

Received August 15, 2019, accepted August 27, 2019, date of publication September 3, 2019, date of current version October 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939232

Polarimetric SAR Data Classification via Reinforcement Learning

MIN WANG¹, ZHIYI WANG¹, CHEN YANG², SHUYUAN YANG², AND YUTENG GAO³

¹Key Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China

²School of Artificial Intelligence, Xidian University, Xi'an 710071, China

³Electrical Engineering Department, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding author: Shuyuan Yang (syyang2009@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61771380, Grant U1730109, Grant 91438103, Grant 61771376, Grant 61703328, Grant 91438201, Grant U1701267, and Grant 61703328, in part by the Equipment Pre-Research Project of the 13th Five-Years Plan under Grant 6140137050206, Grant 414120101026, Grant 6140312010103, Grant 6141A020223, Grant 6141B06160301, and Grant 6141B07090102, in part by the Major Research Plan in Shaanxi Province of China under Grant 2017ZDXM-GY-103 and Grant 017ZDCXL-GY-03-02, in part by the Foundation of the State Key Laboratory of CEMEE under Grant 2017K0202B and Grant 2018K0101B, in part by the Science Basis Research Program in Shaanxi Province of China under Grant 16JK1823 and Grant 2017JM6086.

ABSTRACT Inspired by human's learning characteristic that knowledge is gradually learned little by little, a Spatial-Polarimetric Reinforcement Learning (SPRL) approach is proposed for Polarimetric Synthetic Aperture Radar (PolSAR) data classification, from a new perspective of reinforcement learning. In our method, each pixel has its own "state" and "action", and can modify its "action" based on interactions with the "environment". A spatial-polarimetric "reward" function, is designed from a local neighborhood region to explore both the spatial and polarimetric information for more accurate classification. Thus a self-evolution and model-free classifier can be obtained, which has simple principle and robustness to speckle noises existed in the data. By an interaction with the environment, SPRL can obtain high classification accuracy when only very few labeled pixels are available. Several real PolSAR datasets are used to investigate the effectiveness of the proposed method, and the results show that SPRL is superior to its counterparts.

INDEX TERMS Polarization synthetic aperture radar (PolSAR), data classification, continuous learning, spatial-polarimetric reward, reinforcement learning, environment.

I. INTRODUCTION

Recently Polarimetric Synthetic Aperture Radar (PolSAR) data classification has received increasing research interests. It aims to assign labels for image pixels to indicate their categories [1]. The available PolSAR data classification methods can be mainly classified as unsupervised [2], [3] and supervised [4], [5] ones. Unsupervised approaches work for extracting discriminative features of pixels to cluster them into some predefined categories. The H/α decomposition [6], complex Wishart features [7], model-based polarization decomposition [8], are often adopted in these methods. Supervised approaches form their predictions via a learned mapping from a set of labeled pixels. Various kinds of mapping models, such as support vector machines [9], neural networks [10] and deep learning models [11]–[13] are usually employed. In order to reduce the labeling cost of the

supervised methods, some semi-supervised methods are also developed [14], [15], which explore both labeled and unlabeled samples, to achieve accurate classification with very few labeled data. Most of them utilize unlabeled samples to formulate a Laplacian regularizer, which is of high computation and storage complexity.

Although many approaches have proved their efficiency in PolSAR data classification, most of them belong to traditional unsupervised paradigm and supervised paradigm where the classification is one-off. In other words, pixels are clustered or classified in a forward way, where the labels of pixels are independently predicted and never to be changed afterwards. However, this process is conflict with the biological decision making. In the human's learning, learning is a long-term process and knowledge is gradually learned little by little. Just like the human's decision making, machine classification should also have different status and evolution capability or improve little by little, to achieve a credible and efficient learning of data.

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

A recent trend in machine learning is to integrate naturalistic learning in biological species into learning, such as continuous learning, self-learning and reasoning [16]. Human intelligence indicates a learning agent interacts with a dynamic environment and updates its action policies to maximize long-term rewards. Inspired by it, Reinforcement Learning (RL) has been developed as another paradigm in machine learning [17]. Reinforcement learning can mimic the biological development that at any given moment, a biology receives an array of internal and external stimuli that forms its *state*, with a *behavior* dictating the reaction to each of them. Biology is the *agent* following a temporary behavior, i.e., *action*; stimuli constitutes the *plant* or *system*; the reaction is represented by a numerical feedback called the *reward* [18]. Reinforcement learning refers to an actor or agent that interacts with its environment and modifies its actions, based on stimuli received in response to its *action* [19]. It emphasizes the modification of actions based on interactions with the environment, which has potential applications in data classification [20], [21].

In this paper, we introduce the reinforcement learning into PolSAR data classification, to construct a dynamic and continuous learning classifier based on Spatial-Polarimetric Reinforcement Learning (SPRL). In our work, instead of forming a learned mapping that indicates the correct output for a given input, the labeled pixels are assumed to provide only an indication as to whether an *action* is correct or not, according to a local neighborhood region in the PolSAR images. Moreover, it is well known that PolSAR data are not only polarimetric reflectivities, but also images with certain kinds of spatial organization. Therefore, making full use of both the polarization and spatial information can help to improve the classification accuracy. In the proposed SPRL method, a spatial-polarimetric reward function is designed. Thus, the classifier will label pixels according to their current *states*, and evolves them from a spatial and polarimetric *reward* function, which is derived from the local neighborhood *environment*. By the interaction with *environment*, our proposed SPRL method can correctly predict labels of pixels with low computational complexity.

Compared with the available PolSAR data classification methods, SPRL has the following characteristics: 1) A self-evolution and model-free classifier is proposed under the new paradigm of reinforcement learning, to classify the PolSAR data with very few labeled samples. 2) A spatial - polarimetric reward function is designed from a local neighborhood region. The function can reduce the influence of speckle noises existed in the data on the classification, so resulting in accurate prediction. Several experiments are taken on some real PolSAR datasets, and the results show that SPRL can achieve accurate and rapid classification with a small number of labeled samples. The Overall Accuracy (OA), Average Accuracy (AA) and Kappa Coefficients (KC) of the classification results by SPRL are compared with that of its counterparts, and the results show its superiority.

II. SPATIAL-POLARIMETRIC REINFORCEMENT LEARNING

A. POLSAR DATA CLASSIFICATION VIA REINFORCEMENT LEARNING

In the PolSAR image, each pixel can be expressed by a scattering matrix $\mathbf{S}$:

$$\mathbf{S} = \begin{bmatrix} S_{vv} & S_{vh} \\ S_{hv} & S_{hh} \end{bmatrix} \quad (1)$$

where the subscript represents the transmitted and received polarization mode. For the reciprocity [22], here $S_{hv} = S_{vh}$. Based on the Pauli basis, the polarimetric scattering vector with three unique elements can be expressed as

$$\mathbf{u} = \frac{1}{\sqrt{2}} [S_{vv} + S_{vv}, S_{vv} - S_{hh}, 2S_{vh}]^T \quad (2)$$

Then, we can obtain the polarimetric coherency matrix $\mathbf{T}$:

$$\mathbf{T} = \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^H / n \quad (3)$$

where n is the number of looks, $\mathbf{u}_i$ is the i -th look polarimetric scattering vector and H represents the complex conjugate transpose. Then $\mathbf{T}$ can be written as a 3×3 complex matrix:

$$\mathbf{T} = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix} \quad (4)$$

where $T_{ij} = T_{ji}^H, i \neq j$, and it is considered to contain complete polarimetric information of one pixel [23].

Given a dataset $\mathbf{S} = \mathbf{L} \cup \mathbf{U}$, where $\mathbf{L} = [\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_l]$ is the labeled set and $\mathbf{U} = [\mathbf{T}_{l+1}, \mathbf{T}_{l+2}, \dots, \mathbf{T}_{l+u}]$ is the unlabeled set. The PolSAR data classification aims to classify each pixel in the image into some category that belongs to $\{1, 2, \dots, C\}$, where C is the number of classes, and l, u are the number of labeled and unlabeled samples respectively. In our method, we explore $\mathbf{S}$ to construct a spatial-polarimetric reinforcement classifier, to realize a semi-supervised classification. The flowchart of our proposed SPRL method is shown in Fig.1.

SPRL can be defined by a tuple (S, A, P, R) , where S, A, P, R are the state set, action set, action probability set and reward set respectively. Then we model the environment as a Markov Decision Process (MDP) consisting of a set of discrete states $S = \{s_1, s_2, s_3, \dots, s_u\}$ and actions $A = \{a_1, a_2, a_3, \dots, a_C\}$ that may be executed from each state (C represents the number of actions, u represents the number of unlabeled samples). In our method, the unlabeled pixels of the image is predicted one by one. We denote the "State" s_i as the classification status with a specific set of unlabeled samples. The "Action" refers to assigning labels for one unlabeled sample, so the number of actions is equal to that of categories. The "Probability" is corresponding to the appear probability of each action in the action space, with $P = \{p_1, p_2, \dots, p_C\}$ and $\sum_i p_i = 1$. The "Reward" includes a variety of rewards that take an action from spatial and polarimetric information in some state. The reward function is denoted as $R = \{r_1, r_2, \dots, r_C\}$. Different with the classifier

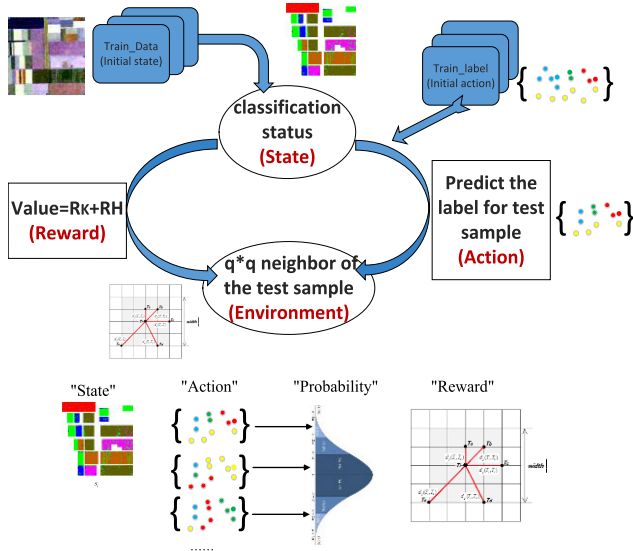


FIGURE 1. Flowchart of our proposed spatial-polarimetric reinforcement classifier.

working in a closed set, SPRL can achieve a self-evolution classification via an interaction with the environment. In the evolution of the classifier, executing action a from state s causes a transition to state s' according to a *transition function* describing $P(s' | s, a)$, and triggers a reward r according to a *reward function*, $r = R(s, a, s')$. Subsequently we can obtain a state-action sequence $s_0 \rightarrow a_0 \rightarrow r_1 \rightarrow s_1 \rightarrow a_1 \rightarrow r_2 \rightarrow s_2 \rightarrow a_2 \dots$ by an interaction with the environment via a reasonable reward function.

B. SPATIAL AND POLARIMETRIC REWARD

In the reinforcement learning, the cumulative return reward is very crucial. Polarimetric SAR data have both spatial consistency and rich polarimetric information. Considering the spatial consistency of polarimetric SAR data in a local region, we explore both the spatial and polarimetric information of data. A $q \times q$ squared neighborhood region of unlabeled sample is used to define a reward function. In order to predict the label of a pixel that locates at the i^{th} row and j^{th} column in the image, we consider its polarimetric coherency matrix $\mathbf{T}_{i,j}$ and its spatial neighbors $\{\mathbf{T}_{m,n}\}$ in the $q \times q$ region ($m \in [i - q/2, i + q/2]$, $n \in [j - q/2, j + q/2]$). In our method, we define a spatial reward function $\mathbf{R}_K$ generated by actions in the action space (labeling the sample as any category) as:

$$\mathbf{R}_K = [k_1, k_2, \dots, k_C], \quad k_l = 2^{g_l} \quad (5)$$

where g_l represents the number of occurrences of the label l in the spatial neighborhood $l \in \{1, 2, \dots, C\}$ at the current state. In our work we initialize each k_l as zero. If there is no labeled samples, the number of occurrences is set as zero. Thus, k_l is an occurrence indicator that represents the reward generated by the label l on the spatial neighborhood. This reward function indicates the return of labeling the pixel as different classes, which is determined by the distribution of

labels in the neighborhood. On the other hand, we define the similarity function between the pixel and its neighbors as:

$$w(\mathbf{T}_{i,j}, \mathbf{T}_{m,n}) = \sum_{r=1}^3 \sum_{t=1}^3 |\mathbf{T}_{i,j}(r, t) - \mathbf{T}_{m,n}(r, t)| \quad (6)$$

Denote the similarity indicator as v_l ($l = 1, 2, \dots, C$) and initialize each v_l as zero. The similarity function is then calculated on a neighborhood region, with

$$v_l = \sum_{m,n} w(\mathbf{T}_{i,j}, \mathbf{T}_{m,n}) \quad \text{if } \text{label}(\mathbf{T}_{m,n}) = l \quad (7)$$

Then another polarimetric reward function $\mathbf{R}_H$ is calculated by various actions in the action space as to polarimetric coherency information:

$$\mathbf{R}_H = [H_1, H_2, \dots, H_C] \quad (8)$$

where

$$H_l = 1 \quad \text{if } v_l < 1 \quad (9)$$

Combining the spatial reward with the polarimetric reward together, we can construct an immediate reward function:

$$r_s^a = \mathbf{R}_K + \mathbf{R}_H \quad (10)$$

which indicates a reward of executing action a from state s . The agent typically has no prior knowledge of the transition and reward functions, and must learn value estimates iteratively as it explores the environment. In order to make the agent's objective to maximize reward, a state value function in state s under a policy π is defined as,

$$v_{t+1}^\pi(s) = \sum_{a \in A} (r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_t(s')) \quad (11)$$

where $\gamma \in [0, 1]$ is a discount factor controlling the agent's degree of preference for immediate rewards over future ones. $P_{ss'}^a$ is the state transition probability from s to s' by executing action a . $v_t(s')$ is the state value under state s' .

Given a small number of labeled pixels and a large number of unlabeled pixels, each unlabeled pixel has C possible actions. As soon as the unlabeled pixel is assigned a specific label, it is defined as the state of the pixel. Executing an action from a state will cause a transition to another state according to a transition function and trigger a reward via a synthetic reward function. In this process, the states of unlabeled samples will update iteratively, which realizes a self-evolution labeling of pixels. At each iteration, we select one action from the action set to assign a label for the current pixel, by maximizing the state value function. The algorithm can be executed iteratively to label all the pixels in the image. That is, when the labels of all the pixels in the image remain unchanged, this iterative process will converge. A detailed description of the spatial-polarimetric reinforcement learning is shown in Table 1.

TABLE 1. Procedure of the SPRL based PolSAR data classification.

Input: Labeled pixels set L ; unlabeled set U ; Discount factor γ ; neighbor size q ; State transition probability $P_{ss'}^a$ from s to s' ;
Initialize: $v_l = 0, k_l = 0$ ($l = 1, 2, \dots, C$);
Repeat the following steps For each unlabeled pixel
1) Calculate the number of occurrences g_l ($l = 1, 2, \dots, C$) in its $q \times q$ spatial neighborhood;
2) Calculate the occurrence indicator $k_l = 2^{g_l}$;
3) Calculate the spatial reward function R_k according to (5);
4) Calculate the similarity indicator v_l in its $q \times q$ spatial neighborhood according to (6) and (7);
5) Calculate the polarimetric reward function R_u according to (8);
6) Calculate the immediate reward function according to (10);
7) Calculate the state value $v_{l+1}^x(s)$ according to (11);
8) Update its label via $\arg \max_s (r_s^x + \gamma \sum_{s' \in S} P_{ss'}^a v_l(s'))$;
Until the labels of unlabeled pixels remain changed

III. EXPERIMENTAL RESULTS

In this section, we test the performance of SPRL on four PolSAR datasets: Flevoland, San Francisco Bay Subimage, Flevoland1 (in the Netherlands) and Xi'an (in China) datasets. Some related methods are used for a comparison, including: 1) some baseline classifiers—K Nearest Neighbor (KNN) [24], Support Vector Machine (SVM) [9] and Wishart [7]), 2) some related and state-of-the-art methods—Semisupervised Spatial-spectral SVM (SS-SVM) [25], Deep Reinforcement Learning (DRL) [26], and 3) some popular method – Convolutional Neural Network (CNN) [13]. For KNN, we determine the optimal number of nearest neighbors k by employing five-fold cross-validation in the range of [1,30]. In SVM and SS-SVM, the Radial Basis Function (RBF) and the one-against-one strategy are used for classification. The width of RBF is determined by five-fold cross-validation in the range of $[1 \times 10^{-5}, 1 \times 10^1]$. The spatial regularizer in SS-SVM is experimentally determined in the set of {0.1 1 10 100} via cross-validation. In DRL and WSDN, the optimal parameters suggested in the literature are adopted. In SPRL we set the same probability in the state transition, that is, $P_{ss'}^a = 1/C$. In the test, different neighborhood size $3 \times 3, 5 \times 5, 7 \times 7, 9 \times 9$ are used, and we find that SPRL can present the best result when the size takes the smallest value. So in our method we set $q = 3$. The value of the discount factor also varies from 0 to 10, with an interval of 0.1. The optimal parameter is also determined via cross-validation. The optimal parameter $\gamma = 1$ is adopted. In the test, we randomly select 0.1% samples from the image, and the rest are taken as the unlabeled samples. In SS-SVM, 20% unlabeled samples are also randomly selected for the semisupervised classification. The Overall Accuracy (OA), Average

Accuracy (AA) and Kappa Coefficient (KC), are used for evaluating the classification results of different methods. Thirty independent experiments are taken and the average accuracies of networks are calculated. All the experiments are realized on a 64GB RAM HP Z840 workstation with dual E5-2630v CPUs and a NVIDIA GeForce GTX TITAN X GPU.

A. RESULTS OF THE FLEVOLAND DATASET

The Flevoland dataset has the size of 750×1024 , which is an L-band four-look fully polarimetric SAR data. This image has a resolution of 12×6 m acquired by NASA/JPL AIRSAR (National Aeronautics and Space Administration/Jet Propulsion Laboratory Airborne Synthetic Aperture Radar), and contains fifteen categories, as shown in the color bar at the lower right corner of Fig.2.

The Pauli RGB pseudo-colored image of this dataset is shown in Fig.2(a), and its ground truth is shown in Fig.2(b). The classification results of KNN [24], SVM [9], Wishart [7], SS-SVM [25], CNN [13], DRL [26] and SPRL for one experiment, are shown in Fig.2(c)-(i) respectively. From the results in Fig.2(c) and Fig.2(d) we can see that the traditional supervised methods (including KNN and SVM) give low classification accuracy for almost all the classes for the very limited number of labeled samples. Wishart classifier can present better results than KNN and SVM, for considering the distribution of polarimetric feature and employing the Bayesian classifier, as shown in Fig.2(e). Because the inter-class differences of this dataset are very small, and the extremely few labeled samples, SS-SVM also suffers from many speckle-like misclassification, as shown in Fig.2(g). Because there are not sufficient labeled samples for training, CNN also produce a lot of errors in Fig.2(g). DRL can generate amounts of valid data by interacting with the agent using the ϵ -greedy strategy, so the classification accuracy is improved remarkably, as shown in Fig.2(h). Compared with these methods, our proposed SPRL can achieve the highest OA, by introducing a spatial-polarimetric reward and continuous trial and error.

To further investigate the efficiency of our proposed method, thirty independent experiments are taken under the same condition, and the average classification results of each class, along with the OA, AA and KC, are shown in Table 2. From the classification results shown in the table we can see that SVM and SS-SVM present relatively worse results than the other methods, because there are not enough training samples to determine a correct boundary. The classification accuracy of DRL is the highest among the comparative methods, for incrementally select the most valid samples via reinforcement learning. However, DRL gives very low accuracy as to the Wheat2 and Building classes, perhaps because there are less samples in these classes compared with other categories. Compared with them, SPRL presents better numerical results as to OA, AA and KC for each class than the other methods.

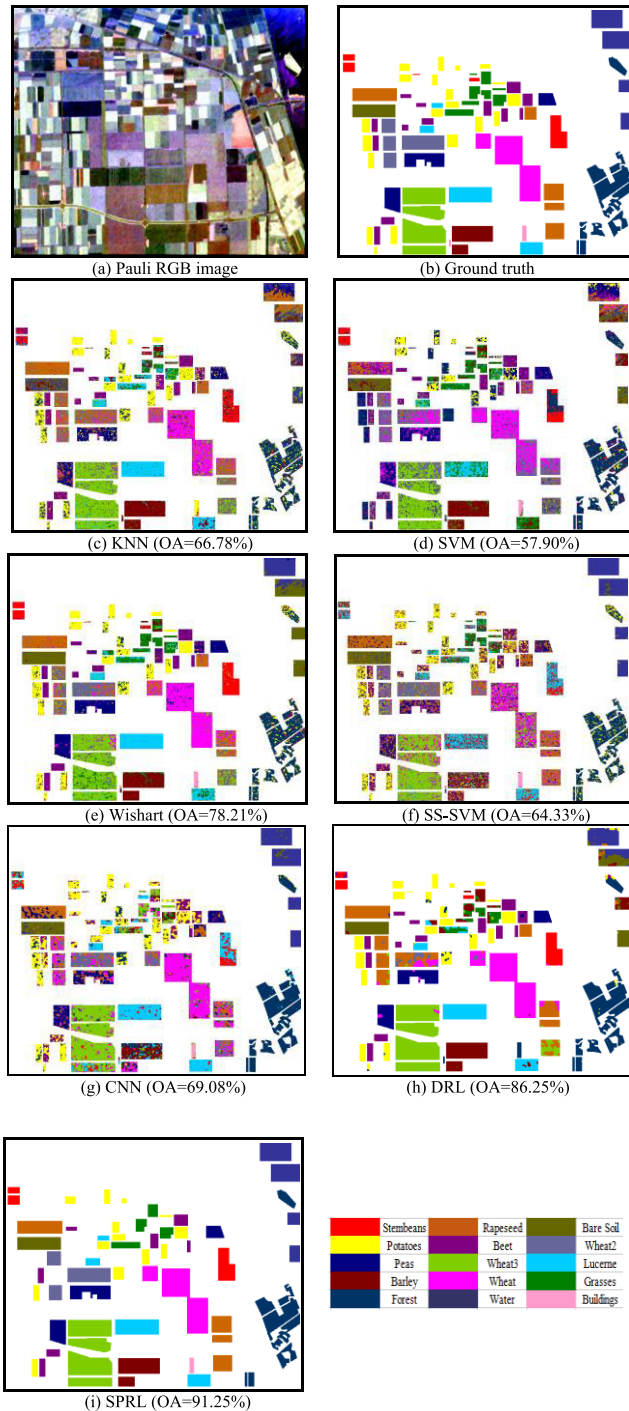


FIGURE 2. Classification results of the flevoland dataset from AIRSAR. (a) Pauli RGB. (b) Ground truth. (c) KNN. (d) SVM. (e) Wishart. (f) SS-SVM. (g) CNN. (h) DRL. (i) SPRL.

B. EXPERIMENTAL RESULTS OF THE SAN FRANCISCO BAY DATASET

This San Francisco Bay image is an L-band four-look fully polarimetric SAR data with four categories. The size of this image is 1300×1300 . The Pauli RGB pseudo-colored image of this dataset is shown in Fig.3(a), and its ground truth is shown in Fig.3(b). The classification results of KNN [24],

TABLE 2. Classification results of flevoland dataset from AIRSAR.

Method Class	KNN	SVM	Wishart	SS-SVM	CNN	DRL	SPRL
Stem beans	67.07	39.84	93.57	37.79	39.20	93.57	93.40
Rapeseed	58.63	34.93	62.72	82.92	46.81	83.59	89.85
Bare Soil	44.78	78.86	97.90	66.46	84.86	98.33	99.30
Potatoes	79.70	64.02	86.83	54.39	74.01	96.60	66.60
Beet	79.17	55.44	84.09	62.85	50.35	98.83	59.90
Wheat2	36.38	57.97	60.67	46.65	40.68	11.85	84.26
Peas	74.58	51.17	93.40	76.09	63.09	96.53	99.75
Wheat3	82.46	64.38	77.14	59.24	80.34	99.19	99.62
Lucerne	86.68	43.19	86.33	63.43	92.57	94.72	99.36
Barley	94.31	25.06	90.60	39.61	46.08	99.82	99.57
Wheat	63.77	53.74	83.13	44.10	76.28	97.73	99.82
Grasses	34.94	52.74	74.40	47.85	13.89	49.70	76.61
Forest	77.64	73.47	80.45	74.20	96.59	97.47	89.35
Water	28.82	96.86	47.81	96.14	95.92	48.15	99.86
Building	0.00	76.70	82.97	79.89	57.03	0.00	84.74
OA(%)	66.63	58.71	78.08	63.80	68.98	84.31	90.05
AA(%)	60.59	57.89	80.13	62.12	63.85	77.74	89.47
KC	0.6363	0.5502	0.7622	0.6053	0.6607	0.8290	0.8925

TABLE 3. Classification results of san francisco bay subimage.

Method Class	KNN	SVM	Wishart	SS-SVM	CNN	DRL	SPRL
Ocean	99.58	99.56	98.74	99.50	99.74	99.82	99.91
Vegetation1	89.24	78.63	92.34	85.07	82.79	98.45	98.33
I-Urban	39.85	38.18	77.19	58.70	61.12	55.83	99.07
NI-Urban	85.44	84.31	54.07	89.94	95.69	98.69	99.78
Vegetation2	52.59	51.77	54.10	62.37	84.99	80.80	99.88
OA(%)	87.85	85.86	84.63	89.90	92.19	94.78	99.65
AA(%)	73.34	70.49	75.29	79.12	84.87	86.72	99.39
KC	0.8151	0.7845	0.7715	0.8356	0.8829	0.9205	0.9948

SVM [9], Wishart [7], SS-SVM [25], CNN [13], DRL [26] and SPRL are shown in Fig.3(c)-(i) respectively. There are five typical categories in this dataset: Ocean, Vegetation1, Vegetation2, Intensive Urban (I-Urban), Non-Intensive Urban (NI-Urban). Among them, the LD urban class has high resemblance to the HD urban class, which increases the difficulty of classification. For the number of land covers decreases, the total classification accuracy of all the methods are better than that of the Flevoland dataset.

As can be seen from Fig.3, the seven methods can classify ocean correctly, but have low classification accuracies as to intensive urban areas and non-intensive urban, especially the Wishart classifier. KNN and SVM have different degrees of misclassification. SS-SVM is superior to SVM for incorporating unlabeled samples in the classifier. DRL and SPRL can achieve high OA for this dataset. The classification accuracy of SPRL is 99.65%, which is higher than the other six methods. The average classification results of different methods are shown in Table 3, which also proves the superiority of our proposed method to its counterparts.

C. EXPERIMENTAL RESULTS OF THE FLEVLAND1 DATASET

The Flevoland1 dataset from RADARSAT-2 has the resolution of 10×5 m, which was obtained at fine quad-mode with 30° incident angles of the airborne sensor in 2008. The size of this subregion is 1200×950 , which has four types of terrains: Urban, Water, Forest and Cropland areas, as shown in the color bar at the lower right corner

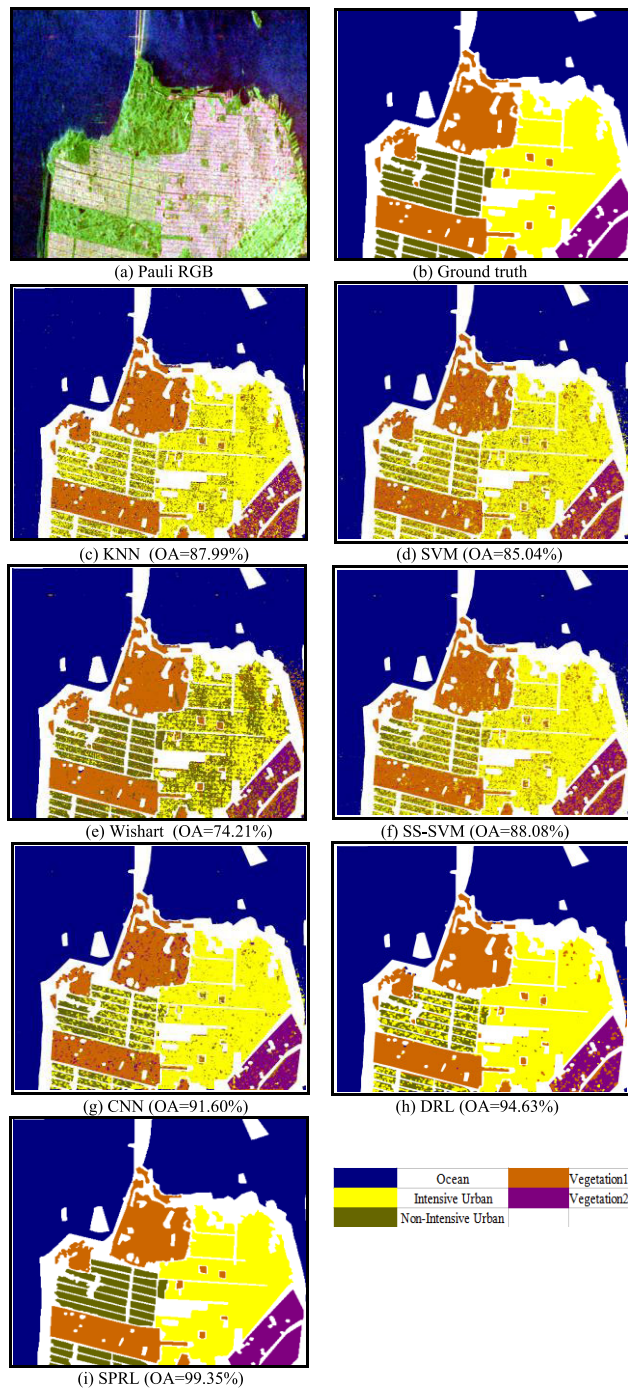


FIGURE 3. Classification results of the San Francisco Bay subimage. (a) Pauli RGB. (b) Ground truth. (c) KNN. (d) SVM. (e) Wishart. (f) SS-SVM. (g) CNN. (h) DRL. (i) SPRL.

of Fig.4. The results of KNN [24], SVM [9], Wishart [7], SS-SVM [25], CNN [13], DRL [26] and SPRL are shown in Fig.4(c)-(i) respectively. From the results we can see that the seven methods present relatively better results for Water. For the urban class, KNN, SVM, SS-SVM and Wishart produce some confusion with the forest. Compared with other methods, CNN achieves more accurate classification for the urban class. Also, we can observe from Fig.4(h) and Fig.4(i)

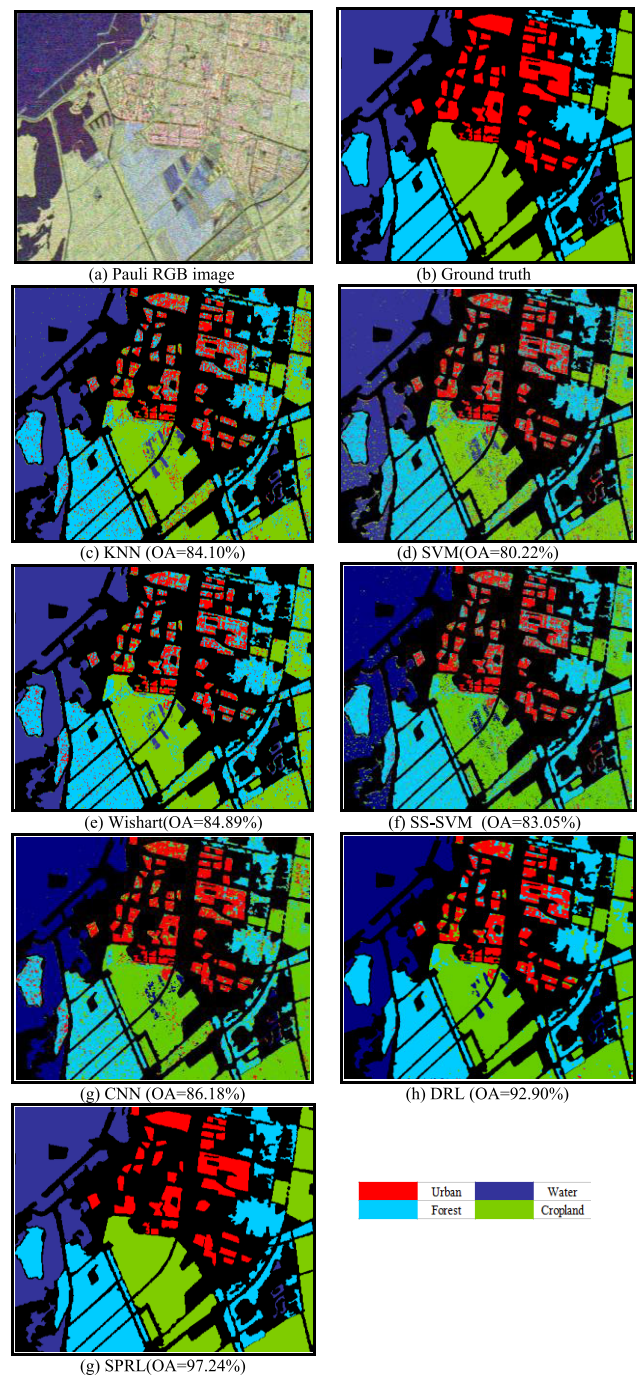


FIGURE 4. Classification results of the dataset of flevoland1. (a) Pauli RGB. (b) Ground truth. (c) KNN. (d) SVM. (e) Wishart. (f) SS-SVM. (g) CNN. (h) DRL. (i) SPRL.

that reinforcement learning can improve the classification accuracy. Even with only 0.1% samples, DRL can achieve high OA (92.90 %), AA(91.02%) and KC(0.9026). For constructing a spatial-polarimetric reward function, our proposed SPRL method has an improvement over DRL (OA is 97.24%, AA is 96.16% and KC is 0.9606).

The average classification results of different methods are shown in Table 4. From it we can also observe an

TABLE 4. Classification results of flevoland1 (in the netherlands) from RADARSAT-2.

Method Class	KNN	SVM	Wishart	SS-SVM	CNN	DRL	SPRL
Urban	60.96	63.07	59.14	63.40	79.89	77.32	89.72
Water	98.72	93.85	98.78	96.89	98.48	99.82	98.93
Forest	83.56	82.81	87.29	87.18	73.87	94.57	97.08
Cropland	81.31	75.64	81.11	78.35	89.59	92.37	99.78
OA(%)	83.81	80.82	84.56	83.26	86.08	92.90	97.39
AA(%)	81.14	78.84	81.58	81.46	85.46	91.02	96.38
KC	0.7785	0.7379	0.7881	0.7660	0.8102	0.9026	0.9647

TABLE 5. classification results of the xi'an (in china) dataset.

Method Class	KNN	SVM	Wishart	SS-SVM	CNN	DRL	SPRL
River	87.90	64.54	94.52	87.46	75.93	89.96	72.60
Urban	83.49	65.43	82.23	72.05	85.17	86.14	93.04
Vegetation	82.18	83.48	83.06	92.08	85.18	85.46	85.10
OA(%)	83.50	74.25	84.49	84.30	83.71	85.38	86.03
AA(%)	84.52	71.15	86.60	83.86	82.09	87.19	83.58
KC	0.7296	0.5742	0.7482	0.7362	0.7317	0.7209	0.7703

improvement of 10% over the comparative methods, as to the overall accuracy.

D. EXPERIMENTAL RESULTS OF THE XI'AN (IN CHINA) DATASET

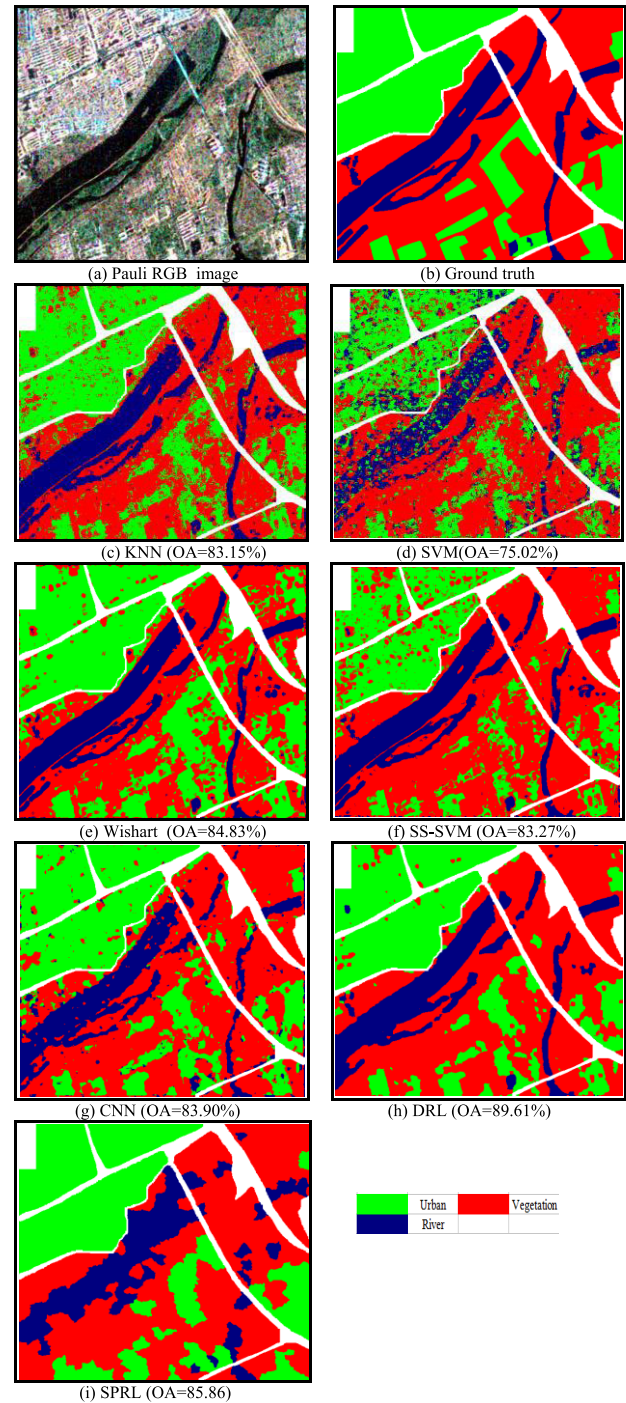
A full polarimetric image of city Xi'an obtained by the Canadian RADARSAT-2 system, with the size 512×512 and resolution 10×5 m, is used. Three categories are labeled in this Xi'an image: Urban, Vegetation and River, as shown in the color bar at the lower right corner of Fig.5. The results of KNN [24], SVM [9], Wishart [7], SS-SVM [25], CNN [13], DRL [26] and SPRL are shown in Fig.5(c)-(i) respectively. In this dataset, the intra-class difference is very large for the coarse labeling of three kinds of land covers.

From Fig.5 we can observe that SVM and CNN produce many misclassification on the river and vegetation classes. Our proposed SPRL makes some misclassification of the river class, for the cumulated errors in the self-evolution.

The average classification accuracy for each class are listed in Table 5. From it we can see that our method have the best classification results in OA, AA and Kappa.

E. RESULTS WITH DIFFERENT NUMBER OF LABELED SAMPLES

In this test, we use different number of labeled sample to evaluate the performance of SPRL on the above four datasets: Flevoland, XI'AN, San Francisco Bay and Flevoland1. The ratio of labeled samples takes 0.01%, 0.05%, 0.1%, 0.5% and 1%. In this test we take some methods with comparative result with that of SPRL for a comparison, including CNN and DRL. The variations of OA with the ratio is calculated and shown in Fig.6, where the result of SPRL, CNN and DRL are denoted as solid line, dotted line and dashed line respectively. From it we can see that OA increases with the increase

**FIGURE 5.** Classification results of the xi'an (in china) dataset. (a) Pauli RGB. (b) Ground truth. (c) KNN. (d) SVM. (e) Wishart. (f) SS-SVM. (g) CNN. (h) DRL. (i) SPRL.

of the proportion of labeled samples. However, the growth rate of the guidelines becomes slow with the increase of the ratio. For the Flevoland1, San Francisco Bay, Flevoland2 and Xi'an datasets, the OA can reach high OAs, even when 0.1% samples are labeled. By a comparison of the three methods, we can observe that our proposed SPRL present better results at different labeling ratios.

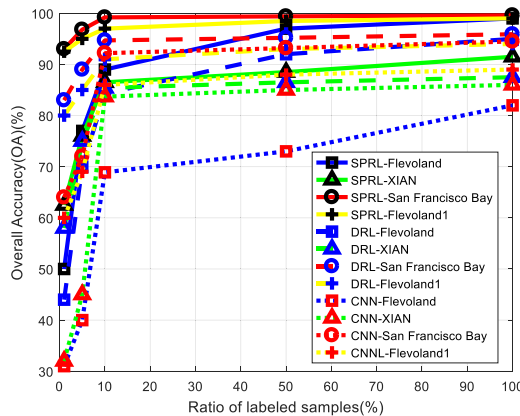


FIGURE 6. Variation of accuracy with different ratio of label samples on different datasets.

IV. CONCLUSION

In this paper, a self-evolution and model-free classifier is proposed under the new paradigm of reinforcement learning, to classify the PolSAR data with very few labeled samples. Compared with the available methods, SPRL has the characteristics of continuous learning and model-free. By utilizing the spatial-polarimetric reward function, SPRL can simultaneously make full use of both the spatial and polarimetric information to achieve accurate classification. Several experiments are performed on some real PolSAR datasets that contain various kinds of land covers, and the results show that it can achieve state-of-the-art classification results with very small number of training samples. However, for the squared shape of local neighborhood region defined in the reward function, the proposed method prone to perform better for images with regular boundaries. In future we will work for constructing an adaptive local region in exploring the spatial-polarimetric information.

REFERENCES

- [1] P. Lanchantin, J. Lapuyade-Lahorgue, and W. Pieczynski, "Unsupervised segmentation of triplet Markov chains hidden with long-memory noise," *Signal Process.*, vol. 88, pp. 1134–1151, May 2008.
- [2] S. Chen, S. Guo, Y. Li, and W. Hong, "Unsupervised classification for hybrid polarimetric SAR data based on scattering mechanisms and Wishart classifier," *Electron. Lett.*, vol. 51, no. 19, pp. 1530–1532, Sep. 2015.
- [3] F. Xue, Y. Lin, W. Hong, S. Chen, and W. Shen, "An improved H/α unsupervised classification method for circular PolSAR images," *IEEE Access.*, vol. 6, pp. 34296–34306, 2018.
- [4] S. L. Durden, J. J. van Zyl, and H. A. Zebker, "Modeling and observation of the radar polarization signature of forested areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 27, no. 3, pp. 290–301, May 1989.
- [5] S. Zhang, B. Hou, L. Jiao, Q. Wu, C. Sun, and W. Xie, "Context-based max-margin for PolSAR image classification," *IEEE Access.*, vol. 5, pp. 24070–24077, 2017.
- [6] P. Lombardo, T. M. Pellizzeri, and A. Tomasuolo, "Classification of polarimetric SAR images of suburban areas using joint annealed segmentation and H/A/α decomposition," in *Proc. IEEE/ISPRS Joint Workshop Remote Sens. Data Fusion Over Urban Areas*, Nov. 2001, pp. 117–121.
- [7] J.-S. Lee, M. R. Grunes, T. L. Ainsworth, L.-J. Du, D. L. Schuler, and S. R. Cloude, "Unsupervised classification using polarimetric decomposition and the complex Wishart classifier," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 5, pp. 2249–2258, Sep. 1999.
- [8] S.-W. Chen, X.-S. Wang, S.-P. Xiao, and M. Sato, "General polarimetric model-based decomposition for coherency matrix," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1843–1855, Mar. 2014.
- [9] S. Fukuda and H. Hirose, "Support vector machine classification of land cover: Application to polarimetric SAR data," in *Proc. IGARSS*, Jul. 2001, pp. 187–189.
- [10] W. Xie, Z. Xie, F. Zhao, and B. Ren, "POLSAR image classification via clustering-WAE classification model," *IEEE Access.*, vol. 6, pp. 40041–40049, 2018.
- [11] L. Zhang, W. Ma, and D. Zhang, "Stacked sparse autoencoder in PolSAR data classification using local spatial information," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 9, pp. 1359–1363, Sep. 2016.
- [12] L. Jiao and F. Liu, "Wishart deep stacking network for fast polar image classification," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3273–3286, Jul. 2016.
- [13] Y. Zhou, H. Wang, F. Xu, and Y.-Q. Jin, "Polarimetric SAR image classification using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1935–1939, Dec. 2016.
- [14] B. Wei, J. Yu, C. Wang, H. Wu, and J. Li, "PolSAR image classification using a semi-supervised classifier based on hypergraph learning," *Remote Sens. Lett.*, vol. 5, no. 4, pp. 386–395, 2014.
- [15] H. Liu, X. Xing, S. Wang, Z. Feng, E. Zhang, S. Yang, B. Hou, and L. Jiao, "Fast semi-supervised classification based on parallel auction graph for polarimetric SAR data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Beijing, China, Jul. 2016, pp. 1528–1531.
- [16] A. Agung and F. Gaol, "Game artificial intelligence based using reinforcement learning," *Procedia Eng.*, vol. 50, no. 9, pp. 555–565, 2012.
- [17] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *IEEE Trans. Neural Netw.*, vol. 9, no. 5, p. 1054, Sep. 1998.
- [18] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: An introduction," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 285–286, 2005.
- [19] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1238–1274, 2013.
- [20] A. S. Polydoros and L. Nalpantidis, "Survey of model-based reinforcement learning: Applications on robotics," *J. Intell. Robot. Syst.*, vol. 86, no. 2, pp. 153–173, 2017.
- [21] T. Gabel and M. Riedmiller, "On a successful application of multi-agent reinforcement learning to operations research benchmarks," in *Proc. IEEE Int. Symp. Approx. Dyn. Program. Reinforcement Learn.*, Apr. 2007, pp. 68–75.
- [22] F. T. Ulaby and C. Elachi, *Radar Polarimetry for Geoscience Applications*. Norwood, MA, USA: Artech House, 1990, p. 376.
- [23] C.-T. Chen, K.-S. Chen, and J.-S. Lee, "The use of fully polarimetric information for the fuzzy neural classification of SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 9, pp. 2089–2100, Sep. 2003.
- [24] L. Yang, S. Yang, P. Jin, and R. Zhang, "Semi-supervised hyperspectral image classification using spatio-spectral Laplacian support vector machine," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 3, pp. 651–655, Mar. 2014.
- [25] M. Tao, F. Zhou, Y. Liu, and Z. Zhang, "Tensorial independent component analysis-based feature extraction for polarimetric SAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2481–2495, May 2015.
- [26] K. Huang, W. Nie, and N. Luo, "Fully polarized SAR imagery classification based on deep reinforcement learning method using multiple polarimetric features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, to be published. doi: 10.1109/JSTARS.2019.2913445.



MIN WANG received the B.S. degree in auto control and the M.S. and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 2000, 2003, and 2005, respectively, where he is currently an Associate Professor with the National Laboratory of Radar Signal Processing. His research interests include intelligent signal and image processing and statistical learning.



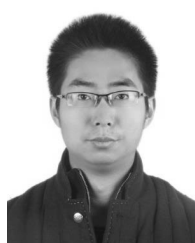
ZHIYI WANG received the B.S. degree in electrical engineering and the M.S. degree in electronics and communication engineering from Xidian University, Xi'an, China, in 2009 and 2014, respectively, where he is currently pursuing the Ph.D. degree. His research interests include intelligent signal processing and image processing.



SHUYUAN YANG received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees in circuit and system from Xidian University, Xi'an, China, in 2000, 2003, and 2005, respectively. Her research interests include intelligent signal processing, machine learning, and image processing. She is an IET Senior Member.



CHEN YANG received the B.S. degree from the Xi'an University of Technology. He is currently pursuing the M.S. degree with the School of Artificial Intelligence, Xidian University. His research interests include machine learning and image processing.



YUTENG GAO received the B.S. and M.S. degrees from Northwestern Polytechnical University, in 2010 and 2013, respectively, where he is currently pursuing the Ph.D. degree with the School of Electronics and Information. His research interests include antenna design and radar imaging.

...