POLSAR Target Recognition Using a Feature Fusion Framework Based on Monogenic Signal and Complex-Valued Nonlocal Network

Feng Li, Min Yi, Chaoqi Zhang, Weijun Yao, Xueyao Hu, and Feifeng Liu

***Abstract*—With the continuous development of synthetic aperture radar (SAR) systems, multipolarization information has been increasingly applied to numerous fields, and automatic target recognition (ATR) in polarimetric SAR (POLSAR) has been recognized as vital problem. The SAR recognition methods can primarily fall into handcrafted feature-based algorithms and deep learning algorithms. The former exhibits excellent interpretability but insufficient generalization; the latter achieves stronger representational ability but relies on a considerable number of samples. To solve above problems, a feature fusion framework is proposed in this article based on monogenic signal and complex-valued nonlocal network (CVNLNet) for POLSAR target recognition. The proposed feature fusion framework effectively uses the complementarity of handcrafted features and deep features, while making up for the disadvantages of single feature-based methods. First, a Mono-BOVW model is proposed based on monogenic signal and bag-of-visual-words (BOVW) model to extract handcrafted features, which can more fully mine the information covered in POLSAR data in multiscale space. Moreover, CVNLNet is built for deep feature extraction to use both the amplitude and phase covered in POLSAR data. Next, a kernel discrimination correlation analysis algorithm is proposed to jointly analyze and transform the two features, so as to remove redundant information while retaining effective and discriminative information. Experiments on the MSTAR dataset and the GOTCHA dataset show that the proposed framework has superior performance on single polarimetric and fully polarimetric datasets.**

***Index Terms*—Complex-valued non-local network (CVNLNet), feature fusion, monogenic signal, polarimetric synthetic aperture radar (POLSAR), target recognition.**

Manuscript received 21 May 2022; revised 11 July 2022; accepted 24 July 2022. Date of publication 28 July 2022; date of current version 21 September 2022. This work was supported in part by the National Key R&D Program of China under Grant 2018YFE0202102, in part by China Postdoctoral Science Foundation under Grant 2021M690412, in part by the Natural Science Foundation of Chongqing, China under Grant cstc2020jcyj-msxmX0812, and in part by Shandong Provincial Natural Science Foundation under Project ZR2021MF134. *(Corresponding author: Xueyao Hu.)*

Feng Li and Xueyao Hu are with the Radar Research Laboratory, Beijing Institute of Technology, Beijing 10081, China, and also with the Beijing Institute of Technology Chongqing Innovation Center, Chongqing 401120, China (e-mail: karl1820@bit.edu.cn; xueyao.hu@qq.com).

Min Yi, Chaoqi Zhang, and Weijun Yao are with the Radar Research Laboratory, Beijing Institute of Technology, Beijing 10081, China (e-mail: minyime@163.com; 3120210773@bit.edu.cn; 1272680027@qq.com).

Feifeng Liu is with the Radar Research Laboratory, Beijing Institute of Technology, Beijing 10081, China, and also with the Advanced Technology Research Institute, Beijing Institute of Technology, Jinan 250300, China (e-mail: feifengliu_bit@bit.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3194551

I. Introduction

SYNTHETIC aperture radar (SAR) plays a vital role in real-time earth observation due to excellent characteristics such as all-day and all-weather, and is widely used in disaster as disaster monitoring [1], environmental protection [2], resource detection [3], meteorological observation [4], and other tasks [5]–[8]. Compared with single polarimetric SAR, the fully polarimetric SAR (POLSAR) system can measure the amplitude of the image, while containing the relative phase between different polarization channels [9]. Accordingly, POLSAR has been widely applied in various earth observation applications (e.g., target detection [10] and terrain classification [11]). Besides, several studies [12]–[14] have found that POLSAR has considerable application potential and value in stationary ground target recognition.

In the field of SAR automatic target recognition (SAR ATR), it is of critical significance to design a set of well-performing feature extraction and classification algorithm [15]. First, extracting the features which can effectively characterize the target is the premise of the subsequent correct classification. Generally, the features could be divided into two types: handcrafted features and deep features [16].

Handcrafted features are extracted from images by experts based on human perception and experience accumulation. They generally have specific physical meanings, such as computer vision features [17]–[21], electromagnetic characteristics [22], polarization characteristics [23], [24], and some special features like the monogenic signal. However, due to speckle noise and others, many vision features may have poor performance when directly transferred to POLSAR images [25]; while electromagnetic or polarization characteristics tend to focus on different specific scattering mechanisms, and generally require a combination of features [23]. The monogenic signal [26], an extended representation of analytic signal in high dimension, has aroused rising attention. Dong et al. [27]–[29] and Zhou et al. [30] used the multiscale components extracted based on the monogenic signal for SAR recognition, and both achieved better recognition accuracy than traditional features. In brief, handcrafted features are interpretable and so less affected by the number of samples; but they over-rely on human experience and lack generalization. For POLSAR data with complex scattering mechanism, it is a challenge to artificially design excellent features.

Unlike handcrafted features, deep networks can automatically learn effective features from data, which also means that it relies on a considerable number of samples. Deep networks could be grouped as real-valued (RV) networks and complex-valued (CV) networks. RV networks are based on RV representations and calculations (e.g., A-ConvNet [31] DCC-CNNs [32] and other networks [33]–[37]), which only use the amplitude while ignoring the phase of SAR data. Therefore, researchers introduce CV networks [38]. Zhang et al. [39] proposed a CV-CNN, Mullissa et al. [40] designed a d PolSARNet, besides, Tan et al. [41] and Zhang et al. [42] used CV-3D-CNN for hierarchical features extraction of POLSAR data. These networks are designed for POLSAR terrain classification rather than POLSAR target recognition. For SAR target recognition, Yu et al. [43] constructed a full convolutional neural network (CV-FCNN), and Scarnati et al. [44] evaluated the performance of several different complex-valued neural network (CVNNs). In brief, deep features show great advantages in the expression of deep abstract semantic and spatial structure; but they generally lack interpretability and are easily affected by the number of samples. However, the sample size of SAR datasets is generally small due to the difficulty of collection. Accordingly, how to apply deep learning robustly to POLSAR target recognition is worth studying in depth.

In order to make up for the deficiencies of a single feature under diverse and complex conditions, researchers have found that a reasonable combination of different types of features can greatly enhance the performance of image processing [45]. The fusion strategies based on multiple handcrafted features [46]–[51] or multiscale deep features [52]–[55] have been confirmed to be effective. Furthermore, some fusion strategies based on the handcrafted features and deep features, which have been indicated to have certain complementary properties, have also emerged [56]–[58]. Jia et al. [59] fused the features from principal component analysis (PCA) and CNN. Zhang et al. [60] fused the features from electromagnetic scattering center and MVGGNet. Feng et al. [61] developed a fusion method based on integration parts model and deep learning algorithm.

Some challenges remain in the fusion process of different types of features.

1) Different features may have different spatial dimensions.
2) The original feature information may be destroyed in the fusion process.
3) With the increase of feature types, the computation increases after fusion.

Accordingly, a suitable feature fusion algorithm is critical in multifeature fusion [62]. The classic fusion algorithms are serial fusion and parallel fusion [63], which are simple to operate but fail to maximize the complementary advantages of features and are prone to redundancy. Thus, the fusion algorithms based on linear transformation are proposed (e.g., PCA [64], linear discriminant analysis [65], canonical correlation analysis [66], discrimination correlation analysis [60]). Furthermore, to analyze the nonlinear relationship between different features, the kernel function [67] is introduced into linear transformation (e.g., KPCA [68], KFDA [69], and KCCA [70]).

As revealed by the above works, it is difficult to characterize the POLSAR target comprehensively and accurately using only handcrafted feature-based algorithms, while only using end-to-end deep learning algorithms is highly susceptible to the number of samples. Inspired by above fusion strategies, we propose an efficient feature fusion framework for POLSAR target recognition to fully use the complementary advantages of handcrafted features and deep features. First, we construct a Mono-BOVW model to extract handcrafted features based on the intrinsic properties of images, which are robust and less affected by the number of samples. The monogenic signal model is capable of using a multiscale space to extract richer information (amplitude, orientation, and phase) from SAR data, whereas its direct use often leads to excessive computation due to the high feature dimension. Thus, we introduce a bag-of-visual-words (BOVW) model to obtain stable low-dimensional features. Meanwhile, we construct a complex-valued nonlocal network (CVNLNet) to extract deep features with stronger representational ability, which uses both the amplitude and phase covered in SAR data. Furthermore, the CV nonlocal block is capable of capturing long-range dependencies. The advantages of the extracted handcrafted features and deep features could just make up for the shortcomings of each other. Lastly, to avoid the redundancy of fusion features while retaining as much effective and highly discriminative information as possible, we propose a fusion algorithm based on kernel discriminant correlation analysis (KDCA). The kernel function is introduced into the DCA algorithm to facilitate linear correlation analysis and dimension reduction of nonlinearly correlated features. Thus, in the fusion process, the features from the same category of targets have the most significant correlation, and the features from different categories have the most significant distinction. On that basis, stronger discriminative and more robust fusion features with lower dimensions can be obtained to increase the accuracy of SAR target recognition. The effectiveness and superiority of the proposed framework are verified on the MSTAR dataset and the GOTCHA dataset.

The main contributions and innovations of the proposed feature fusion framework are elucidated as follows.

1) A Mono-BOVW model is proposed for handcrafted feature extraction.
2) A CVNLNet is proposed for deep feature extraction.
3) A KDCA algorithm is proposed for feature fusion.

The rest of this article is organized as follows. Section II introduces the proposed feature fusion framework for POLSAR target recognition. Section III presents experiments and discussions. Section IV succinctly concludes this article.

II. Feature Fusion Framework Based on Monogenic Signal and Complex-Valued Nonlocal Network

Fig. 1 presents the overall architecture of the proposed feature fusion framework, which mainly comprises three parts, including preprocessing, extraction and fusion of handcrafted features and deep features, as well as classification.

In the first part, for better analysis and classification, the preprocessing operation is conducted by normalizing SAR image [39]. Z-Score function serves as the normalization algorithm, defined as: $x^* = (x-\bar{x})/\sigma$, where x denotes an image, $\bar{x}$ and σ

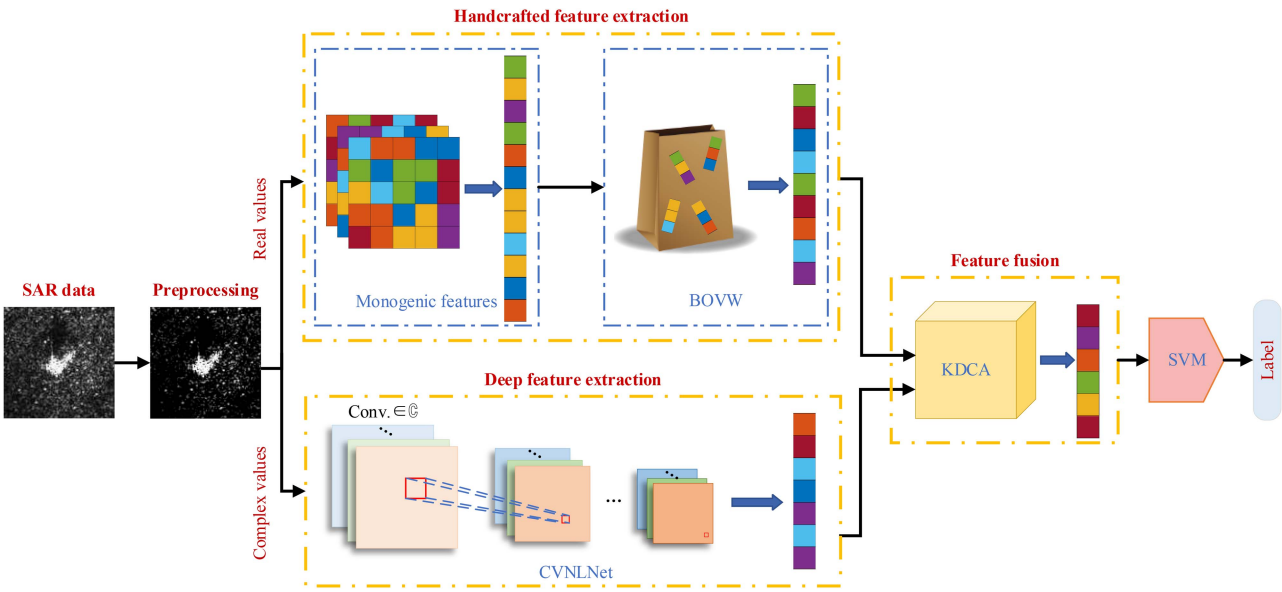

Fig. 1. Overall architecture of the proposed feature fusion framework.

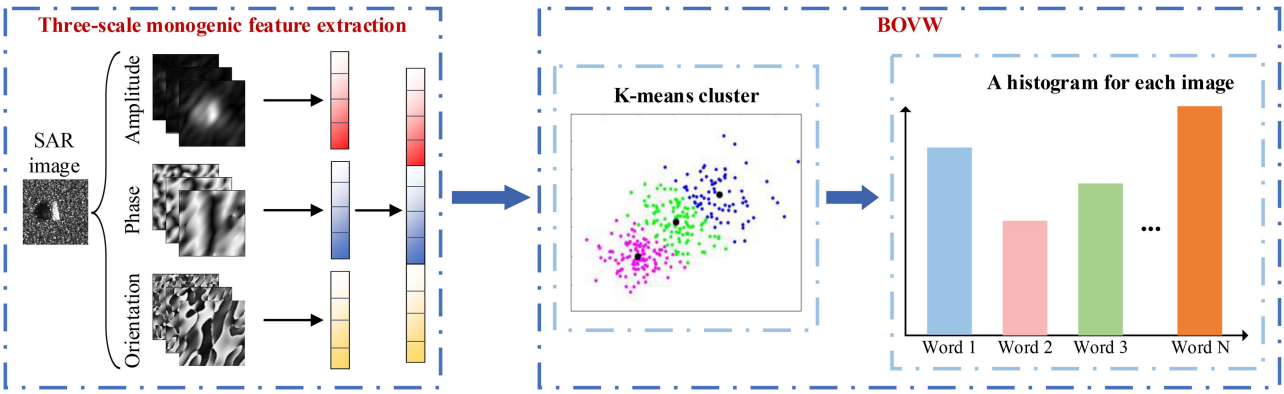

Fig. 2. Architecture of the Mono-BOVW model for handcrafted feature extraction.

denote the mean and the standard deviation of x, respectively. For POLSAR, each channel is independently normalized.

The second part includes three modules: handcrafted feature extraction based on the Mono-BOVW model, deep feature extraction based on CVNLNet, and feature fusion algorithm based on KDCA. First, we use the monogenic components generated by decomposition in different scale spaces to express the target scattering mechanism of SAR images, then extract lower dimensional mid-level semantic features by the BOVW model. Meanwhile, we construct CVNLNet by the inserting CV nonlocal block into the CV residual network (ResNet) to extract deep features. Subsequently, we use the KDCA algorithm to perform correlation analysis and transformation on the extracted handcrafted features and deep features for better fusion.

In the third part, the fusion features are fed into a classifier for training and classification. Since the support vector machine (SVM) classifier achieves high performance in a small number of samples, it serves as the classifier in this article.

A. *Handcrafted Feature Extraction Based on Mono-BOVW*

In the present section, the architecture of the Mono-BOVW model proposed for handcrafted feature extraction is elucidated, as presented in Fig. 2.

First, multiscale monogenic features are extracted from all training images. The monogenic signal is a two-dimensional (2-D) analytical signal, which describes the local amplitude, local orientation, and local phase information of the image in a rotation-invariant manner. It is based on the Riesz transform which is a 2-D extension of the Hilbet transform while retaining the important properties of 1-D analytical information. The Riesz transform spatial kernel function (h_x, h_y) at any point (x, y) in the 2-D signal space could be expressed as follows:

$$(h_x, h_y) = \left(\frac{x}{2\pi ||x||^3}, \frac{y}{2\pi ||y||^3} \right). \tag{1}$$

Since the Fourier spectrum period of the image is infinitely long, it is necessary to extend the input image infinitely by means

of a bandpass filter, and then perform the Riesz transform. This article uses Log-Gabor filter to achieve bandpass filtering. If the input image is I_0, the general form of the 2-D monogenic signal I_M could be expressed as follows:

$$I_M = (I, I_x, I_y) = (I, h_x * I, h_y * I),$$
$$I = I_0 * F^{-1}(G(\omega)) \tag{2}$$

where *I* denotes the extension of I_0; I_x and I_y denote the Riesz transform of *I* in the *x* and *y* direction, respectively. The operator "$*$" denotes convolution, F^{-1} denotes the inverse Fourier transform, $G(\omega)$ denotes the frequency response of the Log-Gabor filter which could be defined as follows:

$$G(\omega) = \exp(-(\log(\omega/\omega_0))^2/2(\log(\sigma/\omega_0))^2),$$
$$\omega_0 = (\lambda_{\min}\mu^{S-1})^{-1} \tag{3}$$

where *S* denotes the scale space of the monogenic signal, ω_0 denotes the central frequency, σ denotes the broadband proportional factor, $\lambda_{\min}$ denotes the minimum wavelength, and μ denotes the wavelength multiplication coefficient. Next, the local amplitude *A*, the local orientation θ, and the local phase *P* of the input image could be defined as follows:

$$A = \sqrt{I^2 + I_x^2 + I_y^2} \tag{4}$$

$$\theta = \arctan(I_y/I_x), \theta \in (-\pi/2, \pi/2] \tag{5}$$

$$P = \arctan\left(|\sqrt{I_x^2 + I_y^2}|, I\right), P \in (-\pi, \pi] \tag{6}$$

where *A*, θ, and *P* contain the local energy information, the local geometric information, and the local structure information.

Based on the S-scale log-Gabor filter, $\{I_M^1, I_M^2, \ldots, I_M^S\}$ denotes the monogenic signal under the condition of different scales, and the corresponding monogenic components are as follows:

$$\left\{\underbrace{A_1, \theta_1, P_1}_{I_M^1}, \underbrace{A_2, \theta_2, P_2}_{I_M^2}, \cdots, \underbrace{A_S, \theta_S, P_S}_{I_M^S}\right\}. \tag{7}$$

When $S = 3$, an SAR image can be characterized as three local amplitude maps, three local orientation maps, and three local phase maps. Then, we expand these feature maps into a long vector to form a monogenic feature vector, as shown in Fig. 2.

Subsequently, the BOVW model is adopted to perform statistical analysis on the distribution of monogenic feature vectors as the input feature descriptors (shown in Fig. 2). First, we make clusters from the descriptors. In the specific implementation, *K*-means is selected as the clustering algorithm. The center of each cluster will serve as a word of the visual dictionary. Next, for each image, a frequency histogram is built according to the visual vocabulary and the frequency of the words contained in this image. Then, the histogram is encoded to form the final feature vector.

B. Deep Feature Extraction Based on CVNLNet

In the present section, a CV network named CVNLNet is proposed to extract deep features, as illustrated in Fig. 3. It mainly includes two parts: CV ResNet and CV nonlocal block. Here, CV nonlocal block, as a separate module, can be inserted into any position in CV ResNet to form CVNLNet (represented by the dotted lines with arrows in Fig. 3). The network performance for different insertion positions in the experiments are discussed to determine the optimal network configuration.

The basic modules in CVNLNet such as convolutional layers, pooling layers, activation layers, and batch normalization layers are substituted with the relevant CV versions. These CV modules are elucidated within papers [38] and [39], in order to exploit both amplitude and phase in the POLSAR data. Notably, in the output layer, the complex features are transformed as real features by calculating the absolute value before the softmax classification which is not applicable to complex values.

1) CV ResNet: The main structure of CVNLNet is a deep CV ResNet. From [71], it is known that ResNet can address the degradation problem, thus achieving higher accuracy from considerably increased depth. Therefore, in order to prevent the gradient disappearance which may occur during the deep network training process, CV residual blocks (convolution blocks and identity blocks) are exploited. Within the CV convolution blocks, the stride of the first convolutional layer and shortcut connection layer is set as 2 to allow the input and output feature maps to have different sizes, inconsistent with that in the CV identity block. CV ResNet comprises three convolutional layers and four residual block groups (with two residual blocks each), as depicted in Fig. 3. In this network, inspired by [43], two convolutional layers are employed for replacing the fully connected layer to prevent overfitting and increase nonlinearity.

2) CV Nonlocal Block: Inspired by [72], a CV version of nonlocal block with high efficiency, as shown in Fig. 3, is presented and applied to our CV ResNet. It is capable of capturing long-range relationships, thus making up for the insufficiency that convolution operations deal with one local area each time [72]. The CV nonlocal block introduces global information by computing a weighted sum of the features at all positions in the feature maps, thereby making the target region more weighted and more prominent, to enhance the recognition performance.

The generic nonlocal operation in the RV domain could be expressed mathematically as follows:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{8}$$

where *x* and *y* respectively denote the input and output feature maps with the same size, *i* denotes a position on the feature map while *j* denotes possible positions of the enumeration, and $C(x)$ is a normalization function. *g* computes an expression for x_j by a 1×1 convolution: $g(x_j) = W_g\ x_j$. *f* computes the relationship between x_I and x_j, which is implemented in this article as an embedded Gaussian function which could be defined in the RV domain as follows:

$$f(x_i, x_j) = \exp(\theta(x_i)\varphi(x_j)^T) \tag{9}$$

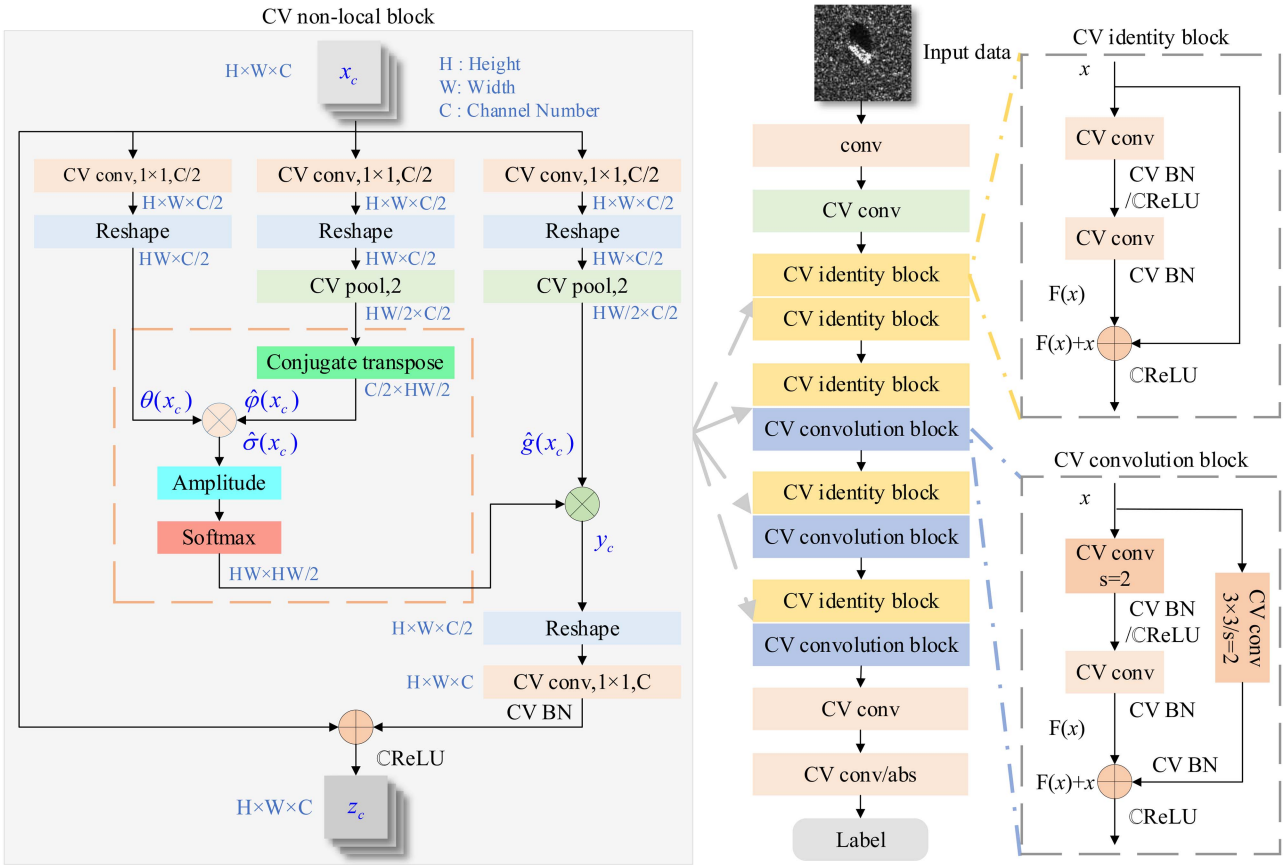

Fig. 3. Architecture of CVNLNet.

Where $\theta(x_j) = W_\theta \ x_j$, $\varphi(x_j) = W_\varphi \ x_j$, calculated by 1×1 convolutions. The superscript T denotes a transpose operation. Correspondingly, the function $C(x)$ is: $C(x) = \sum_{\forall j} f \ (x_I \ , y_j)$. Accordingly, I in (8) at alI i positions is expressed as follows:

$$y = \text{softmax}(\sigma(x))g(x) \tag{10}$$

$$\sigma(x) = \theta(x)\varphi(x)^T \tag{11}$$

where $\theta(x)$, $\varphi(x)$, and $g(x)$ denote the corresponding matrices calculated from x at all positions. A weight matrix is obtained from $\sigma(x)$ with the use of softmax function. Subsequently, this weight matrix is multiplied by the matrix $g(x)$ to obtain the weighted feature map y with the target area highlighted.

However, data are complex (denoted by a subscript c) in CV network. Thus, the conjugate transpose of $\varphi(x_c)$ replaces its transpose in (11), which is improved in the CV domain as follows:

$$\sigma(x_c) = \theta(x_c)\varphi(x_c)^H \tag{12}$$

where the superscript H denotes a conjugate transpose operation. Next, the absolute value of the complex $\sigma(x_c)$ is calculated for softmax. On that basis, (10) is modified as follows:

$$y_c = \text{softmax}(|\sigma_c|)g(x_c) \tag{13}$$

Where $|\sigma_c|$ denotes the absolute value of $\sigma(x_c)$, defined as follows:

$$|\sigma_c| = \sqrt{(\Re(\sigma(x_c))^2 + (\Im(\sigma(x_c))^2}. \tag{14}$$

To decrease computation, a downsampling trick is adopted, which is implemented as a pooling operation. Thereby, $\varphi(x_c)$ in (12) and $g(x_c)$ in (13) are improved as $\hat{\varphi}(x_c)$ and $\hat{g}(x_c)$, which are the corresponding downsampling versions.

According to the above derivation, the general CV nonlocal block in the CV domain could be lastly defined as follows:

$$z_c = W_z y_c + x_c \tag{15}$$

where W_z denotes a weight matrix in the 1×1 convolution. Moreover, to further decrease the computation, the number of channels in the convolutional layers containing W_θ , W_φ, or W_g is decreased to half of that of x_c. Next, the number of channels in the convolutional layer containing W_z is required to match that of x_c. Then, the output of the convolution layer containing W_z is added to the input x_c by a residual connection. Moreover, a $\mathbb{C}$ReLU activation is applied to z_c to enhance nonlinearity.

C. Feature Fusion Algorithm Based on KDCA

To effectively fuse handcrafted features and deep features, a fusion algorithm based on KDCA is developed (shown as Fig. 4). KDCA performs feature selection and dimensionality reduction

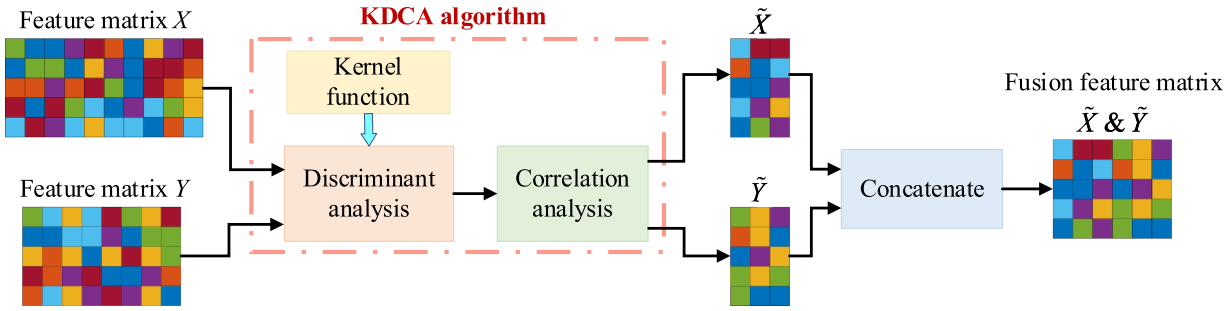

Fig. 4. Architecture of the feature fusion algorithm based on KDCA.

through projective transformation to avoid feature redundancy and fully uses the complementary advantages of the two features. First, the kernel function efficiently maps the original feature space to a higher dimensional feature space during the different feature fusion process, so that the nonlinear relationship is converted into linear relationship for subsequent correlation analysis. Moreover, the effects of different dimensions of the original features are suppressed due to mapping to the same spatial dimension. Then, DCA is capable of maximizing the correlation between features from the same category of targets in two feature sets and eliminating the correlation of features from different categories in the respective feature set.

First, the two features are normalized before fusion to make the scales the same for better fusion. The two feature matrices are assumed as $X \in \mathbb{R}^{p\times n}$ and $Y \in \mathbb{R}^{q\times n}$, where n denotes the number of samples, and p and q denote the dimensions of the two features. Moreover, all samples originate from c separate categories, and n_i denotes the number of samples belonging to the ith category. Thus, X_{ij} and Y_{ij} denote the feature vectors from the jth sample of the ith category. With the feature matrix X as an example, the mean of the ith category and entire feature set could be written as $\bar{X}_I$ and $\bar{X}$, respectively

$$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \tag{16}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i} X_{ij} = \frac{1}{n}\sum_{i=1}^{c} n_i\bar{X}_i. \tag{17}$$

The between-class scattering matrix S_{bx} could be defined as follows:

$$S_{bx} = \sum_{i=1}^{c} n_i(\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T = \Theta_x\Theta_x^T \tag{18}$$

$$\Theta_x = \left[\sqrt{n_1}(\bar{X}_1 - \bar{X}), \sqrt{n_2}(\bar{X}_2 - \bar{X}), \ldots, \sqrt{n_c}(\bar{X}_c - \bar{X})\right]. \tag{19}$$

In this step, considering the nonlinear relationship between different features, it is not ideal to perform linear analysis directly, so we innovatively introduce the kernel function to map the feature F to a linearly separable high-dimensional space using a nonlinear mapping $\Phi(\cdot)$: $\hat{F} = \Phi(F)$. The corresponding kernel function matrix K is constructed as follows:

$$K = \Phi(F)\Phi(F)^T = \hat{F}\cdot\hat{F}^T. \tag{20}$$

Therefore, S_{bx} in (18) is improved by the kernel function as follows:

$$\hat{S}_{bx} = \Phi_x(\Theta_x)\Phi_x(\Theta_x)^T = K_x \tag{21}$$

where $\Phi_x(\cdot)$ denotes a nonlinear mapping for x, and K_x denotes the kernel function matrix of $\Phi_x(\Theta_x)$.

When samples from different categories are strongly discriminative, $\hat{S}_{bx}$ would be a diagonalizable matrix

$$P^T(\hat{S}_{bx})P = \Lambda \tag{22}$$

where Λ denotes the diagonal matrix of real and non-negative eigenvalues which are sorted on the basis of an order of decrease. Moreover, P is a matrix consisting of m most significant eigenvectors corresponding to m largest eigenvalues. And then we can decrease the dimension of X from p to m by the transformation: $W_{bx} = \Phi_x(\Theta_x)P\Lambda^{-1/2}$, which could convert the scattering matrix $\hat{S}_{bx}$ to a identify matrix I: $W_{bx}^T\hat{S}_{bx}W_{bx} = I$. This means that after this transformation, the correlation between different categories is minimized, i.e., the categories are separated. Accordingly, X could be mapped by W_{bx} to X'

$$X' = W_{bx}^T X. \tag{23}$$

Likewise, we could get a transformation W_{by} which decreases the dimension of Y from q to m

$$Y' = W_{by}^T Y. \tag{24}$$

Then, in order that only the features from the same category of targets in two feature sets have nonzero correlation, the between-set covariance matrix $S'_{xy} = X'Y'^T$ need to be diagonalized. The singular value decomposition is adopted for the diagonalization operation

$$S'_{xy} = U\Sigma V^T \rightleftarrows U^T S'_{xy} V = \Sigma \tag{25}$$

where Σ is a diagonal matrix. We let the two transformations, respectively, be $W_{cx} = U\Sigma^{-1/2}$ and $W_{cy} = V\Sigma^{-1/2}$. Next, these two feature matrices can be transformed as follows:

$$\tilde{X} = W_{cx}^T X' = W_{cx}^T W_{bx}^T X \tag{26}$$

$$\tilde{Y} = W_{cy}^T Y' = W_{cy}^T W_{by}^T Y. \tag{27}$$

Through the two-stage transformation derived above, the two feature sets, with the largest distinction between different categories in the same set and the largest correlation of corresponding features between different sets, are lastly obtained.

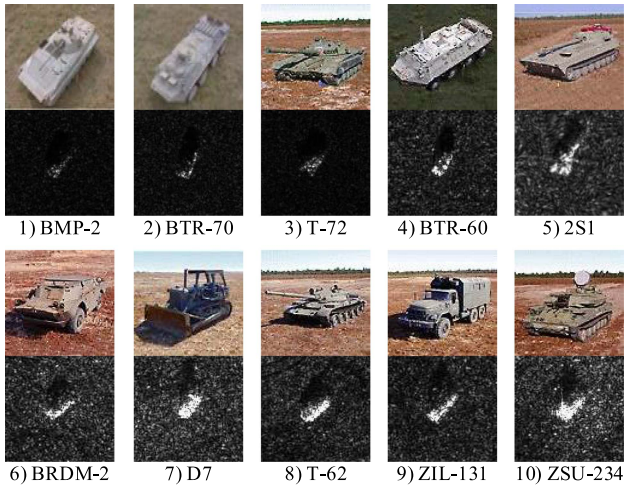

Fig. 5. Optical and SAR images of the ten categories of targets in MSTAR.

Then, the two transformed feature sets are concatenated to form a fusion feature set: $F = \widetilde{X} \ \& \ \tilde{Y}$.

III. Experiments and Discussions

In the present chapter, the effectiveness and superiority of the proposed recognition approach are verified in two public datasets, including the MSTAR dataset (single polarimetric) and the GOTCHA dataset (fully polarimetric), respectively. This article utilizes LIBSVM [73] to build SVM classifier.

A. Experiments on the MSTAR Dataset

1) MSTAR Dataset: The MSTAR dataset is the SAR ground stationary target dataset which are collected through a project funded by the US Air Force Research Laboratory (AFRL) [74], which is extensively employed for SAR target recognition. The dataset comprises ten categories of vehicle target slice images: armored personnel carriers (BTR-60, BTR-70), infantry fighting vehicle (BMP-2), armored scout car (BRDM-2), tanks (T-62, T-72), Bulldozer (D7), truck (ZIL-131), and rocket car (ZSU-234, 2S1). Fig. 5 presents the optical and SAR images of the ten categories of targets. All target images are uniformly cropped to 128×128 for experiment.

In the MSTAR dataset, ten categories of targets at the depression angle of 17° are selected for training, and ten categories of targets at 15° are selected for test. The details of the sample data are listed in Table I.

2) Experimental Setup for Handcrafted Feature Extraction: First, the amplitude of complex data is used to form a gray image in MSTAR. Subsequently, the three-scale monogenic features are extracted from the gray image according to the monogenic signal model, as shown in Fig. 6. In (3), the parameters are specifically set as follows: $S = 3, \sigma = 0.48, \lambda_{\min} = 8, \mu = 2.5$.

Next, the monogenic features extracted from the respective image are arranged as a long feature vector. And 99% of these feature vectors are selected for *K*-means clustering. The parameter *k* in *K*-means clustering algorithm is set as 2280. Then the centers of all clusters obtained are quantified as visual words. Finally, the visual words contained in each image are counted to form a visual word histogram, which is further encoded to obtain a 2279-dimensional handcrafted feature vector.

In order to verify the effectiveness of the proposed Mono-BOVW model, we compare it with the single-scale monogenic signal with BOVW model (single-scale Mono-BOVW) and the multiscale monogenic signal model in terms of the classification overall accuracy (OA) and the feature dimension, as listed in Table II. Accuracy is the most commonly used performance metric in machine learning, defined as follows:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}). \quad (28)$$

where TP denotes the number of the true positives, TN denotes the number of the true negatives, FP denotes the number of the false positives, and FN denotes the number of the false negatives.

TABLE I
Training Dataset and Test Dataset in MSTAR

Category	*Serial*	*Training*	*Test*
		17°	15°
BMP-2	9563	233	195
BTR-70	c71	233	196
T-72	132	232	196
BTR-60	7532	256	195
2S1	b01	299	274
BRDM-2	E-71	298	274
D7	13015	299	274
T-62	A51	299	273
ZIL-131	E12	299	274
ZSU-234	d08	299	274
Total		2747	2425

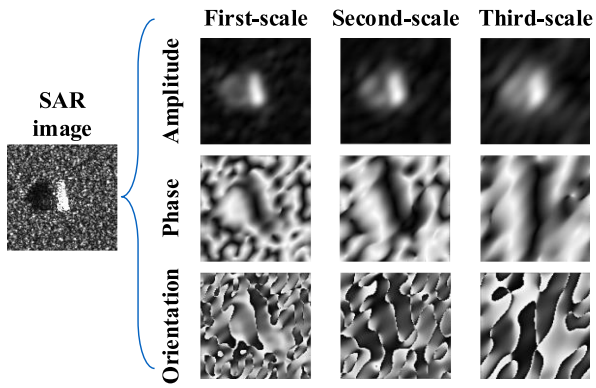

Fig. 6. Three-scale monogenic features of a SAR image.

TABLE II
Comparison of Performance for Mono-BOVW

Model	*Dimension*	*OA (%)*
Mono-BOVW	2279	98.60
Single-scale Mono-BOVW	2279	95.59
Multi-scale monogenic signal	2304	90.97

TABLE III
DETAILED CONFIGURATION OF CVNLNET FOR MSTAR

Layer	*Parameter (Size, Number, Stride)*	*Output*
conv1	7×7, 16, 2	64×64×16
pool	3×3, -, 2	32×32×16
Res1	$\begin{bmatrix}3\times3, 16, 1\\3\times3, 16, 1\end{bmatrix}$, (block), $\begin{bmatrix}3\times3, 16, 1\\3\times3, 16, 1\end{bmatrix}$	32×32×16
Res2	$\begin{bmatrix}3\times3, 32, 2\\3\times3, 32, 1\end{bmatrix}$, (block), $\begin{bmatrix}3\times3, 32, 1\\3\times3, 32, 1\end{bmatrix}$	16×16×32
Res3	$\begin{bmatrix}3\times3, 64, 2\\3\times3, 64, 1\end{bmatrix}$, (block), $\begin{bmatrix}3\times3, 64, 1\\3\times3, 64, 1\end{bmatrix}$	8×8×64
Res4	$\begin{bmatrix}3\times3, 128, 2\\3\times3, 128, 1\end{bmatrix}$, (block), $\begin{bmatrix}3\times3, 128, 1\\3\times3, 128, 1\end{bmatrix}$	4×4×128
conv2	3×3, 128, 1	2×2×128
conv3	2×2, 10, 1, abs, softmax	1×1×10

TABLE IV
CLASSIFICATION RESULTS BASED ON CVNLNET WITH DIFFERENT CONFIGURATIONS

Config.	*Base*	*Res1+*	*Res2+*	*Res3+*	*Res4+*
OA (%)	98.12	**99.50**	98.75	98.58	98.21

The bold entities indicate the best result of the comparison methods.

It can be seen that compared with the single-scale monogenic signal, the multiscale monogenic signal contains more information, so as to achieve a higher accuracy. And the introduction of the BOVW model into the multiscale monogenic signal greatly decreases the feature dimension, and improves the low-level semantic features to the mid-level semantic features with stronger representational ability, which is manifested in that the classification accuracy is greatly increased by 7.63%.

3) Experimental Setup for Deep Feature Extraction: For ease of analysis, experiments are performed in the MSTAR dataset for verifying the effectiveness of CVNLNet and discuss the optimal network model. Table III lists the detailed configuration of CVNLNet with all CV layers. The size of SAR image is 128 × 128 in the input layer with one polarization channel. The first layer refers to a convolutional layer, the second layer refers to an average pooling layer, followed by four stages (Res1, Res2, Res3, and Res4, respectively). Here, Res1 comprises two identity blocks, while Res2, Res3, and Res4 comprise an identity block and a convolution block, with the insertion of a CV non-local block before the last block at any stages. After the four stages, there are two convolutional layers. Notably, complex features should be converted to real features by calculating absolute value before softmax in the output layer. The hyperparameters in CVNLNet are set as follows: the number of epochs is 100, the batch size is 32, cross entropy loss serves as the loss function, and the adam algorithm serves as the optimizer with an initial learning rate of 0.001. Furthermore, dropout layer is employed for the prevention of overfitting.

Table IV lists the accuracy of CVNLNet with different configurations in the MSTAR dataset. Here, Base represents the base network (CV ResNet), and Res1+, Res2+, Res3+, and Res4+ represent CVNLNet with CV nonlocal block inserted into the corresponding stage. It is revealed that, compared to the base network, the accuracy of CVNLNet with CV nonlocal block at any stages is enhanced. However, as the insertion position is further back, the improvement in classification accuracy is smaller. The optimal classification accuracy occurs in Res1+, reaching 99.50%. This result is achieved probably because the deeper the network goes, the smaller the feature maps and the smaller the role of the CV nonlocal block will be. Fig. 7 presents the feature maps before and after the insertion at the Res1 stage. Notably, the target area in the map is highlighted while the background interference and noise are suppressed after inserting the CV nonlocal block, which improves the accuracy of target recognition by 1.38%. Thus, the CV nonlocal block is inserted at the Res1 stage of CV ResNet for constructing the optimal model of CVNLNet.

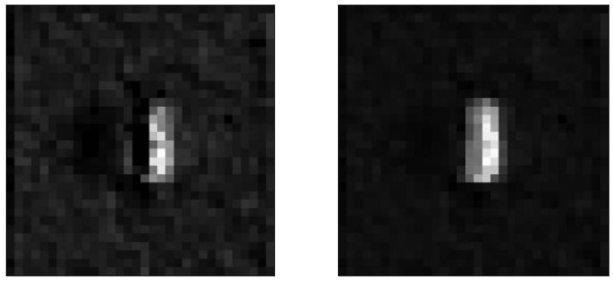

Fig. 7. Feature maps before (a) and after (b) inserting the CV non-local block in Res1 stage.

Next, the deep features are extracted using the optimal CVNLNet model. In general, the input feature maps of the output layer are selected and transformed into a 512-dimensional deep feature vector.

4) Recognition Based on the Feature Fusion Framework: The KDCA algorithm is adopted to fuse the extracted handcrafted feature vector and deep feature vector to form a stronger discriminative fusion feature vector. Next, this vector is fed into the SVM classifier for target classification.

For the KDCA fusion algorithm, a Gaussian kernel is selected as the kernel function, so the handcrafted feature vector and deep feature vector can be analyzed and fused in a suitable high-dimensional space. Lastly, we get a smaller 494-dimensional fusion feature vector, which avoids the redundancy problem in the fusion of the two feature vectors and decreases the subsequent computation. Correspondingly, for the SVM, the Gaussian kernel is selected as the kernel function.

Besides the Mono-BOVW model and CVNLNet, state-of-the-art SAR ATR methods are also cited for comparison, so as to examine the effectiveness and superiority of the proposed feature fusion framework. They include handcrafted feature-based methods, such as moment method [21], attributed scattering center model [22] and joint sparse representation (JSR) of monogenic components [30], as well as end-to-end neural networks, such as A-ConvNet [31], CV-CNN [39], CV-FCNN [43], and RVNLNet with the same architecture as CVNLNet. Furthermore, a fusion framework based on multiple handcrafted features [47], MKSFF-CNN based on fusion of multiscale deep features [55], and FEC based on fusion of handcrafted features and deep features [60] are also used for comparison. Table V

TABLE V
CLASSIFICATION ACCURACY OF DIFFERENT METHODS IN MSTAR

Feature Type		*Method*	*Features*	*Classifier*	*OA (%)*
single feature	handcrafted	Moment [21]	moment features	SVM	96.12
		ASC [22]	attributed scattering centers	matching	96.74
		JSR [30]	monogenic features	SRC	97.90
		Mono-BOVW	monogenic features	SVM	98.60
	deep	A-ConvNet [31]	amplitude values	softmax	98.10
		RVNLNet	amplitude values	softmax	98.35
		CV-CNN [39]	complex values	softmax	98.43
		CV-FCNN [43]	complex values	softmax	98.56
		CVNLNet	complex values	softmax	99.50
fusion feature	handcrafted + deep	Fusion framework [47]	random projection features + scattering centers	matching	99.01
		MKSFF-CNN [55]	amplitude values	softmax	97.40
		FEC [60]	scattering centers + amplitude values	RF	99.22
		Proposed	monogenic features + complex values	SVM	**99.71**

The bold entities indicate the best result of the comparison methods.

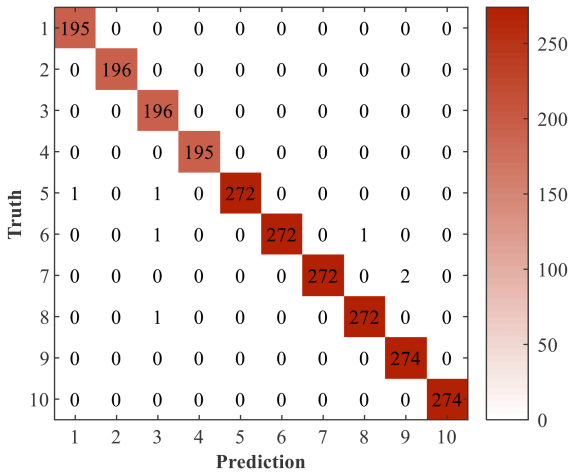

Fig. 8. Confusion matrix of the proposed framework in MSTAR.

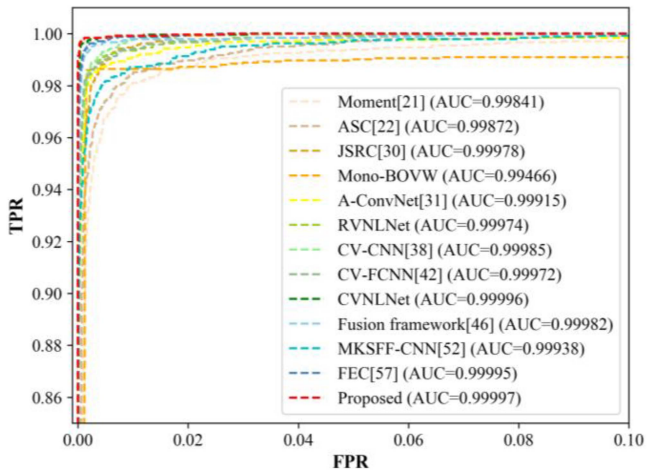

Fig. 9. ROC curves and AUC values of different methods in MSTAR.

lists the classification accuracy of different methods. And the confusion matrix of the proposed feature fusion framework is illustrated in Fig. 8, where each row represents the true category of the samples, each column represents the predicted category, and each cell lists the number of samples predicted as the corresponding category. Moreover, receiver operating characteristic (ROC) curve and area under the ROC curve (Area Under the ROC Curve, AUC) are also used to objectively evaluate the classification performance, as illustrated in Fig. 9. ROC curve is an important evaluation metric in machine learning, which describes the relationship between true positive rate (TPR) and false positive rate (FPR). AUC is usually used to assist ROC to further evaluate the performance (generally say, the larger the AUC, the better the performance of the classification method) [55]. Similar to (28), the specific definitions of TPR and FPR are as follows:

$$\begin{cases} \mathrm{TPR} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \\ \mathrm{FPR} = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN}). \end{cases} \tag{29}$$

Moreover, experiments are performed under the condition of smaller training datasets which are sampled from the original MSTAR dataset, to verify the ability of the proposed feature fusion framework to adapt to small datasets. The sampling proportions of small training datasets in the original MSTAR dataset account for 1/3, 1/5, 1/7, 1/10, respectively, and the corresponding classification accuracy is illustrated in Fig. 10.

As depicted in Table V, the proposed feature fusion framework achieves the highest accuracy of 99.71% in the MSTAR target recognition task. And from the specific classification results illustrated in Fig. 8, it can be seen that in the 10-category dataset containing 2425 samples, only 7 samples in total are misclassified. Moreover, the ROC curves and AUC values in Fig. 9 also show that the classification performance of the proposed feature fusion framework is higher than other methods.

First, as can be seen from Table V and Fig. 9, the proposed feature fusion framework significantly outperforms the single feature-based methods, especially compared with the proposed Mono-BOVW and CVNLNet. On one hand, as depicted in Table V and Fig. 9, compared with CVNLNet which extracts high-level semantic features through automatic learning, the Mono-BOVW model based on artificially designed low-level

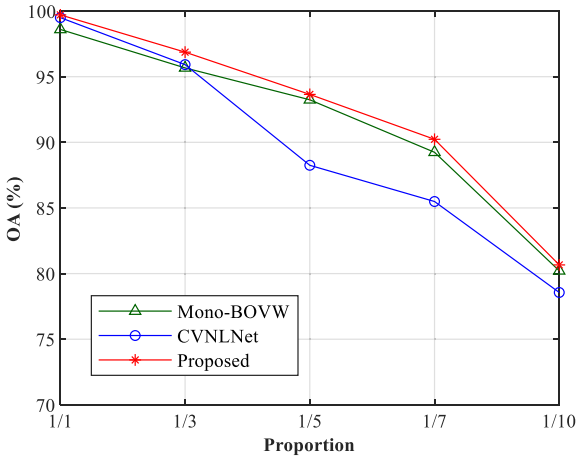

Fig. 10. Accuracy in different size of training datasets in MSTAR.

semantic features may be less adaptable to SAR data. However, the proposed fusion framework based on these two features exploits their complementary characteristics to the maximum extent through KDCA algorithm, so as to obtain stronger discriminative fusion features, which improves the classification accuracy by 1.11% and 0.21% compared with Mono-BOVW and CVNLNet, respectively. On the other hand, as depicted in Fig. 10, the classification accuracy of CVNLNet decreases much faster than that of Mono-BOVW when training samples are greatly reduced. This is because the learning process of the deep neural networks is more easily affected by the number of samples, and it is prone to overfitting when samples are insufficient. Accordingly, the proposed fusion framework can better make up for the deficiency of CVNLNet by introducing handcrafted features (extracted by Mono-BOVW) which are less affected by the number of samples due to its interpretability. Therefore, the classification performance can maintain a certain stability when the number of samples is small. To sum up, the proposed feature fusion framework can maximize the use of the automatic learning and strong representational ability of deep features and combine with the characteristics of handcrafted features, to obtain highly discriminative features which can more comprehensively characterize the target when the samples are sufficient; while the recognition performance can be kept stable to a certain extent, thanks to handcrafted features when the samples are insufficient.

Table V and Fig. 9 also reveals that in terms of fusion feature-based methods, compared to [47] and [55], paper [60] and the proposed framework have higher accuracy and AUC values, which can be attributed to the latter's utilization of complementary advantages of handcrafted features and deep features [60], while the former only fuses handcrafted features or deep features. The proposed feature fusion framework is slightly better than [60], which is due to the more effective use of the phase information of the SAR data through the monogenic signal and CVNLNet. Moreover, the accuracy of CVNLNe is 1.15% higher than that of RVNLNet, because the real-valued images as the input of RVNLNet only contain amplitude, while the complex-valued images as the input of CVNLNet contain both amplitude and phase and are effectively utilized.

TABLE VI
COMPARISON OF PERFORMANCE FOR KDCA

Algorithm	*Dimension*	*Time (s)*	*OA (%)*
KDCA fusion	494	12.11	99.71
DCA fusion	18	1.76	99.50
Serial fusion	2791	34.71	99.54

In addition, in order to verify the excellent performance of the proposed KDCA fusion algorithm, the classical serial fusion algorithm and the advanced DCA fusion algorithm are used for comparison. The feature dimension, the classification accuracy, and the overall classification runtime (representing computational complexity) are listed in Table VI. Obviously, the serial algorithm based on direct connection is too simple, so that the fusion features have redundancy and fail to effectively utilize the complementary advantages of different features, and too high dimension brings too much computation to the classifier. The DCA algorithm gets the least runtime, but the accuracy is not improved compared with CVNLNet. This is because the direct linear transformation of DCA only retains features with the dimension of c-1 (c denotes the number of categories) and lose much effective features, and the gains outweigh the losses. On the contrary, compared with the serial algorithm, the proposed KDCA decreases the dimension of the fusion features from 2791 to 494 by projective transformation, thereby shortening the runtime by 22.6 s; in addition, compared with DCA, KDCA introduces a kernel function used for nonlinear mapping before the discriminant correlation analysis to avoid the transformed fusion feature dimension being too low, so as to retain the effective discriminant information to obtain higher accuracy. To sum up, KDCA can achieve both lower overall computational complexity and higher classification accuracy.

B. Experiments on the GOTCHA Dataset

1) GOTCHA Dataset: The GOTCHA dataset is a fully polarimetric (including HH, HV, VH ,and VV polarization modes) dataset collected by AFRL, which comprises eight complete circular passes (covering the azimuth of 360°) with different depression angles [75]. The scene image comprises numerous calibration targets and ground civilian vehicles. The scene area of interest marked with nine categories of targets is shown in Fig. 11. The optical images and name of nine categories of vehicle targets are presented in Fig. 12.

For imaging, the complete circular aperture (360°) is separated in subaperture which has the azimuth of 4°. Thus, in each of the 8 passes, we can obtain 90 scene images. Then, according to the locations of all targets provided by the GORCHA dataset, nine categories of vehicle target images which have 50 × 50 pixels are selected from the scene image. We select images from pass1, 3, 5, and 7 for training (360 samples per category), and images from pass2, 4, 6, and 8 for test (360 samples per category), as listed in Table VII.

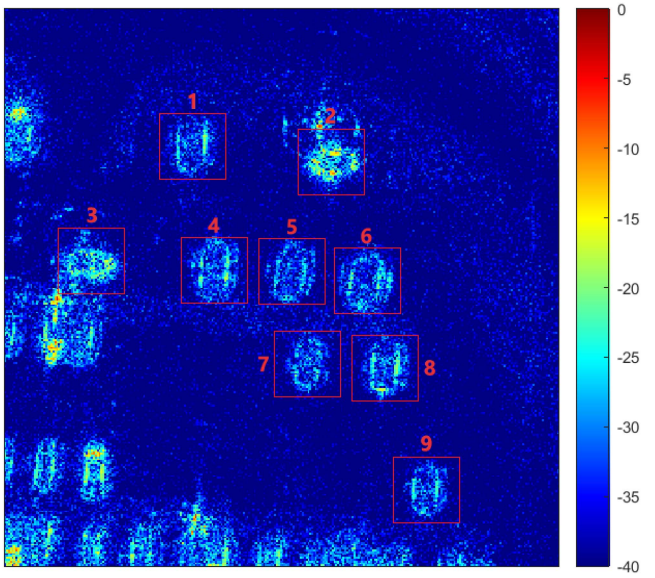

Fig. 11. Scene image (in dB) marked with nine categories of vehicle targets in GOTCHA.

Fig. 12. Optical images and names of nine categories of vehicle targets in GOTCHA.

TABLE VII
DETAILS OF TRAINING SET AND TEST SET IN GOTCHA

Dataset	*Category*	*Pass*	*Number*
Training set	1–9	1, 3, 5, 7	360×9
Test set	1–9	2, 4, 6, 8	360×9

2) Experimental Setup for Feature Extraction: The same as the MSTAR dataset, for the GOTCHA dataset, the Mono-BOVW model is used for handcrafted feature extraction, and CVNLNet (with the CV nonlocal block inserted into the Res1 stage) is used for deep feature extraction. The difference is that, as a fully polarimetric dataset, each image in the GOTCHA dataset has four polarization channels, which can describe the information of the target more comprehensively.

Prior to the Mono-BOVW model, according to the reciprocity principle (HV $\approx$ VH), the three polarization channels (HH, HV, VV) of each complex-valued image are taken as amplitude values to form a gray image. The parameters in the monogenic signal model are set as follows: $S = 3, \sigma = 0.48, \lambda_{\min} = 8, \mu = 2.5$. Then, we select 99% of these extracted monogenic feature vectors for *K*-means clustering (with the parameter *k* set as 3168). According to the BOVW model, a 3167-dimensional handcrafted feature vector is finally obtained.

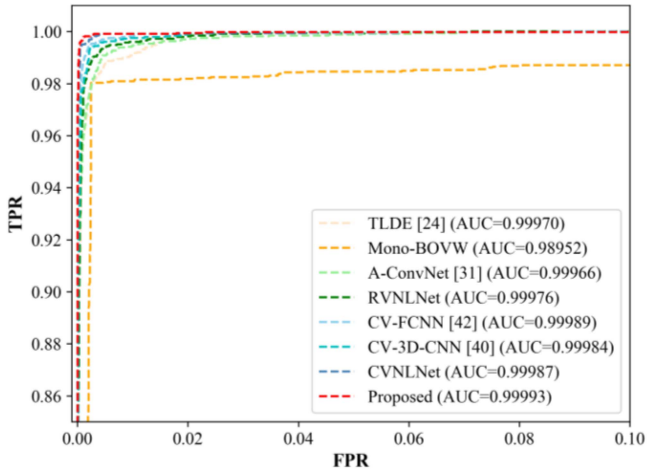

Fig. 13. ROC curves and AUC values of different methods in GOTCHA.

For CVNLNet, the structure is similar to that in the MSTAR dataset. But since the size of the input image in GOTCHA dataset is 50 × 50 (the image is smaller) and has four polarization channels, the size of the convolution kernel in the first convolutional layer of the network is set as 5, and the stride is set as 1. Subsequently, the above network is trained (the hyperparameters are the same as those in MSTAR), and a 512-dimensional deep feature vector is extracted for subsequent fusion.

3) Recognition Based on the Feature Fusion Framework: Similar to the experiments in the MSTAR dataset, a 356-dimensional fusion feature vector, which is extracted by the KDCA fusion algorithm, is fed into SVM classifier for classification, in accordance with the proposed feature fusion framework.

To verify the effectiveness and superiority of the proposed feature fusion in the GOTCHA dataset, in addition to Mono-BOVW and CVNLNet, some other state-of-the-art methods are also used for comparison. For the fully polarimetric SAR data different from single polarimetric SAR data, in order to better utilize the important phase relationship between multichannel data, the commonly used methods mainly include polarization decomposition algorithms and deep neural networks. For example, tensor local discriminant embedding (TLDE) based on multiple polarization decomposition [24], real-valued networks like A-ConvNet [31] and RVNLNet (corresponding to CVNLNet), complex-valued networks like CV-FCNN [43], CV-3D-CNN [41]. The classification accuracy (per class and overall) of these methods are listed in Table VIII, and the ROC curves and AUC values are illustrated in Fig. 13.

As can be seen from Table VIII, the proposed feature fusion framework achieves the highest overall accuracy of 99.63%, and the highest accuracy in almost every category. Fig. 13 also verifies the superior performance of the proposed framework. In particular, the classification accuracy of the proposed feature fusion framework is 1.61% and 0.19% higher than that of

TABLE VIII
CLASSIFICATION RESULTS IN GOTCHA

Accuracy	TLDE	Mono-BOVW	A-ConvNet	RVNLNet	CV-FCNN	CV-3D-CNN	CVNLNet	*proposed*
1	99.44	98.33	97.50	99.72	98.89	99.17	99.17	99.72
2	100.00	100.00	100.00	99.17	100.00	100.00	100.00	100.00
3	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
4	96.11	97.22	96.94	97.78	96.94	99.17	99.44	99.44
5	96.94	97.50	98.06	98.61	99.17	99.17	98.61	99.17
6	96.67	99.17	99.72	99.17	99.72	99.44	100.00	100.00
7	95.56	98.33	96.39	96.39	97.78	99.17	98.61	99.17
8	98.89	96.39	96.11	98.89	100.00	99.44	99.72	99.72
9	98.61	95.28	97.22	96.94	98.33	98.33	99.44	99.44
OA (%)	98.02	98.02	97.99	98.52	98.98	99.32	99.44	**99.63**

The bold entities indicate the best result of the comparison methods.

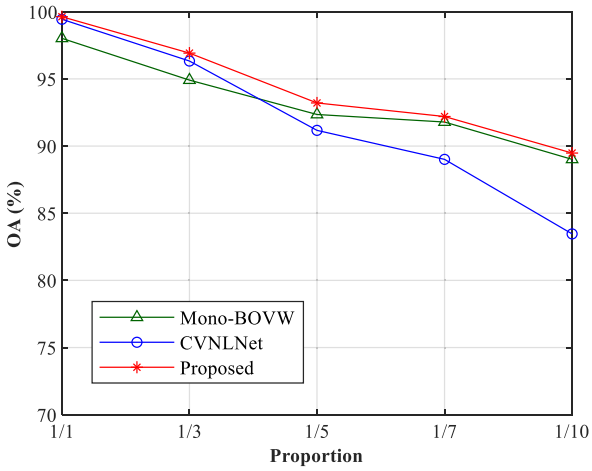

Fig. 14. Accuracy in different size of training datasets in GOTCHA.

Mono-BOVW and CVNLNet, which verifies that it effectively utilizes the complementary features of handcrafted features and deep features to obtain stronger discriminative fusion features. Combining with Fig. 14, it can be seen that when there are enough training samples, deep features play a key role in the framework to achieve higher accuracy, and when there are fewer samples, handcrafted features come into play to avoid the rapid degradation of the performance of the framework. Therefore, the proposed feature fusion framework can always have relatively stable and excellent recognition performance.

In addition, Table VIII also shows that compared with RVNLNet, CVNLNet can fully utilize the phase relationship between different polarization channels in fully polarimetric data, so it can achieve higher accuracy; and compared with other CV networks (CV-FCNN, CV-3D-CNN), CVNLNet performs better because the nonlocal block can make the target region get more attention in the whole image to enhance the ability of the network to extract features.

IV. CONCLUSION

In this article, a feature fusion framework is proposed based on monogenic signal and complex-valued nonlocal network for POLSAR target recognition, which effectively uses the complementary advantages of handcrafted features and deep features to make up for the lack of representation ability of a single feature. First, a Mono-BOVW model is employed to extract robust handcrafted features, and a CVNLNet network is constructed to extract deep features with strong representational ability. Subsequently, the two features are analyzed and transformed based on the proposed KDCA algorithm to form the stronger discriminative fusion features with lower dimension after redundancy removal. In both the single polarimetric MSTAR dataset and the fully polarimetric GOTCHA dataset, the proposed framework achieves a high classification accuracy and exhibits good adaptability to small sample datasets. This article reveals that the proposed feature framework is promising and takes on a critical significance in SAR-ATR.

REFERENCES

[1] L. Landuyt, F. M. B. Van Coillie, B. Vogels, J. Dewelde, and N. E. C. Verhoest, "Towards operational flood monitoring in flanders using Sentinel-1," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11004–11018, Oct. 2021, doi: 10.1109/JSTARS.2021.3121992.

[2] H. Yu, C. Wang, J. Li, and Y. Sui, "Automatic extraction of green tide from GF-3 SAR images based on feature selection and deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10598–10613, Oct. 2021, doi: 10.1109/JSTARS.2021.3118374.

[3] Y. Chen, Y. Tong, and K. Tan, "Coal mining deformation monitoring using SBAS-InSAR and offset tracking: A case study of Yu County, China," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6077–6087, Oct. 2020, doi: 10.1109/JSTARS.2020.3028083.

[4] Y. Zhou, T. Wei, X. Zhu, and M. Collin, "A parcel-based deep-learning classification to map local climate zones from sentinel-2 images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4194–4204, Apr. 2021, doi: 10.1109/JSTARS.2021.3071577.

[5] B. Zhao, Y. Han, H. Wang, L. Tang, X. Liu, and T. Wang, "Robust shadow tracking for video SAR," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 821–825, May 2021.

[6] X. Hu, Y. Li, M. Lu, Y. Wang, and X. Yang, "A multi-carrier-frequency random-transmission chirp sequence for TDM MIMO automotive radar," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3672–3685, Apr. 2019.

[7] J. Ai et al., "Robust CFAR ship detector based on bilateral-trimmed-statistics of complex ocean scenes in SAR imagery: A closed-form solution," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 3, pp. 1872–1890, Jun. 2021.

[8] L. Tang, W. Tang, X. Qu, Y. Han, W. Wang, and B. Zhao, "A scale-aware pyramid network for multi-scale object detection in SAR images," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 973.

[9] J.-S. Lee and E. Pottier, *Polarimetric Radar Imaging: From Basics to Applications*. Boca Raton, FL, USA: CRC Press, 2017.

[10] T. Zhang, J. Ji, X. Li, W. Yu, and H. Xiong, "Ship detection from PolSAR imagery using the complete polarimetric covariance difference matrix," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2824–2839, May 2019.

[11] J. Ai et al., "A fine PolSAR terrain classification algorithm using the texture feature fusion-based improved convolutional autoencoder," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5218714, doi: 10.1109/TGRS.2021.3131986.
[12] L. M. Novak, S. D. Halversen, G. Owirka, and M. Hiett, "Effects of polarization and resolution on SAR ATR," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 33, no. 1, pp. 102–116, Jan. 1997.
[13] C. Clemente, L. Pallotta, I. Proudler, A. De Maio, J. J. Soraghan, and A. Farina, "Pseudo-Zernike-based multi-pass automatic target recognition from multi-channel synthetic aperture radar," *IET Radar, Sonar Navig.*, vol. 9, no. 4, pp. 457–466, 2015.
[14] S. Ohno, S. Kidera, and T. Kirimoto, "Automatic target recognition method based on polsar images with circular polarimetric basis conversion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 3243–3246.
[15] Moussa Amrani, "Research and development of radar target classification algorithms," M.A. thesis, Harbin Inst. Technol., Harbin, China, 2018.
[16] O. Kechagias-Stamatis and N. Aouf, "Automatic target recognition on synthetic aperture radar imagery: A survey," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 3, pp. 56–81, Mar. 2021.
[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.
[19] C. Liu and H. Wechsler, "Gabor feature-based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.
[20] Z. Zhao, Y. Han, T. Xu, X. Li, H. Song, and J. Luo, "A reliable and real-time tracking method with color distribution," *Sensors*, vol. 17, no. 10, 2017, Art. no. 2303.
[21] P. Bolourchi, M. Moradi, H. Demirel, and S. Uysal, "Improved SAR target recognition by selecting moment methods based on fisher score," *Signal, Image Video Process.*, vol. 14, no. 1, pp. 39–47, 2020.
[22] B. Ding, G. Wen, X. Huang, C. Ma, and X. Yang, "Target recognition in synthetic aperture radar images via matching of attributed scattering centers," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3334–3347, Jul. 2017.
[23] R. Touzi, "Polarimetric target scattering decomposition: A review," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 5658–5661.
[24] X. Huang, H. Qiao, B. Zhang, and X. Nie, "Supervised polarimetric SAR image classification using tensor local discriminant embedding," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2966–2979, Jun. 2018.
[25] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: A SIFT-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 453–466, Jan. 2015.
[26] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Trans. Signal Process.*, vol. 49, no. 12, pp. 3136–3144, Dec. 2001.
[27] G. Dong, G. Kuang, N. Wang, L. Zhao, and J. Lu, "SAR target recognition via joint sparse representation of monogenic signal," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 7, pp. 3316–3328, Jul. 2015.
[28] G. Dong and G. Kuang, "Classification on the monogenic scale space: Application to target recognition in SAR image," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2527–2539, Aug. 2015.
[29] G. Dong and G. Kuang, "SAR target recognition via sparse representation of monogenic signal on grassmann manifolds," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 3, pp. 1308–1319, Mar. 2016.
[30] Y. Zhou, Y. Chen, R. Gao, J. Feng, P. Zhao, and L. Wang, "SAR target recognition via joint sparse representation of monogenic components with 2D canonical correlation analysis," *IEEE Access*, vol. 7, pp. 25815–25826, 2019.
[31] 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.
[32] J. Wang, T. Zheng, P. Lei, and X. Bai, "Ground target classification in noisy SAR images using convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4180–4192, Nov. 2018.
[33] R. Min, H. Lan, Z. Cao, and Z. Cui, "A gradually distilled CNN for SAR target recognition," *IEEE Access*, vol. 7, pp. 42190–42200, 2019.
[34] J. Yu, G. Zhou, S. Zhou, and J. Yin, "A lightweight fully convolutional neural network for SAR automatic target recognition," *Remote Sens.*, vol. 13, no. 15, 2021, Art. no. 3029.
[35] L. Zhang, Y. Li, Y. Wang, J. Wang, and T. Long, "Polarimetric HRRP recognition based on ConvLSTM with self-attention," *IEEE Sens. J.*, vol. 21, no. 6, pp. 7884–7898, Mar. 2021.
[36] L. Zhang, C. Han, Y. Wang, Y. Li, and T. Long, "Polarimetric HRRP recognition based on feature-guided transformer model," *Electron. Lett.*, vol. 57, no. 18, pp. 705–707, 2021.
[37] C. Deng, D. Jing, Y. Han, S. Wang, and H. Wang, "FAR-Net: Fast anchor refining for arbitrary-oriented object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 6505805, doi: 10.1109/LGRS.2022.3144513.
[38] C. Trabelsi et al., "Deep complex networks," in *Proc. Int. Conf. Learn. Repr.*, 2018, pp. 1–19.
[39] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin, "Complex-valued convolutional neural network and its application in polarimetric SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7177–7188, Dec. 2017.
[40] A. G. Mullissa, C. Persello, and A. Stein, "PolSARNet: A deep fully convolutional network for polarimetric SAR image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 5300–5309, Dec. 2019.
[41] X. Tan, M. Li, P. Zhang, Y. Wu, and W. Song, "Complex-Valued 3-D convolutional neural network for PolSAR image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1022–1026, Jun. 2020.
[42] P. Zhang et al., "PolSAR image classification using hybrid conditional random fields model based on complex-valued 3-D CNN," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 3, pp. 1713–1730, Jun. 2021.
[43] L. Yu, Y. Hu, X. Xie, Y. Lin, and W. Hong, "Complex-Valued full convolutional neural network for SAR target classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1752–1756, Oct. 2020.
[44] T. Scarnati and B. Lewis, "Complex-valued neural networks for synthetic aperture radar image classification," in *Proc. IEEE Radar Conf.*, 2021, pp. 1–6, doi: 10.1109/RadarConf2147009.2021.9455316.
[45] A. Ismail, X. Gao, and C. Deng, "SAR image classification based on texture feature fusion," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, 2014, pp. 153–156.
[46] Y. Han, C. Deng, Z. Zhang, J. Li, and B. Zhao, "Adaptive feature representation for visual tracking," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 1867–1870.
[47] B. Ding, G. Wen, C. Ma, and X. Yang, "An efficient and robust framework for SAR target recognition by hierarchically fusing global and local features," *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 5983–5995, Dec. 2018.
[48] B. Ding and G. Wen, "Combination of global and local filters for robust SAR target recognition under various extended operating conditions," *Inf. Sci.*, vol. 476, pp. 48–63, 2019.
[49] Y. Han, C. Deng, B. Zhao, and D. Tao, "State-Aware anti-drift object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4075–4086, Aug. 2019.
[50] Y. Han, C. Deng, B. Zhao, and B. Zhao, "Spatial-temporal context-aware tracking," *IEEE Signal Process. Lett.*, vol. 26, no. 3, pp. 500–504, Mar. 2019.
[51] H. Liu, W. Chen, F. Li, and T. Long, "SAR target recognition based on texture feature and contour feature fusion," in *Proc. IET Int. Radar Conf.*, 2020, pp. 571–575.
[52] B. Zhao, B. Zhao, L. Tang, Y. Han, and W. Wang, "Deep spatial-temporal joint feature representation for video object detection," *Sensors*, vol. 18, no. 3, 2018, Art. no. 774.
[53] X. Xu, X. Zhang, and T. Zhang, "Multi-Scale SAR ship classification with convolutional neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4284–4287.
[54] S. Wang, Y. Wang, H. Liu, and Y. Sun, "Attribute-Guided multi-scale prototypical network for few-shot SAR target classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12224–12245, Nov. 2021, doi: 10.1109/JSTARS.2021.3126688.
[55] J. Ai, Y. Mao, Q. Luo, L. Jia, and M. Xing, "SAR target classification using the multikernel-size feature fusion-based convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5214313, doi: 10.1109/TGRS.2021.3106915.
[56] V. Chandrasekhar, L. Jie, O. Morere, H. Goh, and A. Veillard, "A practical guide to CNNs and fisher vectors for image instance retrieval," *Signal Process.*, vol. 128, pp. 426–439, 2016.
[57] Y. Ke, Y. Wang, D. Liang, T. Huang, and Y. Tian, "CNN vs. SIFT for image retrieval: Alternative or complementary?," in *Proc. ACM Multimedia Conf.*, 2016, pp. 407–411, doi: 10.1145/2964284.2967252.
[58] 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.

[59] Z. Jia, D. Guangchang, C. Feng, X. Xiaodan, Q. Chengming, and L. Lin, "A deep learning fusion recognition method based on SAR image data," *Procedia Comput. Sci.*, vol. 147, pp. 533–541, 2019.
[60] J. Zhang, M. Xing, and Y. 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.
[61] S. Feng, K. Ji, L. Zhang, X. Ma, and G. Kuang, "SAR target classification based on integration of ASC parts model and deep learning algorithm," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10213–10225, 2021.
[62] Y. Ding, "Research on high-resolution polarimetric SAR image classification based on multi-feature fusion," MA thesis, Univ. Electron. Sci. Technol. China, Chengdu, Sichuan, China, 2017.
[63] J. Yang, J.-Y. Yang, D. Zhang, and J.-F. Lu, "Feature fusion: Parallel strategy vs. serial strategy," *Pattern Recognit.*, vol. 36, no. 6, pp. 1369–1381, 2003.
[64] G. Turhan-Sayan, "Real time electromagnetic target classification using a novel feature extraction technique with PCA-based fusion," *IEEE Trans. Antennas Propag.*, vol. 53, no. 2, pp. 766–776, Feb. 2005.
[65] G. Bai, W. Jia, and Y. Jin, "Facial expression recognition based on fusion features of lbp and gabor with lda," in *Proc. 2nd Int. Congr. Image Signal Process.*, 2009, pp. 1–5, doi: 10.1109/CISP.2009.5304655.
[66] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognit.*, vol. 38, no. 12, pp. 2437–2448, 2005.
[67] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[68] X. Xiaona, Z. Yue, and P. Xiuqin, "Multimodal recognition fusing ear and profile face based on kpca," in *Proc. 2nd Int. Symp. Syst. Control Aerosp. Astronaut.*, 2008, pp. 1–5, doi: 10.1109/ISSCAA.2008.4776247.
[69] J.-Y. Gan and J. F. Liu, "Fusion and recognition of face and iris feature based on wavelet feature and KFDA," in *Proc. Int. Conf. Wavelet Anal. Pattern Recognit.*, 2009, pp. 47–50.
[70] X. N. Xu, "Feature fusion method based on KCCA and multi-modality recognition fusing ear and face profile information," *J. South China Univ. Technol. (Natural Sci. Ed.)*, vol. 36, no. 9, pp. 117–121, 2008.
[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[72] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[73] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.
[74] T. D. Ross, S. W. Worrell, V. J. Velten, J. C. Mossing, and M. L. Bryant, "Standard SAR ATR evaluation experiments using the MSTAR public release data set," *Proc. SPIE*, vol. 3370, pp. 566–573, 1998.
[75] E. Ertin, C. D. Austin, S. Sharma, R. L. Moses, and L. C. Potter, "GOTCHA experience report: Three-dimensional SAR imaging with complete circular apertures," *Proc. SPIE*, vol. 6568, pp. 9–20, 2007.

Feng Li was born in Jilin, China, in 1978. He received the B.S. degree in electronic engineering from Harbin Engineering University, Harbin, China, in 2009, the M.S. degree in electronic engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2004, and the Ph.D. degree in electronic engineering from Xidian University, Xi'an, China, in 2008.

From 2009 to 2011, he was a Postdoctoral Researcher with the Beijing Institute of Technology, Beijing, China, where he is currently a Lecturer. His major research interests include synthetic aperture radar (SAR) imaging, inverse SAR, and target recognition.

Min Yi was born in Sichuan, China, in 1996. She received the B.S. degree in electronic information science and technology from Nankai University, Tianjin, China, in 2020. She is currently working toward the M.S. degree in information and communication engineering with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China.

Her research interests include synthetic aperture radar target recognition.

Chaoqi Zhang was born in Shan'xi, China, in 1999. He received the B.S. degree degree in electronic information engineering from Xidian University, Xi'an, China, in 2021. He is currently working toward the M.S. degree in information and communication engineering with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China.

His research interests include synthetic aperture radar target recognition.

Weijun Yao was born in Inner Mongolia, China, in 1998. He received the B.S. degree in communication engineering from Jilin University, Jilin, China, in 2019. He is currently working toward the M.S. degree in information and communication engineering with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China.

His research interest focuses on synthetic aperture radar target recognition.

Xueyao Hu was born in Shan' xi, China, in 1990. He received the B.S. degree in detection guidance and control technology from the Nanjing University of Science and Technology, Jiangsu, China, in 2013, and the Ph.D. degree in target detection and recognition from the Beijing Institute of Technology, Beijing, China, in 2020.

He is currently a Postdoctoral Research Associate with the School of Information and Electronics, Beijing Institute of Technology. His research interests include millimeter-wave radar system design, radar signal processing, and sparse recovery.

Feifeng Liu received the B.S. degree in mathematics and the Ph.D. degree in target detection and recognition from the Beijing Institute of Technology, Beijing, China, in 2002 and 2013, respectively.

He joined the Microwave Integrated System Laboratory, University of Birmingham, Birmingham, U.K., as a Visiting Student, in 2010. Since 2013, he has been a Faculty Member with Beijing Institute of Technology. His research interests include imaging technology of SAR and radar signal processing, as well as change detection technology based on space–surface BiSAR system.