Adversarial Network With Higher Order Potential Conditional Random Field for PolSAR Image Classification

Zheng Zhang, Hui Guo, Jingsong Yang, Xianggang Wang, and Yang Du, *Senior Member, IEEE*

***Abstract*—Effective and efficient pixel-level classification of polarimetric synthetic aperture radar (PolSAR) images represents an important step toward their interpretation and knowledge discovery. After the extraction and representation of discriminative features, the related important issues are how to enforce consistency and coherency of labeling using contextual information, and how to make the process computationally efficient. In this article, we propose a method to approach these two issues. The first issue is dealt with from three different levels, namely, first, to combine features at coarse resolution with pixel-wise information for pixel-level classification, we adopt the idea of Unet; second, to reduce labeling inhomogeneity among similar pixels, we relate the PolSAR image with graph theory through the conditional random field (CRF), and develop third-order potentials for a fully connected CRF to account for both higher order effects and long-range interactions; and, third, to utilize adversarial network to improve learning of the label distribution. The efficiency issue is handled by, first, adopting fully convolutional network for the contracting arm of the Unet to speed up hierarchical feature extraction; and, second, for the fully connected third-order potential, by making connection with the pairwise potential, to reduce the computational complexity of the most expensive message passing procedure from quadratic to linear in the number of variables. The effectiveness of the proposed method has been demonstrated by its application to the pixel-level classification of ALOS-2, RADARSAT-2, and ESAR PolSAR images, where its performance has qualitatively and quantitatively shown superiority over several other state-of-the-art models.**

***Index Terms*—Adversarial network, autoencoder (AE), conditional random field (CRF), feature extraction, higher order potential.**

Manuscript received 13 March 2023; revised 1 June 2023 and 11 August 2023; accepted 19 September 2023. Date of publication 9 October 2023; date of current version 28 December 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3903503 and in part by the China National Science Foundation under Grant 41771367. *(Corresponding author: Yang Du.)*

Zheng Zhang, Xianggang Wang, and Yang Du are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: 376187581@qq.com; wangxg_zs@163.com; zjuydu03@zju.edu.cn).

Hui Guo is with the 503 Institute, China Academy of Space Technology, Beijing 100048, China (e-mail: 13661160029@139.com).

Jingsong Yang is with the State Key Laboratory of Satellite Ocean Environment Dynamics, Second Institute of Oceanography, Ministry of Natural Resources, Hangzhou 310012, China, and also with the Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519082, China (e-mail: jsyang@sio.org.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3322344

I. Introduction

Recent years have witnessed the rapid progress of polarimetric synthetic aperture radar (PolSAR) as a major microwave remote sensing technique for Earth observation, urban planning, geological exploration, and environment monitoring. A deluge of Earth system data has been rapidly expanding enabled by the much enriched sensing capabilities. Yet our ability to collect and create geoscience data far outpaces our ability to understand it [1]. To fill the gap, effective and efficient classification of PolSAR images represents an important step toward their interpretation and knowledge discovery, thus, forms a line of active research.

Pixel-level classification or labeling is an important task for image understanding. It is also known as full-scene labeling, or scene parsing in the computer vision literature. High quality pixel labeling enables the delineation and tagging of regions and objects in the image, which becomes increasingly pragmatic and important with the ever improving spatial resolution of SAR images. The related important issues include how to extract and represent features, how to enforce consistency and coherency of labeling using contextual information, and how to make the process computationally efficient and flexible.

The first issue has been addressed satisfactorily with the advance of deep learning and its applications in image processing, so it is not a concern of this article. Briefly speaking, traditional manual feature engineering, such as GLCM has been replaced by automatic feature extraction, and usually in hierarchial manner. For instance, convolutional networks composed of multiple stages can automatically learn hierarchical feature representations. This part of network for feature extraction is usually called the encoder. Using a small region around a pixel for labeling is often insufficient, since the category of a pixel may depend on long-range information as well, a multiscale convolutional network has been proposed to capture large input windows while maintaining minimum number of free parameters. One common choice for the encoder network is the VGG16 classification network, yet its large number of weights is typically pretrained on the large ImageNet object classification dataset [2]. In the work, we choose an approach similar to the fully convolutional network (FCN) for hierarchical feature representations.

We approach the second issue from three different levels. At first level, to combine features at coarse resolution with pixel-wise information for pixel-level classification, we adopt the idea of Unet [3], whose appeal lies in the simultaneous attainment of localization accuracy and contextual scope. Deep, semantic, and coarse-grained feature maps are offered by a contracting subnetwork, whereas shallow, low-level, and fine-grained feature maps resulted from the upsampling expansive subnetwork; skip connections at the same-scale feature maps of the contracting and expansive subnetworks provide the needed information fusion. Such structures have demonstrated effectiveness in their good segmentation/classification performance.

The Unet-like network, although capable of modeling global relationships within an image already, might still be error-prone and not lead to the desired performance. Due to the coherent imaging mechanism of PolSAR systems, the resulting speckle noise degrades the quality of PolSAR images and presents a big challenge to PolSAR image classification task. Moreover, natural scene often contain complex undulating topograph, leading to more complicated special texture structure and spatial arrangement, such as foreshortening, shadow, and layover in the PolSAR image, which adds complication to the classification efforts.

One common way to address this issue is to adopt the idea of ensuring classification label agreement between similar pixels by maximum a posteriori inference in a conditional random field (CRF) defined over pixels or image patches, where its potentials incorporate smoothness terms to maximize label agreement. Conventional CRF models use neighborhood pairwise potentials to account for connections among neighboring pixels or patches, hence, limited their ability to model long-range connections within the image. A recently advanced fully connected CRF [4] establishes pairwise potentials on all pairs of pixels in the image to capture richer semantic information for refined segmentation and labeling. This approach will be adapted to our end in this study to train the whole deep network end-to-end with the usual back-propagation algorithm, avoiding artificial offline postprocessing.

However, the fully connected CRF [4] does not account for higher order potentials. It is observed [5] that pairwise CRFs generally result in oversmoothing the actual object contour, and higher order CRFs are needed for further improvement. Some forms of higher order CRFs have been proposed [5], [9], yet they are based on image patches, without taking into account the effect of long-range connection.

In this work, we consider the case of the third-order potentials in the context of fully connected CRF, bringing together the discriminative power rendered by higher order potentials and long-range interaction. This treatment forms the second level of our approach to the second issue.

However, for image understanding one has to consider the observation level problem, which refers to the situation where an object or its part can be easily classified provided that its segmentation is performed at the right level [7]. The abovementioned two levels in combination may still not adequately address the observation level problem in that there is certain degree of arbitrariness in their segmentation of the image by the resultant segments being too small or too large. Unlike the strategy of [7] to form and prune a segmentation tree, we treat this problem as a problem of learning the true distribution of random labeling field. We design an adversarial network to this end.

The inspiration of our work on adopting adversarial networks comes from the generative adversarial networks (GANs) [8]. GANs are recently advanced powerful framework and have made impressive progresses in image generation, image editing, and representative learning. More recent methods, such as text2image and img2img, also have been inspired by GANs. In general, GANs are powerful for learning generative models through the idea of an adversarial loss that forces the generated images to demonstrate a distribution maximally indistinguishable from the true distribution of true images. The assumption here is that in carrying out the third level of incorporating contextual information, a truly learned distribution of the random label field will more likely to help set the segmentation at the right level so as to avoid the over- or undersegmentation problem, which in turn is beneficial to approach the observation level problem.

On the other hand, keep in mind that one has to deal with several million pixels typically contained in one PolSAR image, the training and inference need to be computational efficient and capable of working with arbitrary image size. For hierarchial feature extraction, we use the FCN so as to capture large input windows while maintaining minimum number of free parameters. For the third-order potential, upon making an assumption about the first pixel severing as separating point of the other two, we are able to preserve the nice properties of the fully connected CRF [4]. Specifically, by modeling the pairwise edge potentials as a linear combination of Gaussian kernels in an arbitrary feature space, and further adopting a mean field approximation to the CRF distribution, a highly efficient inference algorithm is resulted, which allows for reduction of the computational complexity of message passing from quadratic to linear in the number of variables.

The main novelties and contributions of our proposed method are as follows.

1) We proposed an architecture for high performance PolSAR image pixel-level classification by integrating coarse resolution with pixel-wise information in an FCN, which not only capable of dealing with the contextual information at feature level, but also enables the efficient training and inference of large-scale PolSAR images.
2) We developed an efficient algorithm for third-order potential of fully connected CRF to account for higher order effects and long-range interactions. Specifically, the third-order potential can be factored as the product of two pairwise potentials, allowing for the interpretation in terms of Markov properties. The corresponding optimal marginal distribution, which minimizes the Kullback–Leibler divergence, captures the interaction among multiple labels and propagate efficiently, with a linear time complexity in the number of variables during the message passing phase.
3) We presented an adversarial network that learns the true distribution of random labeling fields, enhancing label consistency and coherence at the observation level. Additionally, we systematically dealt with contextual information from these three different levels to maximize its

utilization for PolSAR image classification in an end-to-end trainable network.

The rest of this article is organized as follows. In Section II, we briefly review some related works. The proposed method is described in Section III. In Section IV, we apply the proposed method to three PolSAR images to demonstrate its effectiveness. Comparison is also made to other state-of-the-art methods. Finally, Section V concludes this article. Before passing, we would like to mention that the terms segmentation and (pixel level) classification are used interchangeably in the text.

II. Related Works

Deep learning, as a hierarchical feature learning method, has been widely used in image processing tasks (e.g., [10], [11], [12], [13]). Recent years also have witnessed its numerous application in remote sensing [1], [17], [18], [46]. In contrast to the traditional manual feature engineering method [19], [20], [21], [22], DL allows the automatical feature extraction pertaining to specific tasks. Furthermore, deep learning-based models possess attractive properties, such as unique hierarchical representations and nonlinear expressiveness. These excellent deep learning methods are powerful tools for feature extraction and expression in remote sensing image classification task. The applications of deep learning to remote sensing image classification was initially on unsupervised feature learning using stacked autoencoders (AEs) and restricted boltzmann machines [23], [24], [25], with the aim of learning higher level feature representations from spectral information of optical image or polarimetric feature of the SAR/PolSAR image. This process has been shown to provide better land cover classification accuracy than using raw features. The application of the convolutional neural networks (CNNs) represents a great advance in deep learning assistance to remote sensing image classification [26], [27], [28], with even shallow learned layers achieving significantly better performance than traditional methods, such as random forest or support vector machine classifiers. Modern deep learning classification architectures for remote sensing have grown much larger and more complex compared to earlier networks, where some utilize designs originally developed for computer vision applications, including VGG, Inception, and ResNet, which are tailored with additional classification layers for remote sensing tasks. These advanced architectures have consistently achieved excellent classification accuracy [29], [30]. Graph convolutional networks (GCNs) have been shown to achieve state-of-the-art results in several remote sensing tasks [31], [32] due to their capability of effectively capturing the spatial relationships between pixels in an image, an important aspect in remote sensing tasks since neighboring pixels are often highly correlated. Furthermore, domain adaptation techniques, such as multimodal learning and transfer learning [33], [34] are promising in handling the discrepancies between heterogeneous data, and thus increasing the model's generalizability and robustness.

In the following content, we mainly overview the previous work related to our two focuses, that is, how to enforce consistency and coherency of labeling using contextual information, and how to make the process computationally efficient and flexible.

A. Pixel-Level Classification

The pixel-level classification framework is to assign a label to each pixel of the input image. In recent years, many excellent networks have emerged, such as FCN [35], Unet [3], SegNet [36], PSPNet [37], and DeepLabV3 [38]. Some of the abovementioned models have been applied to PolSAR classification. FCN is the first network structure to deal with PolSAR image classification in the pixel wise way, due to its simplest upsampling operation. In [9], semantic features integrated with sparse and low rank representation utilized for PolSAR classification is proposed. The semantic features are generated by the final layer of the unsupervised FCN. However, this approach does not directly use FCN structure for pixel-level classification. In [39], a sliding window fully convolutional network and sparse coding is proposed, which provides an FCN scheme in a supervised manner for PolSAR image classification. In [40], in order to make full use of the phase information in PolSAR data and reduce the loss of spatial information caused by pooling operation, a complex-valued FCN is proposed for PolSAR image classification. The decoder of SegNet uses pooled indexes calculated in the max pooling step of the corresponding encoder to perform nonlinear upsampling in order to preserve the edge information, which is beneficial to smooth the edge of terrain. PSPNet adopts a spatial pyramid pooling to get multiresolution feature map and concatenates them after upsampling. DeeplabV3 proposes an atrous spatial pyramid pool, which uses atrous convolution. These two networks can obtain multiscale information. In [41], a refined pyramid scene parsing (PSP) network is proposed, which adopts a multilevel features fusion design in its decoder to effectively exploit the features learned from its different encoder branches. The experimental results demonstrate its effectiveness in keeping the boundary information of agricultural area. In [42], a lightweight complex-valued DeepLabV3+ is proposed to deal with insufficient PolSAR data samples. Although the abovementioned method improves the process efficiency in a pixel wise manner for PolSAR image classification, they do not consider the label consistency problem.

It should mentioned that PolSAR semantic segmentation is also an active research area (e.g., [15], [16]). It would be interesting to seek synergy between these two lines of research.

B. Conditional Random Field

CRF regularization is a common method to deal with label consistency problem. A recently advanced fully connected CRF [4] establishes pairwise potentials on all pairs of pixels in the image to capture richer semantic information for refined segmentation and labeling. In [43], a semisupervised term is incorporated into the energy function to improve the classification accuracy of PolSAR images with small samples by using unlabeled terrain. In [44], a hybrid CRFs model based on complex-valued 3-dimensional CNN is proposed, which incorporates a relative entropy pairwise potential and a product of expert potential term to regulate label interaction for the PolSAR image classification. In [45], a high-order potentials generated by superpixels is incoporated in CRF to deeply capture the structure information for PolSAR images. The high-order potentials enforce all pixels in homogeneous clique to take the

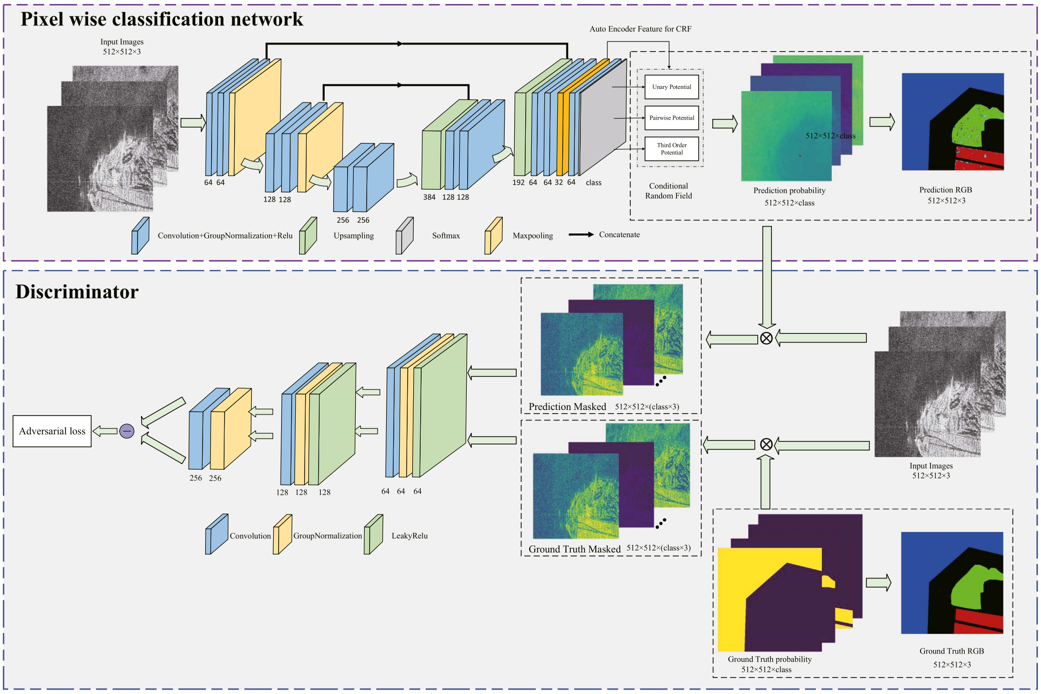


Fig. 1. Adversarial network with third-order potentials for PolSAR image pixel-level classification.

same label. The classification network of the abovementioned methods is still image-level CNN and is independent from CRF training. More importantly, both pairwise and high-order potentials they use are short connected versions that only consider local structures. In short, although the abovementioned methods use CRF to deal with the problem of label consistency, they are still insufficient in the consideration of processing efficiency and the contextual information of label distribution. This is mainly because these CRF methods still address local information, and cannot achieve end-to-end training with feature extraction network.

C. Adversarial Learning

The typical network of adversarial learning is GANs [8], which learn data distributions through adversarial training. Recently, some GAN-based networks have been proposed successively for SAR/PolSAR classification. In [46], to tap into the strength of GANs and its more stable variant deep convolutional GANs, it is proposed of its combination with AE for intermediate feature extraction and reconstruction, and with multiclassifier for better context treatment. The aggregate learned features are fed into a CRFs model, which is trained by the structured support vector machine, to obtain the classification. In [47], to utilize simultaneously statistical features and spatial features, a distribution and structure match auxiliary classifier generative adversarial network is proposed for discriminative feature learning. This is in spirit similar to [46], with the major difference that CRF is not considered in this work. In [54], to enforce the generator for learning useful feature representation, a task-oriented GAN is proposed for PolSAR image classification. Variants of GAN networks in the abovementioned methods have been used to obtain rich statistical features of terrain. The abovementioned methods take GANs as feature extraction network. A cycle-consistent adversarial networks [49] is proposed, which incorporates an adversarial loss to solve the highly underconstrained problem that may arise when mapping image from one domain to another. It brings a new insight to PolSAR classification, that is, the classification problem can be regarded as a mapping from image domain to label domain.

III. Proposed Method

The overall architecture of the proposed classifier is composed of two subnetworks: 1) a forward classifier for pixel level labeling, which consists of a feature extraction network, a CRF, and an AE and 2) a discriminator for adversarial learning, as shown in Fig. 1.

For a PolSAR image, its image features can be expressed $\mathbf{X} \in R^{N_1 \times N_2 \times N_3}$, where N_1 and N_2 represent the height and width of the image, respectively, N_3 the raw feature of the image, and $\mathbf{T} \in R^{N_1 \times N_2 \times C}$ is the corresponding ground truth for a total of C classes/categories. The proposed model will assign a predicted label to each pixel.

A. Raw Feature Extraction From PolSAR Data

For PolSAR radar data, the scattering matrix is a 2×2 complex matrix that describes the polarization properties of a target, based on the complex-valued radar signals received in the H and V polarization channels. It is formed as

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

There are several polarimetric decomposition methods used in PolSAR data analysis, including the Pauli decomposition, Freeman–Durden decomposition, and the Cloude–Pottier decomposition. These decompositions provide different ways of separating the polarimetric scattering mechanisms of the target, based on various assumptions and constraints. In [67], a physically-based assessment of the assumptions of the Freeman–Durden decomposition is provided.

In this work, the Pauli decomposition is adopted, which represents scattering information in terms of three fundamental scattering mechanisms: single bounce, double bounce, and volume scattering. By projecting the scattering matrix onto the Pauli bases, the scattering coefficients associated with each mechanism are obtained. This technique offers several advantages, such as computational efficiency, generating pseudo-RGB images with higher contrast for visualization as demonstrated in Section IV, and bypassing the need for estimating the coherence/covariance matrix through a sliding window, thus eliminating associated estimation errors.

For monostatic scattering case which satisfies the reciprocity condition, given the Pauli bases the measured scattering matrix can be expressed as follows:

$$\mathbf{S} = \frac{\alpha}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \frac{\beta}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + \frac{\gamma}{\sqrt{2}}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{2}$$

where $\alpha = \frac{S_{hh}+S_{vv}}{\sqrt{2}}$, $\beta = \frac{S_{hh}-S_{vv}}{\sqrt{2}}$, $\gamma = \sqrt{2}S_{hv}$. The polarimetric information of $\mathbf{S}$ could be represented by the combination of the intensities $|\alpha|^2$, $|\beta|^2$, $|\gamma|^2$ in a single RGB image.

B. Feature Extraction: From Fully Connected CNN to Fully Convolutional-Based Neural Network

Contracting subnetwork of a Unet structure is used for feature extraction. However, we make modifications to facilitate the computation. In the typical deep learning-based PolSAR image classification network (e.g., [27], [46], [54]), the input is a fixed size patch; the fully connected layers are of fixed size, and this layer converts the two-dimensional feature map of the convolutional layer into a one-dimensional feature vector, thus losing certain amount of spatial information. However, the fully connected layer can be regarded as a special convolutional layer, which enables the network to take input images of arbitrary size, and produce corresponding classification map. Fig. 2 shows a schematic representation of this convolutional transformation. The block diagram of (1) in Fig. 2 shows the model proposed in [27], where the input building patch is of size 16×16. After three layers of convolution layers and two layers of full connection, the output is the probability of five categories. The fully connected layer can be regarded as the processing of the third convolutional layer with a convolution kernel of size 4×4 and stride 1, as shown in the block diagram (2). This process can be readily extended to an image of size $N_1 \times N_2$, as shown in block diagram (3). In this way, only one forward propagation is required, and the input image can be of arbitrary size, allowing more flexible processing.

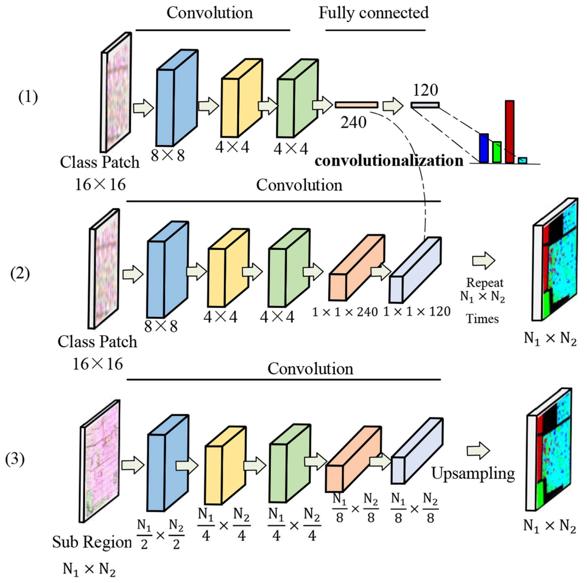


Fig. 2. Transforming fully connected CNN to fully convolutional-based neural network.

C. Contextual Information Through CRF

The labels Y_i of all pixels i in the PolSAR image form a random field $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$. Let $\mathbf{X}$ be all possible input images of size N, then the CRF $P(\mathbf{Y} = y|\mathbf{X})$ is characterized by a Gibbs distribution

$$P(\mathbf{Y} = \mathbf{y}|\mathbf{X}) = \exp(-E(\mathbf{y}|\mathbf{X}))/Z(\mathbf{X}) \tag{3}$$

where $Z(\mathbf{X})$ is the partition function and $E(\mathbf{y}|\mathbf{X})$ is the Gibbs energy of a labeling y. For notational convenience, the conditioning will be dropped in the rest of this article.

The Gibbs energy based on consideration of the unary potential, pairwise potentials, and triplet potentials of CRF is

$$\begin{aligned} \mathrm{E}(\mathbf{y}) = &\sum_i \varphi_u(y_i) + \sum_{i<j} \varphi_p(y_i, y_j) \\ &+ \sum_{i<j<k} \varphi_{\mathrm{t}}(y_i, y_j, y_k) \end{aligned} \tag{4}$$

where the unary potential φ_{u} is the cost associated with assigning label y_i to pixel i, the pairwise potentials φ_{p} are to maintain the consistency of similar feature map classes in adjacent locations and to enhance the smoothness of the estimated class labels. Notice here that we do not explicitly denote the clique c because we intend to adopt the fully connected form in this article. We go beyond the pairwise potential to include the effect of higher order potentials to capture the interaction among multiple labels. For the purpose of illustration, we treat the third-order potential φ_{t} in (4) explicitly.

1) Unary Term: The unary potential $\varphi_u(y_i)$ is the negative log of the likelihood of pixel i being assigned a label y_i. Its discriminative feature depends on many factors including the pixel values, texture, location, appearance model. This complex mapping is represented by a Unet structure in this work, which outputs a feature map with the same size as the original image, as shown in Fig. 1. For the Unet, the contracting path is composed of two blocks, where each block uses two 3×3 convolutional layers, one group normalization layer, one Relu layer, and one 2×2 maxpooling layer; so after each downsampling, the feature map is reduced by half in size, and the number of channels is doubled. Finally, in the bottleneck layer, the feature map is of size $N_1/4 \times N_1/4$. Similarly, the expansive path is also composed of two blocks, where each blcok uses two 3×3 convolutional layer, one group normalization layer, one Relu layer. After the deconvolution operation performed at the first expansion path block, the feature map is doubled in size and the number of channels is reduced by half. The feature map is merged with its counterpart in the contracting path. After two deconvolution operations, the size of the feature map is the same as the original input. The feature map is followed by the Softmax layer for pixel-by-pixel classification.

The convolutional layer contains filter parameters that can be learned. Deeper levels of abstraction can be achieved by stacking more convolutional layers. However, blindly deepening the network will bring issues, such as the Internal Covariate Shift problem [56], [57]. We adopt group normalization layer after the convolutional layer to make the features learned after each update of the network to have a similar distribution, thereby facilitating network convergence and improving generalization capability [57]. The introduction of Relu layer can improve the network's capability of expressing nonlinear features. Max pooling layer is able to not only reduce the number of network parameters to prevent overfitting, but also ensure feature invariance under shifting and rotation. The Softmax Layer performs a pixelwise classification over the feature map having identical size with the original image. The pixelwise Softmax function is as follows:

$$p_k\left(y_i|x_i\right) = \frac{e^{(a_k(x_i))}}{\sum_{k'=1}^{C} e^{(a_{k'}(x_i))}} \tag{5}$$

where $a_k(x_i)$ represents the activation feature at the pixel point x_i, $k \in \{l_1, l_2, \ldots, l_C\}$, C is the number of categories.

2) Pairwise Potentials: The pairwise potentials are given as follows:

$$\varphi_p\left(y_i, y_j\right) = \mu(y_i, y_j) \underbrace{\sum_{m=1}^{M} w^{(m)} g^{(m)}\left(\mathbf{f_i}, \mathbf{f_j}\right)}_{g(\mathbf{f}_i, \mathbf{f}_j)} \tag{6}$$

where each $g^{(m)}$ is a Gaussian kernel, $\mathbf{f_k}$ is the feature vector at pixels k, $w^{(m)}$ are linear connection weights, and $\mu(\cdot)$ is a label compatibility function. Following [4], we use contrast-sensitive two-kernel potentials for the composite kernel $g(\mathbf{f}_i, \mathbf{f}_j)$, defined as follows:

$$g\left(\mathbf{f}_i, \mathbf{f}_j\right) = w^{(1)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|s_i - s_j|^2}{2\theta_\beta^2}\right) + w^{(2)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right) \tag{7}$$

where the first term is the appearance kernel and the second the smoothness kernel. The smoothness kernel solely depends on the pixel pair locations and removes small isolated regions. The dependencies of the appearance kernel are not only the pixel pair locations p_i, but also some form of features s_i at pixel i, for instance, the color vector in [4]. In our approach, we use for s_i the output of an embedded AE stacked between the Unet and CRF. This treatment marks an essential departure from [4], where the color vector pair I_i and I_j are used instead. One should keep in mind that in [4] the ability to use the color vector pair depends on two implicit assumptions: 1) the image is adequately clear of noise; and 2) the subfeature vector (the color vector) is well localized. These assumptions are automatically met in [4] considering the following facts: 1) the images under concern are RGB images; and 2) the only operation applied is the CRF-based filtering. However, in our case, the situation is quite different. The PolSAR images suffer from speckle noises, and the processing using deep learning networks usually compromises localization accuracy to accommodate contextual information. The latter can be alleviated by working with Unet, whereas the issue of speckle noise can be dealt with by either applying denoising preprocessing or using manifold learning. We choose manifold learning by attaching an AE to the last feature map layer (LFML) of the Unet, where the feature dimension is reduced from 64 to 32. This distilled feature $\mathbf{s}$ is used in the appearance kernel. This treatment brings forth the following advantages:

1) it conveys more semantic information without loss of localization accuracy in comparison with the "vanilla" color vector;
2) it leads to dimension reduction of the representation space [65] and, hence, better efficiency than the base case without the AE, i.e., 64-dimensional feature vector has to be used instead;
3) the utility of the AE serves as an extra constraint to the network loss function, hence, assist in the stabilization of the training.

3) Third-Order Potentials: It is observed in [5] that pairwise CRFs tend to be oversmooth and mismatch the actual object contour quite often, and higher order CRFs are needed for further improvement. Here, we consider the case of the third-order potential in a hope to effectively model the interaction among three labels in certain contrast sensitive form. We would like to make it fully connected and computationally efficient. So we shall proceed along the line of [4]. When including the third-order potential, the optimal marginal $Q_i(y_i)$ minimizing the KL-divergence is

$$Q_i\left(y_i\right) = \frac{1}{Z_i} \exp\left\{-\varphi_u\left(y_i\right) - \sum_{j \neq i} E_{Y_j \sim Q_j}\left[\varphi_p\left(y_i, Y_j\right)\right] - \sum_{j \neq k \neq i} E_{Y_j \sim Q_j, Y_k \sim Q_k}\left[\varphi_t\left(y_i, Y_j, Y_k\right)\right]\right\}. \tag{8}$$

In (8), a method of efficiently updating the pairwise potentials has been proposed in [4]. Here, we should focus on the third-order potential. In view of the fact that several assumptions have been made in the fully connected pairwise potential case to allow for the application of facilitating techniques, it is expected that we shall likewise to make some assumptions in the treatment of the third-order potential.

Specifically, in this work, we adopt the approximation that the third-order potential can be factored as

$$\varphi_t(y_i, y_j, y_k) = \varphi_p(y_i, y_j)\,\varphi_p(y_i, y_k). \tag{9}$$

It follows that the term associated with the third-order potential in (8) can be manipulated as follows:

$$\begin{aligned}
&\sum_{j\neq k\neq i} E_{Y_j\sim Q_j, Y_k\sim Q_k}\left[\varphi_t(y_i, Y_j, Y_k)\right]\\
&= \sum_{j\neq k\neq i} E_{Y_j\sim Q_j, Y_k\sim Q_k}\left[\varphi_p(y_i, Y_j)\,\varphi_p(y_i, Y_k)\right]\\
&= \sum_{j\neq k\neq i} E_{Y_j\sim Q_j}\left[\varphi_p(y_i, Y_j)\right] E_{Y_k\sim Q_k}\left[\varphi_p(y_i, Y_k)\right]\\
&= \left(\sum_{j\neq i} E_{Y_j\sim Q_j}\left[\varphi_p(y_i, Y_j)\right]\right)\left(\sum_{k\neq i} E_{Y_k\sim Q_k}\left[\varphi_p(y_i, Y_k)\right]\right)\\
&\quad - \sum_{j\neq i}\left(E_{Y_j\sim Q_j}\left[\varphi_p(y_i, Y_j)\right]\right)^2\\
&= \left(\sum_{j\neq i} E_{Y_j\sim Q_j}[\varphi_p(y_i, Y_j)]\right)^2 - \sum_{j\neq i}\left(E_{Y_j\sim Q_j}[\varphi_p(y_i, Y_j)]\right)^2.
\end{aligned} \tag{10}$$

The subtraction in the next to the last line is due to the fact that when writing the first term in this way to separate the two indices j and k, the repeating terms must be accounted for and subtracted.

The first term can be further evaluated as

$$\begin{aligned}
&\left(\sum_{j\neq i} E_{Y_j\sim Q_j}\left[\varphi_p(y_i, Y_j)\right]\right)^2\\
&= \left(\sum_{l'\in L}\mu(l, l')\sum_{m=1}^{K}\omega^{(m)}\sum_{j\neq i} g^{(m)}(\mathbf{f}_i, \mathbf{f}_j)\,Q_j(l')\right)^2.
\end{aligned} \tag{11}$$

In the above, we use label l for y_i and l' for Y_j to simplify the notation. Although the right side of (11) appears to be quadratic in Q_j, the marginal of any other pixel, the message passing and compatibility transform operations are evident in the expression, and the computationally expensive message passing represented by the term

$$\sum_{j\neq i} g^{(m)}(\mathbf{f}_i, \mathbf{f}_j)\,Q_j(l')$$

can be approximately treated in the same manner as in [4] to allow for reduction of the computational complexity from quadratic to linear in the number of variables.

The analysis of the second term in (10) is somewhat more involved as follows:

$$\begin{aligned}
&\sum_{j\neq i}\left(E_{Y_j\sim Q_j}\left[\varphi_p(y_i, Y_j)\right]\right)^2\\
&= \sum_{j\neq i}\left(\sum_{l'\in L} Q_j(l')\,\mu(l, l')\sum_{m=1}^{K}\omega^{(m)} g^{(m)}(\mathbf{f}_i, \mathbf{f}_j)\right)^2\\
&= \sum_{j\neq i}\sum_{l'\in L}\sum_{l''\in L} Q_j(l')\,Q_j(l'')\,\mu(l, l')\mu(l, l'')\\
&\quad \sum_{m=1}^{K}\sum_{m'=1}^{K}\omega^{(m)}\omega^{(m')} g^{(m)}(\mathbf{f}_i, \mathbf{f}_j) g^{(m')}(\mathbf{f}_i, \mathbf{f}_j)\\
&= \sum_{l'\in L}\sum_{l''\in L}\mu(l, l')\mu(l, l'')\sum_{m=1}^{K}\sum_{m'=1}^{K}\omega^{(m)}\omega^{(m')}\\
&\quad \sum_{j\neq i} g^{(m)}(\mathbf{f}_i, \mathbf{f}_j) g^{(m')}(\mathbf{f}_i, \mathbf{f}_j)\,Q_j(l')\,Q_j(l'')\\
&= \sum_{l'\in L}\sum_{l''\in L}\mu(l, l')\mu(l, l'')\sum_{m=1}^{K}\sum_{m'=1}^{K}\omega^{(m)}\omega^{(m')}\\
&\quad \sum_{j\neq i}\tilde{g}^{(m,m')}(\mathbf{f}_i, \mathbf{f}_j)\tilde{Q}_j(l', l'').
\end{aligned} \tag{12}$$

Here, the message passing part is

$$\sum_{j\neq i} g^{(m)}(\mathbf{f}_i, \mathbf{f}_j) g^{(m')}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l')\,Q_j(l'').$$

This form can be analyzed by adding one layer of abstraction to see its connection with the pairwise potential case by defining

$$\tilde{g}^{(m,m')}(\mathbf{f}_i, \mathbf{f}_j) = g^{(m)}(\mathbf{f}_i, \mathbf{f}_j) g^{(m')}(\mathbf{f}_i, \mathbf{f}_j)$$

we see that $\tilde{g}$ is actually the product of two low-pass filters, and is again a low-pass filter itself. Similarly, let

$$\tilde{Q}_j(l', l'') = Q_j(l')\,Q_j(l'')$$

represent the composite marginals, so the message passing term looks very close to the view of low-pass filtering of the composite marginals, with a bit book keeping similar to the treatment of the pairwise potential, this filtering can be made efficient and again allow for reduction of the computational complexity from quadratic to linear in the number of variables.

Before passing we would like to make some comments. It might be suggested that factorization of the third-order potential in the form of $\varphi_t(y_i, y_j, y_k) = \varphi_p(y_i, y_j)\varphi_p(y_i, y_k)\varphi_p(y_j, y_k)$ seem more balanced among the three labels. Yet we argue that the proposed two-term factorization is conceptually advantageous than the three-term counterpart. It permits interpretation in terms of Markov properties where pixel i serves as the separating point of pixels j and k, hence, make sense to invoke the Hammersley–Clifford theorem to ensure the equivalence between the image process being Markov random field and its distribution being a Gibbs one. In the fully connected context, it means that joint distribution between any pair or triplet is through cliques involved in the factorization.

One might doubt that when using pixel i as the separating point much speciality has been assigned to it. This is actually not the case since when we scan any triplet combination in the image, each pixel will take turn to serve as the separating point. So in the long run the role of each pixel is well balanced.

4) Forward Classifier Loss: The forward classifier loss now consists of two parts: the loss due to CRF and that due to AE. For the fully connected CRF, the exact distribution $P(\mathbf{Y})$ is evaluated using the mean field approximation [4], where an alternative distribution $Q(\mathbf{Y})$ is sought after which is the best among the family of distributions that can be expressed as a product of independent marginals $Q(\mathbf{Y}) = \prod_i Q_i(Y_i)$, in the sense that $Q(\mathbf{Y})$ minimizes the KL-divergence $D(Q||P)$. In view of [4], the MFA can be implemented using convolutional layers to be added to the network, so its loss can be calculated using cross entropy. The combined forward classifier loss is, thus

$$\begin{aligned}\text{Loss}_{G_1} = & -\sum_{k=1}^{C}\sum_{i=1}^{N}\{t_i = l_k\}\,\text{log}p_k\left(y_i|x_i\right)\\ & + \text{MSE}(\text{Enc}(\text{LFML}), \text{Dec}(\text{Enc}(\text{LFML}))\end{aligned} \quad (13)$$

where MSE stands for mean square error.

D. Contextual Information Through Adversarial Learning

The pixel level labeling produces a classification probability map of size $\mathbf{Y} \in R^{N_1 \times N_2 \times C}$. Denote the label distribution as $P_G(\mathbf{Y}; \theta_G)$, where θ_G are the parameter set of the distribution, and the true label distribution as $P_{\text{data}}(\mathbf{Y})$. The intention is to make $P_G(\mathbf{Y}; \theta_G)$ and $P_{\text{data}}(\mathbf{Y})$ as close as possible by adversarial learning. In particular, with the enhanced structure combining the strengths of Unet, CRF, and AE, the capability of the forward classifier to make good labeling is much improved, yet can be further improved with the adversarial learning where a discriminator is designed to challenge the distribution learning.

The architecture of the discriminator is a fully CNN, as shown in Fig. 1. It is composed of 3 blocks and an output layer. Each block uses one 4×4 convolutional layer, one group normalization layer, and one Leaky Relu Layer. As the number of layers deepens, the number of convolutions per block doubles. The convolutional layer uses the stride operation instead of the max pooling operation, so for each block the size of its output feature map is $1/2$ of its input. We adopt the loss function of WGAN as the adversarial term in order to alleviate the unstable numerical gradient that may occur during the training of the forward classifier. It should be mentioned that since the discriminator uses a fully convolutional structure, the network top feature map is two-dimensional. Under the two-dimensional feature map, each element position actually represents a fixed-size perception area of the original image, with the effect of performing sliding window processing on the original image, hence refining the capture of local features. The loss function of the discriminator is defined as follows:

$$\begin{aligned}\text{Loss}_D = & E_{\mathbf{Y}\sim P_r}\left[D\left(\mathbf{X}\bigotimes\mathbf{Y}\right)\right]\\ & - E_{G(\mathbf{X})\sim P_g}\left[D\left(\mathbf{X}\bigotimes G\left(\mathbf{X}\right)\right)\right]\end{aligned} \quad (14)$$

where $\mathbf{X} \bigotimes \mathbf{Y}$ means that each channel of $\mathbf{X}$ is masked by $\mathbf{Y}$, which can be regarded as a conditional term to constrain the model parameters. The discriminator should try to maximize the EM distance between real samples and false samples, that is, to maximize (14).

E. Training

1) Training the Discriminator Model: From the abovementioned analysis, it is seen that the discriminator can be trained by maximizing the EM distance in (14). The input to the discriminator network is either the prediction masked images by the forward classifier or the ground truth masked images, as shown in Fig. 1. The Adam algorithm is used to learn the parameters of the discriminator network.

2) Training the Forward Classifier Model: For the forward classifier, the adversary loss term is added to (13) to form the overall loss

$$\text{Loss}_G = \text{Loss}_{G_1} - E_{G(\mathbf{X})\sim P_g}\left[D\left(\mathbf{X}\bigotimes G\left(\mathbf{X}\right)\right)\right]. \quad (15)$$

This serves to minimize the multiclass cross-entropy loss function and maximize the chance to deceive the discriminator. The Adam algorithm is also used to learn the parameters of the forward classifier network.

IV. Experimental Results and Discussion

In the experiments, we apply the proposed method to pixel-level classification of three PolSAR datasets to demonstrate its effectiveness. The quantitative measures are the overall accuracy (OA) and kappa coefficient (KAPPA) for the overall performance and F_1 score for individual category performance. The three PolSAR datasets are shown Fig. 3. The first image is the San Francisco Bay area with the size of 22 608 × 8080, which was acquired by ALOS-2 in 2015. The second image is Hangzhou area with the size of 6205 × 4904, which was acquired by RADASARSAT-2 on August 7, 2020. The third image is a ESAR L-band data of Oberpfaffenhofen area and the image size is 1300 × 1200. In land cover classification, various standards exist, such as the Coordination of Information on the Environment (CORINE) land cover classification system [69], which classifies land cover into 44 classes based on satellite data, the United States Geological Survey (USGS) land use/land cover classification system [70], which classifies land use and land cover into 24 categories, and the ESA (European Space Agency) Climate Change Initiative (CCI) Land Cover classification [71], which divides land cover into 22 categories. However, the San Francisco Bay Area (ALOS-2 dataset) (see [43], [72]) and Oberpfaffenhofen area (ESAR dataset) (see [73], [74]) classifications in the relevant literature do not strictly adhere to these standards. As such, we choose to follow the common practice of land cover classification in the literature to allow cross comparison. Consequently, for the San Francisco Bay Area dataset, terrain types include *water body, vegetation, high-density urban, and low-density urban*. A 1536 × 1536 region is selected from it as the study area and use ground truth data obtained from [54]. For the Oberpfaffenhofen dataset, types encompass *built-up areas, woodland, and open areas*. For the

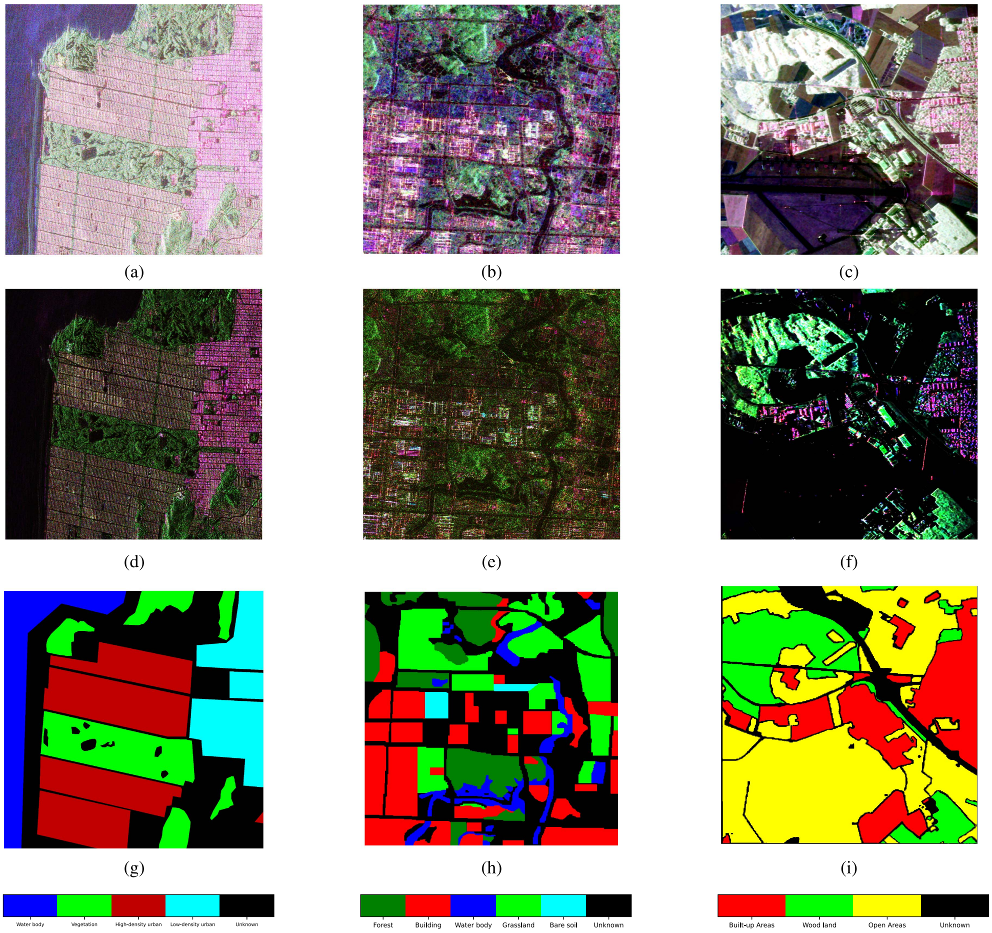


Fig. 3. PolSAR datasets used for the validation of the proposed method. (a)–(c) PauliRGB image of San Francisco bay area (ALOS-2), Hangzhou area (RADARSAT-2), and Oberpfaffenhofen areas (ESAR). (d)–(f) Intensity RGB (HH/HV/VV) image. (g)–(i) Corresponding ground truth.

Algorithm 1: Gradient Descent Training of Adversarial Unet Model.

```
for number of training epoch do
  for $k$ iteration do
    Sample minibatch of $m$ Training PolSAR sample and corresponding ground truth.
        $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)}\}, \{\mathbf{T}^{(1)}, \mathbf{T}^{(2)}, \ldots, \mathbf{T}^{(m)}\}$
    Update the forward classifier by descending its stochastic gradient:
        $\nabla_{\theta_G} \frac{1}{m} L_G$
  end for
  if training epoch==condition then
    for $k$ iteration do
      Sample minibatch of $m$ Prediction sample from the forward classifier and corresponding ground truth.
          $\{\mathbf{G}(\mathbf{X}^{(1)}), \mathbf{G}(\mathbf{X}^{(2)}), \ldots, \mathbf{G}(\mathbf{X}^{(m)})\}, \{\mathbf{T}^{(1)}, \mathbf{T}^{(2)}, \ldots, \mathbf{T}^{(m)}\}$
      Update the forward classifier by descending its stochastic gradient:
          $\nabla_{\theta_D} \frac{1}{m} L_D$
    end for
  end if
end for
```

TABLE I
TRAINING LABEL FOR DIFFERENT CATEGORIES OF DIFFERENT SCENES

Scenes	Category	Pixel number	Sample number
San Francisco	Water	265 225	2652
	Vegetation	361 834	3618
	High-density urban	640 595	6405
	Low-density urban	299 649	2996
Hangzhou	Forest	137 876	6893
	Building	242 264	12 113
	Water	64 925	3246
	Grass	185 939	9296
	Bare soil	14 677	733
Oberpfaffenhofen	Built-up Areas	536 471	5364
	Vegetation	373 561	3735
	Open Areas	1 093 338	10 933

TABLE II
RMS CONTRAST COMPARISON FOR THE ALOS-2, RADARSAT-2, AND ESAR DATASETS

Scenes	Pauli RGB	Intensity RGB
San Francisco bay	0.2270	0.1953
Hangzhou	0.2070	0.1771
Oberpfaffenhofen	0.3210	0.2412

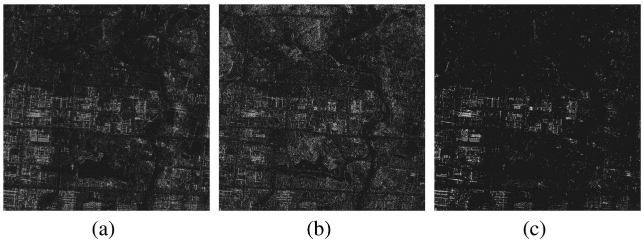


Fig. 4. Different polarization visulization for RADARSAT-2 dataset. (a) HH. RMS contrast: 0.1256. (b) HV. RMS contrast: 0.1483. (c) VV. RMS contrast: 0.1074.

Hangzhou dataset, we select a 1024 × 1024 region from it as the study area, which includes *forest, building, water body, grass land, bare soil*. The above are summarized in Table I.

In Fig. 3, for the three datasets, we show the PauliRGB images and the intensity RGB (HH/HV/VV) images to provide some visual comparison. Yet since visualization is a subjective experience, to provide a more rigorous measure, here we adopt a quantitative measure, namely, the root mean square (RMS) contrast [63] for the purpose, which is defined as $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$, where x_i is a normalized value and $\bar{x}$ is the mean normalized level $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$. The appeal of this definition is that the rms contrast is not dependent on spatial frequency content of the image or the spatial distribution of contrast in the image. The rms contrast values are listed in Table II, where the PauliRGB images uniformly possess higher values than their counterpart HH/HV/VV pseudo-RGB images. To better appreciate the visual effect of the Pauli RGB images, we also provide the original grey level PolSAR images of the RDARSAT-2 dataset, with the corresponding rms contrast value for each polarization (see Fig. 4). The comparison with Fig. 3 clearly demonstrates the improved contrast of the PauliRGB images both visually and quantitatively.

A. Parameter Setting

Each of the three dataset is cropped without overlapping. All the crop sizes are 512×512. Padding is added to ESAR data to fix the overspill issue. The training label is composed of randomly chosen 1% of each terrain type for ALOS-2 and ESAR, 5% for RADARSAT-2. The validation and test labels are composed of randomly chosen 20% and 70% of each terrain type. The labels of training, test, and validation set do not have any overlap. In particular, only the selected labels are used to evaluate the loss function during training. The sample selection method mentioned above is similar to [40], [55].

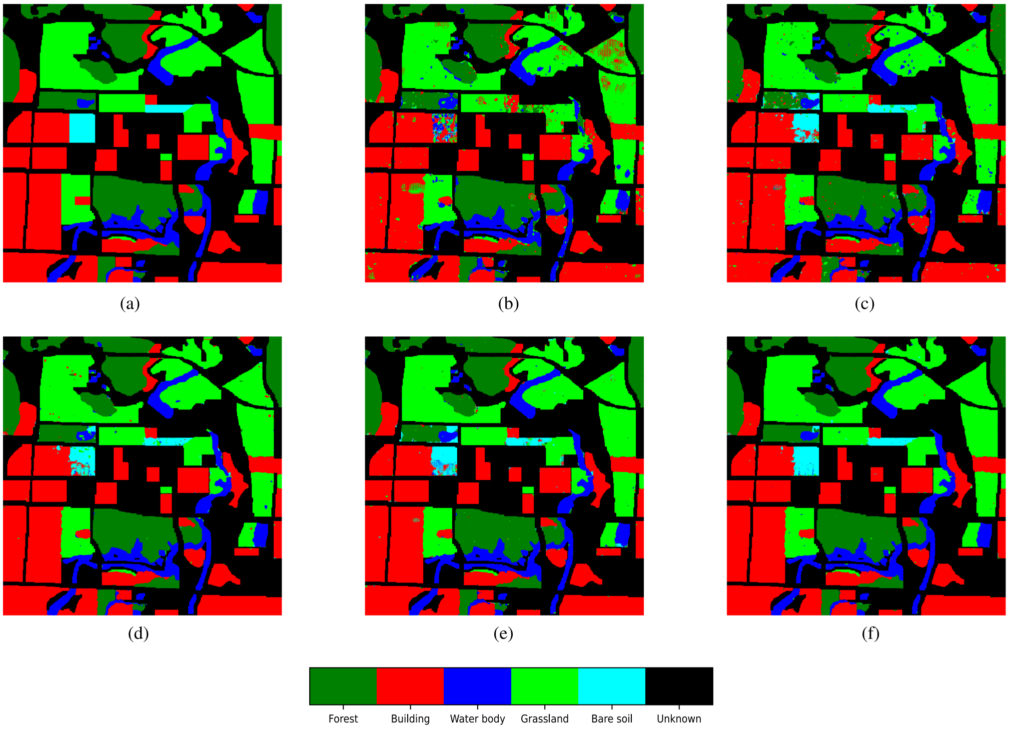
For the network hyperparameters, we set epoch as 500, 300, and 800, respectively, for ALOS-2, RADARSAT-2, and ESAR datasets. The batch size is set as 1 and the learning rate is 0.001 for the forward classifier and 0.0001 for the discriminator. The forward classifier and the discriminator use alternate training for parameter learning. The forward classifier updates the discriminator once after 1 epochs. The Adam algorithm is used for the optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For the CRF, $\theta_\alpha = 160$, $\theta_\beta = 3$, $\theta_\gamma = 3$ as suggested in [4], and the iterations are 5 for the three datasets. The He Initialization is adopted for the network weight initialization [62]. To better appreciate the roles of the third-order potentials in CRF and the adversarial learning, we present four sets of results in this article, the first two with CRF potentials progressing from pairwise to the third order yet both without adversarial learning, termed Unet-CRF2 and Unet-CRF3, respectively, whereas the last two sets incorporates correspondingly the adversarial learning and termed Unet-CRF2-Adv and Unet-CRF3-Adv, respectively.

Performance comparison is also made against the CNN model [27], the FCNs model [35], SegNet model [36], the PSP model [37], DeepLabV3 model [38]. For the CNN model, a patch size of 16×16 is chosen for the dataset. The FCN model uses the FCN-32 structure, that is, it directly restores the original image size by 32 times upsampling from bottleneck layer. The encoder of the SegNet adopts VGG16 model. Both encoders of PSP and DeepLabV3 model are ResNet model. In particular, the dilation rate of atrous convolution for DeepLabV3 in ESAR dataset is set to 6. we use two scales for the ALOS-2 and RADARSAT-2 datasets and their dilation rate are 6,12. To ensure fair comparison, the training set is also composed of randomly chosen 1% of each type for ALOS-2, ESAR, and 5% for RADARSAT-2. The impact of training set percentage on performance is exemplified in Fig. 9 for the three datasets. For the abovementioned models, the learning rate is 0.001, the He Initialization scheme is used for network weight initialization, and the Adam algorithm is used for optimization, where the parameters β_1 and β_2 are 0.9 and 0.999, respectively.

We adopt early-stopping to prevent overfitting problem and save the best model under the validation set in the training process. In addition, five independent experiments with random initialization are conducted for each model in order to reduce performance fluctuations. A coarse calculation of standard deviation using these five results will also shed some light on the degree of fluctuations from experiment to experiment of these models.

TABLE III
QUANTITATIVE PERFORMANCE RESULTS (F_1 SCORE(%), OA(%), KAPPA(%)) FOR CLASSIFICATION OF HANGZHOU DATASET

	Forest	Building	Water body	Grass	Bare soil	OA	Kappa	Parameter
CNN	85.05±0.43	90.64±0.19	78.21±0.39	79.74±0.34	69.48±0.78	84.62±0.20	78.62±0.27	0.18M
FCN-32	89.90±1.26	93.46±0.60	82.25±0.78	89.97±1.50	60.84±5.42	89.64±0.82	85.57±1.14	1.15M
SegNet	95.99±1.78	98.11±0.46	90.77±2.13	96.81±0.84	78.51±9.78	96.11±1.14	94.59±1.58	5.46M
PSPNet	96.10±0.30	97.59±0.07	89.47±0.43	96.92±0.17	98.27±0.28	96.29±0.14	94.85±0.20	2.49M
DeepLabV3	93.97±0.61	97.78±0.47	93.18±0.99	98.08±0.26	92.99±8.15	96.47±0.25	95.11±0.35	12.05M
Unet(baseline)	92.99±0.71	95.54±0.63	80.50±3.13	92.17±2.18	16.23±11.43	91.44±0.78	88.01±1.11	2.21M
Unet-CRF2	95.84±0.33	97.56±0.18	91.98±1.32	96.61±0.26	70.19±6.68	95.76±0.25	94.11±0.35	2.21M
Unet-CRF3	97.22±0.18	98.31±0.16	94.53±0.45	97.91±0.16	86.08±2.05	97.30±0.12	96.25±0.17	2.21M
Unet-CRF2-Adv	97.08±0.50	98.33±0.21	94.19±0.65	97.97±0.15	80.18±1.89	97.11±0.23	95.99±0.32	2.88M
Unet-CRF3-Adv	97.96±0.18	98.91±0.08	95.83±0.19	98.63±0.05	90.89±1.02	98.13±0.08	97.40±0.11	2.88M


(a) (b) (c)

(d) (e) (f)

Fig. 5. Classification results of Hangzhou dataset (RADARSAT-2). (a) Ground truth. (b) Unet. (c) Unet-CRF2. (d) Unet-CRF3. (e) Unet-CRF2-Adv. (f) Unet-CRF3-Adv.

The experimental platform is a custom assembled workstation with GeForce TITAN RTX GPU and 24 GB in-chip memory, Intel Core(TM) i9-9900 K CPU@3.60 GHz, and 64 GB system memory. The operating system is Linux Ubuntu 16.04, and the deep learning framework is tensorflow2.0.

B. Classification Results and Discussions

1) Hangzhou Dataset (RADARSAT-2): The pixel-level classification results for the Hangzhou dataset is shown in Fig. 5 for visual inspection. It is seen that Unet captures the labeling pattern in large, yet suffers from visible misclassification errors across all the five landcover categories, in particular the bare soil category [see Fig. 5(b)]. The incorporation with pairwise potential CRF helps boost performance, for instance, the bare soil patches have been recovered markedly [see Fig. 5(c)]; yet misclassified pixels are scattered almost across every patch of the five category. Further improvement can be observed with the inclusion of CRF3, where the building and forest categories are almost free of misclassification [see Fig. 5(d)]. Both CRF2 and CRF3 can be made more discriminatively powerful when adversarial learning is included, as shown in Fig. 5(e) and (f), respectively.

The classification results are quantitatively provided in Table III, where F_1 score is reported for each category, followed by OA and Kappa for the overall performance. Since five

TABLE IV
QUANTITATIVE PERFORMANCE RESULTS (F_1 SCORE(%), OA(%), KAPPA(%)) FOR CLASSIFICATION OF OBERPFAFFENHOFEN DATASET

	Built-up Areas	Vegetation	Open Areas	OA	Kappa	Parameter
CNN	72.17±5.19	94.66±0.26	83.04±3.18	87.07±1.68	78.00±2.89	0.18M
FCN-32	91.18±0.52	90.54±1.21	95.42±0.54	93.34±0.62	88.91±1.01	1.15M
SegNet	82.92±1.02	85.47±2.08	95.29±0.70	90.15±0.86	83.43±1.35	5.46M
PSPNet	93.50±1.10	91.07±1.88	96.34±0.50	94.63±0.88	90.93±1.52	2.49M
DeepLabV3	89.96±0.07	87.49±1.25	95.70±0.22	92.64±0.50	87.64±0.83	4.18M
Unet(baseline)	93.86±0.20	95.32±0.41	97.65±0.28	96.20±0.26	93.63±0.42	2.21M
Unet-CRF2	96.37±0.40	97.29±0.50	98.63±0.09	97.78±0.22	96.26±0.38	2.21M
Unet-CRF3	97.47±0.27	98.21±0.36	98.93±0.11	98.39±0.14	97.30±0.24	2.21M
Unet-CRF2-Adv	96.88±0.28	97.69±0.42	98.76±0.05	98.06±0.16	96.74±0.28	2.88M
Unet-CRF3-Adv	97.85±0.19	98.55±0.19	99.08±0.07	98.65±0.09	97.74±0.17	2.88M

TABLE V
QUANTITATIVE PERFORMANCE RESULTS (F_1 SCORE(%), OA(%), KAPPA(%)) FOR CLASSIFICATION OF SAN FRANCISCO DATASET

	Water body	Vegetation	High-density urban	Low-density urban	OA	Kappa	Parameter
CNN	98.40±0.59	92.07±0.27	91.18±0.25	85.08±0.93	91.54±0.39	88.04±0.55	0.18M
FCN-32	94.39±0.70	90.71±0.77	94.64±0.37	89.49±0.56	92.71±0.40	89.80±0.55	1.15M
SegNet	94.96±1.88	90.21±3.47	88.06±3.64	82.18±9.77	88.76±3.29	84.29±4.72	5.46M
PSPNet	72.53±1.71	77.35±0.92	83.80±1.11	76.31±1.20	78.80±1.14	70.67±1.54	2.49M
DeepLabV3	94.07±1.34	94.51±1.14	93.48±1.30	83.78±3.52	92.05±1.30	88.82±1.85	4.18M
Unet(baseline)	96.93±1.18	92.06±0.81	95.30±0.91	93.29±1.89	94.44±0.89	92.23±1.24	2.21M
Unet-CRF2	98.91±0.31	95.61±0.47	97.59±0.45	96.02±0.79	97.06±0.36	95.88±0.50	2.21M
Unet-CRF3	99.68±0.17	97.84±0.22	98.89±0.14	98.44±0.41	98.69±0.17	98.17±0.24	2.21M
Unet-CRF2-Adv	99.17±0.17	96.40±0.57	98.05±0.39	96.72±0.80	97.69±0.38	96.65±0.52	2.88M
Unet-CRF3-Adv	99.86±0.05	98.67±0.14	99.34±0.11	99.08±0.16	99.22±0.08	98.91±0.11	2.88M

TABLE VI
TIME EFFICIENCY COMPARISON FOR THE ALOS-2, RADARSAT-2, AND ESAR DATASETS

	ALOS-2		RADARSAT-2		ESAR	
	Train Time(s)	Test Time(s)	Train Time(s)	Test Time(s)	Train Time(s)	Test Time(s)
CNN	2260	852	2351	377	1983	549
FCN-32	14826	17	4210	6	10396	17
SegNet	21682	14	5581	6	5235	18
PSPNet	12044	13	3926	6	10686	19
DeepLabV3	12649	13	1299	6	12369	17
Unet	9430	14	3402	7	8875	17
Unet-CRF2	11405	15	4106	8	7021	30
Unet-CRF3	15569	23	4374	8	27805	27
Unet-CRF2-Adv	11725	24	4399	8	7543	22
Unet-CRF3-Adv	17342	30	4708	8	28534	22

experiments have been independently carried out for each method, the reported value is the mean value with one standard deviation. The progressive performance gains starting from Unet, then to Unet-CRF2, and to Unet-CRF3 and Unet-CRF2-Adv, finally culminating at Unet-CRF3-Adv, are crystal-clear from the table. Results for comparison methods (CNN, FCN-32, SegNet, PSPNet, and DeepLabV3) are also provided. For these methods, except DeepLabV3, each shows at least one weak spot where one category has F_1 score below 90. Of all the models, the proposed Unet-CRF3 and Unet-CRF3-Adv hold the best overall performance (OA and Kappa) and almost every single category (F_1) score except the bail soil category. It is of some interest to watch that for this category, the Unet-line starting from Unet to the four enhanced variants with CRF and/or Adv, the performance is the weakest among the five categories. PSPNet and DeepLabV3, on the other hand, show stronghand in this category, indicating that the ideas behind these two methods are more suitable to bare soil classification, at least for this dataset.

It should be mentioned that the proposed Unet-CRF3 and Unet-CRF3-Adv achieve super performance gains without introducing too much network burden. In fact, their number of parameters is in par to PSPNet, and significantly smaller than SegNet and DeepLabV3.

2) Oberpfaffenhofen Dataset (ESAR): The pixel-level classification results for the Oberpfaffenhofen dataset is shown in Fig. 6. The relative performances among Unet, Unet-CRF2, Unet-CRF3, Unet-CRF2-Adv, and Unet-CRF3-Adv, are closely mimic that of the Hangzhou dataset.

The classification results are quantitatively provided in Table IV. Again, the progressive performance gains starting from Unet, then to Unet-CRF2, and to Unet-CRF3 and Unet-CRF2-Adv, finally culminating at Unet-CRF3-Adv, are manifested from the table.

Fig. 6 also shows that the five Unet related models have more misclassification in the build-up category than the other two terrain types. The built-up area may contain vegetation, bare soil, and so on, which undergoes scattering mechanisms not

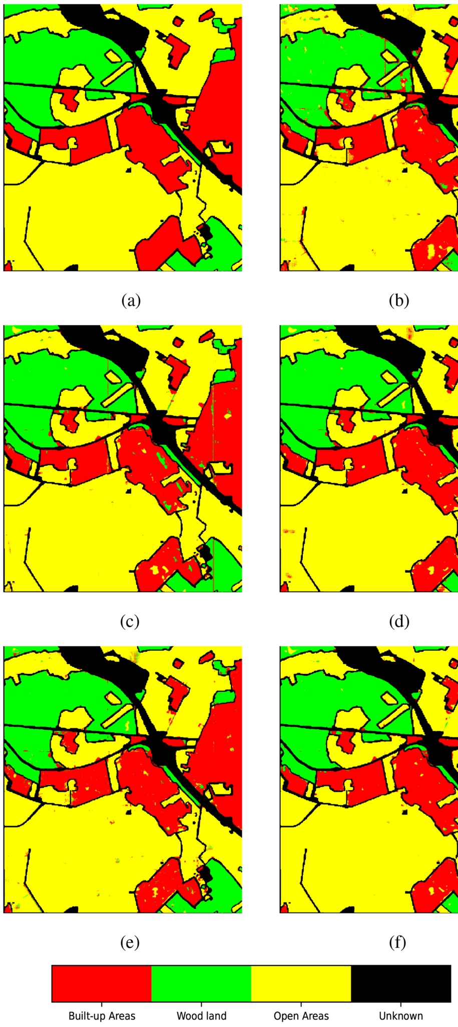


Fig. 6. Classification results of Oberpfaffenhofen dataset (ESAR). (a) Ground truth. (b) Unet. (c) Unet-CRF2. (d) Unet-CRF3. (e) Unet-CRF2-Adv. (f) Unet-CRF3-Adv.

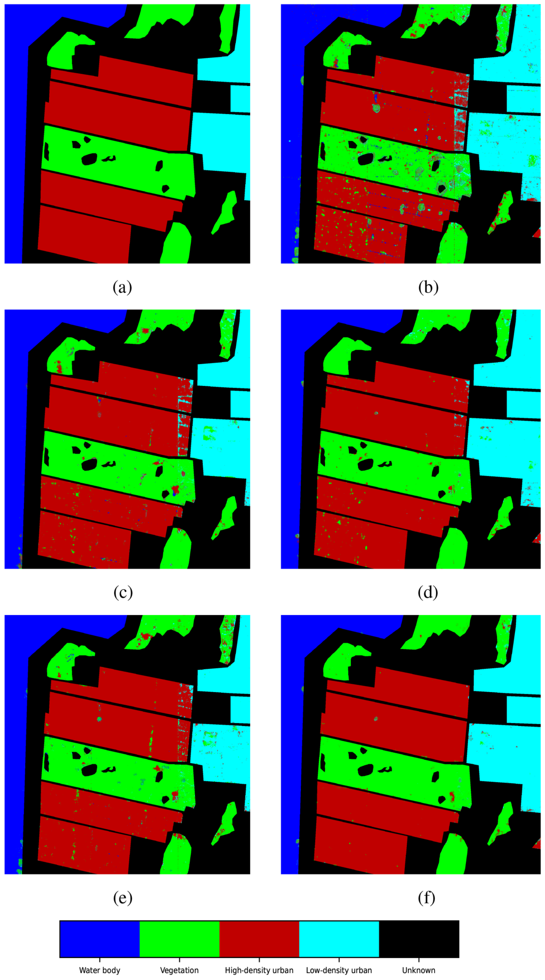


Fig. 7. Classification results of San Francisco Bay area dataset (ALOS-2). (a) Ground truth. (b) Unet. (c) Unet-CRF2. (d) Unet-CRF3. (e) Unet-CRF2-Adv. (f) Unet-CRF3-Adv.

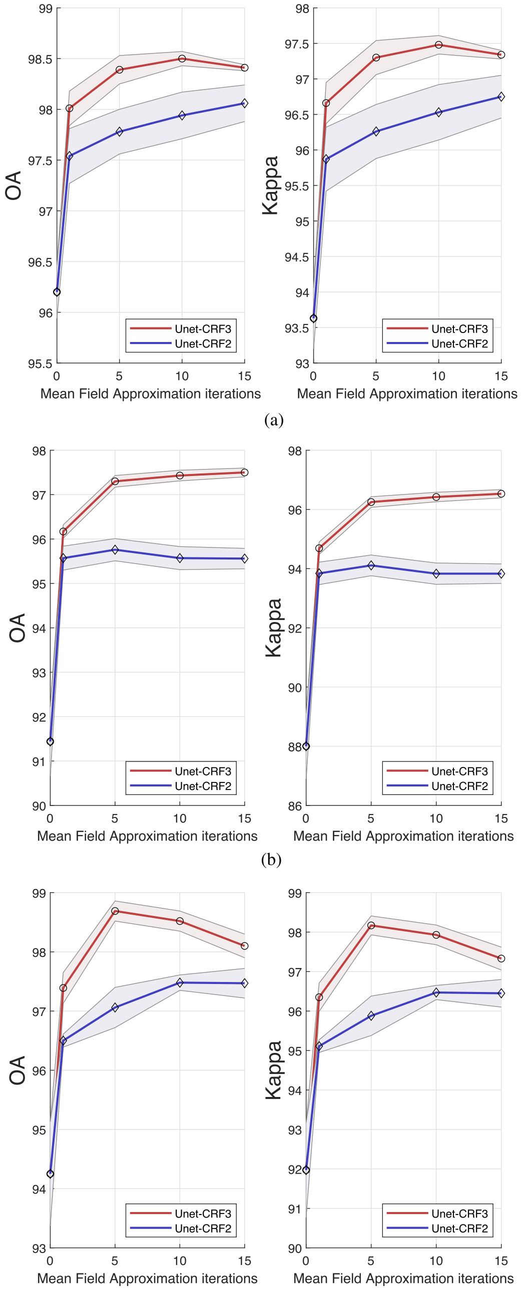
readily distinguishable from that of wood land and open area, hence requiring the deep learning network to have adequate discriminative power. Unet-CRF3 and Unet-CRF3-Adv achieved a F1-score of 97.47% and 97.85%, respectively, in this category, indicating that more accurate modeling of label distribution can help improve the feature expression ability of the network.

Overall, of the 10 models including comparison models (CNN, FCN-32, SegNet, PSPNet, and DeepLabV3), the proposed Unet-CRF3 and Unet-CRF3-Adv hold the best overall performance (OA and Kappa) and every single category (F_1) score, typically with large margin. For instance, Unet-CRF3-Adv has OA score of 98.65%, better than the baseline Unet score of 96.20%, and much better than the highest score of 94.63 among the other five comparison models.

Again, albeit the performance gains of the proposed Unet-CRF3 and Unet-CRF3-Adv models, their numbers of parameters are comparable to other models.

3) San Francisco Bay Area Dataset (ALOS-2): The pixel-level classification results for the San Francisco Bay area dataset is shown in Fig. 7. The relative performances among Unet, Unet-CRF2, Unet-CRF3, Unet-CRF2-Adv, and Unet-CRF3-Adv, are closely mimic that of the other two datasets. However, from Fig. 7(c), it is seen that CRF2 is incapable of handling the severe spillover of the low-density urban category into neighboring high-density urban and vegetation categories. This indicates its difficulty of differentiating the seemly common yet intricate scattering mechanisms for the cases, such as volume scattering and depolarization effect. CRF3, on the other hand, shows remarkable discriminative power. The same comments apply to their adversarial learning counterparts. It also indicates that although higher order CRFs and adversarial learning exert their respective effects in different direction and hard to cross compare them, in this dataset higher order CRFs clearly are more important than the adversarial learning.

The classification results are quantitatively provided in Table V. Again, the progressive performance gains starting from Unet, then to Unet-CRF2, and to Unet-CRF3 and Unet-CRF2-Adv, finally culminating at Unet-CRF3-Adv, are manifested from the table.



(a)

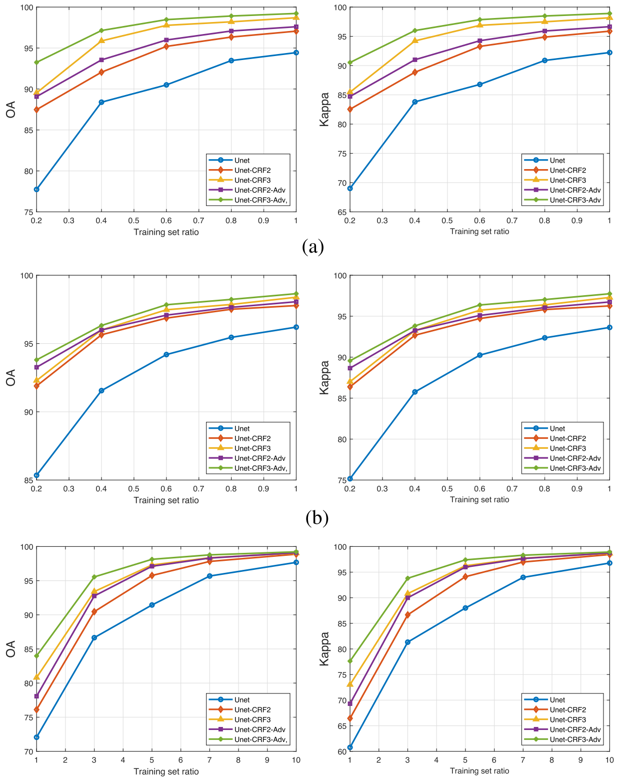
(b)

(c)

Fig. 8. Performances of CRF2 and CRF3 under different iteration number for the three datasets (ESAR, RADARSAT-2, and ALOS-2). (a) Oberpfaffenhofen. (b) Hangzhou. (c) San Francisco.

Of the 10 models including comparison models (CNN, FCN-32, SegNet, PSPNet, and DeepLabV3), the proposed Unet-CRF3 and Unet-CRF3-Adv hold the best overall performance (OA and Kappa) and every single category (F_1) score, with distinguishably large margin. For instance, Unet-CRF3-Adv has OA score of 99.22%, much better than the baseline Unet score of 94.44% and the highest score of 92.05% among the other five comparison models.

Once again, the numbers of parameters of the proposed Unet-CRF3 and Unet-CRF3-Adv models are comparable to other models.

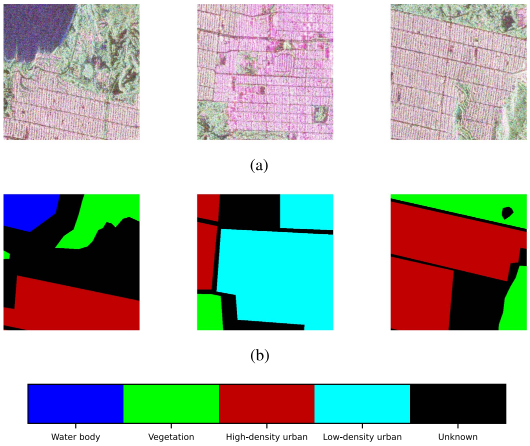


(a)

(b)

(c)

Fig. 9. Model performances under different percentage of training label for the three datasets. (a) ALOS-2 San Francisco. (b) ESAR Oberpfaffenhofen. (c) RADARSAT-2 Hangzhou.

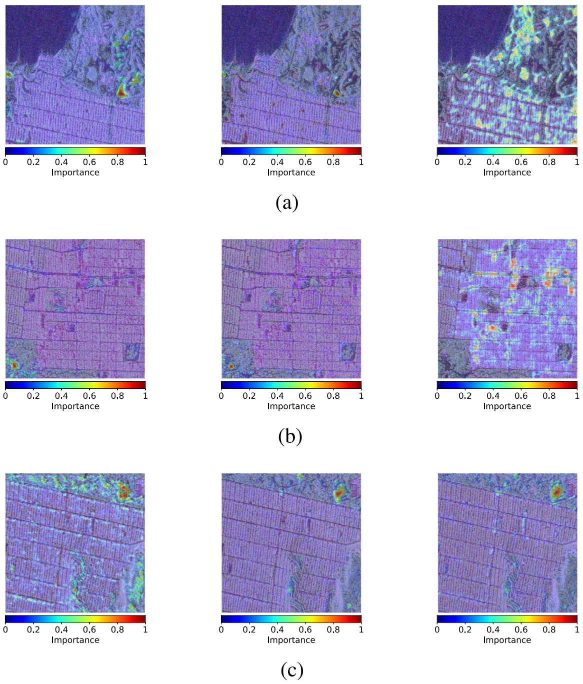


(a)

(b)

Fig. 10. ALOS-2 study area and corresponding ground truth for Seg-Grad-Cam. (a) Original sub image. (b) Ground truth.

4) Time Efficiency: The training time and test time in seconds for the three datasets and for the 10 models are listed in Table VI for comparison. Since the test time is several orders smaller than the training time, we shall focus on the latter. For the Unet related 5 models, it is seen that inclusion of adversarial learning only leads to marginal increase of training time, whereas inclusion of higher order potentials in CRFs leads to longer training time. Specifically, inclusion of CRF3 costs a portion of the



(a)

(b)

(c)

Fig. 11. Seg-Grad-Cam on ALOS-2 San Francisco dataset. From the first to the third column is Unet, Unet-CRF2, Unet-CRF3 model. (a) Activation Map on high-density urban class. (b) Activation Map on low-density urban class. (c) Activation Map on vegetation class.

CRF2 training time for ALOS2 and RADARSAT-2 datasets, yet costs about four times that of CRF2 for ESAR dataset, indicating that for this dataset the difficulty of convergency for the iterative optimization process of the third-order potential. A check of the training curve of Unet-CRF3 model on the ESAR dataset indicates it is more oscillatory than that on the ALOS-2 dataset. This behavior may be attributable to the PolSAR image resolution, as the ESAR dataset has higher resolution, and thus, presents more detailed land cover features; yet the variety of land features is highly complex, hence, CRF3 tends to pick up and activate more pertinent yet scattered spots. Consequently, the model experiences more oscillations during training and requires additional time to achieve optimal accuracy. Due to the intricate interactions among Unet, CRF, and AE, further analysis is challenging. When compared with other five comparison models, the training times of the proposed Unet-CRF3 and Unet-CRF3-Adv are comparable to a majority of the five models for ALOS2 and RADARSAT-2 datasets, and are about twice the training time for the ESAR dataset.

C. Further Analysis

1) Effect of Iteration Number of CRF Updating: The fully connected CRF algorithm is based on a mean field approximation to the CRF distribution, an approximation that is iteratively optimized through a series of message passing steps, each of which updates a single marginal $Q_i(y_i)$. In Fig. 8, the performances of CRF2 and CRF3 under different iteration numbers are examined. To avoid the complication which might be brought forth by adversarial learning, we focus on Unet-CRF2 and Unet-CRF3. In each of the figures, there are three curves, with the middle one representing the mean value of the five independent experiments, and the other two denoting one standard deviation. To better illustrate the effect of iteration number, we preset it to one of the following values: 1, 5, 10, 15. To facilitate the comparison with the baseline Unet as well, we represent the latter to be the case when iteration number $= 0$, since there is no CRF present then.

The following observations are readily made from Fig. 8.

1) The benefit of CRFs is evident when comparing baseline Unet with Unet-CRF2 or Unet-CRF3.
2) CRF3 uniformly outperforms CRF2 across iteration number and datasets, indicating the need to consider higher order potentials in CRFs for better label consistency.
3) CRF3 also tends to have smaller standard deviation, indicating less performance fluctuation from experiment to experiment.
4) The overall performance measure, OA or Kappa, shows similar trend across iteration number, datasets, and CRF order.
5) The performance is not monotonically improving with increasing iteration number, suggesting that too many iteration may lead to overkill of the inhomogeneity and lead to performance degradation instead.

2) Effect of Percentage of Training Data: In the parameter setting, the training label is composed of randomly chosen 1% of each terrain type for ALOS-2 and ESAR, and 5% for RADARSAT-2. Given the satisfactory performance of the proposed methods Unet-CRF3 and Unet-CRF3-Adv, we would like to further investigate how their performance will change if the percentage of training label goes down under 1% for ALOS-2 and ESAR datasets and 5% for RADARSAT-2.

For the ALOS-2 dataset, Fig. 9(a) shows the performances of the five models (Unet and its enhanced four variants with CRF and/or adversarial learning) against varying percentage of training label, which can go as low as 0.2%. The performance improves with more training label available. Relative performance-wise, the progressive OA and Kappa performance gains start from Unet, then to Unet-CRF2, and to Unet-CRF3 and Unet-CRF2-Adv, finally culminating at Unet-CRF3-Adv, in agreement with what have been observed in the previous section. It is worthy of mentioning that even with training label as low as 0.2%, the proposed Unet-CRF3-Adv method can still achieve OA better than 93% and Kappa better than 90%. Similar behaviors can be observed for the ESAR dataset as shown in Fig. 9(b) and for the RADARSAT-2 dataset in Fig. 9(c).

3) Effectiveness of Feature Selection: It is of interest to see how features are selected/extracted for final pixel-level classification. This can be visualized by the utility of heatmap to assess how relevant individual pixels or regions to the classification decision are. We adopt the Seg-gradient-weighted class activation mapping (Grad-CAM) method [64], which is an extension of the popular Grad-CAM) method [66], in order to address the issue of semantic interpretation of image segmentation. Following [64], we generate heatmap using Seg-Grad-CAM for the bottleneck layer (end of the encoder before upsampling) of the Unet due to

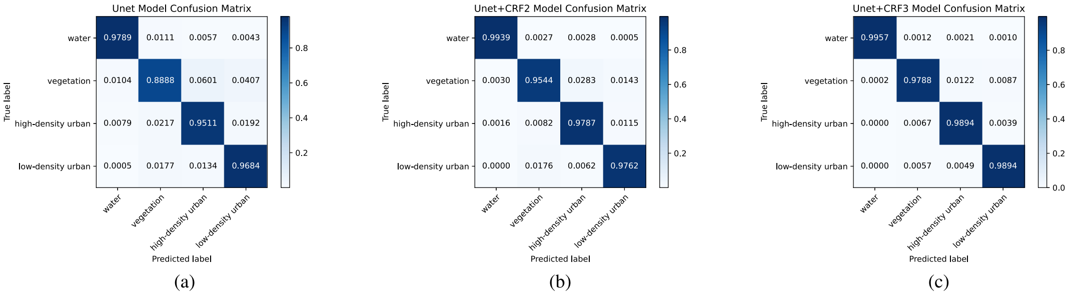


Fig. 12. Unet, Unet-CRF2, and Unet-CRF3 model confusion matrix for ALOS-2 dataset. (a) Unet. (b) Unet-CRF2. (c) Unet-CRF3.

its better informativeness than the layers close to the end of the Unet decoder.

For illustration, we use the dataset of San Francisco Bay area, with the study areas and corresponding ground truth depicted in Fig. 10.

The results are shown in Fig. 11. For the activation map of high-density urban class, it is seen that Unet is the most economical and activates the least, whereas Unet-CRF3 activates a majority of the high-density urban category, and Unet-CRF2 activates some spots in between the two models. Some curious aspect here is that Unet-CRF3 not only activates the high-density urban area marked by the ground truth in Fig. 10, but also areas interspersed within the vegetation/forest region, areas too complex and scattered that have been given up by the ground truth map (the unknown category). It, thus, allows one to appreciate the power of the CRF3 in its capability of picking up and activating the pertinent yet scattered spots. Due to the complex interplay among Unet, CRF, and the AE, meaningful further analysis is elusive, so we turn to check the confusion matrix instead, which clearly indicates the benefit of adding CRF on top of baseline Unet, and furthermore the discriminative power of higher order potentials (see Fig. 12).

V. Conclusion

For pixel level PolSAR image classification, this article proposed a deep learning method to address two important issues of how to enforce consistency and coherency of labeling using contextual information, and how to make the process computationally efficient and flexible.

When applied to the classification of ALOS-2, RADARSAT-2, and ESAR PolSAR images, the proposed method has uniformly demonstrated superior performance to other state-of-the-art models. The results unambiguously indicate the importance of higher order potentials of CRFs and the benefits of adversarial learning.

To extend this work, we envision the following research lines:

1) Working with other types of polarimetric decomposition: regarding the characteristics of PolSAR images, in view of the richer data channel some other researchers have been utilizing (entropy-alpha by Cloude's polarimetric decomposition), our use of the Pauli decomposition may serve as the base for comparison when other types of polarimetric decomposition are adopted. Moreover, to incorporate more physics into the decomposition, we are working on physically rigorously polarimetric decomposition [67] and would like to extend our currently data driven approach in future work.
2) Incorporating data augment technique: This work does not use any data augment technique, whose adoption is expected to further improve the performance.
3) Applying active learning: The learning in this work is of supervised manner, where around 1% labels of each subcategory has to be manually prepared. This is a tedious and error-prone work, an issue which can be mitigated by applying active learning [68].
4) Addressing PolSAR image variability: PolSAR images can be affected by a variety of factors, including speckle noise, complex undulating topography effect, atmospheric conditions, and sensor calibration errors.

A more comprehensive exploration of variability, similar to the approach used in [77] could further enhance the model's robustness and accuracy.

Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful reviews and suggestions.

References

[1] Z. Du et al., "Mapping annual global forest gain from 1983 to 2021 with landsat imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4195–4204, 2023, doi: 10.1109/JSTARS.2023.3267796.

[2] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.

[4] P. Krahenbuhl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2011, pp. 109–117.

[5] P. Kohli, L. Ladicky, and P. H. Torr, "Robust higher order potentials for enforcing label consistency," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 302–324, 2009.

[6] Y. Wang et al., "Self-supervised feature learning with CRF embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2628–2642, May 2019.

[7] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[8] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[9] Y. Wang, C. He, X. Liu, and M. Liao, "A hierarchical fully convolutional network integrated with sparse and low-rank subspace representations for PolSAR imagery classification," *Remote Sens.*, vol. 10, 2018, Art. no. 342.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[14] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[15] W. Wu, H. Li, X. Li, H. Guo, and L. Zhang, "PolSAR image semantic segmentation based on deep transfer learning realizing smooth classification with small training sets," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 977–981, Jun. 2019.

[16] H. Bi, L. Xu, X. Cao, Y. Xue, and Z. Xu, "Polarimetric SAR image semantic segmentation with 3D discrete wavelet transform and Markov random field," *IEEE Trans. Image Process.*, vol. 29, pp. 6601–6614, Jun. 2020.

[17] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[18] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[19] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.

[21] T. S. Lee, "Image representation using 2D Gabor wavelets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 10, pp. 959–971, Oct. 1996.

[22] K. Huang and S. Aviyente, "Wavelet feature selection for image classification," *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1709–1720, Sep. 2008.

[23] Z. Shao, L. Zhang, and L. Wang, "Stacked sparse autoencoder modeling using the synergy of airborne LiDAR and satellite optical and SAR data to map forest above-ground biomass," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 12, pp. 5569–5582, Dec. 2017.

[24] W. Xie et al., "POLSAR image classification via Wishart-AE model or Wishart-CAE model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3604–3615, Aug. 2017.

[25] Y. Chen, X. Zhao, and X. Jia, "Spectral–Spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.

[26] W. Zhao, S. Du, and W. J. Emery, "Object-based convolutional neural network for high-resolution imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3386–3396, Jul. 2017.

[27] 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.

[28] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5517010, doi: 10.1109/TGRS.2021.3124913.

[29] G. Li, L. Li, H. Zhu, X. Liu, and L. Jiao, "Adaptive multiscale deep fusion residual network for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8506–8521, Nov. 2019.

[30] G. Wang, B. Fan, S. Xiang, and C. Pan, "Aggregating rich hierarchical features for scene classification in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4104–4115, Sep. 2017.

[31] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[32] J. Cheng, F. Zhang, D. Xiang, Q. Yin, and Y. Zhou, "PolSAR image classification with multiscale superpixel-based graph convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5209314, doi: 10.1109/TGRS.2021.3079438.

[33] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[34] J. Geng, X. Deng, X. Ma, and W. Jiang, "Transfer learning for SAR image classification via deep joint distribution adaptation networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5377–5392, Aug. 2020.

[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[36] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[38] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation, 2017, *arXiv:1706.05587*. [Online]. Available: https://doi.org/10.48550/arXiv.1706.05587

[39] Y. Li, Y. Chen, G. Liu, and L. Jiao, "A novel deep fully convolutional network for PolSAR image classification," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 1984.

[40] Y. Cao, Y. Wu, P. Zhang, W. Liang, and M. Li, "Pixel-wise PolSAR image classification via a novel complex-valued deep fully convolutional network," *Remote Sens.*, vol. 11, no. 22, 2019, Art. no. 2653.

[41] R. Zhang, J. Chen, L. Feng, S. Li, W. Yang, and D. Guo, "A refined pyramid scene parsing network for polarimetric SAR image semantic segmentation in agricultural areas," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jun. 2022, Art. no. 4014805, doi: 10.1109/LGRS.2021.3086117.

[42] L. Yu et al., "A lightweight complex-valued DeepLabV3+ for semantic segmentation of PolSAR image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 930–943, Jan. 2022, doi: 10.1109/JSTARS.2021.3140101.

[43] H. Bi, J. Sun, and Z. Xu, "A graph-based semisupervised deep learning model for PolSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2116–2132, Apr. 2019.

[44] P. Zhang et al., "PolSAR image classification using hybrid conditional random fields model based on complex-valued 3-D CNN," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 3, pp. 1713–1730, Jun. 2021.

[45] W. Song, Y. Wu, and X. Xiao, "Nonstationary PolSAR image classification by deep-features-based high-order triple discriminative random field," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 8, pp. 1406–1410, Aug. 2021.

[46] Z. Zhang, J. Yang, and Y. Du, "Deep convolutional generative adversarial network with autoencoder for semisupervised SAR image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Sep. 2022, Art. no. 4000405, doi: 10.1109/LGRS.2020.3018186.

[47] Z. L. Ren, B. Hou, Q. Wu, Z. D. Wen, and L. C. Jiao, "A distribution and structure match generative adversarial network for SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3864–3880, Jun. 2020.

[48] F. Liu, L. Jiao, and X. Tang, "Task-oriented GAN for PolSAR image classification and clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2707–2719, Sep. 2019.

[49] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. Int. Conf. Computer Vis.*, 2017, pp. 2223–2232.

[50] K. Heidler, L. C. Mou, C. Baumhoer, A. Dietz, and X. X. Zhu, "HED-UNet: Combined segmentation and edge detection for monitoring the Antarctic coastline," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 4300514, doi: 10.1109/TGRS.2021.3064606.

[51] A. S. Nagi, D. Kumar, D. Sola, and K. A. Scott, "RUF: Effective sea ice floe segmentation using end-to-end RES-UNET-CRF with dual loss," *Remote Sens.*, vol. 13, 2021, Art. no. 2460.

[52] P. Zhang, M. Li, Y. Wu, and H. J. Li, "Hierarchical conditional random fields model for semisupervised SAR image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 9, pp. 4933–4951, Sep. 2015.

[53] Z. S. Sun et al., "SAR image classification using fully connected conditional random fields combined with deep learning and superpixel boundary constraint," *Remote Sens.*, vol. 13, no. 2, 2021, Art. no. 271.

[54] X. Liu, L. Jiao, and F. Liu, "PolSF: PolSAR image dataset on San Francisco," 2019, *arXiv:1912.07259*.

[55] F. Zhao, M. Tian, W. Xie, and H. Liu, "A new parallel dual-channel fully convolutional network via semi-supervised FCM for PolSAR image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4493–4505, Aug. 2020, doi: 10.1109/JSTARS.2020.3014966.

[56] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[57] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[58] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr, "Associative hierarchical CRFs for object class image segmentation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 739–746.

[59] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. Torr, "What, where and how many? Combining object detectors and CRFs," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 424–437.

[60] V. Vineet, J. Warrell, and P. H. Torr, "Filter-based mean-field inference for random fields with higher-order terms and product label-spaces," *Int. J. Comput. Vis.*, vol. 110, no. 3, pp. 290–307, 2014.

[61] A. Arnab, S. Jayasumana, S. Zheng, and P. H. Torr, "Higher order conditional random fields in deep neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 524–540.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[63] E. Peli, "Contrast in complex images," *J. Opt. Soc. Amer.*, vol. 7, no. 10, pp. 2032–2040, 1990.

[64] K. Vinogradova, A. Dibrov, and G. Myers, "Towards interpretable semantic segmentation via gradient-weighted class activation mapping," in *Proc. AAAI Conf. Artif. Intell.*, vol. 4, no. 10, Apr. 2020, pp. 13943–13944.

[65] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[66] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[67] Y. Du, C. Yang, Z. Y. Li, and Q. H. Liu, "Physically based polarimetric volumetric scattering from cylindrically dominated vegetation canopies," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1629–1636, Mar. 2019.

[68] H. X. Bi, F. Xu, Z. Q. Wei, Y. Xue, and Z. B. Xu, "An active deep learning approach for minimally supervised PolSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9378–9395, Nov. 2019.

[69] M. Bossard, J. Feranec, and J. Otahel, "CORINE land cover technical guide: Addendum", European Environment Agency, Copenhagen, Denmark, Tech. Rep. 40, 2000.

[70] J. R. Anderson, *A Land Use and Land Cover Classification System for Use With Remote Sensor Data*, vol. 964. Washington, DC, USA: US Government Printing Office, 1976.

[71] D. Zanaga et al., "ESA Worldcover 10 M 2021 v200," Zenodo, Oct. 28, 2022, doi: 10.5281/zenodo.7254221.

[72] Y. Cao, Y. Wu, M. Li, W. Liang, and X. Hu, "DFAF-Net: A dual-frequency PolSAR image classification network based on frequency-aware attention and adaptive feature fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5224318, doi: 10.1109/TGRS.2022.3152854.

[73] X. Tan, M. Li, P. Zhang, Y. Wu, and W. Song, "Deep triplet complex-valued network for PolSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10179–10196, Dec. 2021.

[74] Z. Wen, Q. Wu, Z. Liu, and Q. Pan, "Polar-spatial feature fusion learning with variational generative-discriminative network for PolSAR Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8914–8927, Nov. 2019.

[75] Y. Cui et al., "Polarimetric multipath convolutional neural network for PolSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5207118, doi: 10.1109/TGRS.2021.3071559.

[76] A. Jamali, M. Mahdianpari, F. Mohammadimanesh, A. Bhattacharya, and S. Homayouni, "PolSAR image classification based on deep convolutional neural networks using wavelet transformation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jun. 2022, Art. no. 4510105, doi: 10.1109/LGRS.2022.3185118.

[77] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

Zheng Zhang received the B.S. degree in electronics and information engineering and the M.S. degree in electronics and communication engineering from Zhengzhou University, Zhengzhou, China, in 2014 and 2017, respectively. He is currently working toward the Ph.D. degree in the field of microwave remote sensing with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China.

His research interests include synthetic aperture radar image interpretation, machine learning, and electromagnetic scattering.

Hui Guo received the B.S. degree in communication engineering and the M.S. degree in signal and information processing from the Harbin Institute of Technology, Harbin, China, in 1995 and 1997, respectively.

She is currently a Professor with the Beijing Institute of Satellite Information Engineering, China Academy of Space Technology, Beijing, China. Her research interests include SAR image processing, space communication, and networks.

Jingsong Yang received the B.S. degree in physics and the M.S. degree in theoretical physics from Zhejiang University, Hangzhou, China, in 1990 and 1996, respectively, and the Ph.D. degree in physical oceanography from the Ocean University of China, Qingdao, China, in 2001.

He is with the Second Institute of Oceanography (SIO), Ministry of Natural Resources (MNR), Hangzhou, China, where he is the Head of the Microwave Marine Remote Sensing, State Key Laboratory of Satellite Ocean Environment Dynamics. He is also an Adjunct Professor and a Doctoral Supervisor with Zhejiang University, Shanghai Jiao Tong University, Shanghai, China, and Hohai University, Nanjing, China. He is also a Senior Member of Southern Marine Science and Engineering Guangdong Laboratory, Zhuhai, China. He has more than 20 years of experience in microwave marine remote sensing. He has been a Principal Investigator and a participant of more than 20 research projects, and authored or coauthored more than 100 scientific articles in peer-reviewed journals and international conference proceedings. His research interests include microwave marine remote sensing, data fusion, image processing, and satellite oceanography.

Xianggang Wang, photograph and biography not available at the time of publication.

Yang Du (Senior Member, IEEE) received the B.E. degree in precision instrumentation from Tsinghua University, Beijing, China, in 1991, the M.S. degree in electrical engineering from the University of Massachusetts Dartmouth, Dartmouth, MA, USA, in 1997, and the joint M.S. degree in mathematics and the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2001 and 2003, respectively.

In 1993, he joined Ericsson China Limited, Beijing. In 2007, he joined the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA, as a Visiting Scientist. In 2012, he joined the Department of Atmospheric Science, Texas A&M University, College Station, TX, USA, as a Research Scholar. He is currently a Professor with Zhejiang University, Hangzhou, China, and also the Associate Director of the Innovative Institute of Electromagnetic Information and Electronic Integration. His research interests include microwave remote sensing, image processing and analysis, statistical signal processing, and deep learning.

Dr. Du was a Member of IEEE James H. Mulligan, Jr., Education Medal Committee. He is an Associate Editor for IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and was the Deputy Editor of IEEE GRSS REMOTE SENSING CODE LIBRARY.