Data*, Oct. 2003, pp. 92–99.
- [3] A. Chang, Y. Eo, S. Kim, Y. Kim, and Y. Kim, "Canopy-cover thematic-map generation for military map products using remote sensing data in inaccessible areas," *Landscape Ecol. Eng.*, vol. 7, no. 2, pp. 263–274, Jul. 2011.
- [4] R. D. Hudson and J. W. Hudson, "The military applications of remote sensing by infrared," *Proc. IEEE*, vol. 63, no. 1, pp. 104–128, Jan. 1975.
- [5] S. Kodors, A. Rausis, A. Ratkevics, J. Zvirgzds, A. Teilans, and I. Ansons, "Real estate monitoring system based on remote sensing and image recognition technologies," *Procedia Comput. Sci.*, vol. 104, pp. 460–467, Jan. 2017.
- [6] D. Liu, L. He, and L. Carin, "Airport detection in large aerial optical imagery," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, pp. 761–764.
- [7] Y. Li, M. Li, F. Li, X. Sun, and W. Liu, "Real-time interactive object extraction system for high resolution remote sensing images based on parallel computing architecture," in *Proc. 18th Int. Conf. Geoinform.*, Jun. 2010, pp. 1–6.
- [8] P. Druyts, W. Mees, D. Borghys, C. Perneel, A. Marc, and J. L. Valero, "Semi-automatic help for aerial region analysis," in *Proc. SAHARA Project*, 1999, pp. 1–5.
- [9] B. Guindon, "Computer-based aerial image understanding: A review and assessment of its application to planimetric information extraction from very high resolution satellite images," *Can. J. Remote Sens.*, vol. 23, no. 1, pp. 38–47, Mar. 1997.
- [10] G. Fu, H. Zhao, C. Li, and L. Shi, "Road detection from optical remote sensing imagery using circular projection matching and tracking strategy," *J. Indian Soc. Remote Sens.*, vol. 41, no. 4, pp. 819–831, Dec. 2013.
- [11] S. D. Mayunga, D. J. Coleman, and Y. Zhang, "Semi-automatic building extraction in dense urban settlement areas from high-resolution satellite images," *Surv. Rev.*, vol. 42, no. 315, pp. 50–61, Jan. 2010.
- [12] R. Hulik, M. Spänel, P. Smrz, and Z. Materna, "Continuous plane detection in point-cloud data based on 3D Hough transform," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 86–97, Jan. 2014.
- [13] W. Wang, J. Sun, H. Rui, and S. Mao, "Knowledge-based bridge detection from SAR images," *J. Syst. Eng. Electron.*, vol. 20, no. 5, pp. 929–936, Oct. 2009.
- [14] X. Li, S. Zhang, X. Pan, P. Dale, and R. Cropp, "Straight road edge detection from high-resolution remote sensing images based on the ridgelet transform with the revised parallel-beam radon transform," *Int. J. Remote Sens.*, vol. 31, no. 19, pp. 5041–5059, Oct. 2010.
- [15] X. Xu, C. Liu, and X. Li, "Building damage detection based on single high-resolution remote sensing imagery," in *Proc. Int. Conf. Autom. Control Artif. Intell. (ACAI)*, Mar. 2012, pp. 618–621.
- [16] M. Barzohar and D. B. Cooper, "Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp. 707–721, Jul. 1996.
- [17] P. Zhong and R. Wang, "A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3978–3988, Dec. 2007.
- [18] B. Sirmacek and C. Unsalan, "Road detection from remotely sensed images using color features," in *Proc. 5th Int. Conf. Recent Adv. Space Technol. (RAST)*, Jun. 2011, pp. 112–115, doi: 10.1109/RAST.2011.5966802.
- [19] A. R. Zamir and M. Shah, "Image geo-localization based on MultipleNearest neighbor feature matching Using Generalized graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1546–1558, Aug. 2014.

- [20] B. Jin, Y. Cong, W. Zhou, and G. Wang, "A new method for detection of ship docked in harbor in high resolution remote sensing image," in *Proc. IEEE Int. Conf. Prog. Informat. Comput.*, May 2014, pp. 341–344.
- [21] W. Xin, B. Wang, and L. Zhang, "Airport detection in remote sensing images based on visual attention," in *Proc. Neural Inf. 18th Int. Conf. (ICONIP)*, 2011, pp. 475–484.
- [22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [23] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3446–3456, Sep. 2010.
- [24] Z. Li, D. Yang, and Z. Chen, "Multi-layer sparse coding based ship detection for remote sensing images," in *Proc. IEEE Int. Conf. Reuse Integr.*, Aug. 2015, pp. 122–125.
- [25] Z. Song, H. Sui, and Y. Wang, "Automatic ship detection for optical satellite images based on visual attention model and LBP," in *Proc. IEEE Workshop Electron., Comput. Appl.*, May 2014, pp. 722–725.
- [26] F. Yang, Q. Xu, F. Gao, and L. Hu, "Ship detection from optical satellite images based on visual search mechanism," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 3679–3682.
- [27] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.
- [28] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.
- [29] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2002, pp. 1–8.
- [30] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," in *Proc. Artificial Intelligence and Statistics (PMLR)*, May 2016, pp. 370–378.
- [31] S.-R. Zhou, J.-P. Yin, and J.-M. Zhang, "Local binary pattern (LBP) and local phase quantization (LBQ) based on Gabor filter for face representation," *Neurocomputing*, vol. 116, pp. 260–264, Sep. 2013.
- [32] M. Amrani, F. Jiang, Y. Xu, S. Liu, and S. Zhang, "SAR-oriented visual saliency model and directed acyclic graph support vector metric based target classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3794–3810, Oct. 2018.
- [33] M. Amrani and F. Jiang, "Deep feature extraction and combination for synthetic aperture radar target classification," *J. Appl. Remote Sens.*, vol. 11, no. 4, Oct. 2017, Art. no. 042616.
- [34] M. Amrani, K. Yang, D. Zhao, X. Fan, and F. Jiang, "An efficient feature selection for SAR target classification," in *Proc. Pacific Rim Conf. Multimedia*. Cham, Switzerland: Springer, 2017, pp. 68–78.



Senior Member of the Chinese Institute of Electronics.

**LONG SUN** was born in Anhui, China. He is currently pursuing the Ph.D. degree with the National Laboratory of Radar Signal Processing, Xidian University. He is also a Researcher-Level Senior Engineer with the 38th Research Institute, China Electronics Technology Group Corporation. His main research interests include new radar system technology, remote sensing military and civilian application technology, target characteristic analysis, and target recognition technology. He is a



the 38th Research Institute, China Electronics Technology Group Corporation. His research interests include deep learning, intelligent interpretation of remote sensing images, and autonomous driving. He is a Professional Member of the China Computer Federation.

**JIE CHEN** (Member, IEEE) received the Ph.D. degree in computer applied technology from the Hefei Institutes of Physical Science, Chinese Academy of Sciences, University of Science and Technology of China, in 2019. He is currently a Lecturer and a Master Supervisor with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Electronics and Information Engineering, Anhui University, China, and a Postdoctoral Fellow with



processing, array signal processing, communication signal processing, blind signal processing, and radar imaging technique.

**DAZHENG FENG** (Member, IEEE) received the M.S. degree from Xi'an Jiaotong University, Xi'an, China, in 1986, and the Ph.D. degree in electronic engineering from Xidian University, Xi'an, in 1996. He worked with the National Laboratory of Radar Signal Processing, Xidian University, where he is currently a Professor. He has published more than 100 journal articles. His research interests include adaptive signal processing, intelligence and brain information processing,



radar (SAR) signal processing. His research interests include SAR, inverted synthetic aperture radar (ISAR), sparse signal processing, and microwave remote sensing. He serves as an Associate Editor for radar remote sensing of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

**MENGDAO XING** (Fellow, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 1997 and 2002, respectively. He is currently a Professor with the National Laboratory of Radar Signal Processing, Xidian University, where he holds the appointment of the Associate Dean of the Academy of Advanced Interdisciplinary Research. He has authored or coauthored more than 200 refereed scientific journal articles and two books about synthetic aperture

• • •