Progress In Electromagnetics Research B, Vol. 83, 111–128, 2019

Land Cover Classification for Polarimetric SAR Image Using Convolutional Neural Network and Superpixel

Yilu Ma, Yuehua Li*, and Li Zhu

Abstract—The classification algorithms of polarimetric synthetic aperture radar (PolSAR) images are generally composed of the feature extractors that transform the raw data into discriminative representations, followed by trainable classifiers. Traditional approaches always suffer from the hand-designed features and misclassification of boundary pixels. Following the great success of convolutional neural network (CNN), a novel data-driven classification framework based on the fusion of CNN and superpixel algorithm is presented in this paper. First, the region-based complex-valued network utilizes both the intensity and phase information to predict the label of each pixel and constructs the label map based on spatial relations. Second, superpixel generating algorithm is adopted to produce the superpixel representation of the Pauli decomposition image, and the contour information which reflects the boundary of each category is preserved. Finally, the original label map and contour information are fused to make the decision of each pixel, outputting the final label map. Experimental results on public datasets illustrate that the proposed method can automatically learn the intrinsic features from the PolSAR image for classification purpose. Besides, the fusion of the superpixel features can effectively correct the misclassification of the boundary and singular pixels, thus achieving superior performance.

1. INTRODUCTION

As synthetic aperture radar (SAR) could operate in all-weather conditions and generate high resolution images, it has been applied in diverse applications from vehicle recognition [1] to sea ice monitoring [2, 3], and agricultural crop identification [4]. In the last decade, land cover classification using PolSAR images has attracted increasing attention of scholars and experts, and many classification schemes for PolSAR images have been reported. Based on the complex Wishart distribution, Lee and Grunes [5] developed a Wishart classifier for multi-look PolSAR data, which is a milestone in the research of PolSAR data classification. Gao et al. [6] proposed modified mixture models to reduce the modeling error of heterogeneous regions and designed two maximum likelihood classifiers to improve the overall accuracy of land cover classification. Classifiers based on the assumption of other distributions, such as Guassian distribution [7], K-distribution [8], and G-distribution [9] had also been reported. Chen [10] et al. presented a method based on polarimetric scattering similarity, which used the major and minor scattering mechanisms to increase the recognition rate. Recently, Wang et al. [11] presented a segmentation method for the fully polarimetric synthetic aperture radar (PolSAR) data by coupling the cluster analysis in the tensor space and the Markov random field (MRF) framework. In [12], Shang and Hirose proposed a Poincare-Sphere-Parameter space based quaternion neural network for land classification, which can be used for complicated terrains.

All of these previously mentioned approaches performed classification and identification on per-pixel basis. The class label to a pixel was completely predicted independently, and the influences

Received 21 November 2018, Accepted 8 February 2019, Scheduled 1 March 2019

* Corresponding author: Yuehua Li (nlglyh2013@sina.cn).

The authors are with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.

of the local neighborhood information had been ignored. It is generally known that polarimetric scattering information extracted from the PolSAR image is affected by speckle noise, which could deteriorate the results. Therefore, region-based methods are advantageous because they are able to reduce the computation demand by working on regions instead of pixels, help the optimization procedure converge more effectively to the global solution, and alleviate problems with noisy imagery by using region statistics instead of individual pixel values [13]. In [14], Cao et al. proposed the agglomerative hierarchical clustering technique, which segmented the image into a number of regions by clustering over a polarimetric decomposition data space and then produced the final segmentation labels. Wu et al. [15] first segmented the image into square regions and then utilized a region-based Wishart MRF framework to classify the image samples. Yu et al. [13] incorporated region growing model and an MRF edge strength model, and proposed a region-based unsupervised segmentation and classification algorithm. Another PolSAR classification technique based on multilayer auto-encoder was proposed in [16], which used RGB images to produce superpixels to integrate contextual information of neighborhood.

Convolutional neural network (CNN) has been well known in computer version areas [17] and achieved a number of breakthroughs in image classification task [18, 19]. Li et al. [19] designed a customized convolutional neural network with shallow layers to classify lung image patches with interstitial lung disease (ILD). Several studies had adapted this effective approach to SAR imagery interpretation and achieved superior performances in target recognition and terrain surface classification. In [20], Chen et al. presented an all convolutional neural network (A-CNN) for SAR target recognition, which only utilized convolutional layers to avoid overfitting. This method achieved an outstanding accuracy of 99% for MSTAR 10-class datasets. In [21], Zhou et al. employed a deep CNN to achieve 92% recognition accuracy for PolSAR image classification on Flevoland benchmark data, which was considerably better than that of the previous state of the art. A drawback of most of the traditional SAR image classification methods is that the researchers only consider about amplitude information. However, phase information is unique to SAR image, and it is a significant component for SAR image classification task. Hirose [22] pioneered that complex-valued neural networks can be more effective in various areas, such as image reconstruction and land-surface classification, as both the amplitude and phase information can be included in the complex-valued data, thus providing more details for processing. In [23], Zhang et al. proposed a complex-valued CNN (CV-CNN) for PolSAR image classification and achieved better recognition accuracy on Flevoland benchmark data than conventional classification methods.

Although the above methods have achieved considerable results, several problems remain to be concerned when performing the SAR image classification task. 1) For the region-based methods, the label map is constructed by the identification of the patches, which are usually extracted by a sliding window. Thus the misclassification of the patches located at the boundary, which consist of several categories, always degrades the recognition accuracy. 2) Different from the traditional classification tasks, such as face recognition and vehicle identification, it is observed that PolSAR terrain image patches are more texture-like that have no distinct structures. Hence deep layers in CNN could cause the extracted features to be too abstract. 3) Successful supervised training of ConvNets usually needs large amount of labeled data. Therefore, overfitting is another potential problem when training large neural network with many parameters, especially for PolSAR image classification with a limited number of training samples.

Hence, a new fusion model, based on the CV-CNN and superpixel segmentation, is presented for PolSAR image classification in this paper. Rather than defining a set of features, hierarchic representations of SAR image patches are automatically constructed by complex-valued convolutional layer and complex-valued deconvolutional layer. The complex-valued coherency matrix, which contains both amplitude and phase information, is taken as input of the proposed network. In particular, we incorporate random neural node drop-out and shallow convolutional layer architecture to reduce the number of parameters in the CNN model to avoid the over-fitting problem. On the other hand, the superpixel generating algorithm is applied to the Pauli decomposition image to generate the oversegmentation representation, and the contour information of each region is preserved. Then, the contour information is adopted to correct the singular pixels and the misclassified pixels near the boundary by the majority vote of each region. A threshold value is used to control the strength of the optimization. The experimental results show that the proposed neural network achieves considerable classification accuracy with features learned automatically from PolSAR image. Besides, the contour extracted from the superpixel representation is able to preserve the boundary of each class. Thus, the fusion model effectively reduces the misclassification of the boundary pixels and improves the recognition accuracy.

The remainder of this paper is organized as follows. Section 2 introduces preliminaries about PolSAR data and the components used in the recognition framework. The subsequent Section 3 formulates detailed process of the proposed fusion recognition framework, including data preprocessing, specific architecture of the neural network, and fusion operation. Section 4 presents the experimental results and analysis on the PolSAR datasets, including Flevoland dataset, Flevoland benchmark dataset, Oberpfaffenhofen dataset and San Francisco dataset. Finally, the discussion and conclusion are given in Sections 5.

2. RELATED WORK

The related theories used in our work are briefly introduced in this section, including the definition of the complex-valued input, the introduction of building blocks used in the network, and the superpixel generating algorithm.

2.1. Pauli Decomposition of PolSAR Data

The PolSAR characterizes the targets with fully polarized radar wave. In the natural image, each pixel in the PolSAR is considered as the interactions of correlated coherent interference process. Usually, each resolution cell of the basic single-look complex format is represented as a 2 × 2 complex scattering matrix.

$$S = \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix} \tag{1}$$

where S_{HH} denotes the scattering coefficient of horizontal transmitting and horizontal receiving, S_{HV} the scattering coefficient of horizontal transmitting and vertical receiving, and S_{VV} the scattering coefficient of vertical transmitting and vertical receiving [23]. In particular, S_{HV} and S_{VH} are assumed identical in the monostatic case.

According to [26], S can be represented as:

$$S = a\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + b\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + c\frac{1}{\sqrt{2}}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{2}$$

where

$$a = \frac{(S_{HH} + S_{VV})}{\sqrt{2}}, \quad b = \frac{(S_{HH} - S_{VV})}{\sqrt{2}}, \quad c = \sqrt{2}S_{HV} \tag{3}$$

The polarimetric scattering information can be regarded as a 3-D scattering vector $\boldsymbol{k}$, and $\boldsymbol{k}$ can be represented as:

$$\mathbf{k} = \tfrac{1}{\sqrt{2}}[S_{HH} + S_{VV} \quad S_{HH} - S_{VV} \quad 2S_{HV}]^T \tag{4}$$

Then the coherency matrix of PolSAR data is obtained as follows:

$$T = \mathbf{k}\mathbf{k}^H = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix} = \begin{bmatrix} 2|a|^2 & 2ab^* & 2ac^* \\ 2a^*b & 2|b|^2 & 2bc* \\ 2a^*c & 2b^*c & 2|c|^2 \end{bmatrix} \tag{5}$$

where $\cdot^H$ denotes the conjugation transpose operation.

For the multilook case, $T = \frac{1}{n}\sum\limits_{i=1}^{n}\mathbf{k}_i\mathbf{k}_i^H$, and n is the number of looks. It is obvious that the coherency matrix T is a Hermitian matrix whose diagonal elements are real numbers, while off-diagonal elements are complex-valued. Elements $\{T_{11}, T_{12}, T_{13}, T_{22}, T_{23}, T_{33}\}$ in the upper triangular part of the coherency matrix are adopted as the input samples.

2.2. Definition of the Related Neural Network

Convolutional): In a typical convolution process, the input data convolve multiple learnable filters (convolutional kernels) in parallel. The convolutional results are fed to nonlinear activation functions, such as sigmoid or rectified linear unit (ReLU), to generate the same number of feature maps as convolutional filters. The hidden units in the same feature map share identical weights, and they can be regarded as feature extractors, which detect specific features at each position of the previous layer. So, each feature map represents a unique feature at a different position of the previous layer. The details of the complex-valued convolutional process can be presented as follows.

In the convolution layer, the previous layer's input feature maps $O_i^{l-1}(i = 1, \dots I)$ are connected to all output feature maps $O_j^l(j = 1, \dots, J)$, where $O_i^{l-1}(x, y)$ is the unit at the position (x, y) of the i-th input feature map, and $O_j^l(x, y)$ is the unit at the position (x, y) of the j-th output feature map. Then, $O_j^l(x, y)$ can be calculated by:

$$O_j^l(x,y) = g(V_j^l(x,y)) \tag{6}$$

$$V_j^l(x,y) = \sum_{i=1}^{I} f_{ji}^l * O_i^{l-1}(x,y) + b_j^l = \sum_{i=1}^{I}\sum_{v=0}^{F-1}\sum_{u=0}^{F-1} f_{ji}^l(u,v) \cdot O_i^{l-1}(x-u, y-v) + b_j^l \tag{7}$$

where $g(\cdot)$ denotes the nonlinear activation function, $V_j^l(x, y)$ the translated feature map after filters, F the size of the filter, f_{ji}^l the bank of filters (convolutional kernel), and b_j the bias of the j-th output feature map. The hyperparameters in a convolutional layer include the number of feature maps (J), filter size (F), stride (S), and zero-padding (P). The stride means the intervals to apply the filters to the input image. Zero-padding is a common technique to preserve the spatial size of the feature maps. If the input is composed of I feature maps with size $W_1 \times H_1$, then the output will be J feature maps with size $W_2 \times H_2$, where $W_2 = (W_1 - F + 2P)/S + 1$ and $H_2 = (H_1 - F + 2P)/S + 1$. A common view of the convolution filters is that higher layer tends to have more feature maps than the lower layer, and stride 1 and using small filters usually leads to a better performance [20].

Deconvolution): The deconvolutional layer uses a decoder-only model to synthesize the feature maps [24]. Considering a single deconvolutional network layer applied to feature maps O^l which is composed of J channels $(O_1^l, \dots, O_J^l)$, and each channel can be represented as a linear sum of K_1 latent feature maps z_k convolved with filters f_{kj}:

$$O_j^l = \sum_{k=1}^{K_1} z_k * f_{kj} \tag{8}$$

If the size of O_j^l is $W \times H$ and the filters is of size $F \times F$, then the latent feature maps are $(W + F - 1) \times (H + F - 1)$ in size. A sparsity regularization term is introduced on z_k to find the unique resolution for the under-determined system, and the overall cost function can be presented as:

$$L(O^l) = \frac{\lambda}{2}\sum_{j=1}^{J}\left\|\sum_{k=1}^{K_1} z_k * f_{kj} - O_j^l\right\|_2^2 + \sum_{k=1}^{K_1}|z_k|^p \tag{9}$$

where J represents the number of input feature maps, and K_1 represents the number of latent feature maps. $|\cdot|^p$ denotes the sparse p-norm on the vectorial version of the matrix, and $p = 1$ typically. λ is the tradeoff that balances the contribution of reconstruction and the sparsity. Unlike traditional neural network, the filters are learnt by alternately optimizing the filters and feature maps: keep the filters fixed when minimizing $L(O^l)$ over feature maps and keep the feature maps fixed when minimizing $L(O^l)$ over filters.

Softmax): When dealing with multiclass classification problems, the softmax classifier composed of one or more fully connected layers is often used in the final output layer. The feature maps are flatted to one-dimensional vector before feeding to the fully connected layer at first, and the output of a fully connected layer is a $C \times 1$ vector, where C represents the number of classes. The softmax function used to perform the complex-valued multiclass logistic regression is represented as:

$$S_i = \frac{\mathrm{e}^{a_i}}{\sum_c e^{a_c}} \tag{10}$$

where a_i represents the i-th element in the vector, and S_i represents the probability of i-th class. Given a labeled training dataset of N samples $\{(x_i, y_i), i = 1, \ldots, N\}$, where y_i is the true complex-valued label. Then, the cross-entropy loss function is represented as:

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log S_i \tag{11}$$

In complex-valued domain [23], the loss function is represented as:

$$L = -\frac{1}{2N}\sum_{i=1}^{N} [\log(\frac{\mathrm{e}^{\Re[a_i]}}{\sum_c e^{\Re[a_c]}}) + \log(\frac{\mathrm{e}^{\Im[a_i]}}{\sum_c e^{\Im[a_c]}})] \tag{12}$$

where $\Re[a_i]$ and $\Im[a_i]$ represent the real and imaginary parts of a_i, respectively.

The trainable weight matrix could be optimized to increase the probability of the correct class label by minimizing the loss function. An L_2 regularization term is usually adopted in the loss function to prevent overfitting in practice.

2.3. Superpixel Algorithm

The idea of superpixel was originally proposed by Ren and Malik [25] in 2003. By computing the intra-region similarity and inter-region similarity, including brightness, texture, and contour energy, all the neighboring similar pixels are connected to construct a perceptually consistent unit, named superpixel. The superpixel representation of a neural image can be regarded as a segmentation of the image, and the pixels in each region are homogeneous. Usually, the complexity of the processed image can be greatly reduced, as a small number of superpixels are generated by iterative clustering to represent the numerous pixels of the original image. Several superpixel methods, including Mean Shift, SEEDS (Superpixels Extracted via Energy-Driven Sampling), and SLIC (simple linear iterative cluster), have been widely applied to computer vision areas, such as image segmentation, pose estimation, target tracking, and object detection. Compared with traditional pixel-based techniques, SLIC has several advantages which are adopted in this study:

1) The image is firstly transformed into a Lab color model. Therefore, the superpixel can be applied to segmenting not only the RGB image but also the gray image.

2) The generated superpixels are compact and arrayed. The boundary and neighboring information of pixels are preserved. The pixel-based techniques can be easily transformed and applied to the superpixel representation.

3) The complexity and consuming time are greatly reduced, while the superpixel can produce a more descriptive and effective representation than traditional pixel-based techniques.

3. METHODS

In this section, the proposed classification method for polarimetric SAR data is presented in detail. First, as the normalization is essential for the CV-CNN, the data preparation for the PolSAR image is briefly introduced. Second, we present the specific configuration of the proposed network. Then, the description of the whole classification model, which is based on the fusion of superpixel method and the convolutional network, is presented.

3.1. Data Preparation

For the region-based PolSAR image classification, the label of each pixel is identified by a local patch, which is defined by a neighborhood window of size $W \times H$. Therefore, not only the polarimetric characteristic and intensity information are captured, but also the relationship between a pixel and its neighbors is adopted to contribute to the feature extraction. For a large PolSAR image, a sliding widow is utilized to generate a number of patch images, which can be adopted for training and evaluation. As introduced in Subsection 2.1, the patch image is formed with six channels.

In order to achieve a better performance, preprocessing for the patch image is necessary. The zero-mean normalization is applied to the raw data. For instance, for each channel, the average μ and standard deviation σ are first calculated, and then all samples of each channel are normalized. The formula can be presented as follows:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} T_i \tag{13}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (T_i - \mu)\overline{(T_i - \mu)}}{n}} \tag{14}$$

$$T_{nor} = \frac{T - \mu}{\sigma} \tag{15}$$

where $\overline{(T_i - \mu)}$ denotes the conjugate.

3.2. Configuration of the Proposed Neural Network

As shown in Fig. 1, the architecture of the proposed neural network can be presented as a cascading of convolutional neural networks, followed by several fully connected layers. Different from traditional target images with large scale, the generated PolSAR images consist of no distinct structures such as corners. Deep layers in CNN would lost the advantage on feature extraction and make the extracted features too abstract. The combination of convolutional layer and deconvolutional layer has been proved effective in adaptive feature learning [24] and image reconstruction [27, 29]. Therefore, a convolutional layer is firstly adopted to take the advantage of spatial relationship between pixels and to learn the low-level feature representations. Then, the deconvolutional layer, which can be regarded as a decoder, is utilized to reconstruct the high-level feature representation for classification. The output feature is then reshaped into a vector and feed to the following fully connected layer. The output of the fully connected layers is regarded as the corresponding labels.

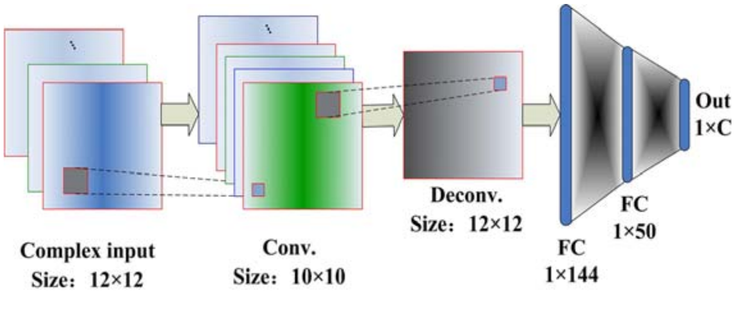


Figure 1. Structure of proposed network.

The specific arrangement of each layer is shown in Fig. 2. As the input image consists of six channels with the size of 12×12, the input layer is fixed to the size $12 \times 12 \times 6$. In the first convolution layer, the input image is filtered by 16 convolution filters of size $3 \times 3 \times 6$ with stride 1. The convolutional results are activated by ReLU, producing 16 feature maps with the size of 10×10. The second layer deconvolves the feature maps with 16 filters of $3 \times 3 \times 6$ with stride 1, activated by ReLU and reconstructing a feature map with size 12×12. The feature map is then reshaped to a vector with dimension of 144 and fed to the first fully connected layer. The dropout technique is used in this layer, and the probability is fixed to 0.5. The ReLU is also selected as the activation function of the fully connected layer. The result is

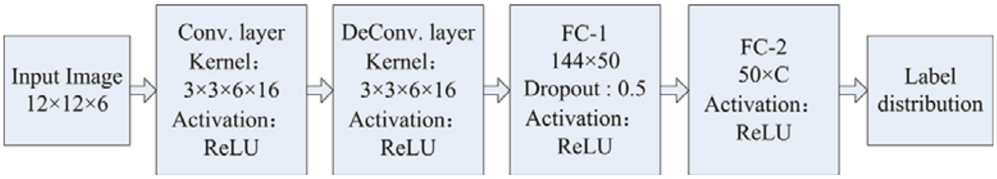


Figure 2. Arrangement of each layer.

fed to the second fully connected layer, and the output layer contains the same number of units as the classification classes C. For example, C is 15 for Flevoland dataset.

As in most deep learning algorithms, all the weight and bias are learned by minimizing the loss function. Usually, it is not possible to compute the global minimum of the loss function directly. Several approaches have been studied to optimize the learning process [28], such as stochastic gradient descent (SGD) and momentum. In this study, the parameters are learned by SGD to minimize the loss function in backpropagation. By computing the error gradient of parameters, the parameters can be updated with the following rule:

$$w^{k+1} = w^k - \eta(\partial L/\partial w^k) \tag{16}$$

$$b^{k+1} = b^k - \eta(\partial L/\partial b^k) \tag{17}$$

where η is the learning rate, and k is the number of iterations.

3.3. Classification with Information Fusion

As described above, the patch images, which are adopted as training and testing samples, are generated by the sliding window. The convolutional network is able to take the advantage of the spatial structure and produce robust and accurate result. Meanwhile, it must be noticed that the patch image extracted at the boundary, which is composed of pixels from several categories, will confuse the pre-trained network and cause misclassification. On the other hand, because of the stride of the sliding window, the predicted label usually needs to be repeated in a small block. The misclassification of single pixel will lead to the error of the neighborhood.

To solve these problems, a novel PolSAR image classification method using convolutional network and superpixel is introduced in this section. The detailed classification process is illustrated in Fig. 3. The proposed recognition framework is composed of two branches: the extraction of superpixel structure and the prediction based on region-based CNN. Then, the region and contour information are adopted

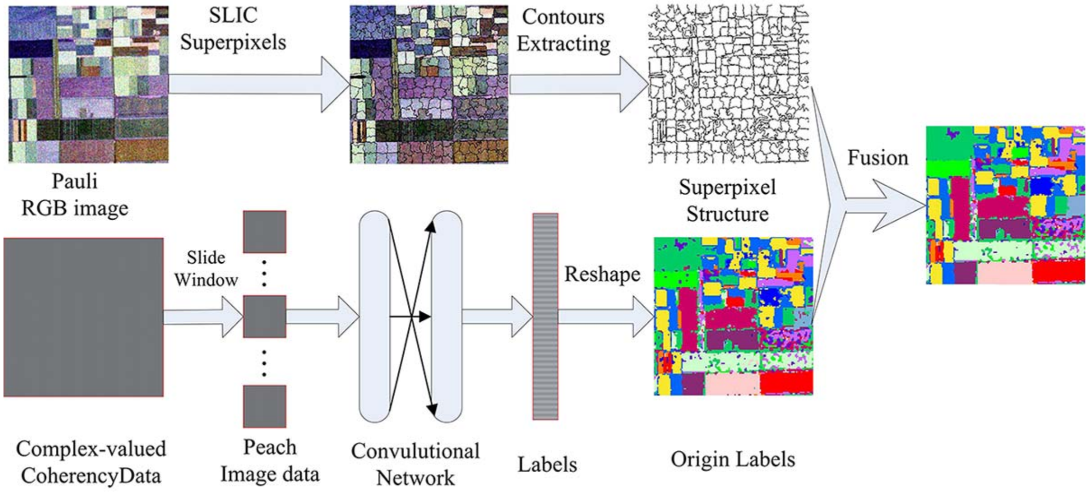


Figure 3. Structure of proposed framework.

to correct the original predicted labels, outputting the integrated result. The introduction of each part is presented as follows:

Superpixel: First, to oversegment the image, Pauli decomposition is applied to form the RGB image. Then, the SLIC algorithm is applied to the RGB image to generate the superpixel representation. Each region is a homogeneous area. Small superpixels are combined to form a large superpixel, and the contour information of each superpixel is retained.

Convolutional neural network: As described in Section 2, the input data (complex-valued coherency matrix) consists of six channels, represented as $\{T_{11}, T_{12}, T_{13}, T_{22}, T_{23}, T_{33}\}$. First, a sliding window is applied to extract patch images of each channel. The corresponding label of the central pixel in each patch is also extracted to form the label vector. A small part of these labeled images are adopted to train the convolutional network, and the rest images are utilized to evaluate the performance of the trained network. After the training of the neural network, the whole patch image dataset is fed to the trained convolutional network, producing the corresponding label of each pixel. The predicted label vector is then reshaped to reconstruct the 2-D label map. Part of the label map is show in Fig. 3, which is presented as "Origin Labels".

Fusion: It can be observed from Fig. 3 that the label map directly generated by the convolutional network suffers from the misclassification of the boundary pixels. Furthermore, there are some singular pixels within some regions. On the other hand, comparing the superpixel representation with the RGB image, it is obvious that the pixels in one region are homogeneous. Besides, the boundary of each category is retained by the superpixels from the comparison of RGB image. Thus, the contour structure information provided by superpixel method is then applied to correct the misclassification of these pixels.

Assuming that the oversegmentation is constructed by the SLIC method, the procedure is summarized as follows:

1. Adopt the contour structure to divide the origin label map into R regions $\{G_1, \cdots, G_R\}$. Let M_r be the number of pixels in region G_r.

2. Compute the number of pixels belonging to each category in G_r and express the result as $Num\,(c, q_c)$, where c denotes the label, and q_c denotes the number of pixels with label c, $M_r = \sum\limits_{c=1}^{C} q_c$.

3. Compute the max ratio p_{cr} for pixels of each category in G_r, where $p_{cr} = \max\limits_{i=1}^{C}(\frac{q_i}{M_r})$.

4. Select a threshold value T_s. If $p_{cr} \geq T_s$, replace the other labels with c, else, keep the origin labels of each pixel.

4. RESULTS AND ANALYSIS

In this section, experiments on several PolSAR datasets are performed to verify the effectiveness of the proposed recognition framework. (1) Flevoland dataset: This PolSAR dataset of Flevoland is a subset of an L-band four-look image acquired by Airborne SAR (AIRSAR). It measures (750 × 1024) pixels which consist of 15 categories. (2) Oberpfaffenhofen dataset: This dataset is measured by electronically steered array radar (ESAR) over Oberpfaffenhofen in Germany. It covers a size of 1300 × 1200 and contains 3 categories. (3) Flevoland Benchmark dataset: This PolSAR data is the benchmark dataset of an AIRSAR data obtained over Flevoland. It covers a size of (1020 × 1024) pixels and contains 14 categories. (4) San Francisco dataset: This PolSAR is acquired by Airborne SAR (AIRSAR), covering a size of (900 × 1024) pixels and containing about 5 categories. The overall accuracy (OA) and confusion matrix are used to evaluate the performance of the proposed methods. Moreover, the CV-CNN model and RV-CNN model are used to compare with the performance of the proposed fusion model.

4.1. Experiments on Flevoland Dataset

The first experiment was performed on the dataset of Flevoland, which is an agriculture area of the Netherlands. An RGB image (750 × 1024) formed with the intensities from Pauli decomposition is shown in Fig. 4(a), and its ground truth map is shown in Fig. 4(b). There are in total 15 identified classes including stem beans, peas, forest, lucerne, three types of wheat, beet, potatoes, bare soil, grass,

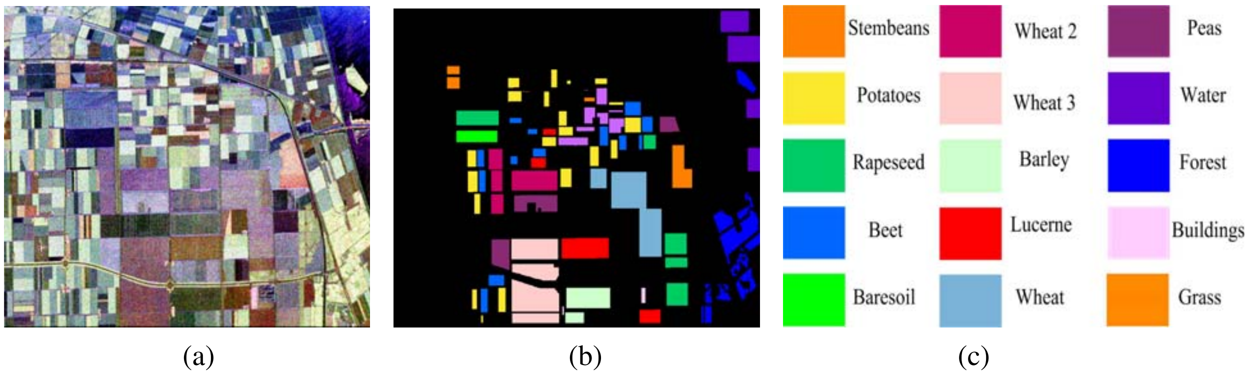


Figure 4. Flevoland dataset. (a) Pauli RGB composition. (b) Ground truth map; (c) Legends of Ground truth map.

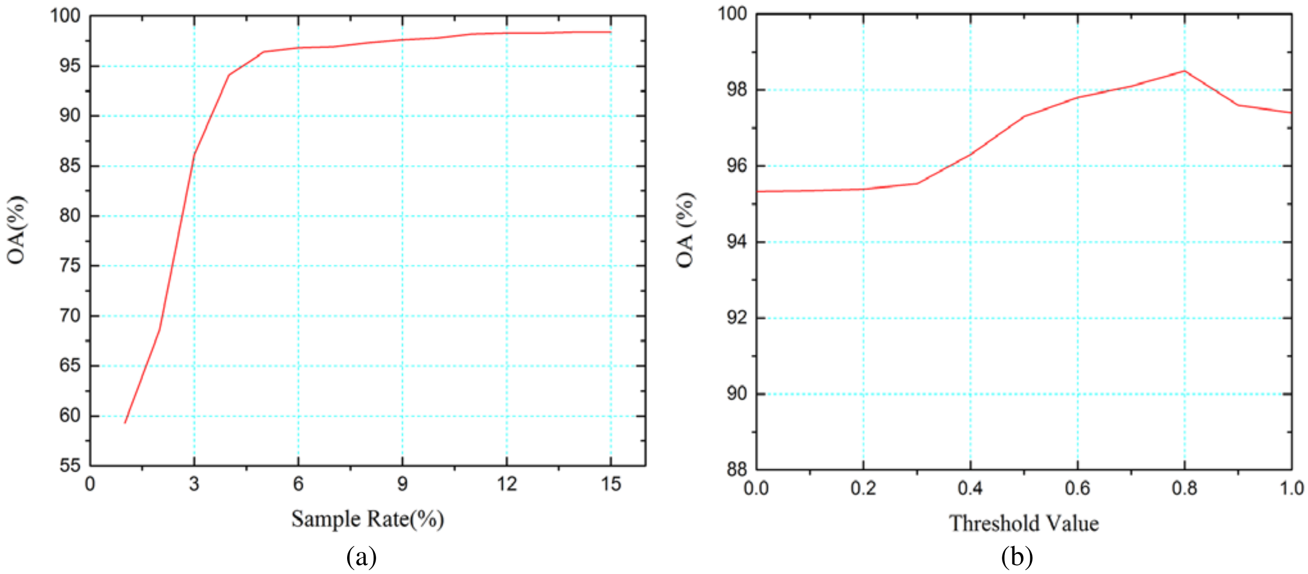


Figure 5. Classification results on variable coefficients. (a) OA of variable sample rates. (b) OA of variable threshold value.

rapeseed, barley, water, and a small number of buildings. The legends of the ground truth are shown in Fig. 4(c).

For the labeled pixels near the boundary which is selected as the center of the sliding window (12 × 12), a zero-padding strategy is applied to those pixels outside the image region. According to previous research [24], a lower sample rate is generally acceptable for the classification of small number of classes. On the other hand, the structure of the proposed network is not complex. Thus, a range of 1%–15% for sample rate is utilized to evaluate the performance, and the results are shown in Fig. 5(a). As can be seen from Fig. 5(a), the classification accuracy increases rapidly when the sampling rate increases from 1% to 5%. The accuracy becomes stable at about 97% when the sampling rate is greater than 10%. Therefore, 10% sampling rate is suitable for the Flevoland case, and 14000 samples are selected as training samples in this research.

Another experiment is conducted to analyze the sensitivity on variable threshold value T_s, and the results are shown in Fig. 5(b). When $T_s = 0$, each region is filled up with the label which takes up the biggest share in that region. The performance depends mainly on the validity of the superpixels. Hence, the OA is little lower than the other cases. When $T_s = 1$, the label of each pixel is entirely identified by the proposed network, and the misclassification directly degrades the OA, especially for the misclassification of the boundary pixels. As can be seen form Fig. 5(b), the range of 0.5 to 0.9 is acceptable for the decision fusion model, and the threshold value T_s is fixed 0.8 for the Flevoland case.

As illustrated in Fig. 2, the filter size is 3×3 for both convolutional layer and deconvolutional layer. Hyperparameters are chosen as follows: the learning rate η is 0.5; the probability of dropout is 0.5; the batch size is 100 with 50 training epochs. The initial number of the superpixels is fixed to 1200, and the threshold value T_s is fixed as 0.8.

Figure 6(a) shows the origin classification result of the proposed neural network. Figs. 6(b) and 6(c) show the results of the fusion model and the areas with the ground truth. Fig. 7(a) shows the final classification of the comparing CV-CNN for the whole dataset, and Fig. 7(b) shows the results of the areas with the ground truth. From Fig. 6, it can be found that the classification results are in good agreement with the ground truth map at the first sight. Compared with Figs. 6(a) and 7(a), from which the speckles can be easily observed, we can find that most of the singular pixels within the regions have been corrected with the superpixel information in Fig. 6(b). Furthermore, the boundary of Fig. 6(b) is clearer than Figs. 6(a) and 7(a), especially for smooth areas such as forest, bare soil, and water.

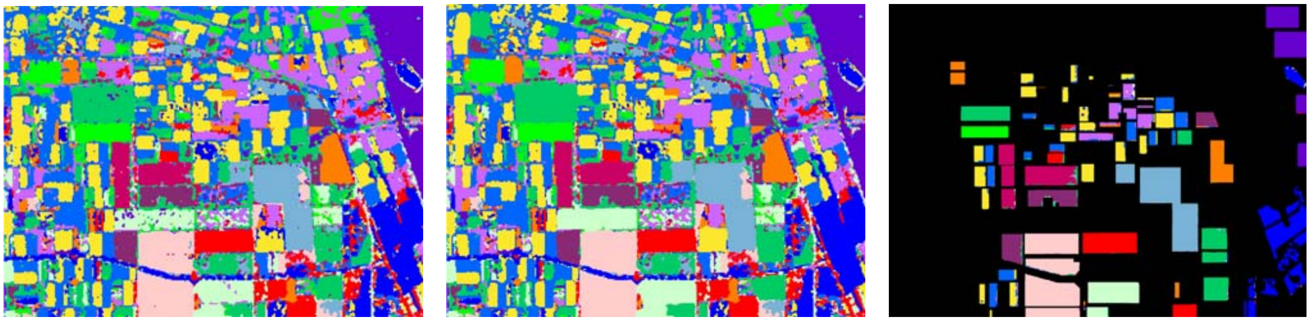

(a) (b) (c)

Figure 6. Classification results of the proposed algorithm on the Flevoland dataset. (a) Result of the proposed network. (b) Final result of the fusion framework. (c) Final result overlaid with the ground truth map.

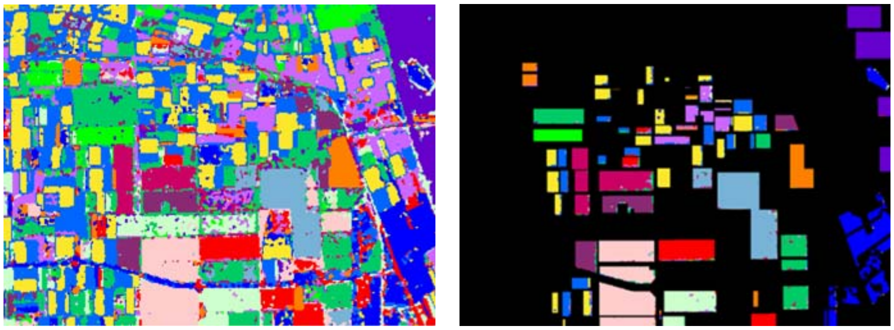

(a) (b)

Figure 7. Classification results of the CV-CNN algorithm on the Flevoland dataset. (a) Result of the CV-CNN. (b) CV-CNN result overlaid with the ground truth map.

The detailed classification results compared with the ground truth are listed in Table 1, and the confusion matrix of the fusion model is shown in Table 2. The OAs of CV-CNN and the proposed network are 96.4% and 96.7%, and is 98.3% for the fusion model. Note that a considerable improvement of 1.9% in terms of classification accuracy has been achieved by combining the hierarchical information of network with superpixels structure information.

From Table 2, it can be seen that the majority classes have a correct rate higher than 90%, except for building areas. It shows that the continuous areas, which have enough training samples, are easy

Table 1. OA of each category for the whole dataset.

Class	CV-CNN	Proposed network	Fusion model
Stem beans	97.8	94.1	**98.1**
Peas	97.2	**97.4**	96.9
Forest	96.9	96.8	**98.5**
Lucerne	95.2	98.1	**98.2**
Wheat	94.8	95.0	**97.7**
Beet	96.3	**97.2**	96.5
Potatoes	94.8	95.3	**97.1**
Bare soil	98.2	98.4	**100**
Grass	89.7	**94.3**	90.1
Rapeseed	92.6	92.1	**92.5**
Barley	95	95.5	**98.7**
Wheat2	91.7	91.6	**92.8**
Wheat3	97.1	96.6	**98.4**
Water	97.7	98.5	**100**
Buildings	85.5	83.3	**88.6**
OA	96.4	96.7	**98.3**

Table 2. Confusion Matrix of the proposed recognition framework.

%	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	98.1	0.1	0.0	0.2	0.0	1.2	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0
2	1.0	96.9	0.0	0.0	0.5	0.2	0.0	0.0	0.1	1.3	0.0	0.0	0.0	0.0	0.0
3	0.1	0.0	98.5	0.0	0.1	0.0	0.6	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0
4	0.1	0.0	0.0	98.2	0.0	0.5	0.2	0.0	0.0	0.5	0.1	0.4	0.0	0.0	0.0
5	0.0	0.2	0.0	0.0	97.7	0.1	0.0	0.1	0.3	0.7	0.0	0.8	0.0	0.1	0.0
6	0.2	0.0	0.0	0.2	0.0	96.5	0.9	0.2	0.2	1.5	0.3	0.0	0.0	0.0	0.0
7	0.1	0.0	0.4	0.1	0.1	1.6	97.1	0.0	0.0	0..1	0.1	0.0	0.5	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	1.4	0.8	0.0	0.7	0.0	3.9	1.6	0.0	90.1	0.0	1.2	0.0	0.0	0.3	0.0
10	0.1	0.9	0.1	0.1	3.3	1.0	0.1	0.0	0.5	92.5	0.1	1.0	0.2	0.1	0.0
11	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	1.1	0.1	98.7	0.0	0.0	0.0	0.0
12	0.0	0.4	0.0	0.0	2.0	0.0	0.0	0.4	0.2	4.1	0.0	92.8	0.0	0.1	0.0
13	0.0	0.0	0.3	0.0	0.6	0.0	0.0	0.0	0.0	0.2	0.4	0.1	98.4	0.0	0.0
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0
15	11.1	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	88.6

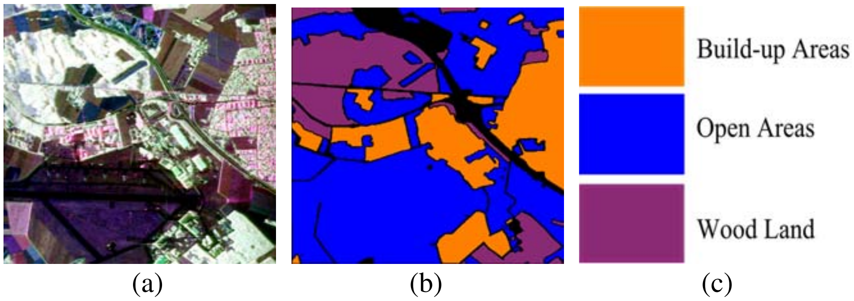
1 stem beans; 2 peas; 3 forest; 4 lucerne; 5 wheat; 6 beet; 7 potatoes;
8 bare soil; 9 grass; 10 rapeseed; 11 barley; 12 wheat 2; 13 wheat3; 14 water; 15 buildings

to obtain higher accuracy. Less training samples usually lead to the misclassification with others, such as building areas and grass. The fusion model has obtained considerable improvement of the majority classes, especially for the large and homogeneous areas, such as Barley, Bare soil, and Water.

Through the analysis on Fig. 6, Fig. 7, Table 1, and Table 2, the reason that the proposed framework obtained a higher accuracy can be summarized as follows: 1) The proposed network utilized complex-valued input data, which consists of both the intensity and the phase information of SAR images. As introduced in [24], the phase distribution varies obviously with the categories but stable for the same category in different areas. Thus, the complex-valued neural network can obtain more accurate feature representations form the raw data and improve the recognition accuracy. 2) The terrain region extracted by the sliding window is more texture-like and contains less spatial structure features. Too many layers would produce too abstract features and decrease the recognition accuracy. The proposed network constructed the feature representations from the convolutional result without pooling operation which would reduce the size of feature maps and abandon some useful information, thus generating reliable features based on the spatial structure. 3) The fusion model adopted the superpixel information to correct the singular pixels within the terrain region. Beside, as can be seen from Fig. 3, the contours of the superpixels are in good agreement with the boundary of each region. Hence, using the contour of each superpixel to redetermine the labels of each region can effectively reduce the misclassified boundary pixels. The use of the threshold value also prevents the poor superpixel from further reducing the recognition performance.

4.2. Experiments on Oberpfaffenhofen Dataset

Figure 8(a) shows the Pauli RGB image of the Oberpfaffenhofen in Germany, which is of the size 1300 × 1200. The ground truth and the corresponding legend are shown in Figs. 8(b) and 8(c). There are mainly three categories, including open areas, wood land, and build-up areas. 14000 labeled pixels, about 1%, are randomly selected as the centers to generate the training samples. The sliding window of the size 12 × 12 is utilized to extracted the patch images, and the corresponding labels of the center pixels are used to guide the training of the proposed network. Parameters of the whole framework are set as follows: the learning rate η is 0.8; the probability of dropout is 0.5; and the batch size is 100 with 50 training epochs. The initial number of superpixels is fixed to 600, and the threshold value T_s is fixed at 0.7.



(a) (b) (c)

Figure 8. Oberpfaffenhofen dataset. (a) Pauli RGB composition. (b) Ground truth map. (c) Legend of Ground truth map.

Figure 9(a) illustrates the results of CV-CNN, and the results of the proposed fusion model are shown in Fig. 9(b). Compared with Fig. 8(b), Fig. 9(b) shows that the proposed fusion model performs better in majority areas. It can be seen that this SAR image is more complicated than Flevoland, as the boundary is irregular, and there are similarities between the build-up areas and the wood land. Thus, several regions are obviously affected by the misclassification, especially the wood land areas and build-up areas. Comparing Fig. 9(b) with 9(a), it can be observed that the result of the fusion model matches much better with the ground truth, and the labeled regions are more homogeneous than the CV-CNN model. Fig. 9(c) illustrates the superpixel presentation for part of the RGB image, the contour information of each region, and the final results of the regions. We can find that the contour information matches well with the boundary of each region for each category, and the optimized classification result map is in better agreement with ground truth map than CV-CNN.

The detailed results of the methods are shown in Table 3. We can find that a considerable improvement of the classification accuracy for build-up areas has been achieved by the optimization of the superpixels, as a number of the misclassified pixels have been corrected by the vote operation. Compared with CV-CNN, the proposed fusion model achieves about 2% gain in this case.

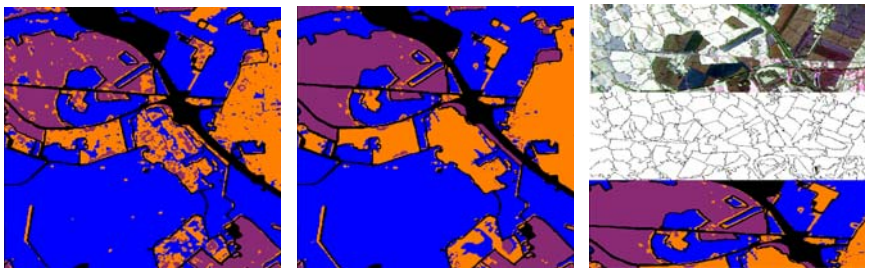

(a) (b) (c)

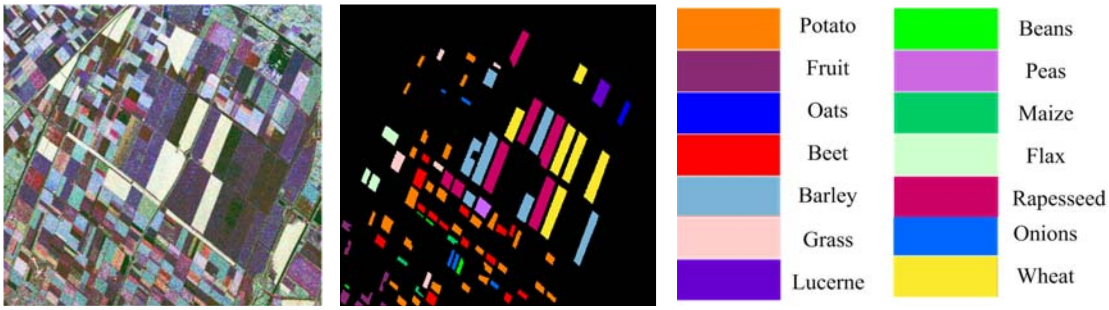
Figure 9. Oberpfaffenhofen dataset. (a) Results of CV-CNN. (b) Results of the proposed fusion model. (c) Superpixel presentation for part image.

Table 3. Confusion matrix of the oberpfaffenhofen dataset.

Methods	Fusion Model			CV-CNN		
Categories	Open	Wood	Build-up	Open	Wood	Build-up
Open	95.9	0.4	3.7	94.6	0.2	5.2
Wood	2.8	92.3	4.9	0.7	92.1	7.2
Build-up	2.2	1.5	96.3	5.5	3.2	91.3
OA	**95.6**			93.2		

4.3. Experiment on Flevoland Benchmark Dataset

Another experiment has been conducted to evaluate the performance of the proposed fusion model on the Flevoland Benchmark dataset, which is of the size 1020×1024. The PolSAR RGB image is shown in Fig. 10(a), and the ground truth map is shown in Fig. 10(b) along with the legend of labels in Fig. 10(c), which is collected from [24]. 14000 patch images are randomly selected as the training samples. All the samples are generated by a sliding window, which is of the size 12×12. The parameters used in the recognition task are set as follows: the learning rate η is 0.5; the probability of dropout is 0.5; and the batch size is 100 with 80 training epochs. The initial number of the superpixels is fixed at 1200, and the threshold value T_s is fixed at 0.7.



(a) (b) (c)

Figure 10. Flevoland Benchmark dataset. (a) Pauli RGB composition. (b) Ground truth map. (c) Legend of Ground truth map.

Figure 11 shows the final classification results of comparison methods (RV-CNN and CV-CNN) for the whole dataset, where Figs. 11(a) and 11(b) represent the results of RV-CNN, and Figs. 11(c) and 11(d) represent the results of CV-CNN, respectively. The final result of the proposed fusion model is shown in Fig. 12. As shown in Fig. 12(b), the label map is in good agreement with the ground truth

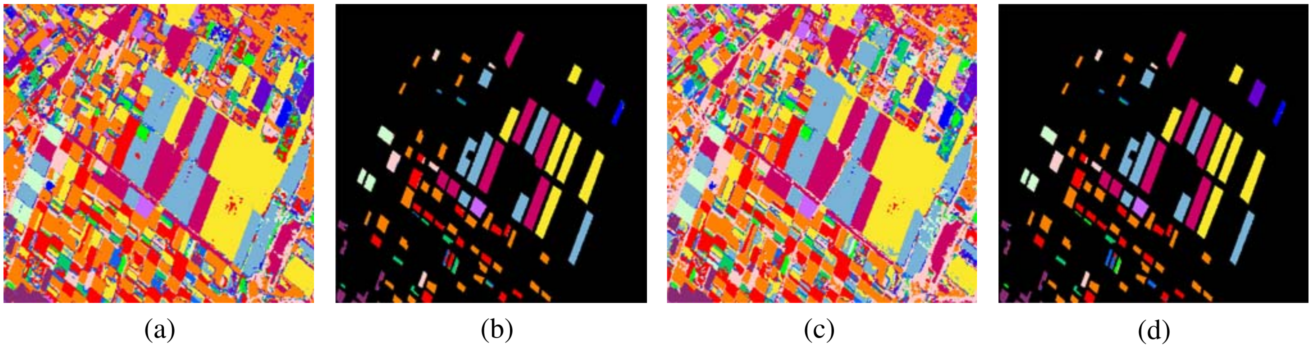


Figure 11. Flevoland Benchmark dataset. (a) Classification result of RV-CNN. (b) RV-CNN overlaid with ground truth. (c) Classification result of CV-CNN. (d) CV-CNN overlaid with ground truth.

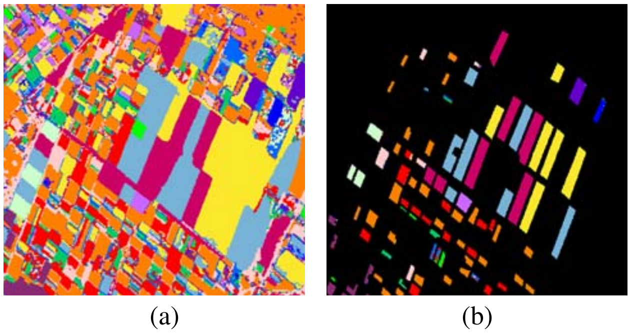


Figure 12. Flevoland Benchmark dataset. (a) Classification result of the proposed Fusion model. (b) Fusion model overlaid with ground truth.

Table 4. Classification results of the Flevoland Benchmark dataset.

%	RV-CNN	CV-CNN	Fusion model
Potato	97.9	**99.8**	99.6
Fruit	96.2	**98.3**	97.2
Oats	96.2	98.1	**99.0**
Beet	96.4	96.2	**98.4**
Barley	99.6	99.6	**99.8**
Onions	70.8	93.2	**94.8**
Wheat	99.0	**99.9**	99.4
Beans	84.6	**90.5**	88.9
Peas	94.4	98.3	**100**
Maize	90.2	**98.5**	92.6
Flax	92.9	**96.6**	96.3
Rapeseed	98.8	99.4	**99.9**
Grass	92.2	96.6	**96.8**
Lucerne	90.9	98.2	**99.5**
OA	97.1	98.7	**99.0**

image. For the comparison method, most of the pixels also match well with the ground truth image. The CV-CNN is more robust than RV-CNN in the labeled regions form the comparison of Fig. 11(b) and Fig. 11(d). A reasonable explanation is that the combination of phase information compensates for the deficiency of the traditional feature representations. Therefore, it is more convenient to obtain higher recognition accuracy. However, the misclassification (singular pixels and the boundary) can be obviously observed in Figs. 11(a) and 11(c). Comparing Fig. 12(a) with Fig. 11(a), it can be seen that some misclassified pixels have been corrected by the neighbors within the superpixels, especially for the categories with large samples, such as wheat, barley, and rapeseed. The superior performance obviously presents the advantage of the fusion model. On the other hand, for the regions where the q_c is smaller than the fixed threshold value T_s, the original labels have been retained to prevent further misclassification.

The classification results of the benchmark dataset for each category are shown in Table 4. It can be seen that the proposed fusion model has obtained better performance in majority categories than RV-CNN and CV-CNN. Compared with RV-CNN, we can find that with the use of intensity and phase information, CV-CNN and the proposed model could effectively reduce the misclassification of the confusing categories, such as onions and beans. Besides, the fusion model could help to correct the misclassified pixels, thus improving the OA.

Table 5. Confusion matrix of the Flevoland benchmark dataset.

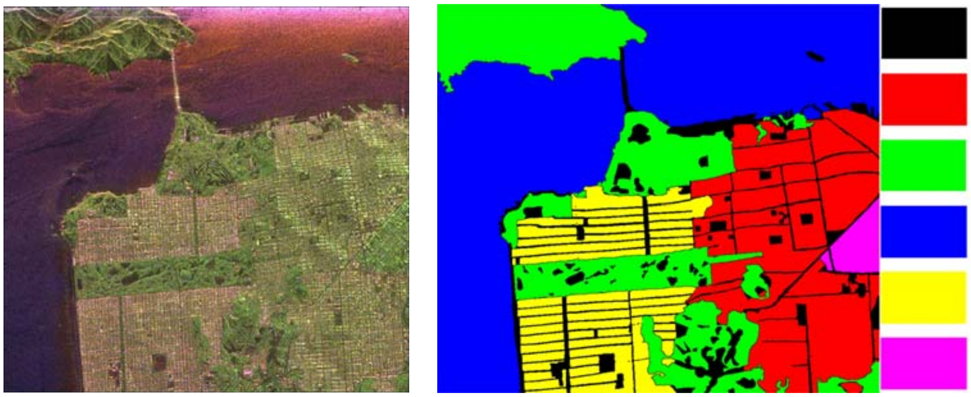
%	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	99.6	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.7	97.2	0.0	1.0	0.6	0.0	0.0	0.0	0.4	0.0	0.0	0.1	0.0	0.0
3	0.0	0.0	94.0	0.6	1.5	1.1	0.0	0.0	0.0	2.8	0.0	0.0	2.0	0.0
4	0.5	0.0	0.0	98.4	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
5	0.0	0.0	0.0	0.0	99.7	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1
6	0.0	0.0	0.0	0	0.6	88.8	1.3	0.3	0.0	8.9	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	99.4	0.0	0.0	0.0	0.0	0.5	0.0	0.1
8	0.0	0.0	0.3	1.1	0.0	3.0	0.0	88.9	0.0	0	0	1.6	0.3	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.9	0.0	0.0	0.1	0.0	0.0
10	0.0	0.0	0.0	6.4	0.0	0.9	0.0	0.0	0.0	92.6	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.9	0.0	0.0	0.0	1.7	0.0	0.0	94.3	0.6	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.9	0.0	0.0
13	1.1	0.0	0.0	10.0	0.0	2.0	7.9	0.2	1.9	0.6	0.0	1.6	72.9	1.8
14	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.3	0.0	99.5

1 Potato; 2 Fruit; 3 Oats; 4 Beet; 5 barley; 6 Onions; 7 wheat; 8 Beans;
9 Peas; 10 Maize; 11 Flax; 12 Rapeseed; 13 Grass; 14 Lucerne

Table 5 lists the confusion matrix of the proposed method. It can be seen from Table 5 that accuracies of all the categories are higher than 90%, and some of them are close to 100%, which shows that the proposed framework has achieved considerable performance.

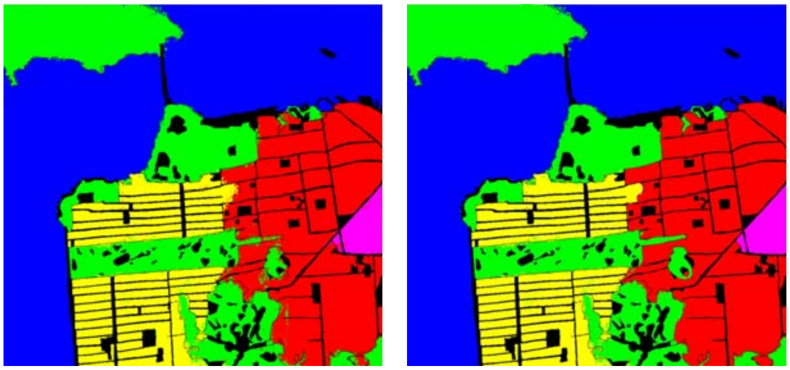
4.4. Experiment on San Francisco Dataset

This experiment was carried out on the dataset of San Francisco. An RGB image (900 × 1024) formed with the intensities from Pauli decomposition is shown in Fig. 13(a), and its ground truth map is shown in Fig. 13(b). There are in total 5 identified classes including vegetation, sea, developed urban, low-density urban, and high-density urban. The legend of the ground truth is shown in Fig. 13(c). 14000 patch images are randomly selected as the training samples. All the samples are generated by a sliding window, which is of the size 12 × 12. The parameters used in the recognition task are set as follows:



(a) (b)

Figure 13. San Francisco dataset. (a) Pauli RGB composition. (b) Ground truth map and Legend.



(a) (b)

Figure 14. San Francisco dataset. (a) Original classification labels; (b) Fusion classification labels.

the learning rate η is 0.5; the probability of dropout is 0.5; and the batch size is 100 with 80 training epochs. The initial number of the superpixels is fixed at 400, and the threshold value T_s is fixed at 0.7.

Figure 14 shows the classification results of the proposed original framework and the fusion model. It can be clearly seen from the ground truth map that the areas on the graph are uniformly continuous. Thus, the proposed framework has obtained a remarkable original overall classification rate (about 97.62%), especially for the sea area (about 99.76%) and the high-density urban area (about 99.07%). An obvious improvement can be seen from Fig. 14(b) that majority of the misclassified pixels near the boundary between high-density urban and vegetation have been corrected after the fusion model, and the overall classification rate is about 98.91%.

In addition, the average computation time of the four datasets for each component (generating original label, superpixel, generating fusion label) is listed in Table 6. The devices for the testing experiment are as follows: Memory: 16G; GPU: 1050 Ti; System: Win7. The training time is ignored as it costs several hours to generate a trained model. As shown in Table 6, it takes similar time to

Table 6. Comparison of average computation time on four datasets.

Generating original labels	Original labels	Superpixels	Fusion	All time
RV-CNN	47.6 s	-	-	47.6 s
Proposed Model	52.7 s	80.3 s	1.2 s	134.2 s

generate original labels for RV-CNN and the proposed complex model. The time cost for superpixel is related to the image size. Typically, as listed in table, the average time to extract superpixel information for the four data sets is about 80.3 s, and the fusion time is about 1.2 s. Generally speaking, as the SAR image terrain recognition is not a real-time task, it is reasonable to utilize the superpixel information to correct the original label and improve the recognition accuracy.

5. CONCLUSIONS

This paper proposes a new network based on complex-valued CNN for PolSAR image classification. The patch images extracted from raw coherency matrix are fed to the input layer. Then, the proposed network extracts nonlinear relationship between the input samples automatically. The superpixel generating algorithm (SLIC) is used to produce oversegmentation representation of RGB image. The contour information is adopted to retain the relations between pixels. A threshold value is used to balance the strength of the final decision fusion, which is able to reduce the effect of speckle noise. Therefore, better consistency is observed in final results. The proposed network can be regarded as a feature extractor which projects the original low-level input data onto a high-level feature space and extracts the inner concepts embedded in the complex-valued PolSAR data. Its advantage is that we can use a small number of parameters to automatically characterize the complex structure of the PolSAR data. With the use of dropout operation, it is able to avoid the overfitting, which is important for learning with limited training samples. The combination of convolutional layer and deconvolutional layer is effective for constructing the relationships between the neighbor pixels. On the other hand, the superpixel representation shows its advantage for matching the boundary of the Pauli decomposition image, which is essential to correct the misclassified pixels near the boundary. Different from traditional post-processing methods which use fixed window to optimize the label map, the contour information follows the structure of each category which means that it can be used for terrain classification which has complex and irregular structures, such as Oberpfaffenhofen. By this way, the impacts of complex terrains and noise on classification will be alleviated. It is still worth noting that with a suitable threshold value, the fusion framework is able to inherit the advantage from both network and superpixel algorithm. Therefore, the final label map produced by the fusion framework not only has high accuracy for each pixel, but also could correct the misclassified pixels near the boundary effectively. The future work includes improving the proposed method using better parameters selection and adaptive threshold value and applying it to the terrain classification.

REFERENCES

1. Jiao, L. and F. Liu, "Wishart deep stacking network for fast PolSAR image classification," *IEEE Transactions on Image Processing*, Vol. 25, 3273–3286, 2016.
2. Scheuchl, B., D. Flett, R. Caves, and I. Cumming, "Potential of RADARSAT-2 data for operational sea ice monitoring," *Canadian Journal of Remote Sensing*, Vol. 30, No. 3, 448–461, 2004.
3. Wang, L., K. A. Scott, L. Xu, et al., "Sea ice concentration estimation during melt from dual-pol SAR scenes using deep convolutional neural networks: A case study," *IEEE Transactions on Geoscience & Remote Sensing*, Vol. 54, No. 8, 4524–4533, 2016.
4. Freeman, A., J. Villasenor, J. D. Klein, et al., "On the use of multi-frequency and polarimetric radar backscatter features for classification of agricultural crops," *International Journal of Remote Sensing*, Vol. 15, No. 9, 14, 1994.
5. Lee, J. S. and M. R. Grunes, "Classification of multi-look polarimetric SAR data based on complex Wishart distribution," *National Telesystems Conference, IEEE*, 1992.
6. Gao, W., J. Yang, and W. Ma, "Land cover classification for polarimetric SAR images based on mixture models," *Remote Sensing*, Vol. 6, No. 5, 3770–3790, 2014.
7. Rignot, E. and R. Chellappa, "Segmentation of polarimetric synthetic aperture radar data," *IEEE Transactions on Image Processing*, Vol. 1, No. 3, 281–300, 1992.
8. Lee, J. S., D. L. Schuler, R. H. Lang, et al., "K-distribution for multi-look processed polarimetric SAR imagery," *International Geoscience & Remote Sensing Symposium, IEEE*, 1994.

9. Freitas, C. C., A. C. Frery, and A. H. Correia, "The polarimetric distribution for sar data analysis," *Environmetrics*, Vol. 16, No. 1, 13–31, 2010.
10. Chen, Q., G. Kuang, J. Li, et al., "Unsupervised land cover/land use classification using PolSAR imagery based on scattering similarity," *IEEE Transactions on Geoscience & Remote Sensing*, Vol. 51, No. 3, 1817–1825, 2013.
11. Wang, Y., C. Han, and F. Tupin, "PolSAR data segmentation by combining tensor space cluster analysis and Markovian framework," *IEEE Geoscience & Remote Sensing Letters*, Vol. 7, No. 1, 210–214, 2010.
12. Shang, F. and A. Hirose, "Quaternion neural-network-based PolSAR land classification in Poincare-sphere-parameter space," *IEEE Transactions on Geoscience & Remote Sensing*, Vol. 52, No. 9, 5693–5703, 2014.
13. Yu, P., A. K. Qin, and D. A. Clausi, "Unsupervised polarimetric SAR image segmentation and classification using region growing with edge penalty," *IEEE Transactions on Geoscience & Remote Sensing*, Vol. 50, No. 4, 1302–1317, 2012.
14. Cao, F., W. Hong, Y. Wu, and E. Pottier, "An unsupervised segmentation with an adaptive number of clusters using the SPAN/H/α/A space and the complex Wishart clustering for fully polarimetric SAR data analysis," *IEEE Transactions on Geoscience & Remote Sensing*, Vol. 45, No. 11, 3454–3467, 2007.
15. Wu, Y., K. Ji, W. Yu, and Y. Su, "Region-based classification of polarimetric SAR images using Wishart MRF," *IEEE Geoscience & Remote Sensing Letters*, Vol. 5, No. 4, 668–672, 2008.
16. Hou, B., H. Kou, and L. Jiao, "Classification of polarimetric SAR images using multilayer autoencoders and superpixels," *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, Vol. 9, No. 7, 3072–3081, 2017.
17. Krizhevsky, A., I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Proc. Adv. Neural Inf. Process. Syst*, 1097–1105, 2012.
18. He, K., X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," 2015.
19. Li, Q., W. Cai, X. Wang, et al., "Medical image classification with convolutional neural network," *International Conference on Control Automation Robotics and Vision. IEEE*, 844–848, 2014.
20. Chen, S., H. Wang, F. Xu, and Y.-Q. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Transactions on Geoscience & Remote Sensing*, Vol. 54, No. 8, 4806–4817, 2016.
21. Zhou, Y., H. Wang, F. Xu, and Y.-Q. Jin, "Polarimetric SAR image classification using deep convolutional neural networks," *IEEE Geoscience & Remote Sensing Letters*, Vol. 13, No. 12, 1935–1939, 2016.
22. Hirose, A., *Complex-Valued Neural Networks: Advances and Applications*, Wiley, 2013.
23. Zhang, Z., H. Wang, F. Xu, et al., "Complex-valued convolutional neural network and its application in polarimetric SAR image classification," *IEEE Transactions on Geoscience & Remote Sensing*, Vol. 99, 1–12, 2017.
24. Zeiler, M. D., G. W. Taylor, R. Fergus, et al., "Adaptive deconvolutional networks for mid and high level featurelearning," *International Conference on Computer Vision*, 2018–2025, 2011.
25. Ren, X. and J. Malik, "Learning a classification model for segmentation," *International Conference on Computer Vision*, Vol. 1, 10–17, 2003.
26. Cloude, S. R. and E. Pottier, "A review of target decomposition theorems in radar polarimetry," *IEEE Transactions on Geoscience & Remote Sensing*, Vol. 34, No. 2, 498–518, 1996.
27. Zhu, B., J. Z. Liu, S. F. Cauley, et al., "Image reconstruction by domain-transform manifold learning," *Nature*, 2017.
28. Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, Vol. 86, No. 11, 2278–2324, 1998.
29. Zhang, Y., W. Miao, Z. Lin, H. Gao, and S. Shi, "Millimeter-wave InSAR image reconstruction approach by total variation regularized matrix completion," *Remote Sens*, Vol. 10, 1053, 2018.