A Feature Fusion-Net Using Deep Spatial Context Encoder and Nonstationary Joint Statistical Model for High-Resolution SAR Image Classification

Wenkai Liang, Yan Wu, Member, IEEE, Ming Li, Member, IEEE, and Yice Cao

Abstract—The nonstationary and non-Gaussian distribution of the high-resolution (HR) synthetic aperture radar (SAR) image provides much valuable information. However, the current methods, especially deep learning models, directly learn spatial features from HR SAR data while ignoring global statistical information. Combining the local spatial features and global statistical properties of HR SAR images is urgently needed to capture complete HR SAR characteristics. In this paper, a feature fusion network (Fusion-Net) using both deep spatial context encoder and nonstationary joint statistical model (NS-JSM) is proposed for the first time. Fusion-Net realizes the fusion description of local spatial and global statistical features in an end-to-end supervised classification framework. First, a deep spatial context encoder network (DSCEN) is designed based on multiscale group convolution (MSGC) module and channel attention (CA) module. The DSCEN expands the scope of context information extraction with few parameters and increases the interaction between high-level feature channels. Then, the NS-JSM is adopted to capture the unique SAR statistical information. Specifically, the SAR image is transformed into the Gabor wavelet domain. The produced sub-band magnitudes and phases are modeled by the log-normal and uniform distribution. The covariance matrix (CM) is calculated for mapped sub-band data to capture the interscale and intrascale nonstationary correlation. Finally, the group compression and smooth normalization units are introduced into Fusion-Net to fuse the statistical features and spatial features, which not only exploits the complementary information between different features but also optimizes the fusion feature representation. Experiments on four real HR SAR images validate the superiority of the proposed method over other related algorithms.

Index Terms—Convolutional neural networks, feature fusion, high-resolution (HR) synthetic aperture radar (SAR) images, image classification, nonstationary joint statistical model (NS-JSM).

I. INTRODUCTION

SYNTHETIC aperture radar (SAR) imaging system has the ability to observe the Earth surface without the constraints of illumination and cloud coverage [1]. It has become a very significant source of ground information in the field of modern remote sensing. SAR land cover classification is an important step in a variety of SAR image interpretations and applications, such as agricultural monitoring, urban planning, and damage assessment [2], [3]. With the development of new generation SAR sensors, e.g., TerraSAR-X [4], Gaofen-3 [5], and airborne SAR, large amounts of high-resolution (HR) SAR images have become available. Although the HR SAR image can provide sufficient detailed information of ground objects, it also presents more complex backscattering and spatial layout hard to deal with. Thus, pushing toward the novel HR SAR classification methods is urgently needed.

For SAR image classification, effective feature extraction or feature learning is essential. In most cases, the discriminant ability of the features determines the quality of SAR image classification. The dominant discrimination of single-polarized SAR images is the amplitude or intensity information. Thus, mining effective feature representations from the spatial context of a pixel plays a crucial role in the classification decision. According to the current research trend of feature extraction, the mainstream methods for SAR image classification can be roughly categorized as data-driven deep learning methods and statistical analysis methods.

To depict the content of SAR images, most traditional methods rely on extracting intensity [6] and texture information [7]. Some work also focuses on designing more discriminant handcrafted feature descriptors for SAR images, such as SAR histogram of oriented gradients (SAR-HOG) features [8] and covariance of textural features (CoTF) [9]. Compared with traditional methods, many studies have proved the ability of deep neural networks (DNNs) [10] to automatically extract discriminative features of SAR images and achieve remarkable results with limited labeled data. Geng et al. [11] proposed a deep supervised and contractive neural network (DSCNN) whose inputs are the combination of GLCM, Gabor, and HOG features for high-level feature learning. Chen et al. [12] proposed an all-convolutional network (A-ConvNet) for SAR target recognition, which removes the fully connected layer and adds a dropout layer to prevent overfitting. Sun et al. [13] proposed a ConvCRF model to combine fully connected CRF MAP modeling with convolutional representation layers, which realizes comprehensive feature representation with global and local feature information of SAR image. A greedy hierarchical convolutional neural network (GHCNN) is developed to realize an efficient patch-based classification for single-polarized SAR

1558-0644 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
images [14]. In [15], a modified VGGNet (MVGG-Net) that transfers pretraining parameters from the ImageNet dataset is proposed for extracting deep features of SAR amplitude images. Besides, there are some attempts to use statistical methods to encode CNN features to improve classification accuracy. Liu et al. [16] proposed a statistical CNN (SCNN) for SAR land-cover classification, which characterizes the distributions of CNN features by the mean and variance statistics. Liang et al. [17] integrated the covariance-based second-order statistics of CNN features, verifying that high-order statistics can improve the ability of CNN to distinguish various SAR land covers. In addition, a new Fully Convolutional Network (FCN) architecture [18] was proposed that can be trained in an end-to-end scheme and is specifically designed for the classification of wetland complexes using PolSAR imagery.

Due to the unique coherent imaging mechanism, the statistical properties of SAR images provide valuable information. The parametric approach is to postulate a given mathematical distribution for the statistical modeling, which has been intensively studied for SAR feature extraction due to simplicity and applicability [19]. Some non-Gaussian parametric models have been employed to extract statistical features of backscatters from different land covers, such as Rayleigh [20], Gamma [21], K [22], log-normal (LN) [23], Weibull [24], Fisher [25], and generalized Gamma [19]. As the resolution increases, the emergence of heterogeneous regions makes the modeling of HR SAR images still challenging. To accurately describe the statistical properties of HR SAR images, two extended methods have been proposed. One is to use the basic probability model combination to build a stronger model to describe the SAR statistical properties. Many mixed statistical models, such as the Gamma mixture model [26], generalized Gamma mixture model [27], and LN mixture model [28] have been proposed. These models are then used in a Bayesian framework such as Markov random fields [29] to implement classification. The other is to extend SAR images to complex value domains through transform domain methods such as Gabor transform [30], Wavelet transform [31], and Contourlet transform [32]. By statistical modeling of multidimensional complex value sub-bands, the statistical features are more discriminant. Karine et al. [33] used Weibull or Gamma distributions to model the dual-tree complex wavelet transform (DT-CWT) sub-bands of SAR images and stacked the statistical parameters obtained by each sub-band to form the statistical feature vectors. In [34], the author established the statistical dependence between DT-CWT sub-bands by introducing the Copula model [35] and used the multivariate copula parameters constructed as the statistical features. Then, the above features are fed into a classifier such as the Softmax [36] or the sparse representation [37] for classification.

There are a few works [16], [38], [39] that try to fuse multiple features to improve the classification accuracy for SAR images. They focus on the fusion of deep features and other primary features such as polarization features and wavelet features. The effectiveness of statistical properties of the SAR image was ignored in these methods. Also, because each part of these fusion methods is individually learned, it cannot benefit from end-to-end learning. As a result, we recognize that the fusion description of local spatial and global statistical features in an end-to-end supervised classification framework can capture complete HR SAR characteristics.

To realize this goal, three important challenges remain. First, HR SAR images show complex structural and geometrical features, and the traditional CNNs only use the single-scale convolution blocks that limit the scope of spatial information extraction. Transfer learning makes it easier for the deep model to learn the discriminative features from SAR images, but it ignores the inherent imaging mechanism differences between different datasets. In addition, deep models such as VGGNet [40] and ResNet [41] contain a huge amount of parameters, which will increase the computational burden and memory consumption. More practically, the development of a more effective lightweight network is an inevitable requirement for intelligent SAR processing systems. Second, through the statistical modeling of multiple wavelet sub-bands of the SAR image and establishing the dependence between the sub-bands, the expression ability of statistical features can be effectively improved. The common way is to use the copula model to jointly model the dependence between multiple sub-band distributions. However, the copula model usually needs to calculate a closed-form Kullback–Leibler divergence in the similarity measure. The parameter optimization process will be very time-consuming. Thus, another challenge is designing a more efficient statistical feature representation scheme for HR SAR images, and it can be effectively integrated with CNN features in linear space. Third, to fuse spatial and statistical features, the most direct way is to concatenate two types of features in a proportional weight parameter. However, the determination of the proportional parameter requires tedious experiments and is difficult to put into practice. Hence, the third challenge is how to exploit spatial and statistical information more effectively.

To address the aforementioned challenges, we propose a novel HR SAR image classification framework (named Fusion-Net) in this article. The proposed framework contains the following three modules: a spatial feature extraction module, a statistical feature extraction module, and a feature fusion module. First, inspired by the convolutional block attention module [42], group convolution [43], and dilated convolution [44], a new deep spatial context encoder network (DSCEN) is proposed to extract spatial features from SAR images with a small amount of labeled data. Second, the covariance matrix (CM) is introduced to describe the non-stationary correlation of multidimensional statistical wavelet sub-bands for the HR SAR image. This method can fully capture the statistical properties of the HR SAR image and form a discriminant statistical feature descriptor. Finally, considering the complementarity of spatial and statistical features, Fusion-Net is proposed to fuse two types of features and the complete model is trained in an end-to-end manner. Compared with other SAR image classification methods, the proposed method can combine the advantages of local spatial features and global statistical features, and the multifeature information fusion in a unified training process can boost the robustness
of Fusion-Net for various land covers. The main contributions of this article are listed as follows.

1) A new DSCEN with a lightweight structure is proposed to extract spatial features from SAR images. Our DSCEN consists of the multiscale group convolution (MSGC) module and channel attention (CA) module. The MSGC module can expand the scope of context information extraction in the spatial domain with few parameters. The CA module, located at the last layer of the network, is used to increase the interaction between high-level feature channels. Consequently, the DSCEN is able to capture multiscale contextual information in a higher performance and can be effectively trained with the limited SAR labeled data, resulting in a more competitive spatial feature representation.

2) The nonstationary joint statistical model (NS-JSM) is first adopted to capture multidimensional scattering statistics in the Gabor wavelet domain of SAR image and form a more distinguishable global statistical feature. Specifically, the NS-JSM uses different distributions to model the magnitudes and phases of Gabor wavelet sub-bands and then uses the CM to form the compact global statistical features for the mapped data in cumulative distribution function (CDF) space. The obtained statistical descriptor can not only capture statistical dependence and nonstationary correlation of SAR images but also suppress the influence of noise.

3) A feature fusion network (Fusion-Net) based on group compression and smooth normalization is constructed to fuse spatial and statistical features and optimize the fusion feature representation. Fusion-Net not only utilizes the complementary information of spatial and statistical features but also merges the global statistical information into the network to participate in end-to-end training. Thus, the feature representation ability and classification performance of Fusion-Net are improved.

The rest of this article is organized as follows. Section II reviews some related works on CNN and the statistical distribution of SAR images. Section III first introduces the deep spatial feature extraction-based DSCEN, then presents the statistical feature extraction-based NS-JSM, and further proposes the feature Fusion-Net model. The experiments and results are presented in Section IV. Finally, some concluding remarks are drawn in Section V.

II. RELATED WORK

A. CNN for SAR Image Classification

The single-polarization SAR image classification method is usually to extract the patch centered on the pixel to assist the classification of the central pixel. Thanks to the efficiency with local connections, shared weights, and shift-invariance, CNNs have been applied to SAR image patches to exploit the spatial features. To achieve the pixel-level classification, the pixels from the SAR image need to extract their local patches as the input of CNN. Given the SAR image $X$ and the ground-truth $Y$ of size $m \times n$, where $m$ and $n$ are the height and width of the SAR image, respectively. $N = m \times n$ denotes the number of image pixels. $x_i, i = 1, \ldots, N$ represents the $i$th pixel, and $y_i$ stands for the pixel label. First, all the labeled pixels together with their local patches $X_i$ with the size of $s \times s$ are extracted to form the samples. Then, training, validation, and test samples are constructed for training and evaluating the CNN model. The basic components of the CNN model contain a convolutional layer, nonlinear activation layer, and pooling layer. Generally, we treat a combination of a convolutional, activation, and pooling layer as a convolution block. Specifically, the major operations performed in the convolutional block can be represented as

$$F^l = \text{pool}(\sigma(F^{l-1} \otimes W^l + b^l))$$  \hspace{1cm} (1)

where $F^{l-1}$ is the input feature map of the $l$th layer, and $W^l$ and $b^l$ are the weights and bias of the $l$th layer, respectively. The input $F^0$ of the first layer of CNN is the patch $X_i$. $\sigma(\cdot)$ denotes the activation function, which can be sigmoid, tanh, or ReLU [45]. pool(·) is the pooling operation for abstracting the feature maps.

The successive convolutional blocks that are stacked together can extract the high-level features. Then, the output feature maps are flattened into a 1-D vector, and a fully connected classification is performed. In the end, the softmax function is connected on the last layer to form the class conditional probability $p_i^c$ of each sample, which is defined as

$$p_i^c = \frac{\exp(z_{i}^c)}{\sum_{c=1}^{C} \exp(z_{i}^c)}, \hspace{1cm} c = 1, \ldots, C$$  \hspace{1cm} (2)

where $z_{i}^c$ denotes the weighted sum of inputs to the $c$th unit on the output layer, and $C$ is the total number of classes. To optimize the CNN model, the cross-entropy loss function [46] is adopted as the learning objective, which is defined as

$$L = -\sum_{c} c_i \log(p_i^c).$$  \hspace{1cm} (3)

The mini-batch gradient descent algorithm is used to optimize the parameters of CNNs. After the model completed the training, the label of the test sample can be selected according to the maximum probability as

$$\hat{y}_i = \max_{c=1, \ldots, C} p_i^c.$$  \hspace{1cm} (4)

Due to the limited available training samples of SAR images, some optimization trick modules can be used to speed up the training procedure and prevent overfitting. Generally, Batch-normalization [47] is connected to the convolutional layer to accelerate model convergence by preventing gradient vanishing. Data augmentation and dropout [48] are used to prevent overfitting and further enhance network performance.

B. Statistical Models for SAR Images

SAR is an active microwave imaging system that emits electromagnetic waves and receives backscattered echo signals. Due to the coherent scattering processes at each pixel, the
parametric scattering models of SAR images can be expressed as follows:

\[ X = X_{re} + jX_{im} = A e^{j\theta} = \sum_{k=1}^{K} A_k e^{j\theta_k} \]  

where \( K \) is the number of discrete scatterers, and \( A_k \) and \( \theta_k \) are the amplitude and phase of the \( k \)th scatterer, respectively. \( X_{re} \) and \( X_{im} \) are the decomposition of \( X \) in its real and imaginary parts. With the Gaussian assumption of \( X_{re} \) and \( X_{im} \), the amplitude \( A \) follows a Rayleigh distribution, and the intensity \( I = A^2 \) has a negative exponential probability density function (pdf) [20]. In many practical cases, the statistical distribution of SAR images exhibits non-Gaussian behavior. Some prior hypothesis models such as Gamma, K distribution are proposed to fit the distribution of SAR data. Additionally, there are some empirical distribution models such as LN, Weibull, and Fisher distributions that are obtained by experimental analysis on actual SAR images.

In SAR image segmentation or classification tasks, modeling only the amplitude or intensity of SAR images may not be enough for statistical features to have sufficient discriminant ability. To fully explore the statistical properties of the SAR image, a feasible way is to extend the SAR image to the wavelet domain to enhance the discriminative of the statistical features. In general, there is no specific statistical model for signals in the complex domain of SAR images. It is commonly accepted that the coefficients are highly non-Gaussian and exhibit heavy-tails [49]. Note that parameter estimation is a key issue for the use of statistical models in SAR image processing. Accuracy of the model solution and complexity of the estimation has a great impact on the results and their usage [19]. Based on the above analysis, the combination of wavelet decomposition and simple pdfs mentioned above to capture the statistical properties of SAR images can not only improve the computational efficiency but also improve the discriminant ability, which may be a powerful and promising way.

III. PROPOSED METHODS

In this article, a novel end-to-end feature Fusion-Net framework is proposed to make full use of the complementarity among spatial and statistical information. The proposed framework is illustrated in Fig. 1, which consists of the following three steps: 1) spatial feature extraction using the DSCEN; 2) statistical feature extraction using the NS-JSM; and 3) spatial-statistical feature fusion and classification using Fusion-Net. Relevant details of each part are introduced in Sections III-A–III-C.

A. Deep Spatial Feature Extraction

Contextual information, which reflects the underlying spatial dependencies between the central pixel and its surroundings, is pivotal to identifying SAR ground objects. Some classical CNNs, such as VGGNet [40], and ResNet [41], exploit spatial context features by successive stacking standard convolution blocks. However, both the size of the input patch and the complexity of CNN should be considered when applying deep CNNs to the SAR image processing task. On the one hand, these networks contain too many pooling layers, which will overly contract the feature space of the small SAR patch, thus affecting the classification accuracy. On the other hand, the limited SAR labeled data is not enough to support the fully supervised very deep CNN training. In addition, the computational burden and memory consumption of a large CNN model are often faced with many practical limitations. Based on the above analysis, a new lightweight DSCEN model is designed for SAR feature extraction. As shown in Fig. 1, the MSGC module and CA module are proposed in DSCEN to generate spatial feature representation for the SAR image.

1) Multiscale Group Convolutional Module: Generally, the deeper layers in CNNs contain the larger receptive field on the input image, which can learn more extensive spatial context features. Fig. 2(a) shows a standard convolutional module usually used in CNN, which is composed of two consecutive 3 × 3 convolutional layers. There are two imperfections that need attention. One is that a single standard convolution block
to increase the receptive field while reducing the parameters. Dilated convolution is to insert zeros in convolutional kernels. The main idea of we can see the difference between the standard convolution, G-Conv, and DG-Conv. From the top of Fig. 3, context features. Fig. 3 shows a comparison of the standard convolution, G-Conv, and DG-Conv. From the top of Fig. 3, context features. The dilate convolution and group convolution are applied to the MSGC module. The top branch of the module is a standard convolutional module, which is used to perform dimension transformation and increase the nonlinearity. The bottom branch contains two different types of feature extractors to capture the channel-wise dependencies. Note that the shared MLP can ensure that the model does not highlight the importance of some salient feature maps, failing to distinguish complex scenes. Inspired by the attention mechanism [42], our goal is to enhance the feature representation power by CA module. Usually, the CA module is applied behind each convolution module. However, it ignores the characteristics of different layers and the influence of the CA module on the HR SAR image. Based on insights about CNN properties [50], the output feature maps at the end of CNN contain high-level semantic information as well as sufficient spatial information. Zhao et al. [54] proved that there is no need to use the CA module for low-level features, because there are almost no semantic distinctions among different channels of low-level features. As for the HR SAR image, if we add the CA module at each layer in DSCEN, some strong response feature maps caused by noise may dominate in the shallow layer. These interferences can be accumulated through the network and propagated to the deep layer, thereby weakening the discriminative ability of DSCEN. Also, since DSCEN belongs to a lightweight model, adding too many CA modules will increase the number of parameters and computation. Based on the above analysis, we apply the CA module in the last layer of DSCEN, which can efficiently pay more attention to the meaningful class-specific information and reduce the computational burden. Moreover, the effects of different positions of the CA module are compared in the experiments and we empirically verify that adding the CA module at the end of the DSCEN is an optimal choice for the current task.

To learn effective channel-wise weights, we first aggregate the spatial dimension of the input feature map. Similar to [42], the global average pooling (GAP) and global max pooling (GMP) operators are applied to aggregate spatial information of the input features. According to the characteristics of GAP and GMP, GAP can obtain the global context-aware information of the object for promoting classification performance, while GMP can capture the most distinctive object features to infer finer channel-wise attention. Thus, the use of the GAP and GMP methods together is conducive to learning better feature expression. Then, two feature descriptors are fed into a shared multilayer perceptron (MLP) with one hidden layer (where the number of hidden layer units is FN/re, re is the reduction ratio) to capture the channel-wise dependencies. Note that the shared MLP can ensure that the model does across the channels in different partitions, and it can be regarded as structured sparse. Compared with standard convolution, group convolution can reduce the number of model parameters and may also achieve better performance. Besides, we further compress the number of feature channels to half of the input in both G-Conv and DG-Conv layers. A simple concatenation layer followed by the BN and ReLU operators is utilized to aggregate multiscale features. Finally, the proposed MSCB module can achieve a significant performance boost while keeping the model complexity to a minimum.

2) Channel Attention Module: HR SAR images contain complex structural and geometrical information, which may lead to high intraclass variance and low interclass difference. The commonly used CNNs consider that each channel feature map in one layer has the same importance. Thus, it may not be able to highlight the importance of some salient feature maps, failing to distinguish complex scenes. Inspired by the attention mechanism [42], our goal is to enhance the feature representation power by CA module. Usually, the CA module is applied behind each convolution module. However, it ignores the characteristics of different layers and the influence of the CA module on the HR SAR image. Based on insights about CNN properties [50], the output feature maps at the end of CNN contain high-level semantic information as well as sufficient spatial information. Zhao et al. [54] proved that there is no need to use the CA module for low-level features, because there are almost no semantic distinctions among different channels of low-level features. As for the HR SAR image, if we add the CA module at each layer in DSCEN, some strong response feature maps caused by noise may dominate in the shallow layer. These interferences can be accumulated through the network and propagated to the deep layer, thereby weakening the discriminative ability of DSCEN. Also, since DSCEN belongs to a lightweight model, adding too many CA modules will increase the number of parameters and computation. Based on the above analysis, we apply the CA module in the last layer of DSCEN, which can efficiently pay more attention to the meaningful class-specific information and reduce the computational burden. Moreover, the effects of different positions of the CA module are compared in the experiments and we empirically verify that adding the CA module at the end of the DSCEN is an optimal choice for the current task.

To learn effective channel-wise weights, we first aggregate the spatial dimension of the input feature map. Similar to [42], the global average pooling (GAP) and global max pooling (GMP) operators are applied to aggregate spatial information of the input features. According to the characteristics of GAP and GMP, GAP can obtain the global context-aware information of the object for promoting classification performance, while GMP can capture the most distinctive object features to infer finer channel-wise attention. Thus, the use of the GAP and GMP methods together is conducive to learning better feature expression. Then, two feature descriptors are fed into a shared multilayer perceptron (MLP) with one hidden layer (where the number of hidden layer units is FN/re, re is the reduction ratio) to capture the channel-wise dependencies. Note that the shared MLP can ensure that the model does
not introduce additional parameters. Finally, we merge the output feature vectors by using an element-wise addition and a sigmoid function. The obtained CA vector can be computed as
\[
M_c(F) = \text{sigm}(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F)))
\]
where \text{sigm} denotes the sigmoid function, and MLP is a shared multilayer perceptron. The flowchart is shown in the CA module of Fig. 1. Finally, we can obtain the more discriminative output features by employing a scale layer to re-weight the high-level features with the CA vector. It should also be noted that through end-to-end training, the network is capable of adaptive learning the weights of the feature maps, thereby focusing on important features and suppressing useless ones more efficiently.

3) Network Architecture: Based on the MSGC module and CA module, we construct the DSCEN model to extract the spatial features from the SAR image patch. As shown in Fig. 1, the proposed DSCEN model contains four MSGC modules and one CA module. The number of output channels of each MSGC module is 16, 32, 64, and 128, respectively. At the end of each MSGC module, a max-pooling layer is followed to abstract the feature maps. In addition, dropout is added after each pooling layer to prevent overfitting. Given an input image patch \( X_i \) centered on the pixel \( x_i \), whose spatial feature \( F_{\text{Spatial}} \) can be conducted multiple MSGC module-based convolutions and one CA module-based recalibration in DSCEN. The formula for calculating the spatial feature \( F_{\text{Spatial}} \) is described as follows:
\[
F_{\text{Spatial}} = M_c(\text{MSGC}_4(\text{MSGC}_3(\text{MSGC}_2(\text{MSGC}_1(X_i))))) \tag{7}
\]
where MSGC represents an MSGC operation. When the spatial feature representation is obtained, it can be imported to the feature fusion network for classification.

B. Statistical Feature Extraction

High-order scattering statistics of the SAR image provide much valuable information for data representation. However, CNN-based methods rarely consider and exploit the statistical properties of the SAR image. It is necessary to merge the statistical information into CNN to improve the feature representation. Due to the lack of multidimensional information on a single pixel for the single-polarization SAR image, the extracted pixel-based statistical features may not have sufficient discriminative ability. Considering that HR SAR data provides more information in the spatial domain, we can utilize the statistical characteristics of the pixel neighborhood to compensate for the lack of spectral information. Furthermore, Karine et al. [33] proposed to extend the SAR image wavelet sub-bands and represent each sub-band by a parametric probabilistic model to enhance the representation ability of statistical features. Following the same philosophy, we extend the single-polarization HR SAR image to the Gabor wavelet domain to model it statistically and make the statistical feature more discriminative. Notably, Dong et al. [55] proved that introducing the global statistics from different polarization channels of multipolarization SAR data patches can further improve the expression ability of statistical features. Another feasible way is to replace the wavelet sub-bands with multidimensional polarization features as the extended features, which will be considered as our future research work for multipolarization HR SAR images.

Considering that the Gabor filter has scale and direction selection characteristics, it is compatible with the direction sensitivity of SAR images. Therefore, the Gabor filter is chosen to transform the SAR image into the complex wavelet domain. Given an SAR image \( I \) and the Gabor filter \( G_{u,v} \), the Gabor wavelet sub-bands can be computed as follows:
\[
O_{u,v}(z) = I(z) \ast G_{u,v}(z) \tag{8}
\]
where \( \ast \) is the convolution operator, \( z = (z_1,z_2) \) denotes the coordinates in the spatial domain, \( v = 1,\ldots,V \) and \( u = 1,\ldots,U \) represent the scale and direction. \( U \) and \( V \) are the number of scales and directions of Gabor filters, respectively. \( O_{u,v}(z) \) is the complex number result. Due to the fact that the phase of the wavelet is also important discriminative information, we extract the magnitude and phase of the Gabor wavelet simultaneously. Then, the magnitude \( M_{u,v}(z) \) and the phase \( P_{u,v}(z) \) of the Gabor filter output are computed as
\[
M_{u,v}(z) = \left( \text{Re}(O_{u,v}(z)) \right)^2 + \left( \text{Im}(O_{u,v}(z)) \right)^2 \tag{9}
\]
\[
P_{u,v}(z) = \arctan \left( \frac{\text{Im}(O_{u,v}(z))}{\text{Re}(O_{u,v}(z))} \right) \tag{10}
\]

For the convenience of subsequent representation, the magnitude and phase sub-bands are vectorized and organized into the observation matrix \( M_{\text{mag}} \in \mathbb{R}^{n \times d} \) and \( M_{\text{pha}} \in \mathbb{R}^{n \times d} \), respectively. Each column of \( M_{\text{mag}} \) or \( M_{\text{pha}} \) is the observation of the magnitude variable \( m_j \) or phase variable \( a_j \), \( j = 1,\ldots,d \). Here, each column corresponds to a sub-band. \( n \) is the total number of variables in a sub-band, and \( d = U \times V \) represents the number of sub-bands.

For the observations of a sub-band, it is customary to make an underlying assumption based on a specific distribution and then compute the distribution parameters to build the feature vectors. Supposing the observations in \( j \)-th column of \( M_{\text{mag}} \) or \( M_{\text{pha}} \) obey a certain marginal distribution which CDF is \( F_j(r|\theta) \) and PDF is \( f_j(r|\theta) \), then we can use the observations of variables to estimate the distribution \( f_j(r_j|\theta) \) and parameter \( \theta_j \) by maximum likelihood or moments. Yu et al. [51] concluded that non-Gaussian pdfs are suitable to model the statistical behavior of the magnitude of Gabor wavelet sub-bands. Considering simplicity and applicability, the LN distribution was used to fit the distribution of the magnitude sub-bands, and the uniform distribution is adopted to model the distribution of the phase sub-bands in our work.

Afterward, inspired by Li et al. [52], the observations of each column are projected to its corresponding CDF space by using CDF \( F_j(r_j|\theta_j) \) histogram. The CDF space of the amplitude sub-band can be expressed as \( M_F \in \mathbb{R}^{n \times d} \)
\[
M_F = [F_1, F_2, \ldots, F_d] = \begin{bmatrix} F_{1,1} & F_{1,2} & \cdots & F_{1,d} \\ F_{2,1} & F_{2,2} & \cdots & F_{2,d} \\ \cdots & \cdots & \cdots & \cdots \\ F_{n,1} & F_{n,2} & \cdots & F_{n,d} \end{bmatrix} \tag{11}
\]
where \( F_i = [F_{1,i}, F_{2,i}, \ldots, F_{n,i}] \) is the detailed CDF vector for each column.

The most direct way is to construct the statistical descriptor by utilizing whole statistics. However, this strategy will lead to a very high-dimensional vector while ignoring the statistical dependence and nonstationary correlation between sub-bands. To overcome this limit, the NS-JSM is used to describe these sub-bands statistics, which can form a more compact and robust statistical descriptor. To conveniently calculate the CM, we change \( M_F \) into the following form, \( \bar{M}_F = (M_F)^T = [Z_1, Z_2, \ldots, Z_n] \), where \( Z_l = [F_{1,l}, F_{2,l}, \ldots, F_{d,l}]^T, l = 1, \ldots, n \). Finally, the magnitude-based statistical feature can be represented by \( d \times d \) CM of the sub-bands statistics

\[
C_{\text{mag}} = \frac{1}{n-1} \sum_{l=1}^{n} (Z_l - \mu)(Z_l - \mu)^T
\]  

(12)

where \( \mu \) is the mean of the feature vectors \( Z_l, l = 1, \ldots, d \).

Through the above calculation steps, we can also obtain the CM \( C_{\text{pha}} \) corresponding to the phase-based statistical feature.

There are two advantages of the NS-JSM: 1) CM builds the dependence and nonstationarity between two different sub-bands, which can fuse complementary information coming from different sub-bands and form the more compact and discriminative feature and 2) there is an average filter during bands, which can fuse complementary information coming as feature descriptor of the SAR image patch can be expressed based on magnitude and phase, we flatten and concatenate the two types of covariance features. Since the CM here is symmetric, we only use the upper triangular part of the magnitude CM and the lower triangular part of phase CM. Notably, the CM usually resides on the Riemannian manifold of the SPD matrix [53]. Logarithmic transformation is applied to map the CM into the linear space, and the final statistical feature descriptor of the SAR image patch can be expressed as

\[
F_{\text{Statistical}} = \text{triu}(\log(C_{\text{mag}})) || \text{tril}(\log(C_{\text{pha}}))
\]

(13)

where || denotes the operation of concatenating. The scheme of the statistical feature extraction process is shown in Fig. 1.

C. Spatial-Statistical Feature Fusion-Net

After obtaining the spatial and statistical features, it is important to effectively fuse them for classification. A commonly used method is to utilize a weighted strategy to fuse the two types of features. However, it required a large number of experiments to find the optimal weight parameter. Also, this strategy cannot benefit from end-to-end learning to obtain more robust performance. To utilize the complementary information between the two types of features, we propose to train a two-layer perceptron feature fusion network (Fusion-Net) that can embed the statistical features into the spatial features in nonlinear feature space. In Fusion-Net, we use a sparsely connected network based on group convolution in the first layer for dimension reduction on the two features. Then, we adopt a fully connected network in the second layer to further fuse and optimize features so as to enhance feature discrimination. We define the fusion scheme as follows:

\[
F_{\text{Fuse}} = \text{sigm}(W_2(\text{sigm}(GConv(W_1, F_{\text{Spatial}}))))
\]

\[
\times ||\text{sigm}(GConv(W_1', F_{\text{Statistical}})))
\]

(14)

where \( W_1 \) and \( W_1' \) are the weights of the first sparsely connected layer, respectively. \( W_2 \) is the weight of the second fully connected layer. GConv denotes the group convolution. The proposed Fusion-Net has the following advantages. First, since spatial and statistical features have high dimensions, the sparsely connected layer can achieve dimensional reduction with fewer parameters. It can also suppress useless information in each feature before fusion. Second, the sigmoid function as a smooth normalization mechanism can transform the features into a relatively consistent space. This choice can prevent either feature from becoming dominant, thus encouraging the contribution from both features. Third, compared to the fully connected network, Fusion-Net has fewer parameters to learn complementary information, so it can further prevent overfitting.

Finally, the fusional feature vectors are directly input to the softmax function to generate the predicted labels. In the training stage, the cross-entropy loss is adopted as the objective function. The parameters of the proposed method are trained in an end-to-end manner through iterative methods. Thus, the CNN-based spatial information and the global statistical information can interact during the training process, and the classification performance can be significantly improved.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we evaluate the performance of the proposed method for HR SAR image classification. The dataset description, detailed experimental setup, and experimental results with reasonable analysis are presented below.

A. Description of the Datasets

Four real HR SAR images obtained from different sensors were used to validate the effectiveness of the proposed method. These four data are HR and contain complicated structural and geometrical scene information. For each dataset, the ground truth images are generated by manual annotation according to the associated optical image, which can be found in Google Earth. The first HR SAR image [56] was acquired by the TerraSAR-X satellite over the scene of Lillestroem, Norway, in 2013. It has 2675 \( \times \) 1450 pixels in size with an HH-polar imaging mode. The acquisition mode of the data is staring spotlight and the resolution of the image is 0.5 m. This scene contains five kinds of ground objects: Water, residential, roads, woodland, and open land. The original image and the ground truth image are shown in Fig. 8(a) and (b).

The second dataset was collected by the Chinese Gaofen-3 satellite in Guangdong Province, China, in 2017. The image size is 2600 \( \times \) 4500 pixels, and the spatial resolution is 0.75 m. The image is HH-polarization data with the sliding spotlight mode. This data consists of seven classes: Mountains, water, building, roads, woodland, and open land. The original image and the ground truth are shown in Fig. 9(a) and (b).
The third dataset was from the area of Shaanxi province, China, collected by a Chinese airborne sensor in 2016. It was provided by the China Electronics Technology Group Corporation (CETC) Institute. This data has a size of 1800 × 3000 with a spatial resolution of 0.3 m. The image is HH-polarization data with spotlight mode. Seven classes are considered for the experiment: open land, roads, rivers, runway, woodland, residential, and commercial. The original image and the ground truth image are shown in Fig. 10(a) and (b).

The fourth dataset [57] was acquired by an X-band F-SAR sensor of the German Aerospace Center in 1989. The dataset was from the Bavaria region in Germany, whose spatial resolution is 0.67 m. It contains 6187 × 4278 pixels and the polarization of the data is VV-polar mode. Four classes of interest were considered: water, residential, vegetation, and open land. The original image and the ground truth image are shown in Fig. 11(a) and (b), respectively.

B. Experimental Setup and Evaluation Metrics

The proposed classification method consists of three parts: spatial feature extraction using DSCEN, statistical feature extraction using NS-JSM, spatial-statistical feature fusion, and classification using Fusion-Net. First, we set the public parameters and training strategies involved in model training. Then, the detailed parameter determination for each module is discussed in the subsequent part. During the training phase, the unique loss function is optimized by the Adam optimizer [58] with a constant learning rate of 0.001. The mini-batch size is set to 100 and the number of epochs is set as 150. In our DSCEN, all the convolutional weights are initialized from Gaussian distributions with zero mean and a standard deviation of 0.01, and no bias term. Moreover, the dropout ratio is set to 0.2.

Following the setting in [16], [59], and [60], all the labeled pixels together with their neighborhood patches are first extracted to form the samples. The 300 samples of each category are sampled according to the ground truth. Then, these samples are divided into the training and validation, accounting for 70% and 30%. The other samples are used for testing. To avoid the overlap phenomenon between samples, the overlap threshold is set to 50% (if the overlap area is more than 50% between training and test samples, the test samples are not involved in the evaluation). The patch spatial size $s$ is set to 64 in our work. The amplitudes of the HR SAR image are normalized in the range of 0–1 by max–min normalization and used as the input. In addition, data augmentation including rotation and flipping were applied for training samples, which can increase the number of training samples to eight times the original.

Notably, the parameter determination and analysis of the TerraSAR-X SAR image are discussed in the subsequent ablation study. The specific trend analysis of the other three HR SAR images is the same as the TerraSAR-X SAR image. Here, we hope to avoid parameter tuning for each dataset and apply the same optimization model to other datasets. It can effectively verify the generalization performance of the model while reducing time consumption for actual application scenarios. To reduce the influence of random initialization, each experiment was run five times independently. The overall accuracy (OA), average accuracy (AA), kappa coefficient ($\kappa$), and class-specific accuracy are used for evaluating the classification results. All experiments were conducted by MATLAB 2014a on the platform of a computer with I7 3.2-GHz CPU and 32-GB memory.

C. Analysis of DSCEN

In the proposed DSCEN, the feature number of the network, the depth of the network, the effect of the MSGC module, and the effect of the CA module were discussed in detail as follows. When analyzing the effect of a particular module, we vary this parameter while fixing all others.

1) Effect of Different Combinations of the Depth and Feature Number of DSCEN: The depth and the feature number of DSCEN are two vital parameters that impact the classification performance and computational complexity. Thus, we conduct experiments to test the effect of different combinations of the depth and feature number of DSCEN. In DSCEN, the number of MSGC modules determines the depth of the network. Meanwhile, the feature number in each MSGC module mainly affects the total parameters of DSCEN. Since there is a pooling layer behind each MSGC module, the size of the feature map will be further compressed as the number of layers deepens. Considering the small size of the SAR patch, we train and test on five types of DSCEN models with depths ranging from 1 to 5. We follow roughly the rule that the feature number in an MSGC module is twice that of the previous MSGC module. We vary the feature number of features of the first MSGC block in each compared DSCEN and set its range to 8, 16, 32, 64, and 128. Fig. 4 shows the accuracy comparisons of different combinations of the DSCEN depth and feature number of DSCEN. Due to space issues, the feature number in Fig. 4 only represents the size of the first MSGC module. For DSCEN with different layers, it is actually a combination of model parameters. For example, when the feature number of the first MSGC module of the five-layer DSCEN is 8, the model parameter is set to 8–16–32–64–128.
From the bottom to the top in Fig. 4, we observe that the accuracy tends to increase gradually as the network deepens. The best performance can be achieved by setting the network depth to four and the model parameters to 16–32–64–128. When the network depth reaches the fifth layer, there is no further increase in accuracy. This is because the SAR image patch size is small; the feature space is overly contracted and causes too much resolution loss, thus affecting the classification accuracy. From the left to the right in Fig. 4, we can see that the classification performance of DSCEN with fewer parameters is limited. This is mainly because too few features in CNN may not express sufficient discriminability. When the feature number of the first MSGC block is increased to 16, DSCEN obtains better accuracy. As the number of features further increases, the classification accuracy gradually declines. This is because the DSCEN model with more parameters may not be sufficiently trained due to limited training samples. In addition, too many feature channels will also lead to overfitting and increase the computational complexity of the model. Thus, we use the four-layer DSCEN with parameters set to 16–32–64–128 as the default setting in our experiments.

2) Effect of MSGC Module: There are three important components in the MSGC module that need to be compared, namely, multiscale convolution dilate convolution, and group convolution. Therefore, we define seven convolutional blocks to evaluate the effectiveness of the MSGC module. Specifically, we fixed the network at four layers and only used different convolutional blocks instead of the MSGC module for comparison. We used $M - 1$ to indicate that only a standard $5 \times 5$ convolutional layer was adopted in the block. The $M - 2$ was used to denote that two standard $3 \times 3$ convolution layers were applied. We use a $3 \times 3$ compressed convolution layer and a $5 \times 5$ compressed convolution layer to replace the bottom convolution block in the MSGC module, which is defined as $M - 3$. The main difference between it and the MSGC module is that no dilate convolution and group convolution is used. Furthermore, a $3 \times 3$ dilate convolutional layer replaces the $5 \times 5$ convolution layer in $M - 3$ as $M - 4$. It is similar to the MSGC block, but here the grouping coefficient $G$ is set to 1. $M - 5$ to $M - 7$ was constructed mainly to verify the performance of group convolution. The main difference is that they used grouping coefficients $G$ of 2, 4, and 8 in the module, respectively. The experimental results are shown in Table I.

From Table I, we can see that the $M - 6$ module with the grouping coefficient $G = 4$ can obtain the best classification results. Compared with the standard $M - 1$ and $M - 2$ blocks, we can conclude that introduction of multiscale convolution can achieve better performance, mainly because it expands the receptive field and captures more contextual information. Then, comparing the $M - 3$ and $M - 4$ models, it can be observed that the introduction of dilate convolution can basically keep the accuracy unchanged while reducing the number of parameters. In models $M - 5$ to $M - 7$, the parameter amount is further reduced and the classification performance is improved by introducing group convolution. This is because the sparse property of group convolution may reduce feature redundancy and improve the feature learning ability of the MSGC module. In summary, we use the $M - 6$ module as the default setting of DSCEN. It can significantly improve the classification results while reducing the number of model parameters.

3) Effect of CA Module: In this section, we first discuss the effectiveness of CA modules in different positions of DSCEN. Specifically, we used “CA_Stage_1,” “CA_Stage_2,” “CA_Stage_3,” and “CA_Stage_4” to indicate that the CA module is only placed behind the MSGC block1, MSGC block2, MSGC block3, and MSGC block4, respectively. The symbol “CA_Stage_All” represents that it adds the CA module behind each MSGC block in DSCEN. Also, the symbol “No-CA” indicates that the CA module is not used in the DSCEN model. Table II shows the OA, AA, kappa coefficients, training time and model size for different cases. First, we can conclude that adding the CA module can obtain better classification performance. The reason is that the CA module can selectively enhance useful features and restrain useless features, thereby improving the recognition accuracy of DSCEN. Second, it can be seen that when the CA module is placed on the last layer of DSCEN alone (Corresponds to CA_Stage_4), the highest classification accuracy can be obtained. When the CA module is placed behind the MSGC block1, there is only a small improvement in accuracy. As the CA module is gradually placed in the deeper layer, the classification accuracy also increases. This is possible because the high-level features express complex semantic features with more channels, which enables the CA module to perform feature recalibration more effectively. Besides, we observe that the accuracy increase of CA_Stage_All is lower than that of CA_Stage_4, which
TABLE III
CLASSIFICATION ACCURACY OF DIFFERENT REDUCTION RATIO re IN CA MODULE

| CA block | OA     | AA     | Kappa  |
|----------|--------|--------|--------|
| No-CA    | 0.8823 | 0.8814 | 0.8359 |
| CA( re=1 ) | 0.8898 | 0.8866 | 0.8459 |
| CA( re=4 ) | 0.8900 | 0.8872 | 0.8462 |
| CA( re=16 ) | 0.8858 | 0.8850 | 0.8407 |
| CA( re=32 ) | 0.8826 | 0.8846 | 0.8366 |
| CA( re=64 ) | 0.8875 | 0.8859 | 0.8428 |
| CA( re=128 ) | 0.8883 | 0.8867 | 0.8439 |

TABLE IV
COMPARISON RESULT OF DIFFERENT AUGMENTATION TYPES

| Augmentation Type | OA     | AA     | Kappa  |
|-------------------|--------|--------|--------|
| None              | 0.8511 | 0.8435 | 0.8267 |
| Flip              | 0.8818 | 0.8870 | 0.8347 |
| Rotation          | 0.8826 | 0.8806 | 0.8364 |
| Flip & Rotation   | 0.8900 | 0.8872 | 0.8462 |

shows that the benefits of CA blocks at different stages may be incompatible. This may be due to the accumulation of strong response features caused by noise at the output of the network, resulting in ambiguous features between some categories. Thus, we apply the CA module in the last layer of DSCEN as the default setting, which can achieve better accuracy and save more computation.

Later, we compare the performance of DSCEN without the CA module to the version with the CA module for different reduction ratios re ∈ {1, 4, 16, 32, 64, 128}. The experimental results are reported in Table III. We can observe that the CA module with a reduction ratio re = 4 can achieve the best performance. Intuitively, we can see that the CA module with different reduction ratios can produce better accuracy than without the CA module in DSCEN. This explicitly demonstrates that the CA module can boost the classification performance of our model. Moreover, the results in Table III show that the accuracy does not gradually increase as re increases. The reason is that excessive compression of global feature descriptors may not capture better feature channel-wise interaction relationships.

4) Effect of Data Augmentation: We conducted experiments to test the effectiveness of DSCEN by whether to introduce data augmentation. Table IV reports the classification accuracies obtained with different augmentation ways, including flip and rotation. We observe that the classification accuracy of DSCEN can be further improved by introducing data augmentation operations of rotation and flipping. The accuracy of OA and AA improved by almost 4% when CNN training was performed using a combination of samples generated by flipping and rotation. The possible reason is that CNN itself is a kind of “black box” mechanism, and it does not understand the unique imaging characteristics of SAR images. Although rotation and flipping may not be able to generate the real SAR data, these data provide some pseudo-coherent images with different postures. This may introduce some potential and useful information to optimize filter parameters and improve the performance of DSCEN.

D. Analysis of NS-JSM

1) Effect of Modeling of the Gabor Wavelet Sub-Bands: We tested the effect of statistical modeling of the Gabor wavelet sub-bands on the classification results. This section has different degrees of freedom such as the choice of scales and directions of Gabor filters and whether to use phase features. Here, we varied the scales and directions of Gabor filters to 4 × 4, 6 × 6, 8 × 8, 10 × 10, and 12 × 12 to observe the classification results. Meanwhile, we used the symbols “NS-JSM-M” to indicate that only the magnitude (where LN distribution is used for magnitude modeling) of the Gabor wavelet sub-band is constructed for statistical features. Similarly, the symbol “NS-JSM-MP” represents that the magnitude and phase of Gabor sub-bands are used to construct the statistical features by the NS-JSM.

Fig. 5. Effect of NS-JSM with different settings on the OA.

2) Effect of Different Statistical Models: To verify the fitting ability of the statistical modeling, we compared the classification accuracy with and without modeling the Gabor wavelet sub-bands. Also, we use the symbols “Mag” to represent only the amplitude features are used to evaluate the accuracy. The symbol “Mag-Pha” denotes that the amplitude and phase features of the Gabor sub-band are used to construct statistical features at the same time. The comparison results are presented in Table V. First, we can see that by modeling the amplitude of the sub-bands, there is an accuracy improvement.
of about 4.4% in OA, which shows that by projecting the sub-bands coefficients into the CDF space, the noise coefficients in the sub-band may be smoothed so that the more robust performance can be obtained. Then, we observed that by jointly modeling the amplitude and phase of the Gabor sub-bands, NS-JSM has an improvement of about 1.2% in OA and AA compared to the nonmodeling solution. This indicates that phase information provides a very important contribution to this method.

To illustrate the discriminative performance of the NS-JSM, we present the amplitude and phase covariance matrices corresponding to five different types of image patches from TerraSAR-X images in Fig. 6. It can be seen that the covariance features of different categories have different manifestations. Therefore, to a certain extent, the use of global statistical covariance features can distinguish the category attributes of different image patches.

E. Analysis of Fusion-Net

1) Effect of Node Number and Activation of Fusion-Net: To evaluate the performances of the proposed Fusion-Net, we conduct experiments to test the effect of different node numbers embedded in the neural network. To simplify the experiment, the number of nodes in the sparsely connected layer of the first layer and the number of nodes in the fully connected layer of the second layer in Fusion-Net are set to be equal. Then, we selected the number of nodes as 32, 64, 128, 256, and 512 for comparison. The results in Table VI show that setting the number of nodes to 128 can produce the best classification results. When the number of nodes is reduced, the feature representation ability of the converter and the fusion layer will be weakened. As the number of nodes increases, the accuracy does not increase further, but it increases the number of parameters.

In addition, to illustrate the effectiveness of smooth normalization, we test the classification performance of the sigmoid and ReLU activation in Fusion-Net. Table VII shows the comparison of our experiments. It is clearly seen that sigmoid activation yield better classification results than ReLU activation. The possible reason is that the feature value range of ReLU is from 0 to infinite. Thus, there may be some outliers that affect the performance of the fusional feature. The sigmoid activation, as a normalization mechanism, can transform the characteristics into a relatively consistent space with a range of zero to one. It can maintain more detailed information, so that the Fusion-Net can mine complementary information more effectively, thus improving classification accuracy.

2) Comparison of Different Fusion Schemes: To validate the proposed feature fusion scheme, we compare it with two other empirical fusion schemes including probability fusion and feature fusion. For probability fusion, we adopt the weight fusion to fuse the class probability maps of the DSCEN and the NS-JSM. The optimal weight is selected by trial and error. For feature fusion, we directly concatenate the output features of the DSCEN and the NS-JSM. Then, the fusion feature is fed into the softmax classifier for classification. Table VIII shows the comparison of our experiments. From Table VIII, we can see that our Fusion-Net has about a 1% improvement in OA compared to the empirical fusion method. The results indicate that the proposed fusion scheme can effectively fuse
the spatial and statistical features, and benefit from the end-to-end training manner, the prediction accuracy of the entire model can be further refined.

3) Feature Visualization: To further illustrate the ability of our Fusion-Net to extract complementary information between spatial and statistical features, Fig. 7 visualizes the distribution of three different features, which are spatial, statistical features, and fusion features, respectively. The distribution is generated by using the t-distributed stochastic neighbor embedding (t-SNE) [61] algorithm on 1000 patches (200 patches per category) randomly selected from the test data. It can be observed that the feature distribution of our fusion network has fewer overlaps than the other two types of features. For the output features of the Fusion-Net, the distance between different classes of features becomes larger. The results show that our Fusion-Net can achieve even superior feature recognition compared with the single spatial or statistical feature.

F. Experiments Comparisons

To demonstrate the superiority of the proposed method, we compare it with several related methods for HR SAR image classification. These comparison approaches are divided into three groups, including traditional feature extraction, statistical feature extraction, and feature extraction based on deep learning. The details of these comparisons are described as follows.

1) CoTF [9]: CM of the magnitude of Gabor filter responses is calculated. The matrix logarithm operation is applied to map the CM to the Euclidean space. The upper triangular part of the CM was used as the feature vector.

2) Statistical Dictionary [33]: For a fair comparison, the Gabor filter instead of DT-CWT is implemented on the SAR image. Then, the produced complex sub-bands magnitudes are modeled by the lognormal model. The obtained statistical parameters are concatenated to build a feature vector for each SAR sample.

3) Statistical CNN [16]: SCNN contains three convolutional layers, and the feature numbers are 12, 32, and 64, respectively. Then, a GAP and a global variance pooling layer are applied to the output of the last convolutional layer to capture statistical features for classification.

4) MFFN-CPMN [17]: A four-layer multiscale unsupervised feature fusion network (MFFN) is constructed. The channel numbers of each layer are set to 200. Furthermore, the covariance pooling manifold network (CMPN) is utilized to encode the global second-order statistics of SAR patches over the MFFN feature.

5) MVGG-Net [15]: We migrated the first four convolution blocks of the original VGG16-Net and added a fully connected layer containing 256 neurons to classify SAR images. This method is used to verify the performance of transfer learning and compare it to the proposed DSCEN model on the SAR image classification task.

6) ConvCRF [13]: Following the setting in [13], the CNN model is used for the unary potential to build the preliminary labeling condition. The fully conditional random field with superpixel boundary constraint is further adopted for postprocessing of classification.

7) SAR-FCN [18]: The SAR-FCN network is composed of an encoding module and a decoding module. Following the settings of SAR-FCN, the feature numbers of the four encoding layers are set to 64, 128, 128, and 256. The feature numbers of the decoding layer are the same as the encoding layer at the corresponding scale. Finally, a $1 \times 1$ convolutional layer and Softmax layer are used to project the feature map into the desired classification.

Moreover, the other three comparison methods are NS-JSM, DSCEN, and Fusion-Net proposed in this article. To ensure the fairness of comparison, comparison methods (1) and (2) also adopted the Softmax classifier to achieve the classification. For the four SAR datasets from different sensors, the comparison method and the proposed method all adopted the same model structure and parameter setting as described in Sections IV-F1–IV-F7, which can effectively verify the stability and generalization performance of the model.

8) Experiments With TerraSAR-X SAR Image: In this section, experiments are conducted on the TerraSAR-X SAR image to compare the classification result with different methods. The compared classification results of AA, OA, and kappa coefficient are reported in Table IX. From Table IX, it can be seen that the proposed Fusion-Net achieves the highest accuracy among the comparison methods and produces better classification results. The testing OA, AA, and kappa of our approach can reach 90.95%, 90.55%, and 0.8727%, respectively. For traditional feature extraction methods, our statistical features based on the NS-JSM have obtained better
classification accuracy than CoTF and SD. It indicates that our statistical features can effectively mine the scattering statistics of SAR image patches, making the formed descriptors more robust to various land covers. Compared with SCNN and MVGG-Net, our DSCEN has about a 4% improvement in classification accuracy, which proves that the proposed lightweight DSCEN can learn more effective discriminative feature representation. As for ConvCRF, we can see that the accuracy of woodland and open land is very close to our DSCEN. However, our DSCEN does not consider the CRF-based postprocessing steps, which shows that our algorithm can better suppress the influence of speckles in homogeneous areas. As for SAR-FCN, the OA and AA are less than 4% in comparison with DSCEN, which is due to the limited SAR labeled data that makes the SAR-FCN mode unable to be effectively trained. Also, because the Fusion-Net combines the complementarity of spatial and statistical features, it further improves the classification accuracy. For heterogeneous areas such as residential and texture areas such as woodland and open land, the Fusion-Net has achieved an accuracy improvement of about 2% to 4% compared to the DSCEN that uses a single type of feature classification. In summary, the Fusion-Net shows that the joint consideration of the spatial and statistical features can improve the classification performance of HR SAR images.

Fig. 8 depicts the classification result maps of each compared method on the TerraSAR-X image. As is shown in Fig. 8, the CoTF and SD produced serious misclassifications in the road area. Our statistical features show better recognition ability in open land and woodland areas than the other three traditional methods. This implies that our statistical features can suppress the influence of noise or shadows to a certain extent. Moreover, we can see that these deep learning-based methods can roughly identify all categories. But the classification map of our DSCEN has fewer isolated misclassification points, especially in residential and open land areas. Finally, compared with the ground truth, it can be concluded that the proposed Fusion-Net produces the optimal visual effect appearance.

9) Experiments With Gaofen-3 SAR Image: Classification results of each approach on the Gaofen-3 data are reported in Table X. As can be observed, the testing OA, AA, and kappa of our approach are 92.25%, 93.88%, and 0.8991%, respectively. The proposed Fusion-Net yields the highest classification accuracies than other approaches, which proves the rationality of the joint consideration of spatial and statistical
Fig. 9. Classification maps of Gaofen-3 SAR image with different methods. (a) Original SAR image. (b) Ground truth. (c) CoTF. (d) SD. (e) SCNN. (f) MFFN-CPMN. (g) MVGG-Net. (h) ConvCRF. (i) SAR-FCN. (j) NS-JSM. (k) DSCEN. (l) Fusion-Net.

TABLE X
CLASSIFICATION PERFORMANCE OF GAOFEN-3 SAR IMAGE WITH DIFFERENT METHODS

| Class      | CoTF  | SD    | SCNN  | MFFN-CPMN | MVGG-Net | ConvCRF | SAR-FCN | NS-JSM | DSCEN | Fusion-Net |
|------------|-------|-------|-------|-----------|----------|---------|---------|--------|-------|------------|
| Mountain   | 83.08 | 66.48 | 84.60 | 83.39     | 82.42    | 86.68   | 84.38   | 87.23  | 89.75 | 89.75      |
| Water      | 91.40 | 91.96 | 94.44 | 95.07     | 94.06    | 95.85   | 95.62   | 92.67  | 96.45 | 96.36      |
| Building   | 88.81 | 73.75 | 79.85 | 82.77     | 81.63    | 85.64   | 80.86   | 86.94  | 82.78 | 89.62      |
| Roads      | 93.65 | 86.75 | 92.26 | 94.58     | 95.48    | 95.54   | 95.70   | 94.83  | 96.84 | 98.94      |
| Woodland   | 92.16 | 86.12 | 88.20 | 85.93     | 83.70    | 79.60   | 87.37   | 95.51  | 88.84 | 94.37      |
| Open land  | 81.14 | 74.81 | 88.48 | 88.52     | 91.60    | 94.67   | 91.45   | 89.52  | 94.56 | 94.29      |
| OA         | 87.89 | 78.01 | 86.36 | 86.82     | 86.06    | 88.39   | 87.05   | 89.67  | 89.10 | 92.25      |
| AA         | 88.37 | 79.98 | 87.97 | 88.38     | 88.25    | 89.66   | 89.28   | 91.11  | 90.96 | 93.88      |
| Kappa      | 0.8422| 0.7199| 0.8241| 0.8293    | 0.8201   | 0.8493  | 0.8328  | 0.8661 | 0.8585| 0.8991      |

features. Comparing the CoTF and SD models, we can see that NS-JSM obtains better classification performance, especially in mountains and open land areas. This shows that NS-JSM can suppress speckle by modeling the statistical properties and nonstationary correlation of Gabor wavelet sub-bands of the SAR image. Compared with SCNN, MFFN-CPMN, MVGG-Net, and SAR-FCN, the proposed DSCEN performs better than these deep learning methods by 2% in terms of OA, which indicates that DSCEN can obtain a more discriminative spatial feature representation while maintaining lightweight. Notably, it can be observed that the classification accuracy of NS-JSM on the Gaofen-3 SAR data is even better than other deep learning-based methods, which implies that our NS-JSM can perform reasonably well with less training data. An interesting finding is that methods based on second-order statistics such as CoTF and NS-JSM have achieved more than 92% accuracy in woodland, which proves that second-order statistics can help improve the accuracy of objects with complex textures. Besides, it can be seen from the class-specific accuracy that statistical and spatial feature methods have their advantages in the identification of different objects. By using Fusion-Net to fuse the two types of features, it produces the best classification accuracy, which proves that it is urgently needed to combine statistical and spatial features to process HR SAR image classification tasks.

Fig. 9 shows the classification result maps of each compared method on the Gaofen-3 SAR image. The SD method contains more misclassified points, especially in mountains and building areas. It can be seen that NS-JSM has fewer misclassified pixels in the mountain and open land area than CoTF. The possible reason is that our statistical features also introduce phase information, which is effective for identifying the irregularly textured mountain and open land regions in Gaofen-3 data. In buildings and woodland areas than other deep learning methods. On the contrary, our DSCEN method is better at extracting narrow road features and water and open land categories that contain homogeneous areas. Finally, compared with the ground truth, the proposed Fusion-Net can maintain the minimum noise classifications and has a better visual effect, which verifies the effectiveness of multifeature fusion for improving SAR image classification. Here, we did not further consider the CRF postprocessing steps, and it can
TABLE XI
CLASSIFICATION PERFORMANCE OF AIRBORNE SAR IMAGE WITH DIFFERENT METHODS

| Class      | CoTF | SD  | SCNN | MFFN-CPMN | MVGG-Net | Conv-CRF | SAR-FCN | NS-JSM | DSCEN | Fusion-Net |
|------------|------|-----|------|-----------|----------|----------|---------|--------|-------|------------|
| Open land  | 80.61| 80.59| 83.93| 86.87     | 88.79    | 90.13    | 88.52   | 83.97  | 90.16 | 91.48      |
| Road       | 80.00| 70.38| 80.68| 84.81     | 83.07    | 84.98    | 87.02   | 86.35  | 87.85 | 90.68      |
| Water      | 97.04| 93.48| 96.42| 95.67     | 97.89    | 95.28    | 98.37   | 97.93  | 98.10 | 98.66      |
| Runway     | 96.28| 95.96| 99.02| 98.81     | 97.90    | 98.60    | 98.14   | 99.31  | 98.66 | 99.63      |
| Woodland   | 88.00| 78.44| 85.23| 85.65     | 85.84    | 84.90    | 86.23   | 87.40  | 90.17 | 88.51      |
| Residential| 91.78| 86.16| 86.07| 88.56     | 90.65    | 88.74    | 85.56   | 90.71  | 88.84 | 91.56      |
| Commercial | 98.06| 98.42| 97.94| 98.24     | 98.79    | 99.39    | 99.26   | 98.94  | 98.82 | 99.24      |
| OA         | 83.68| 81.62| 85.27| 87.62     | 89.06    | 89.82    | 88.84   | 86.11  | 90.77 | 91.64      |
| AA         | 90.64| 86.20| 89.90| 91.36     | 91.85    | 91.73    | 91.88   | 92.09  | 93.44 | 94.25      |
| Kappa      | 0.715| 0.6783| 0.7741| 0.7738    | 0.7975   | 0.8094   | 0.7935  | 0.7524 | 0.8273| 0.8417     |

Fig. 10. Classification maps of Airborne SAR image with different methods. (a) Original SAR image. (b) Ground truth. (c) CoTF. (d) SD. (e) SCNN. (f) MFFN-CPMN. (g) MVGG-Net. (h) ConvCRF. (i) SAR-FCN. (j) NS-JSM. (k) DSCEN. (l) Fusion-Net.

TABLE XII
CLASSIFICATION PERFORMANCE OF F-SAR IMAGE WITH DIFFERENT METHODS

| Class      | CoTF | SD  | SCNN | MFFN-CPMN | MVGG-Net | Conv-CRF | SAR-FCN | NS-JSM | DSCEN | Fusion-Net |
|------------|------|-----|------|-----------|----------|----------|---------|--------|-------|------------|
| Water      | 97.71| 94.44| 96.32| 94.72     | 93.82    | 93.83    | 91.92   | 98.00  | 96.33 | 98.20      |
| Residential| 93.73| 90.15| 93.70| 92.41     | 93.34    | 93.71    | 91.02   | 96.22  | 92.24 | 95.40      |
| Vegetation | 94.84| 92.01| 96.08| 92.63     | 95.39    | 96.41    | 95.05   | 94.58  | 97.55 | 97.78      |
| Open land  | 94.13| 95.60| 96.75| 95.06     | 96.91    | 98.18    | 97.46   | 97.62  | 97.99 | 97.80      |
| OA         | 94.54| 92.86| 96.01| 93.32     | 95.59    | 96.59    | 95.26   | 97.25  | 97.07 | 97.53      |
| AA         | 95.11| 93.02| 95.71| 93.70     | 94.87    | 95.53    | 93.71   | 96.27  | 96.03 | 97.30      |
| Kappa      | 90.34| 87.57| 92.97| 88.39     | 92.26    | 93.96    | 91.66   | 91.71  | 94.77 | 95.59      |

10) Experiments With Airborne SAR Image: Classification results of each approach on the Airborne SAR data are shown in Table X. As is shown in Table XI, the OA, AA, and kappa of the Fusion-Net model are 91.64%, 94.25%, and 0.8417%, respectively. It is seen from the compared results that the proposed Fusion-Net achieves the highest classification accuracies. Airborne SAR data contains more categories, but there is a serious class imbalance, where open land occupies most of the pixels in the image. Therefore, it is the most challenging task to classify objects that show a narrow structural appearance and objects that show an extremely complex texture distribution. Compared with CoTF and SD, NS-JSM has more than a 3.5% improvement in the kappa coefficient. This indicates that introducing statistical modeling and CM into the Gabor wavelet sub-bands of the SAR image can obtain more discriminative statistical features. Second, we can be used as a follow-up solution to further improve the spatial labeling consistency.
Fig. 11 shows the classification result maps by using different methods on the F-SAR image. First, it can be observed that SD and MFFN-CPMN have many isolated misclassified pixels in residential and vegetation areas. At the same time, SD did not fully detect the residential scenes in some small areas at the bottom of the image. As for SCNN, MVGG-Net, and SAR-FCN, they have a small number of misclassified pixels in all categories. Due to consideration of spatial label correlation, ConvCRF obtained better label consistency in homogeneous areas such as vegetation and open land. Compared with all other models, the proposed Fusion-Net produces the best visual effects, especially where the categories are adjacent, and the boundaries are clearer. Therefore,
Fusion-Net can greatly improve classification performance for HR SAR images.

V. CONCLUSION

In this article, a novel end-to-end Fusion-Net classification model is proposed for HR SAR images, which aims to embed the statistical features into deep spatial features objects in the end-to-end representation learning. In our model, the proposed DSCEN can extract multiscale spatial features while keeping the fewer model parameters amounts. The NS-JSM can fully mine the statistical properties of the magnitudes and phases of the Gabor wavelet sub-bands of the HR SAR image, and form a more compact and robust statistical descriptor. The proposed Fusion-Net can take full advantage of the complementary information of spatial features and statistical features to make the entire classification model achieve a significant accuracy improvement. Experimental results on four SAR images demonstrate that the proposed Fusion-Net yields much higher accuracies and better visual appearance than other related approaches.

In the future, data augmentation such as Mixup [62] or self-supervised learning such as contrastive learning [63] will be considered to enhance the feature representation ability of DSCEN. Besides, instead of using the empirical statistical models, we intend to consider adaptively modeling high-order scattering statistics by deep learning to increase the classification capabilities of statistical models. Another future research direction is to extend our Fusion-Net for multipolarized HR SAR data, which can learn complementary ground information provided by different polarization channels to further improve the classification accuracy.

REFERENCES

[1] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papanastassiou, “A tutorial on synthetic aperture radar,” IEEE Geosci. Remote Sens. Mag., vol. 1, no. 1, pp. 6–43, Apr. 2013.

[2] C. Rossi and E. Ertan, “Paddy-rice monitoring using TanDEM-X,” IEEE Trans. Geosci. Remote Sens., vol. 53, no. 2, pp. 900–910, Feb. 2015.

[3] M. Satake, T. Kobayashi, J. Uemoto, T. Umehara, and S. Kojima, “Damage estimation of the great East Japan earthquake by NICT airborne SAR (PI-SAR2),” Int. Soc. Opt. Eng., vol. 53, no. 60, pp. 1190–1191, 2012.

[4] H. Breit, T. Fritz, U. Balss, M. Lachaise, A. Niedermeier, and M. Vonavka, “TerrasAR-X SAR processing and products,” IEEE Trans. Geosci. Remote Sens., vol. 48, no. 2, pp. 727–740, Feb. 2010.

[5] G. Xingfa and T. Xudong, “Overview of China Earth observation satellite programs [space agencies],” IEEE Geosci. Remote Sens. Mag., vol. 3, no. 3, pp. 113–129, Sep. 2015.

[6] T. Esch, A. Schenk, T. Ullmann, M. Thiel, A. Roth, and S. Dech, “Characterization of land cover types in TerraSAR-X images by combined analysis of speckle statistics and intensity information,” IEEE Trans. Geosci. Remote Sens., vol. 49, no. 6, pp. 1911–1925, Jun. 2011.

[7] C. O. Dumitru and M. Datcu, “Information content of very high resolution SAR images: Study of feature extraction and imaging parameters,” IEEE Trans. Geosci. Remote Sens., vol. 51, no. 8, pp. 4591–4610, Aug. 2013.

[8] S. Song, B. Xu, and J. Yang, “SAR target recognition via supervised discriminative dictionary learning and sparse representation of the SAR-HOG feature,” Remote Sens., vol. 8, no. 8, p. 683, 2016.

[9] D. Guan, D. Xiang, X. Tang, L. Wang, and G. Kuang, “Covariance of textural features: A new feature descriptor for SAR image classification,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 12, no. 10, pp. 3932–3942, Oct. 2019.

[10] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[11] J. Geng, H. Wang, J. Fan, and X. Ma, “Deep supervised and contractive neural network for SAR image classification,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 4, pp. 2442–2459, Apr. 2017.

[12] S. Chen, H. Wang, F. Xu, and Y. Q. Jin, “Target classification using the deep convolutional networks for SAR images,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 8, pp. 4806–4817, Aug. 2016.

[13] Z. 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, p. 271, Jan. 2021.

[14] Z. Sun, J. Li, P. Liu, W. Cao, T. Yu, and X. Gu, “SAR image classification using greedy hierarchical learning with unsupervised stacked CAEs,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 7, pp. 5721–5739, Jul. 2021.

[15] J. Zhang, M. Xing, and X. Xie, “FEC: A feature fusion framework for SAR target recognition based on electromagnetic scattering features and deep CNN features,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 3, pp. 2174–2187, Mar. 2021.

[16] X. Liu, C. He, Q. Zhang, and M. Liao, “Statistical convolutional neural network for land-cover classification from SAR images,” IEEE Geosci. Remote Sens. Lett., vol. 17, no. 9, pp. 1548–1552, Sep. 2020.

[17] W. Liang, Y. Wu, M. Li, Y. Cao, and X. Hu, “High-resolution SAR image classification using multi-scale deep feature fusion and covariance pooling manifold network,” Remote Sens., vol.13, no. 328, 2020.

[18] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Meijer, “A new fully connally convolutional neural network for semantic segmentation of polarimetric sar imagery in complex land cover ecosystem,” ISPRS J. Photogramm. Remote Sens., vol. 151, pp. 223–236, May 2019.

[19] H. C. Li, W. Hong, Y. R. Wu, and P. Z. Fan, “On the empirical-statistical modeling of SAR images with generalized Gamma distribution,” IEEE J. Sel. Topics Signal Process., vol. 5, no. 3, pp. 386–397, Jun. 2011.

[20] F. Argenti, A. Lapini, L. Alparone, and T. Bianchi, “A tutorial on speckle reduction in synthetic aperture radar images,” IEEE Geosci. Remote Sens. Mag., vol. 1, no. 3, pp. 6–35, Apr. 2013.

[21] J. Martín-de-Nicolás, M.-P. Jarabo-Amores, D. Mata-Moya, N. del-Rey-Maestre, and J.-L. Bárcea-Humanes, “Statistical analysis of SAR sea clutter for classification purposes,” Remote Sens., vol. 6, no. 10, pp. 9379–9411, Sep. 2014.

[22] J. Hao, “Amplitude distribution of composite terrain radar clutter and the K-distribution,” IEEE Trans. Antennas Propag., vol. AP-32, no. 10, pp. 1049–1062, Oct. 1984.

[23] G. V. Trunk and S. F. George, “Detection of targets in non-Gaussian sea clutter,” IEEE Trans. Aerosp. Electron. Syst., vol. AES-6, no. 5, pp. 620–628, Sep. 1970.

[24] M. Sekine and Y. Mao, Weibull Radar Clutter. London, U.K.: IEEE Press, 1990.

[25] L. Bombrun and J. M. Beaulieu, Fisher distribution for texture modeling of polarimetric SAR data,” IEEE Geosci. Remote Sens. Lett., vol. 5, no. 3, pp. 512–516, Jul. 2008.

[26] J. M. Nicolas and F. Tupin, “Gamma mixture modeled with ‘second kind statistics’ : Application to SAR image processing,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., vol. 4, Jun. 2002, pp. 2489–2491.

[27] H.-C. Li, V. A. Krylov, P.-Z. Fan, J. Jerubia, and W. J. Emery, “Unsupervised learning of generalized Gamma mixture model with application in statistical modeling of high-resolution SAR images,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 4, pp. 2153–2170, Apr. 2016.

[28] X. Zhou, R. Peng, and C. Wang, “A two-component K-lognormal mixture model and its parameter estimation method,” IEEE Trans. Geosci. Remote Sens., vol. 53, no. 5, pp. 2640–2651, May 2015.

[29] W. Song, M. Li, P. Zhang, Y. Wu, X. Tan, and L. An, “Mixture WG I-MRF model for PolSAR image classification,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 2, pp. 905–920, Feb. 2018.

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

[31] P. Zhang, M. Li, Y. Wu, M. Liu, F. Wang, and L. Gan, “SAR image multiclass segmentation using a multiscale TMF model in wavelet domain,” IEEE Geosci. Remote Sens. Lett., vol. 9, no. 6, pp. 1099–1103, Nov. 2012.

[32] M. Golparvar, M. S. Helfroush, and H. Danyali, “Nonsubsampled contourlet transform-based conditional random field for SAR images segmentation,” Signal Process., vol. 174, Sep. 2020, Art. no. 107623.

[33] A. Karine, A. Toumi, A. Khenchaf, and M. E. Hassouni, “Target recognition in radar images using weighted statistical dictionary-based sparse representation,” IEEE Geosci. Remote Sens. Lett., vol. 14, no. 12, pp. 2403–2407, Dec. 2017.
J. S. Bridle, “Training stochastic model recognition algorithms as
J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face
C. He, M. Tu, D. Xiong, and M. Liao, “Nonlinear manifold learning
A. Karine, A. Toumi, A. Khenchaf, and M. E. Hassouni, “Multivariate
T. Zhao and X. Wu, “Pyramid feature attention network for saliency
K. Simonyan and A. Zisserman, “Very deep convolutional networks for
K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for
S. Woo, J. Park, J. Lee, and I. Kweon, “CBAM: Convolutional block
A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier nonlinearities
G. Huang, S. Liu, D. Xiong, and M. Liao, “Nonlinear manifold learning
J. Ai, R. Tian, Q. Luo, J. Jin, and B. Tang, “Multi-scale rotation-invariant
K. He, M. Tu, D. Xiong, and M. Liao, “Nonlinear manifold learning integrated with fully convolutional networks for PolSAR image classification,” Remote Sens., vol. 12, no. 4, p. 655, Feb. 2020.
J. Ai, R. Tian, Q. Luo, J. Jin, and B. Tang, “Multi-scale rotation-invariant Haar-like feature integrated CNN-based ship detection algorithm of multiple-target environment in SAR imagery,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 12, pp. 10070–10087, Dec. 2019.
K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Represent., 2014, pp. 1–14.
K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
H. Dong, X. Xu, H.-G. Sui, and J.-Y. Liu, “Copula-based joint statistical model for polarimetric features and its application in PolSAR image classification,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 10, pp. 5777–5789, Oct. 2017.
TerraSAR-X Dataset. Accessed: Oct. 2018. [Online]. Available: https://www.intelligence-airbusds.com/en/5750-image-gallery-results?search=gallery&keyword=TerraSAR
F-SAR Dataset. Accessed: Oct. 2007. [Online]. Available: https://www.dlr.de/hr/en/desktopdefault.aspx/tabid-4698/7782_read-12248/