

# C-CNN: Contourlet Convolutional Neural Networks

Mengkun Liu<sup>1b</sup>, Licheng Jiao<sup>1b</sup>, *Fellow, IEEE*, Xu Liu<sup>1b</sup>, *Member, IEEE*, Lingling Li, *Member, IEEE*, Fang Liu<sup>1b</sup>, *Senior Member, IEEE*, and Shuyuan Yang<sup>1b</sup>, *Senior Member, IEEE*

**Abstract**—Extracting effective features is always a challenging problem for texture classification because of the uncertainty of scales and the clutter of textural patterns. For texture classification, spectral analysis is traditionally employed in the frequency domain. Recent studies have shown the potential of convolutional neural networks (CNNs) when dealing with the texture classification task in the spatial domain. In this article, we try combining both approaches in different domains for more abundant information and proposed a novel network architecture named contourlet CNN (C-CNN). The network aims to learn sparse and effective feature representations for images. First, the contourlet transform is applied to get the spectral features from an image. Second, the spatial-spectral feature fusion strategy is designed to incorporate the spectral features into CNN architecture. Third, the statistical features are integrated into the network by the statistical feature fusion. Finally, the results are obtained by classifying the fusion features. We also investigated the behavior of the parameters in contourlet decomposition. Experiments on the widely used three texture data sets (kth-tips2-b, DTD, and CURET) and five remote sensing data sets (UCM, WHU-RS, AID, RSSCN7, and NWPU-RESISC45) demonstrate that the proposed approach outperforms several well-known classification methods in terms of classification accuracy with fewer trainable parameters.

**Index Terms**—Contourlet transform, multiscale, remote sensing scene classification, texture classification.

## I. INTRODUCTION

TEXTURE is defined as smoothness or roughness of the surface of an object, which contains important information about the structural arrangement of surfaces and their relationships to the surrounding environment [1], such as the

grain of the wood and the pattern of fabric with repeated structure. The texture is a key visual cue for various image analysis applications, such as image segmentation, image classification, and texture synthesis. It plays a significant role in biomedical image analysis, remote sensing, object recognition [2], and so on. Texture classification, a process of assigning an unknown texture to a known set of texture class, has been an active research topic due to challenges about the uncertainty of scales, lighting conditions, and the clutter of textural patterns; even textures within the same class can vary greatly.

Traditional texture classification generally consists of two main stages: feature extraction and classification. With the pioneering work of Julesz [3], early texture feature extraction methods are based on statistics [1], [4], [5], geometry [6], [7], structure [8]–[12], model [13], [14], transform in signal processing [15]–[18], and the multiresolution analysis [19]–[30]. In the classification stage, the extracted features are then classified by various classifiers, such as SVM [31]–[34], learning vector quantization (LVQ) [35], Bayes [36], and K-nearest-neighbor (K-NN) classifier [37]–[39].

### A. Texture Feature Extraction

1) *Statistics-Based Methods*: The statistics-based methods provide measures of distribution properties of pixels intuitively, such as gray-level co-occurrence matrices (GLCMs) [1], which represent co-occurrences of the pixel intensities at given angles and distances. The gray-tone spatial-dependence matrices (GTSDMs), based on statistics that summarize the relative frequency distribution [1], provide an efficient texture classification with two decision rules: a piecewise linear decision rule and a min-max decision rule. The similar work in [4] aimed to measure the statistical properties of pixel pairs at several distances with the statistical feature matrix; the major property that differs from [1] is the matrix size depends on the maximum distance rather than the number of gray levels. The features such as the Fourier power spectrum, first-order statistics of gray-level differences, and second-order gray-level statistics are extracted as texture representation [5].

2) *Geometry-Based Methods*: The geometry approaches represent textures with the properties of geometrical constraints. Blostein and Ahuja [6] integrated the texture elements with the corresponding scene layout, which identifies the texture elements and recovers the orientation of texture simultaneously by constructing a multiscale region detector in Laplacian-of-Gaussian scale space. They overcame the challenges in practice that the texture is occluded partially or in fine-scale. Stevens [7] proposed the valid constraints to compute surface orientation and distance via a visual analysis

Manuscript received October 8, 2019; revised April 2, 2020; accepted June 28, 2020. Date of publication July 21, 2020; date of current version June 2, 2021. This work was supported in part by the State Key Program of National Natural Science of China under Grant 61836009, in part by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Project 61621005, in part by the Major Research Plan of the National Natural Science Foundation of China under Grant 91438201, Grant 91438103, and Grant 61801124, in part by the National Natural Science Foundation of China under Grant U1701267, Grant 61871310, Grant 61573267, and Grant 61906150, in part by the Fund for Foreign Scholars in University Research and Teaching Program's 111 Project under Grant B07048, in part by the Program for Cheung Kong Scholars and Innovative Research Team in University under Grant IRT 15R53, in part by the S&T Innovation Project from the Chinese Ministry of Education, in part by the National Science Basic Research Plan in Shaanxi Province of China under Grant 2019JQ-659, and in part by the China Postdoctoral Fund under Grant 2019M663641. (Corresponding author: Licheng Jiao.)

The authors are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an 710071, China, and also with the Joint International Research Laboratory of Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: lchjiao@mail.xidian.edu.cn; mengkunliu31@163.com).

Color versions of one or more of the figures in this article are available online at <https://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TNNLS.2020.3007412

2162-237X © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.

See <https://www.ieee.org/publications/rights/index.html> for more information.

that uses the surface shape and contour of the texture. It can be decomposed into two modules: the contour generators reconstruction and the surface relation determination, which separates the problem of the projective geometry from the intrinsic geometry.

3) *Structure-Based Methods*: The structure-based methods find the local or global properties of textures with a specified descriptor. Traditional texture descriptors, such as local binary patterns (LBPs) [8] and its extensions [10]–[12], are used for capturing the main properties of the texture. The LBP describes the texture as a two-level version of the texture spectrum [9]. The complementary information for effective texture representation is provided by a simple contrast measure [10], but it is still a rotation variant. Later, Ojala [11] improved the LBP further with a multiresolution gray-scale and rotation invariant. The completed robust LBP (CRLBP) has been developed for improving the robustness to noise and illumination variants [12]. The volume LBP (VLBP) is proposed as a dynamic texture descriptor in the temporal domain, combining motion and appearance. The completed LBP (CLBP) [37] can effectively represent the missing information in the LBP with center pixel and a local difference sign-magnitude transform. These methods describe the local or global properties of textures, but it is difficult to study the heredity or dependence of pixels between texture scales.

4) *Model-Based Methods*: The model-based methods describe texture with estimated parameters of the model. It is under the hypothesis that texture is formed on a distributed model with the interaction of correlated random variables in the spatial domain. Common model-based approaches include the Gaussian Markov random field model [13], simultaneous autoregressive model [14], random field model [40], [41], fractal model [42], and so on, whereas the computation burden is significantly increased when the model is complex.

5) *Signal Processing-Based Methods*: The traditional feature extraction methods in signal processing also achieved satisfactory results. Super and Bovik [15] measured the local values of the fractal dimension by multiple Gabor filters. The spatial frequency filter and the orientation channel filter are introduced to the texture analysis [16]. The multiscale Hurst parameters are used as features for texture classification, which is comparable with the Gabor features [17]. The rotation-invariant texture features are generated via the circular neighborhoods and 1-D discrete Fourier transforms [18].

6) *Multiresolution Analysis-Based Methods*: Instead of describing the texture features in the former case, the multiresolution analysis-based approaches describe the energy distribution in the spectral domain. It can efficiently capture the intrinsic geometrical structures of texture. In addition, more precise texture information is obtained by compact energies. The multiscale transforms mainly include the wavelet transform [19], [31], [34], [39], the shearlet transform [20]–[22], the ridgelet transform [23], the curvelet transform [24], the contourlet transform [25]–[27], and the brushlet transform [28]–[30], and the multiresolution and orientated representations of these transforms are suitable to describe texture with repeated structure.

## B. Texture Classification

Traditional methods performs well in classification, such as SVM [31]–[34], [43] and K-NN classifier [37]–[39]. However, the extracted handcrafted features of the involved algorithms rely heavily on the experience, and the classifier is separated from feature extraction in two stages. In recent years, CNN has achieved remarkable results in many fields as an end-to-end method [44]–[54].

With the recent studies on deep learning [55], [56], convolutional neural networks (CNNs), a biologically inspired multistage structure, has enjoyed significant performance on some tasks, such as image classification [57], semantic segmentation [58], [59], and object detection [60]. It utilizes the features extracted by networks directly rather than carefully designed. The methods based on CNN are considerably better than the previous state-of-the-art methods.

CNN is first introduced in texture classification where feature extraction and classification are processed in the same four-layer network [61]. Later, CNN is applied to lung image classification [44] and forest species recognition [45] in 2014 since both tasks can be viewed as the texture classification problems. However, these approaches are not specifically tailored for textures.

For texture classification, CNNs cannot be handled very well. The reason is that the repeated patterns have lower complexity in texture, while CNN is appropriate to extract high-level semantic features. Besides, parameters and computational complexity are greatly increased. Under such circumstances, the improved texture classification on CNN is specifically addressed in an end-to-end fashion. In [46]–[49], network architectures are modified by considering texture characteristics. FV-CNN [46], a valid texture descriptor, is obtained by the Fisher vector pooling of a CNN filter bank. Texture CNN (T-CNN) is developed [47] on AlexNet with an energy measure, namely, a new energy layer is derived from the last convolution layer by averaging the output in each feature map. It is similar to an energy response of a filter bank. The approach improved classification accuracy slightly while reduced complexity with fewer parameters. Wavelet CNN [48], [49] is proposed to integrate the Haar wavelet decomposition as a spectral analysis into CNN. Based on the insight that the pooling and the convolution can be thought of as a limited form of spectral analysis, thus, it captures both types of features well under a single model. Wavelet CNN [49] uses projection shortcuts [50] to maximize the retention of information flowing over the network, and dense connection [51] is applied for adapting changes to dimensions. Similar to the energy layer in T-CNN, global average pooling (GAP) [52] is used rather than a fully connected layer in order to prevent overfitting. Wavelet CNN achieves better accuracy with fewer parameters than CNNs but unable to outperform Fisher vector-CNN (FV-CNN) [46].

From the abovementioned CNN-based methods for texture classification, three main challenges can be summarized as follows.

- 1) Texture images usually include the local or global properties of uncertainty scales and various patterns. Feature

representation within a fixed scale may not be robust and discriminative. Thus, it is vital to study the heredity or dependence of pixels between texture scales.

- 2) In addition to spatial information, texture image contains rich spectral information. Therefore, how to mine and make full use of both the information will be very helpful to the effective representation and classification.
- 3) High-level semantic feature extracted by CNN is not appropriate for texture representation since the repeated patterns in texture have lower complexity. Moreover, the computation burden is significantly increased when the model is complex.

### C. Motivation

The image classification is to identify which class a given sample belongs to. The CNN-based classifier is employed in the proposed method. Given an input signal  $x(u)$ ,  $u \in \mathbb{R}^n$ , with  $n = 2$  for images, a general convolutional network function in the specific layer  $x_j$  is defined as

$$\begin{aligned} x_j &= \rho W_j x_{j-1} \\ &= \rho \sum_k w_{j,k_j}(k) * x_{j-1}(k) \end{aligned} \quad (1)$$

where  $j(0 < j \leq J)$  is the network depth, and  $W_j$  is a linear operator which includes the bias for more concise expression.  $\rho$  denotes the nonlinearity transforms, such as sigmoid or ReLu.  $k_j$  represents a channel index. The symbol  $*$  is the convolution operator.

For most classification architectures,  $W_j$  is covariant to translations. Data augmentation and deep architecture are common tricks to provide geometric transformations. The pooling operation can keep consistent expression to small translation. Nevertheless, CNNs lack the ability of geometric transformation to some extent.

In order to solve the abovementioned problems, the contourlet CNNs (C-CNNs) are proposed in this article, which is inspired by [49] and employs wavelet as a spectral analysis into CNN. The contourlet transform is a kind of multiscale geometric analysis tool and has the substantial advantages of locality and directionality. It can be introduced into neural networks to enhance the ability of geometric transformations. Meanwhile, it avoids the network falling into the local optimum too early. The introduction of multiscale geometric filter banks increases the interpretability and controllability of the network, which enhances the robustness of the scale and directionality.

Compared with the wavelet transform, the contourlet transform offers better directionality, anisotropy, spatial locality, and the bandpass property, which is consistent with the main characteristics of the human visual system (HVS) (see Section II-A). It can effectively capture the geometry of contours with the edge singularity features, whereas wavelet captures point singularity features [53]. Thus, the contourlet transform has a more sparse representation. The details concerning the sparsity comparison are shown in Section II-A. In addition, the wavelet transform has limited directionality, which only captures horizontal, vertical, and diagonal details.

However, the contourlet transform is a new extension to the wavelet transform and provides more abundant directional information of texture. Therefore, the C-CNN based on the contourlet transform [54], [62] and CNN provides multidirectional and multiscale information effectively. The proposed method mainly focuses on three aspects.

- 1) *Obtaining Spectral Features From Different Scales:* We define layers to perform multiresolution analysis by the contourlet transform in the CNN framework, which allows forward propagation to learn more effective texture features. The role of prior knowledge and higher quality information are dominant in effective texture representation with multiscale and multidirection from the contourlet transform. The information in both spatial and spectral domains can improve the approximation ability of the network, which is conducive to escaping from saddle point in network optimization.
- 2) *Enhancing Features Representation:* In order to use the spectral features adequately and represent features effectively, the multifeature fusion strategies are used in the framework.
- 3) *Lightweight Structure:* Due to the tight frame with the sparse factor of contourlet and simple backbone network, C-CNN achieves comparable accuracy with few trainable parameters. Besides, some weights in the contourlet kernel can be interpreted to make the network from the black box to the gray box.

Fig. 2 shows an overview of the C-CNN. It combines the spectral analysis by the contourlet transform with CNN. To summarize, the main contributions of this article are as follows.

- 1) A novel network architecture, named C-CNN, is proposed, which combines the information in the spectral domain and the spatial domain.
- 2) The contourlet transform is integrated into CNN in order to mine features in the spectral domain and make the network more compact.
- 3) Multifeature fusion strategies are designed to obtain better features, which includes spatial-spectral feature fusion (SSFF) and statistical feature fusion (SSF).
- 4) We analyzed the sparsity, theoretically, of contourlet and the parameters of contourlet decomposition in C-CNN.

The remainder of this article is organized as follows. Section II describes the proposed CNN-based network. Experiments and results discussion is presented in Section III. Section IV summarizes our work.

## II. CONTOURLET CONVOLUTIONAL NEURAL NETWORKS

For texture classification, it is vital for effective texture representation against the challenges that the uncertainty of scales and the clutter of textural patterns. Section II-A explains why the contourlet transform is a sparse representation rather than a wavelet. Section II-B describes the contourlet transform combined with CNN and feature representations in a multiscale and multidirection way. The implementation details of the proposed method are given in Section II-C.



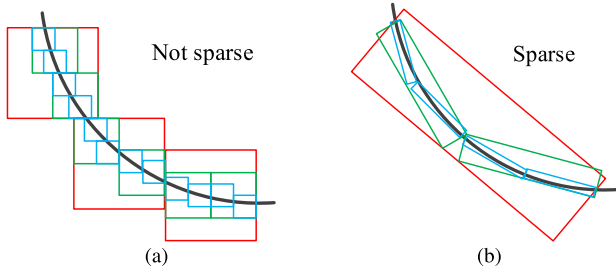


Fig. 1. Comparison of the contour representations with (a) wavelet and (b) contourlet. (a) Wavelet has square supports that can only capture point discontinuities. (b) Contourlet has elongated supports that can capture linear segments of contours and, thus, can effectively represent a smooth contour with fewer coefficients.

### A. Sparsity of Contourlet

The receptive fields of the visual cortex in mammals can be characterized as being localized, oriented, and bandpass [55], which are similar to the basis functions of contourlet transforms. The characters of the receptive fields enable the HVS to “capture” key information in natural scenes with the least number of visual neurons, comparable to the most sparse representation of natural scenes [63].

According to the research results of physiologists on the HVS and the statistical model of natural images, the “optimal” image representation method should have the following characteristics [64].

- 1) *Multiresolution*: The image can be approximated continuously from coarse to fine resolution, which is equivalent to “bandpass.”
- 2) *Locality*: The representation basis should be local in both spatial and spectral domains.
- 3) *Directionality*: The basis should be directional and not limited to three directions of the separable 2-D wavelets.

Fig. 1 shows the representations of approximating curves with wavelet and contourlet. The wavelet approximate curves by point, and the basis of wavelet has square support intervals of different sizes at different resolutions. The number of nonzero wavelet coefficients increases exponentially when the scale becomes finer, and a large number of nonnegligible coefficients appeared; thus, the curve cannot be “sparsely” expressed.

As one of the signal representation tools, contourlet is similar to the edge in object natural image. It can make full use of geometric regularity to obtain the sparse representation. The contourlet transform, an extension of the wavelet transform, can describe a smooth contour as line segments with much fewer coefficients [see Fig. 1(b)]. However, there are largely inefficient decompositions as points in the wavelet transform [see Fig. 1(a)]. Thus, the representation of the image with the contourlet transform is sparser since contours are basic units of image. The sparsity of contourlet is not limited to visual quality [see Fig. 1] but also in terms of the approximation rate. Next, we sketch briefly the approximation error and refer to [65] for details.

Given a 2-D piecewise smooth function  $f$ , which is  $C^2$  (twice continuously differentiable) away from  $C^2$  discontinuity

curves, a general series expansion of  $f$  can be defined by

$$f = \sum_{i=1}^{\infty} \alpha_i \phi_i \quad (2)$$

with a basis  $\{\phi_i\}_{i=1}^{\infty}$ , such as Fourier or wavelet. The efficiency of an expansion can be measured by the nonlinear approximation (NLA) [66], which is defined as

$$\hat{f}_M^B = \sum_{i \in I_M} \alpha_i \phi_i \quad (3)$$

where  $\hat{f}_M^B$  is the best  $M$ -term approximation of function  $f$  with the basis  $B$ , and  $M$  is the number of the most significant coefficients. For the  $M$ -largest  $|\alpha_i|$ ,  $I_M$  is the indexes of the  $M$ -term coefficients.  $\hat{f}_M^B$  reflects the sparsity of the expansion by  $\{\phi_i\}_{i=1}^{\infty}$ . Thus, the best  $M$ -term approximation error can be written as  $\|f - \hat{f}_M^B\|_{L_2}^2$  in  $L_2$ -norm square. The  $M$ -term approximation error by the contourlet frame satisfies

$$\|f - \hat{f}_M^{(\text{contourlet})}\|_{L_2}^2 \leq C(\log M)^3 M^{-2} \quad (4)$$

where  $C$  is a constant, the  $M$ -term approximation error of curvelet [67] is similar to contourlet, in order to compare the approximation performance of different basis concisely, the notation

$$A \asymp B \quad (5)$$

is introduced to mean that the two expressions are equivalent within multiplicative constants. Namely, there are constants  $C_1, C_2 > 0$  that make  $C_1 A \leq B \leq C_2 A$  valid. The squared error of the  $M$ -term expansion with different basis satisfies

$$\|f - \hat{f}_M^{(\text{fourier})}\|_{L_2}^2 \asymp M^{-1/2} \quad (6)$$

$$\|f - \hat{f}_M^{(\text{wavelet})}\|_{L_2}^2 \asymp M^{-1} \quad (7)$$

$$\|f - \hat{f}_M^{(\text{contourlet})}\|_{L_2}^2 \asymp M^{-2}. \quad (8)$$

Equations (6) and (7) show that the decay rates of Fourier and wavelet are  $O(M^{-1/2})$  [66] and  $O(M^{-1})$  [53], [68], respectively. Contourlet has the decay rate of  $O(M^{-2})$  in (8). The optimal approximation rate is  $M^{-2}$  according to [66]. In this case, the contourlet expansion can be seen as the optimal approximation rate. It compacts the signal  $f$  into a few coefficients and achieves the optimal sparse representation.

Moreover, the properties of directionality, localization, multiresolution, and anisotropy that arise in contourlet proved strictly in [65]. Contourlet has elongated support intervals at various scales and directions with more sparse representation. Different directional contours in texture images can be described owing to an abundant directional selectivity. Therefore, contourlet should be a better choice than a wavelet in feature representation of texture images.

### B. Contourlet Convolutional Neural Networks Model

As a geometrical multiresolution analysis, the contourlet transform can efficiently represent texture image with the basis of contours. It consists of an arbitrary number of directional bands at each level. The contourlet transform is formed by a tight frame with only a small redundancy factor. Therefore, we integrated the contourlet transform into CNN named

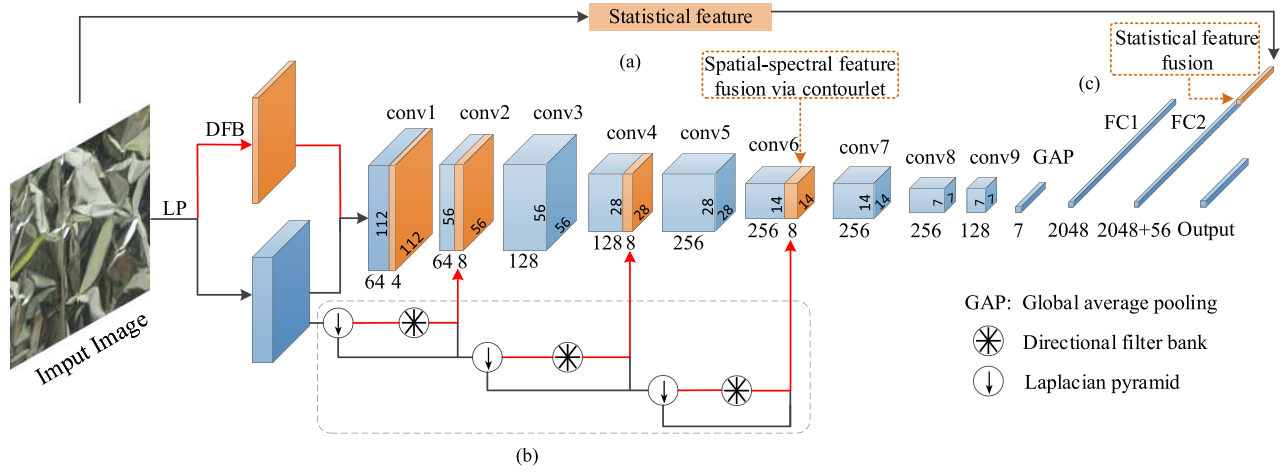


Fig. 2. C-CNN architecture. It includes three parts as follows. Backbone: mainly includes nine convolution layers, one GAP layer followed by three fully connected layers. Multiscale decomposition (in dotted box): the red line represents the flow direction of the high-frequency components that are concatenated to corresponding feature maps on CNN with SSFF. Statistical feature extraction: the mean and variance are calculated in each contourlet subband coefficient of the input image and connected with FC2 to enhance feature representation and rotation invariant.

C-CNN [see Fig. 2], which can enhance the ability of feature representation and improve the approximation ability of the network. Thus, the C-CNN provides a fast convergence speed with sparse and effective feature representation, and the better local optima are obtained.

In general, convolution layers in CNN can be denoted as  $Y = (X * \mathbf{w})$ , where  $X$  and  $Y$  are the input and output of the convolution layer, respectively.  $\mathbf{w}$  represents the weights including the bias. The symbol  $*$  is the convolution operator. Average pooling can be expressed as convolution by means of average filter and corresponding downsampling. Fujieda [49] considered CNN as parts of multiresolution analysis and represents convolution and pooling with a unified equation

$$Y = (X * \mathbf{k}) \downarrow p$$

$$\mathbf{k} = \begin{cases} \mathbf{w} \text{ with } p = 1, & \text{convolution} \\ \mathbf{p} \text{ with } p > 1, & \text{pooling} \\ \mathbf{w} * \mathbf{p} \text{ with } p > 1, & \text{convolution + pooling} \end{cases} \quad (9)$$

where  $\mathbf{p}$  represents the averaging filter in pooling layers, and  $p$  is the stride of downsampling. Based on (7) in [49], given an input  $X$ ,  $X_{l,i}$  represents the low-pass component in level  $i$ ; thus,  $X_{l,0} = X$ . The output of the contourlet decomposition in level  $i$  can be described by the following equations:

$$X_{l,i+1}, X_{h,i+1} = (X_{l,i} * F_{LP}) \downarrow p$$

$$X_{h\_bds,i+1} = (X_{h,i+1} * F_{DFB}) \quad (10)$$

where  $F_{LP}$  and  $F_{DFB}$  are the Laplacian pyramid (LP) and directional filter bank (DFB), respectively.  $p$  denotes the interlaced downsampling factor. The symbol  $\downarrow$  is downsampling operator that resembles the pooling operation in CNN. The subscripts  $l$  and  $h$  of  $X$  represent the low- and high-pass components, respectively. The subscript  $h\_bds$  denotes the bandpass directional subbands. CNN ignores the high-pass components. Therefore, the spectral information is integrated into CNN, as shown in Fig. 2, where additional spectral information is highlighted and generated by the contourlet transform.

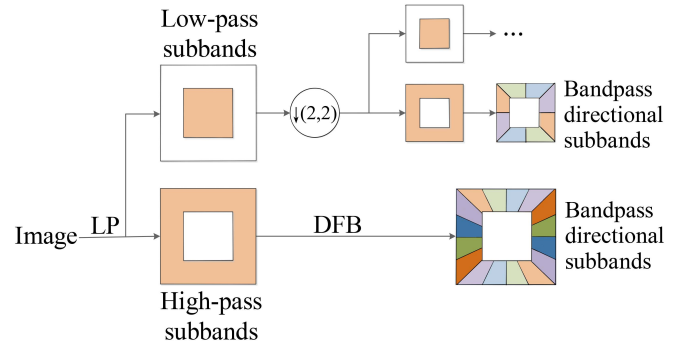


Fig. 3. PDFB of contourlet decomposition. It is achieved by LP and DFB for low-pass subbands iteratively. The input image is first decomposed by an LP filter to low- and high-pass subbands; then, the high-pass subbands are decomposed into  $2^i$  directional subspaces through the DFB.

The well-trained filters in CNN with different orientations can capture edges in different directions in the spatial domain. Similarly, traditional filters, such as contourlet filters, have natural advantages for edge representation with different scales and directions in the spectral domain, which is neglected in CNN.

Fig. 3 shows the contourlet transform decomposed independently with multiscale and multidirection by iteratively adopting pyramidal DFB (PDFB) on the low-pass image. For example, the dotted box in Fig. 2 implements the three-level contourlet decomposition by using PDFB iteratively. The PDFB is a cascade of a LP [69] and a directional filter bank (DFB) [70]. The input image is first decomposed by LP filter to low- and high-pass subbands. The low-pass subbands are generated with 2-D low-pass analysis filtering and interlaced downsampling. The high-pass subbands are obtained by subtracting the low-pass component from the original image, where the low-pass component is the same size as the original image after applying upsampling and low-pass synthesis filtering. High-pass subbands

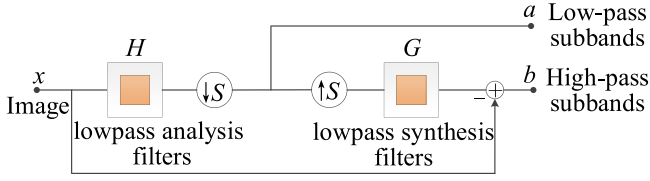


Fig. 4. One level of LP decomposition. First, the original image  $x$  is used to generate downsampled low-pass subbands  $a$  through a low-pass analysis filters  $H$  and a sampling matrix  $S$ . The high-pass subbands  $b$  are then computed as the difference between the original  $x$  and the prediction with a sampling matrix  $S$  followed by a low-pass synthesis filters  $G$ .

are decomposed into  $2^l$  directional subspaces through the directional filters ( $l$  is an arbitrary positive integer); then, multiresolution and multidirection decomposition is achieved by performing the abovementioned steps on low-pass subbands iteratively. Thus, abundant spectral information is obtained via the contourlet transform from the original image.

The details of LP and DFB are shown in the following.

1) *Laplacian Pyramid*: The LP is proposed to obtain multiscale decomposition by Burt and Adelson [69]. The LP decomposition at each level produces a low-pass component downsampled from the original signal and the residual signal between the original and the prediction, generating a bandpass component. Fig. 4 shows the decomposition process.

Where  $H$  and  $G$  denote the low-pass analysis and synthesis filters, respectively.  $S$  is called the sampling matrix. The decomposition can be iterated on the low-pass signal obtained by downsampling. Sampling matrices are used in multidimensional filter banks. For instance, given the input signal  $x$ , the downsampling version of signal  $x$  is obtained by an integer sampling matrix  $S$ , resulting in  $Sx$ , and the output  $a$  is a coarse approximation.  $b$  is a difference between the original signal and the prediction. In addition, the effect that “scrambled” frequencies generated in the wavelet filter banks is avoided by only downsampling the low-pass channels.

2) *Directional Filter Bank*: Bamberger and Smith [70] introduced an oriented 2-D filter bank to reconstruct the original signal with a minimum sample representation. The DFB is generated by  $l$ -level binary tree decomposition in the 2-D frequency domain, resulting in  $2^l$  wedge-shaped subbands, as shown in Fig. 5. The frequency domain is divided into  $2^3 = 8$  directional subbands where  $l = 3$ , of which subbands 0–3 and 4–7 correspond to the vertical and horizontal details, respectively.

Do and Vetterli [71] employed a  $2^l$  parallel channel filter bank to represent a  $l$ -level binary tree with multirate identities. The parallel channel filter bank constructed from equivalent filters and sampling matrices. The equivalent filters are denoted by  $G_k^{(l)}$ , and  $k$  is the subband’s index, as shown in Fig. 5 where  $0 \leq k \leq 2^l$ . Let  $\text{diag}(a_1, \dots, a_n)$  represent a diagonal matrix and  $a_n$  be the  $i$ th diagonal element. The corresponding sampling matrices are formed by

$$S_k^{(l)} = \begin{cases} \text{diag}(2^{l-1}, 2), & 0 \leq k \leq 2^{l-1} \\ \text{diag}(2, 2^{l-1}), & 2^{l-1} \leq k \leq 2^l. \end{cases} \quad (11)$$

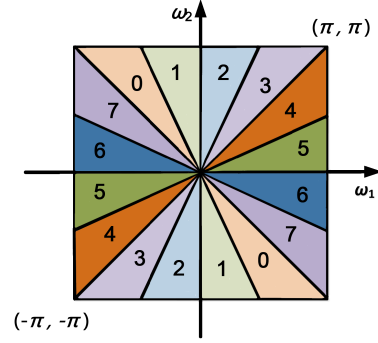


Fig. 5. Frequency partitioning with eight-direction wedge-shaped subbands.

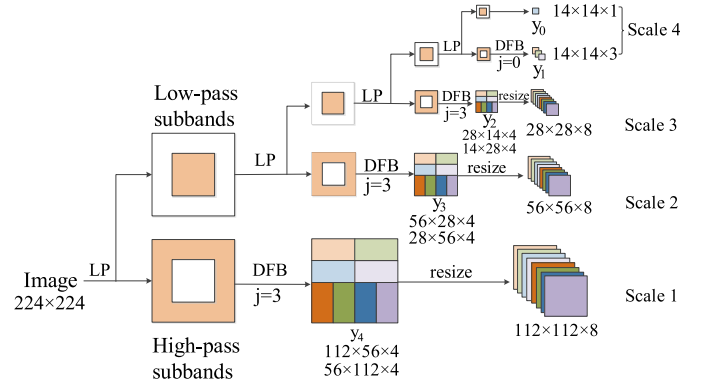


Fig. 6. Contourlet transform of an input image with size of  $224 \times 224$ . The decomposition level of LP is 4. The numbers of DFB are  $[0, 3, 3, 3]$  from coarse to fine scale, where 0 denoted a 2-D wavelet decomposition. The numbers below coefficient images denote width  $\times$  height  $\times$  channels of the output.

Given the discrete signal  $y[n]$  in  $l^2(\mathbb{Z}^2)$ , the family  $\{g_k^{(l)}[n - S_k^{(l)}m]\}_{0 \leq k \leq 2^l, m \in \mathbb{Z}^2}$  ( $m$  is the location index) provides a basis for  $y[n]$ , which is generated by translating the impulse responses of  $G_k^{(l)}$  and demonstrates the directional and localized properties.

In order to obtain the corresponding coefficient images with the contourlet transform, we use PDFB to decompose the image iteratively, as shown in Fig. 6. The details are described as follows.

Given an input image  $x \in \mathbb{R}^{n \times n}$  ( $n = 224$ ),  $l = 1, \dots, L$ .  $L$  is the decomposition level of LP that represents scales, and  $k \in [0, 2^l]$  is the subband’s index in each decomposition level  $j$  of DFB that represents directions. Use (10) to obtain a group of coefficients denoted  $C_k^l(x, y)$ . Therefore, different scales  $(n/2^j) \times (n/2^j)$  ( $112 \times 112, 56 \times 56, 28 \times 28, 14 \times 14$ ) are obtained after LP decomposition with  $L = 4$ . The numbers of DFB are  $[0, 3, 3, 3]$  from coarse to fine scale, where 0 denoted a 2-D wavelet decomposition, resulting a coarse approximation  $y_0$  and three bandpass directional subbands  $y_1$  in scale 4. The eight bandpass directional subbands are obtained from scale1 to scale3, respectively. When the decomposition level of DFB is even, the shape of bandpass subbands is square, and it is rectangular when the decomposition level is odd. Thus, the decomposed images need to be processed before concatenation. For more details of multifeature fusion, refer to

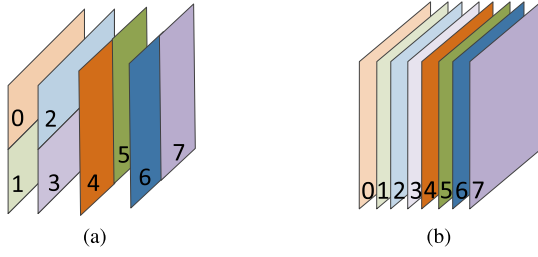


Fig. 7. Process of eight coefficients images before fusion. (a) Splicing the elongated coefficients images into squares. (b) Resizing the elongated decomposed images into squares.

Section II-B3. The processed coefficient maps from scale1 to scale4 correspond to the spectral feature maps of orange in conv1, conv2, conv4, and conv6 in Fig. 2, respectively.

In C-CNN, we utilize a  $3 \times 3$  convolution operation with  $1 \times 1$  padding and stride of 2 rather than pooling operation. It is vital to realize that the role of convolution layers on CNN is very similar to filter banks in traditional texture analysis. However, the spatial features extracted from CNN could be less dominant in the texture analysis [49]. Therefore, we want to get more spectral information of texture.

3) *Multifeature Fusion*: In order to make full use of spectral features, SSFF via contourlet is applied to obtain more abundant spectral information. In addition, statistical features are flattened and concatenated to the FC layer for the purpose of increasing the distinction between classes. More details are as follows.

a) *Spatial-spectral feature fusion via contourlet*: During the feature fusion, it needs to be considered how to concatenate the feature maps obtained by CNN and the contourlet decomposition images in corresponding sizes since contourlet has elongated supports. For example, when the number of DFB decomposition levels is 3, resulting in  $2^3 = 8$  wedge-shaped subbands in the frequency domain, as shown in Fig. 5. The image is then decomposed into eight subbands with rectangle-shaped [see Fig. 6]. In order to concatenate contourlet decomposition images to the network, the decomposed image should be treated as square before concatenation. Fig. 7 shows the two strategies to process the decomposed images. Contrast experiments are carried out by two strategies: splicing the feature maps of several elongated images into squares [see Fig. 7(a)] and resizing the elongated decomposed images directly to its corresponding sizes [see Fig. 7(b)]. We adopted the latter that is in line with our intuition: the spatial and spectral features in texture images, the two different but related features, can alternatively assist each other to obtain better local optima. For texture image decomposition with contourlet, where the coefficients are large and the edges and textures are sensitive, each pixel corresponds to a spatial coordinate. Therefore, the resized high-frequency coefficient maps can be consistent with spatial feature maps in pixelwise fashion. The experiment [see Section III-A3] shows the effectiveness of the SSFF via contourlet though it is easy to implement.

b) *Statistical feature fusion*: Given an image  $I$ , we can obtain different coefficients  $C_n(I)$  of various scales and directions with the contourlet decomposition. The subscript  $n$  is

the index of coefficients. The mean  $\mu_n$  and variance  $\sigma_n$  are chosen as the texture features of decomposed images; thus, the statistical texture feature  $f$  is formulated as follows:

$$f = (\mu_1, \sigma_1, \mu_2, \sigma_2, \dots, \mu_{2^l}, \sigma_{2^l})$$

$$\mu_n = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H |C_n(x, y)|$$

$$\sigma_n = \sqrt{\frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H (|C_n(x, y)| - \mu_n)^2} \quad (12)$$

where  $l = 1, 2, \dots, L$  represents the number of decomposition level.  $C_n(x, y)$  is the  $n$ th contourlet subband coefficients of an image. The width and height of the  $n$ th subband are  $W$  and  $H$  separately.

Then, the statistical texture feature  $f$  is concatenated to the features in FC2 layer; thus, the concatenated features  $a_k$  can be expressed as

$$a_k = f \oplus FC2 \quad (13)$$

where the symbol  $\oplus$  denotes features concatenation.

4) *Loss Function and Update Rule*: The softmax loss is applied to compute the multinomial logistic loss of the softmax using a single sample

$$L = -\log \frac{e^{a_j}}{\sum_{k=1}^T e^{a_k}} \quad (14)$$

where  $a_k$  is the input and  $T$  denotes the number of classes. We achieve stochastic optimization with Adam (adaptive moment estimation) [72], which is generated by combining the algorithm of Momentum [73] with RMSProp [74]. The update rule of Adam is

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} \quad (15)$$

where  $\theta_t$  is the parameter at time step  $t$  and  $\eta$  is the learning rate.  $\hat{m}_t = (m_t/1 - \beta_1^t)$  and  $\hat{v}_t = (v_t/1 - \beta_2^t)$  are bias-corrected estimates of the first ( $m_t$ ) and second raw moments ( $v_t$ ) of the gradients, respectively. The hyperparameters  $\beta_1$  and  $\beta_2$  denote the exponential decay rates with the constraint:  $0 \leq \beta_1$  and  $\beta_2 < 1$ .

### C. Implementation of C-CNN

The proposed C-CNN architecture is derived from AlexNet as can be found from Fig. 2. It consists of nine convolution layers, one GAP layer followed by three fully connected layers. The  $3 \times 3$  kernels and  $1 \times 1$  padding are applied in convolution layers, which guarantees the size of the feature maps. We utilize the stride of 2 in convolution layers instead of pooling. The learning procedure is listed in Algorithm 1.  $k$  is the subband's index in  $F_{DFB}$  and the decomposition level  $L = 4$ . Softmax loss and Adam algorithm are employed. The default values in Adam are  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 10^{-8}$ . As explained in [75] that the average pooling improved the gradient flow in texture generation rather than max pooling, the GAP layer was employed in our network. The invariance size of feature maps is ensured by the two-pixel



**Algorithm 1** Learning Procedure of C-CNN

**Input:** Training dataset:  $\mathbf{x} = \{x_n | n = 1, 2, \dots, N\}$ , and their corresponding labels  $\mathbf{y} = \{y_n | n = 1, 2, \dots, N\}$ ; the number of categories  $T$ .

**Output:** Classification result  $\hat{\mathbf{y}} = \{\hat{y}_n | n = 1, 2, \dots, N\}$ .

- 1: **Preparation:** The backbone network can be simply expressed as  $[conv1, \dots, conv9, GAP, FC1, FC2, FC3]$ ,  $nlevels = [0, 3, 3, 3]$ ,  $l = 1, \dots, L$ ,  $F_{LP} = \text{'maxflat'}$ ,  $F_{DFB} = \text{'dmaxflat7'}$ .
- 2: **Begin**
- 3: **for**  $n = 1$  to  $N$  **do**
- 4:   Given an input image  $x_n$ , using Eq.(10) to obtain a group of coefficients denoted  $C_k^l(x, y)$ .  $k$  is the subband's index in each decomposition level;
- 5:   Calculate the statistical feature  $f$  of  $C_k^l(x, y)$  by Eq.(12);
- 6:   **if**  $height(C_k^l(x, y)) \neq width(C_k^l(x, y))$  **then**
- 7:     resize the  $C_k^l(x, y)$  to the size of  $M^2$ , where  $M = \max(height(C_k^l(x, y)), width(C_k^l(x, y)))$ .
- 8:   **end if**
- 9:   Cascade the feature maps as  $conv1 \oplus C_k^1(x, y)$ ,  $conv2 \oplus C_k^2(x, y)$ ,  $conv4 \oplus C_k^3(x, y)$ ,  $conv6 \oplus C_k^4(x, y)$ ;
- 10:   Concatenate statistical feature to  $FC2$ :  $f \oplus FC2$ ;
- 11:   Minimize Eq.(14) and update the parameters by Eq.(15) until convergence.
- 12: **end for**
- 13: **End**

stride in the convolutional layer with  $1 \times 1$  padding. The size of the output layer is equal to the number of classes. Some regular operations are adopted in our networks, such as batch normalization and dropout. We implement our network on Caffe, and the memory consumption of our model is 52.9 MB.

We concatenated the feature maps obtained by CNN and the contourlet decomposition images in corresponding sizes. The contourlet decomposition is achieved by using filters “maxflat” for the pyramidal decomposition and “dmaxflat7” for the directional decomposition. The numbers of DFB decomposition levels at each pyramidal level are  $[0, 3, 3, 3]$  from coarse to fine-scale, where 0 denoted a 2-D wavelet decomposition. The interlaced downsampling factor  $p = 2$ .

In order to make full use of the spectral information in the decomposed images, we use the SSFF and SSF. SSFF strengthens the correlation between feature maps in different domains. Statistical features are commonly used in traditional texture classification, which is a good representation of texture in a coarse degree. Thus, we calculate the mean and variance of each contourlet subband coefficient. Then, these features (dimension is 56) are connected to the FC layer. It enhanced feature representation, rotation invariant, and intraclass disparity to some extent.

### III. VALIDATION AND GENERALIZATION

In this section, the proposed C-CNN is evaluated on eight benchmark data sets: three texture data sets for image texture classification (see Section III-A) and five remote sensing data

TABLE I  
CLASSIFICATION RESULTS (ACCURACY %) FOR NETWORKS

Methods	Strategy	kth-tips2-b	DTD
AlexNet [56]	not pre-trained	48.3 $\pm$ 1.4	22.7 $\pm$ 1.3
T-CNN [47]	not pre-trained	48.7 $\pm$ 1.3	27.8 $\pm$ 1.2
	pre-trained	73.2 $\pm$ 2.2	55.8 $\pm$ 0.8
Shearlet [20]	pre-trained	62.3 $\pm$ 0.8	21.6 $\pm$ 0.9
Wavelet CNN [49]	not pre-trained	63.7 $\pm$ 2.3	35.6 $\pm$ 0.7
	pre-trained	74.2 $\pm$ 1.2	59.8 $\pm$ 0.9
Contourlet CNN	not pre-trained	<b>70.31</b> $\pm$ 1.6	<b>40.62</b> $\pm$ 1.7
	pre-trained	<b>79.25</b> $\pm$ 1.7	<b>61.03</b> $\pm$ 1.0

sets for remote sensing scene classification (see Section III-B). We also check the influence of parameters in C-CNN and evaluate the effectiveness of the proposed network components.

#### A. Image Texture Classification

We evaluated C-CNN on three benchmark data sets: kth-tips2-b [76], DTD [77], and CURET [78]. The kth-tips2-b data set is composed of 11 classes of texture images with one training sample and three testing samples. Each sample contains  $4 \times 3 \times 9 = 108$  texture images under four illumination conditions, three poses, and nine scales with the size of  $200 \times 200$ . The results are averaged in the four train-val-test splits on the kth-tips2-b evaluation. The describable texture data sets (DTDs) are composed of 47 classes of 120 images, providing ten splits into an equal size of 40 training images, 40 validation images, and 40 test images. For this data set, we apply the same evaluation as is used in the kth-tips2-b data set, which averages the results on the ten train-val-test splits.

The Columbia-Utrecht Reflection and Texture (CURET) data set collects 61 classes of real-world surfaces that cover a wide range of geometric and photometric properties; 205 samples per class with a size of  $640 \times 480$ . We use 92 images for each class: 46 for training and the rest for testing and cropped the region ( $200 \times 200$ ) as in [37]. The results are averaged over the random ten splits. All the images of the data sets are resized into  $224 \times 224$ .

1) *Networks From Scratch and Pretrained:* We applied C-CNN to these two widely used texture data sets: kth-tips2-b and DTD. The classification results are reported in Table I. The strategy of “pretrained” is the method on ImageNet. Table I demonstrates that the accuracy of C-CNN reaches 70.31% and 40.62% on the kth-tips2-b data set and the DTD data set, respectively. The proposed method is compared with AlexNet [56], the spectral method with shearlet transform [20], T-CNN [47], and Wavelet CNN [49]. The comparison results indicate that the proposed method describing textures at multiple scales and directions has great potential for texture classification.

The proposed method is evaluated on the CURET data set. Table II reports the average classification accuracy of different methods. It can be found that C-CNN has comparable accuracy (99.7%) with the state of the art (99.8%) [77]. Overall, the remarkable behaviors achieved on three public texture data sets indicate the superior discriminative capability of



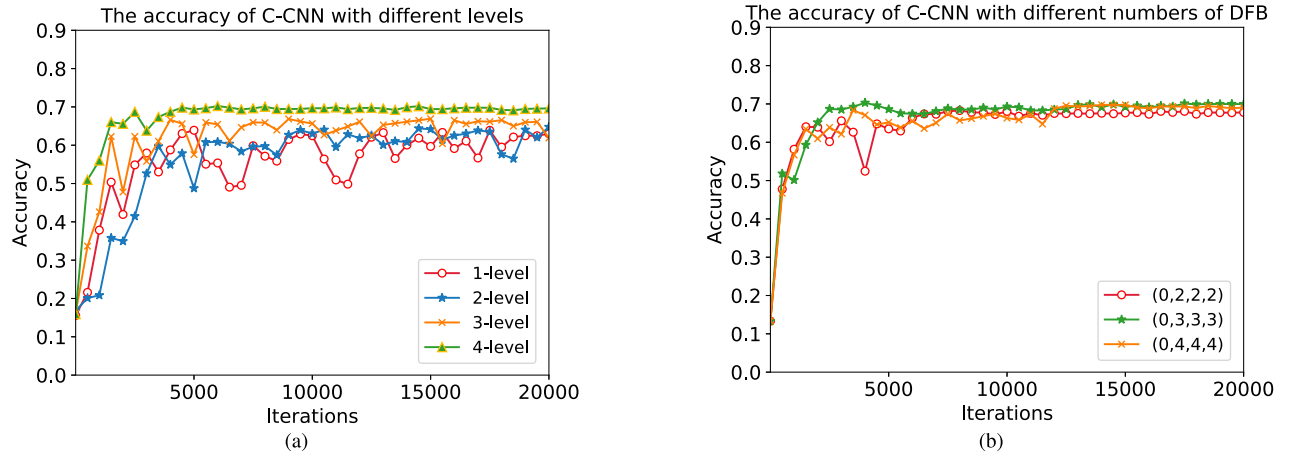


Fig. 8. C-CNN parameters analysis. (a) Accuracy of C-CNN with different contourlet decomposition levels. (b) Accuracy of C-CNN with different numbers of DFB.

TABLE II  
RESULTS (ACCURACY %) ON THE CURET DATA SET

Methods	Accuracy	Methods	Accuracy
LDTP [79]	91.54	AlexNet[56]	99.4
RALBGC [80]	93.82	T-CNN-1[47]	99.0
LCCMSP [81]	94.45	T-CNN-3[47]	99.5
ARCS-LBP [82]	94.72	IFV+ DeCAF[77]	<b>99.8</b>
LDEP [83]	95.23	IFV[77]	99.5
Wavelet CNN[49]	99.4	Contourlet CNN	99.7

TABLE III  
COMPARISON OF THE CLASSIFICATION RESULTS  
(ACCURACY %) WITH C-CNN

Methods	Accuracy	Methods	Accuracy
1-level C-CNN	63.91 $\pm$ 2.4	different numbers of DFB	
2-level C-CNN	64.78 $\pm$ 1.3	(0,2,2,2)	68.28 $\pm$ 1.2
3-level C-CNN	66.90 $\pm$ 2.1	(0,3,3,3)	<b>70.31<math>\pm</math>1.6</b>
4-level C-CNN	<b>70.31<math>\pm</math>1.6</b>	(0,4,4,4)	69.77 $\pm$ 1.1

C-CNN. Furthermore, the C-CNN combines spectral features and spatial features at multiscales with relatively sparse representation, which increased the interpretability of the network.

2) *Network Parameter Analysis*: In the texture images, contourlet decomposition captures the spectral information at different scales and directions. Accordingly, we investigated the behavior of our C-CNN on the kth-tips2-b data set with various decomposition levels and various numbers of the DFB in Table III.

In Fig. 8(a), we can see that the accuracies improved with an increase in the decomposition levels and the fluctuation gradually stabilized down. The results trained from scratch show that the proposed approach performs best with four-level C-CNN. It is consistent with our intuition: the more the decomposition features extracted, the more the representative information is generated. In terms of the optimal numbers of DFBs, [0, 3, 3, 3] is superior to the others, as shown in Fig. 8(b).

TABLE IV  
CLASSIFICATION RESULTS (ACCURACY %) OF SEVERAL VARIANTS  
WITH C-CNN TRAINED FROM SCRATCH

origin	SSFF	SSF	accuracy
✓			64.70 $\pm$ 2.3
✓	✓		67.06 $\pm$ 1.9
✓	✓	✓	70.31 $\pm$ 1.6

3) *Ablation Study*: In order to evaluate the effectiveness of the components in the proposed network, we experiment with several variants of the network [see Table IV] on the kth-tips2-b data set. The “origin” splices the elongated decomposed images into its corresponding sizes since the contourlet has elongated supports. SSFF via contourlet represents the elongated decomposed images that are directly resized into its corresponding sizes. SSF denotes the additional texture features of decomposed images. For more details of statistical features, refer to Section II-B3.

The “origin” version of C-CNN achieved an accuracy of 0.84 on additional data sets of kth-tips2-b that contains four classes of 36 images with no pretrained. In addition, we evaluated the performance of C-CNN in remote sensing scene classification.

### B. Remote Sensing Scene Classification

We also evaluated C-CNN on another five benchmark data sets of remote sensing: UC Merced data set (UCM) [84], WHU-RS data set [85], AID data set [86], RSSCN7 data set [87], and NWPU-RESISC45 [88]. The UCM data set is collected from the USGS National Map Urban Area Imagery in various urban areas with the pixel resolution of 1 ft. It is composed of 21 classes of 100 images with a size of  $256 \times 256$  pixels; 80 training images per category are randomly selected and the remaining 20 images for the testing set. The results are averaged on the ten train-val-test splits in random on the UCM evaluation.

The satellite images in the WHU-RS data set are exported from Google Earth with high resolution up to 0.5 m.

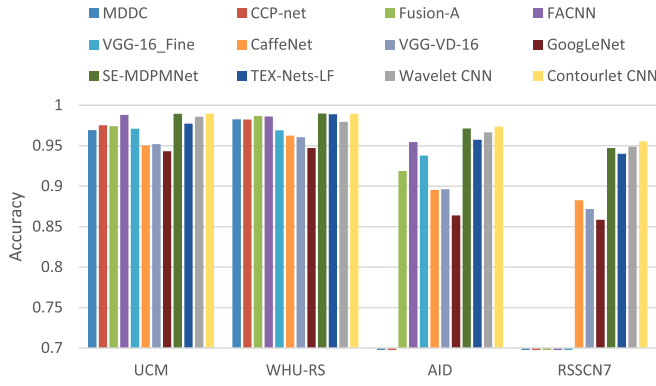


Fig. 9. Results of the accuracy on the four data sets.

It composed of 19 classes of 50 images with a size of  $600 \times 600$  pixels. For this data set, we apply the same evaluation as was used in the UCM data set, which averages the results on the ten train-val-test splits; 30 training images per category are randomly selected and the remaining 20 images for the testing set.

The AID data set is collected from the Google Earth imagery, which contains 10000 images within 30 classes. It is in a multisource condition with a size of  $600 \times 600$ . Unlike the single-source images data sets, the AID is challenging since images are extracted at different times and seasons with various resolutions. The RSSCN7 data set is also collected from Google Earth, which includes seven categories of 400 images with a size of  $400 \times 400$  pixels. Images are sampled at four different scales with 100 images per scale in each class under varying seasons and weathers; 50% samples are fixed for training both in the RSSCN7 and AID data sets. We average the results on the ten train-val-test splits and resize all the images of the data sets to  $224 \times 224$ .

The NWPU-RESISC45 data set is created by Northwestern Polytechnical University (NWPU) from Google Earth, which is used for remote sensing image scene classification (RESISC). This data set includes 31500 images, covering 45 scene classes of 700 images with a size of  $256 \times 256$ . The spatial resolution of these images varies from about 30–0.2 m per pixel for most scene classes. The data set is challenging since images are captured at different weather conditions and imaging angles from different sensors. The training ratios are 10% and 20%, respectively. The overall accuracy is averaged on the ten train-val-test splits.

The visual classification results on the four remote sensing scene data sets are shown in Fig. 9, and the quantitative results are listed in Table V.

The comparison methods are considered as follows.

- 1) *MDDC* [89]: The multiscale deeply described correlations method incorporates appearance and spatial features at different scales to obtain an accurate classification of the land-use scene images.
- 2) *CCP-Net* [90]: It proposes an end-to-end trainable CNNs that generate a spatial-rotation-invariant representation with “concentric circle pooling.”

TABLE V  
CLASSIFICATION RESULTS (ACCURACY %) FOR NETWORKS

Methods	UCM	WHU-RS	AID	RSSCN7
MDDC [89]	96.92±0.57	98.27±0.53	-	-
CCP-net [90]	97.52±0.97	98.23±0.40	-	-
Fusion-A [91]	97.42±1.79	98.65±0.43	91.87±0.36	-
FACNN [92]	98.81±0.24	98.61±0.44	95.45±0.11	-
VGG-16_Fine [92]	97.10±0.51	96.91±0.21	93.78±0.36	-
CaffeNet [86]	95.02±0.81	96.24±0.56	89.53±0.31	88.25±0.62
VGG-VD-16 [86]	95.21±1.20	96.05±0.91	89.64±0.36	87.18±0.94
GoogLeNet [86]	94.31±0.89	94.71±1.33	86.39±0.55	85.84±0.92
SE-MDPMNet [93]	98.95±0.12	<b>98.97±0.24</b>	97.14±0.15	94.71±0.15
TEX-Nets-LF [94]	97.72±0.54	98.88±0.49	95.73±0.16	94.0±0.57
Wavelet CNN [49]	98.58±0.13	97.96±0.90	96.65±0.24	94.89±0.67
Contourlet CNN	<b>98.97±0.21</b>	98.95±0.82	<b>97.36±0.45</b>	<b>95.54±0.71</b>

- 3) *Fusion-A* [91]: It builds the fusion by addition strategy on high-level features extracted from CaffeNet, VGG-VD-16, and GoogLeNet in [86].
- 4) *FACNN* [92]: It achieves the remote sensing scene classification in the feature aggregation CNN, which describes the intermediate features with convolutional features encoding module and the progressive aggregation strategy.
- 5) *VGG-16\_Fine* [92]: It employs the VGG-16 trained on ImageNet to initialize the weight parameters of the network.
- 6) *CaffeNet* [86]: It comprises five convolutional layers, each followed by a pooling layer and three fully connected layers.
- 7) *VGG-VD-16* [86]: It composed of 13 convolutional layers and three fully connected layers.
- 8) *GoogLeNet* [86]: It is a 22-layer architecture with the inception modules.
- 9) *SE-MDPMNet* [93]: It extracts effective multiscale features with the dilated convolution, multidilation pooling module, and channel attention in MobileNet V2 that maintains high accuracy.
- 10) *TEX-Nets-LF* [94]: It uses the late fusion strategy on two-stream deep architectures, combining the texture and color information as complementary.
- 11) *Wavelet CNN* [49]: It integrates a spectral analysis of wavelet into CNNs based on the insight that the pooling and the convolution can be thought of as a limited form of the spectral analysis.

It shows that C-CNN achieves comparable performance on three data sets compared with other studied approaches, while it is inferior to the SE-MDPMNet [93] of 0.02% on the WHU-RS data set. The reasons for competitive performance are as follows. First, the multiscale and multidirectional features of contourlet can be associated with a biological interpretation of HVS. Second, the remote sensing images are mainly composed of geometrical and morphological features, which are consistent with the characteristics of multiscale geometry. Thus, we have embedded *a priori* knowledge into CNNs to refine the network. Third, wavelet CNN [49] employs the

TABLE VI  
CLASSIFICATION RESULTS (ACCURACY %)   
ON THE NWPU-RESISC45 DATA SET

Methods	Training ratios	
	10%	20%
unsupervised feature learning methods		
BoVW+SPM [88]	27.83 $\pm$ 0.61	32.96 $\pm$ 0.47
LLC [88]	38.81 $\pm$ 0.23	40.03 $\pm$ 0.34
BoVW [88]	41.72 $\pm$ 0.21	44.97 $\pm$ 0.28
CNN-based methods		
AlexNet [88]	76.69 $\pm$ 0.21	79.85 $\pm$ 0.13
VGGNet-16 [88]	76.47 $\pm$ 0.18	79.79 $\pm$ 0.15
GoogLeNet [88]	76.19 $\pm$ 0.38	78.48 $\pm$ 0.26
LASC-CNN [95]	81.37	84.30
BoCF(VGGNet-16) [96]	82.65 $\pm$ 0.31	84.32 $\pm$ 0.17
Attention GANs [97]	<b>86.11</b> $\pm$ 0.22	89.44 $\pm$ 0.18
Two-stream fusion [98]	85.02 $\pm$ 0.25	87.01 $\pm$ 0.19
Wavelet CNN [49]	84.69 $\pm$ 0.44	88.84 $\pm$ 0.38
Contourlet CNN	85.93 $\pm$ 0.51	<b>89.57</b> $\pm$ 0.45

constant parameters of filters, which is better than the one with single CNN. The contourlet has sparser feature representation than wavelet, and the additional statistical features enhanced feature representation, as well as increased intraclass disparity to some extent. Thus, the effective performances are obtained with the singularity and directionality of the contours.

Table VI reports the comparative results on the NWPU-RESISC45 data set, which includes the representative approaches, namely, unsupervised feature learning methods and CNN-based methods. Among them, BoVW + SPM [88], LLC [88], and BoVW [88] are unsupervised feature learning methods. AlexNet [88], VGGNet-16 [88], GoogLeNet [88], LASC-CNN [95], BoCF(VGGNet-16) [96], Attention GANs [97], two-stream fusion [98], Wavelet CNN [49], and the proposed method are CNN-based methods. As can be seen in Table VI, the CNN-based methods perform much better than unsupervised feature learning methods under the training ratios of 10% and 20%. In addition, the proposed method surpasses most of the CNN-based methods that deal with the scene classification task in the spatial domain. The proposed method achieves an accuracy of 85.93  $\pm$  0.51 (10% training samples) and 89.57  $\pm$  0.45 (20% training samples). The improvements may come from the combination of the spatial and spectral features. This is due to the fact that the information of two domains has supplied complementary components, and the statistical features can enhance the discrimination of global scene representation. Besides, compared with other CNN-based methods, the proposed model is more compact, which has the advantage of relative shadow structure in small network size.

#### IV. CONCLUSION

In this article, C-CNN is presented to extract and integrate spatial and spectral features for texture classification. The contourlet transform provides more multidirectional and multi-scale information effectively in the spectral domain. It is novel

that the spectral analysis is integrated into CNN with statistical properties for texture sparse representation. The correlation between feature maps in different domains is strengthened by the SSFF via contourlet at different decomposition levels. The additional statistical features enhanced the discriminative feature representation and rotation invariant.

The theoretical analysis and complete experiments on C-CNN are discussed in this article. The proposed method can achieve comparable results against state-of-the-art methods on three texture data sets and five remote sensing data sets with effectiveness and reliability. In terms of the model generality, a good feature representation can be obtained without pre-training; thus, the model can be employed in other related visual tasks. However, there are still limitations on the adaptive selection of the contourlet decomposition basis on different tasks. The contourlet kernel is integrated into the network explicitly. In future studies, we plan to incorporate a guide attention strategy into our model to make full use of spectral information.

#### REFERENCES

- [1] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
- [2] L. Liu and P. Fieguth, "Texture classification from random features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 574–586, Mar. 2012.
- [3] B. Julesz, "Visual pattern discrimination," *IEEE Trans. Inf. Theory*, vol. IT-8, no. 2, pp. 84–92, Feb. 1962.
- [4] C.-M. Wu and Y.-C. Chen, "Statistical feature matrix for texture analysis," *CVGIP, Graph. Models Image Process.*, vol. 54, no. 5, pp. 407–419, Sep. 1992.
- [5] J. S. Weszka, C. R. Dyer, and A. Rosenfeld, "A comparative study of texture measures for terrain classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, no. 4, pp. 269–285, Apr. 1976.
- [6] D. Blostein and N. Ahuja, "Shape from texture: Integrating texture-element extraction and surface estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 12, pp. 1233–1251, Dec. 1989.
- [7] K. A. Stevens, "Surface perception from local analysis of texture," AI Lab., Massachusetts Inst. Technol. Cambridge, Cambridge, MA, USA, AI-Tech. Rep.-512, 1980.
- [8] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proc. 12th Int. Conf. Pattern Recognit.*, vol. 1, Oct. 1994, pp. 582–585.
- [9] L. Wang and D.-C. He, "Texture classification using texture spectrum," *Pattern Recognit.*, vol. 23, no. 8, pp. 905–910, Jan. 1990.
- [10] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996.
- [11] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
- [12] Y. Zhao, W. Jia, R.-X. Hu, and H. Min, "Completed robust local binary pattern for texture classification," *Neurocomputing*, vol. 106, pp. 68–76, Apr. 2013.
- [13] R. Chellappa and S. Chatterjee, "Classification of textures using Gaussian Markov random fields," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 4, pp. 959–963, Aug. 1985.
- [14] A. Khotanzad and R. L. Kashyap, "Feature selection for texture recognition based on image synthesis," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-17, no. 6, pp. 1087–1095, Nov. 1987.
- [15] B. J. Super and A. C. Bovik, "Localized measurement of image fractal dimension using Gabor filters," *J. Vis. Commun. Image Represent.*, vol. 2, no. 2, pp. 114–128, Jun. 1991.
- [16] J. M. Coggins and A. K. Jain, "A spatial filtering approach to texture analysis," *Pattern Recognit. Lett.*, vol. 3, no. 3, pp. 195–203, May 1985.



- [17] L. M. Kaplan, "Extended fractal analysis for texture classification and segmentation," *IEEE Trans. Image Process.*, vol. 8, no. 11, pp. 1572–1585, Nov. 1999.
- [18] H. Arof and F. Deravi, "Circular neighbourhood and 1-D DFT features for texture classification and segmentation," *IEE Proc. Vis., Image, Signal Process.*, vol. 145, no. 3, pp. 167–172, 1998.
- [19] Q. Sun, B. Hou, and L.-C. Jiao, "Automatic texture segmentation based on wavelet-domain hidden Markov tree," in *Proc. Iberoamerican Congr. Pattern Recognit.* Springer, 2005, pp. 470–480.
- [20] K. G. Krishnan, V. P. T., and R. Abinaya, "Performance analysis of texture classification techniques using shearlet transform," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2016, pp. 1408–1412.
- [21] Y. Dong, D. Tao, X. Li, J. Ma, and J. Pu, "Texture classification and retrieval using shearlets and linear regression," *IEEE Trans. Cybern.*, vol. 45, no. 3, pp. 358–369, Mar. 2015.
- [22] C. Vivek and S. Audithan, "Classification of colour textures using shearlets," in *Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES)*, Feb. 2014, pp. 1–5.
- [23] K. Huang and S. Aiyente, "Rotation invariant texture classification with ridgelet transform and Fourier transform," in *Proc. Int. Conf. Image Process.*, Oct. 2006, pp. 2141–2144.
- [24] Y. Shang, Y.-H. Diao, and C.-M. Li, "Rotation invariant texture classification algorithm based on curvelet transform and SVM," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 5, Jul. 2008, pp. 3032–3036.
- [25] Z. Xiangbin, "Texture classification based on contourlet and support vector machines," in *Proc. Int. Colloq. Comput., Commun., Control, Manage. (ISECS)*, vol. 2, Aug. 2009, pp. 521–524.
- [26] Y. Hu, B. Hou, S. Wang, and L. Jiao, "Texture classification via stationary-wavelet based contourlet transform," in *Proc. Int. Workshop Intell. Comput. Pattern Anal. Synth.* Springer, 2006, pp. 485–494.
- [27] Y. Sha, L. Cong, Q. Sun, and L. Jiao, "Unsupervised image segmentation using contourlet domain hidden Markov trees model," in *Proc. Int. Conf. Image Anal. Recognit.* Springer, 2005, pp. 32–39.
- [28] X. Q. Shang, G. X. Song, and B. Hou, "Content based texture image classification," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 3, Nov. 2003, pp. 1309–1313.
- [29] T. Shan, X. Zhang, and L. Jiao, "A brushlet-based feature set applied to texture classification," in *Proc. Int. Conf. Comput. Inf. Sci.* Springer, 2004, pp. 1175–1180.
- [30] L. Cong, Y. Sha, B. Hou, and L. Jiao, "Unsupervised image segmentation based on the anisotropic texture information," in *Proc. 3rd Int. Conf. Image Graph. (ICIG)*, 2004, pp. 124–127.
- [31] L. Chen and H. Man, "Hybrid IMM/SVM approach for wavelet-domain probabilistic model based texture classification," *IEE Proc. Vis., Image, Signal Process.*, vol. 152, no. 6, pp. 724–730, 2005.
- [32] K. I. Kim, K. Jung, S. H. Park, and H. J. Kim, "Support vector machines for texture classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 11, pp. 1542–1550, Nov. 2002.
- [33] S. Liu, H. Yi, L.-T. Chia, and D. Rajan, "Adaptive hierarchical multi-class SVM classifier for texture-based image classification," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2005, pp. 1–4.
- [34] I. Turkoglu and E. Avci, "Comparison of wavelet-SVM and wavelet-adaptive network based fuzzy inference system for texture classification," *Digit. Signal Process.*, vol. 18, no. 1, pp. 15–24, Jan. 2008.
- [35] E. P. Lam, "Texture classification using wavelet decomposition," in *Proc. IEEE Int. Conf. Syst. Syst. Eng.*, Jun. 2008, pp. 1–5.
- [36] Z. Sun and K. Jia, "Road surface condition classification based on color and texture information," in *Proc. 9th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Oct. 2013, pp. 137–140.
- [37] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1657–1663, Jun. 2010.
- [38] G. Girish and J. K. Dash, "Adaptive fuzzy local binary pattern for texture classification," in *Proc. 2nd Int. Conf. Man Mach. Interfacing (MAMI)*, Dec. 2017, pp. 1–5.
- [39] G. S. Raghtate and S. S. Salankar, "Comparison of classification methods with second order statistical analysis and wavelet transform for texture image classification," in *Proc. Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Dec. 2015, pp. 312–317.
- [40] R. Paget, I. D. Longstaff, and B. Lovell, "Texture classification using nonparametric Markov random fields," in *Proc. 13th Int. Conf. Digit. Signal Process.*, vol. 1, 1997, pp. 67–70.
- [41] X. Bai and K. Wang, "Research on classification of wood surface texture based on Markov random field," in *Proc. 2nd IEEE Conf. Ind. Electron. Appl.*, May 2007, pp. 664–668.
- [42] H. Potlapalli and R. C. Luo, "Fractal-based classification of natural textures," *IEEE Trans. Ind. Electron.*, vol. 45, no. 1, pp. 142–150, Feb. 1998.
- [43] L. Jiao, L. Bo, and L. Wang, "Fast sparse approximation for least squares support vector machine," *IEEE Trans. Neural Netw.*, vol. 18, no. 3, pp. 685–697, May 2007.
- [44] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *Proc. 13th Int. Conf. Control Autom. Robot. Vis. (ICARCV)*, Dec. 2014, pp. 844–848.
- [45] L. G. Hafemann, L. S. Oliveira, and P. Cavalin, "Forest species recognition using deep convolutional neural networks," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1103–1107.
- [46] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3828–3836.
- [47] V. Andrearczyk and P. F. Whelan, "Using filter banks in convolutional neural networks for texture classification," *Pattern Recognit. Lett.*, vol. 84, pp. 63–69, Dec. 2016.
- [48] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks for texture classification," 2017, *arXiv:1707.07394*. [Online]. Available: <http://arxiv.org/abs/1707.07394>
- [49] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks," 2018, *arXiv:1805.08620*. [Online]. Available: <http://arxiv.org/abs/1805.08620>
- [50] 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.
- [51] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
- [52] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. Int. Conf. Learn. Representations*, 2014.
- [53] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 1999.
- [54] M. Do and M. Vetterli, *Contourlets, Beyond Wavelets*. New York, NY, USA: Academic, 2003.
- [55] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat," *J. Neurophysiol.*, vol. 28, no. 2, pp. 229–289, Mar. 1965.
- [56] I. Sutskever, G. E. Hinton, and A. Krizhevsky, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [57] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–11.
- [58] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
- [59] 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.
- [60] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
- [61] F. H. C. Tivive and A. Bouzerdoum, "Texture classification using convolutional neural networks," in *Proc. IEEE Region Conf. (TENCON)*, Nov. 2006, pp. 1–4.
- [62] M. N. Do and M. Vetterli, "Contourlets: A directional multiresolution image representation," in *Proc. Int. Conf. Image Process.*, vol. 1, Sep. 2002, pp. I-357–I-360.
- [63] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, Jun. 1996.
- [64] D. L. Donoho and A. G. Flesia, "Can recent innovations in harmonic analysis 'explain' key findings in natural image statistics?" *Netw., Comput. Neural Syst.*, vol. 12, no. 3, pp. 371–393, Jan. 2001.
- [65] M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2091–2106, Dec. 2005.
- [66] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2435–2476, Oct. 1998.

- [67] E. J. Candès and D. L. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise  $C^2$  singularities," *Commun. Pure Appl. Math. A. J. Issued by Courant Inst. Math. Sci.*, vol. 57, no. 2, pp. 219–266, 2004.
- [68] L. Jiao, J. Pan, and Y. Fang, "Multiwavelet neural network and its approximation properties," *IEEE Trans. Neural Netw.*, vol. 12, no. 5, pp. 1060–1066, Sep. 2001.
- [69] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.
- [70] R. H. Bamberger and M. J. T. Smith, "A filter bank for the directional decomposition of images: Theory and design," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 882–893, Apr. 1992.
- [71] M. N. Do and M. Vetterli, "Contourlets: A new directional multiresolution image representation," in *Proc. Conf. Rec. 36th Asilomar Conf. Signals, Syst. Comput.*, vol. 1, Nov. 2002, pp. 497–501.
- [72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd Int. Conf. Learn. Representations*, 2015, pp. 1–15.
- [73] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Netw.*, vol. 12, no. 1, pp. 145–151, Jan. 1999.
- [74] T. Tieleman and G. Hinton, "Lecture 6.5-RMSPROP: Divide the gradient by a running average of its recent magnitude," *COURSERA, Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.
- [75] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 262–270.
- [76] E. Hayman, B. Caputo, M. Fritz, and J.-O. Eklundh, "On the significance of real-world conditions for material classification," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2004, pp. 253–266.
- [77] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3606–3613.
- [78] K. J. Dana, B. van Ginneken, S. K. Nayar, and J. J. Koenderink, "Reflectance and texture of real-world surfaces," *ACM Trans. Graph.*, vol. 18, no. 1, pp. 1–34, Jan. 1999.
- [79] I. El Khadiri, A. Chahi, Y. El Merabet, Y. Ruichek, and R. Touahni, "Local directional ternary pattern: A new texture descriptor for texture classification," *Comput. Vis. Image Understand.*, vol. 169, pp. 14–27, Apr. 2018.
- [80] I. El Khadiri, M. Kas, Y. El Merabet, Y. Ruichek, and R. Touahni, "Repulsive-and-attractive local binary gradient contours: New and efficient feature descriptors for texture classification," *Inf. Sci.*, vol. 467, pp. 634–653, Oct. 2018.
- [81] Y. El Merabet and Y. Ruichek, "Local concave-and-convex micro-structure patterns for texture classification," *Pattern Recognit.*, vol. 76, pp. 303–322, Apr. 2018.
- [82] Y. El Merabet, Y. Ruichek, and A. El Idrissi, "Attractive-and-repulsive center-symmetric local binary patterns for texture classification," *Eng. Appl. Artif. Intell.*, vol. 78, pp. 158–172, Feb. 2019.
- [83] Y. Dong *et al.*, "Locally directional and extremal pattern for texture classification," *IEEE Access*, vol. 7, pp. 87931–87942, 2019.
- [84] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th Int. Conf. Adv. Geographic Inf. Syst. (SIGSPATIAL GIS)*, 2010, pp. 270–279.
- [85] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," in *Proc. 7th Symp.-100 Years (ISPRS TC)*, vol. 38, 2010, pp. 298–303.
- [86] G.-S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
- [87] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.
- [88] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
- [89] K. Qi, C. Yang, Q. Guan, H. Wu, and J. Gong, "A multiscale deeply described correlations-based model for land-use scene classification," *Remote Sens.*, vol. 9, no. 9, p. 917, 2017.
- [90] K. Qi, Q. Guan, C. Yang, F. Peng, S. Shen, and H. Wu, "Concentric circle pooling in deep convolutional networks for remote sensing scene classification," *Remote Sens.*, vol. 10, no. 6, p. 934, 2018.
- [91] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.
- [92] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.
- [93] B. Zhang, Y. Zhang, and S. Wang, "A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2636–2653, Aug. 2019.
- [94] R. M. Anwer, F. S. Khan, J. van de Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 74–85, Apr. 2018.
- [95] B. Yuan, S. Li, and N. Li, "Multiscale deep features learning for land-use scene recognition," *J. Appl. Remote Sens.*, vol. 12, no. 1, p. 1, Feb. 2018.
- [96] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.
- [97] Y. Yu, X. Li, and F. Liu, "Attention GANs: Unsupervised deep feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 519–531, Jan. 2020.
- [98] Y. Yu and F. Liu, "Dense connectivity based two-stream deep feature fusion framework for aerial scene classification," *Remote Sens.*, vol. 10, no. 7, p. 1158, 2018.



**Mengkun Liu** received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 2014 and 2017, respectively. She is currently pursuing the Ph.D. degree with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence Xidian University, Xi'an, China.

Her research interests include machine learning and image processing.



**Licheng Jiao** (Fellow, IEEE) received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

Since 1992, he has been a Professor with the School of Artificial Intelligence, Xidian University, Xi'an, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. His research interests include image

processing, natural computation, machine learning, and intelligent information processing.

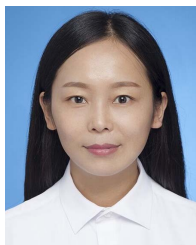
Dr. Jiao is a member of the IEEE Xi'an Section Execution Committee. He is also the Chairman of the Awards and Recognition Committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, a Councilor of the Chinese Institute of Electronics, a Committee Member of the Chinese Committee of Neural Networks, and an Expert of the Academic Degrees Committee of the State Council.



**Xu Liu** (Member, IEEE) received the B.Sc. degree in mathematics and applied mathematics from the North University of China, Taiyuan, China, in 2013, and the Ph.D. degree from Xidian University, Xi'an, China, in 2019.

He is currently a Post-Doctoral Researcher with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University. His research interests include machine learning and image processing.

Dr. Liu is also a member of the IEEE Geoscience and Remote Sensing Society.



**Lingling Li** (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2011 and 2017, respectively.

From 2013 to 2014, she was an Exchange Ph.D. Student with the Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU), Leioa, Spain. She is currently a Post-Doctoral Researcher with the School of Artificial Intelligence, Xidian University. Her current research interests include quantum evolutionary opti-

mization, machine learning, and deep learning.



**Shuyuan Yang** (Senior Member, IEEE) received the B.A. degree in electrical engineering and the M.S. and Ph.D. degrees in circuit and system from Xidian University, Xi'an, China, in 2000, 2003, and 2005, respectively.

She has been a Professor with the School of Artificial Intelligence, Xidian University. Her research interests include machine learning and multiscale geometric analysis.



**Fang Liu** (Senior Member, IEEE) received the B.S. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.S. degree in computer science and technology from Xidian University, Xi'an, in 1995.

She is currently a Professor with the School of Computer Science, Xidian University. Her research interests include signal and image processing, synthetic aperture radar image processing, multiscale geometry analysis, learning theory and algorithms, optimization problems, and data mining.