New SAR target recognition based on YOLO and very deep multi-canonical correlation analysis

Moussa Amrani, Abdelatif Bey and Abdenour Amamra
Ecole Militaire Polytechnique, Chahid Abderrahmane Taleb, Algiers, Algeria

ABSTRACT
Synthetic Aperture Radar (SAR) images are prone to be contaminated by noise, which makes it very difficult to perform target recognition in SAR images. Inspired by great success of very deep convolutional neural networks (CNNs), this paper proposes a robust feature extraction method for SAR image target classification by adaptively fusing effective features from different CNN layers. First, YOLOv4 network is fine-tuned to detect the targets from the respective MF SAR target images. Second, a very deep CNN is trained from scratch on the moving and stationary target acquisition and recognition (MSTAR) database by using small filters throughout the whole net to reduce the speckle noise. Besides, using small-size convolution filters decreases the number of parameters in each layer and, therefore, reduces computation cost as the CNN goes deeper. The resulting CNN model is capable of extracting very deep features from the target images without performing any noise filtering or pre-processing techniques. Third, our approach proposes to use the multi-canonical correlation analysis (MCCA) to adaptively learn CNN features from different layers such that the resulting representations are highly linearly correlated and therefore can achieve better classification accuracy even if a simple linear support vector machine is used. Experimental results on the MSTAR dataset demonstrate that the proposed method outperforms the state-of-the-art methods.

1. Introduction
Synthetic aperture radar (SAR) is a very high-resolution airborne and space-borne remote sensing system for imaging distant targets on a terrain, which can operate proficiently in all-weather day-and-night conditions and generate images of extremely high resolution. A SAR system sends electromagnetic pulses from radar mounted on a moving platform to a fixed particular area of interest on the target and combines the returned signals coherently to achieve a very high-resolution depiction of the scene.

SAR has been substantially employed for many applications, such as surveillance, reconnaissance, and classification. However, speckle greatly disrupts SAR image readability, which makes defining discriminative and descriptive features a hard task. To address this problem, several feature processing methods have been advanced and...
utilized to understand the target from SAR images, such as geometric descriptors and transform-domain coefficients (Olson and Huttenlocher 1997). Although the above-mentioned methods may have some advantages, most of these methods are hand-designed and relatively simple (Amrani et al. 2017). Besides, they failed to achieve the promising classification performance (i.e. accuracy) (Amrani et al., 2017, September). Recently, deep convolutional neural networks (DCNNs) have been referenced by several authors to design classification algorithms for SAR images (Chen et al. 2016; Srivastava et al. 2014). A deep neural network is an artificial network with multiple hidden layers between the input and the output. In Chen et al. (2016), the authors proposed A-ConvNets, which consists of sparsely connected layers to reduce the number of free parameters from over-fitting. After implementing data augmentation, the network is trained using mini-batch stochastic gradient descent with momentum and back-propagation algorithm. Then, a soft-max layer is applied to output a probability distribution over class labels. One of the uncertain points of this method is that the convolutional layers suffer less from over-fitting because they have smaller number of parameters compared to the number of activations. Therefore, adding dropout to convolutional layers slows down the training (Srivastava et al. 2014). In addition, they employed cropped patches of $88 \times 88$ from the original SAR images in the training phase (i.e. as data augmentation) to deal with the translation invariance of DCNN for SAR-ATR system (Furukawa 2017). The other well-known deep learning-based algorithm (Chen and Wang 2014) focuses on solving the classification problem by learning randomly sampled image patches using unsupervised sparse auto-encoder instead of using the classical back-propagation algorithm. Then, a single layer of convolutional neural network is used to automatically learn features from SAR images. These feature maps are then adopted to train the final soft-max classifier, which results in a little low classification accuracy (Du et al. 2016). The research in Amrani and Ji (2017) proposes to select deep features for SAR target classification task, in which the top layers of CNNs contain more semantic information and describe the global features of the images, whereas the intermediate layers describe the local features, and the bottom layers contain more low-level information for the description of texture, edges, etc. However, the authors used large kernels in the first and the second layers, which increase the number of parameters and have less discriminative decision functions. Moreover, they adopted a pre-processing method to remove some noise from target images and metric learning for feature selection to adjust the accuracy level, which is time consuming.

More recently, very deep CNNs using small-size convolution filters have been used by Ciresan et al. (Ciresan et al. 2011). However, their networks are significantly less deep and they have not been evaluated on SAR images. Goodfellow et al. (Goodfellow et al. 2013) applied very deep ConvNets (11 weight layers) to street number recognition tasks and have depicted that increasing the depth of the network produces better performance. Szegedy et al. (Szegedy et al. 2014), an ILSVRC-2014 top-performing classification task, has developed independently a network based on very deep ConvNets (22 weight layers) and small convolution filters. However, their network is more complex and the spatial resolution of the feature maps in the first layers is sharply reduced to decrease the computation amount (Simonyan and Zisserman 2014). C.P. Schwegmann et al. (Schwegmann et al. 2016) have introduced very deep features for ship discrimination in SAR images. When
trained on a small database, their proposed very deep high network can provide better classification performance than conventional DCNNs.

Compared to the handcrafted and deep features, our approach uses very deep features to have more powerful discriminative and robust representation abilities, and multi-canonical correlation analysis (MCCA) algorithm to maximize the correlation among the feature sets, remove irrelevant features, and overcome the curse of dimensionality issue. The contributions of the proposal method mainly include two aspects.

- YOLOv4 is used for SAR target detection. Besides, a very deep network is proposed for SAR image target classification, which uses very small receptive small filter sizes throughout the whole net to reduce the speckle noise that degrades the quality of SAR images, and therefore achieves better performance as we go deeper.

- Multi-Canonical Correlation Analysis (MCCA) is proposed to adaptively select and fuse CNN features from different layers and such that the resulting representations are highly linearly correlated and therefore remove irrelevant features, speed up the training task, and improve the classification accuracy. As a result, we come up with particularly more accurate CNN-SAR architecture, which achieves the state-of-the-art accuracy on MSATR SAR target recognition tasks even when adopted as a relatively simple pipeline.

The rest of this paper is organized as follows. Section 2 describes the framework of the proposed method and introduces the feature fusion method using MCCA. The experimental results are presented in Section 3. Finally, Section 4 gives the concluding remarks of the paper.

2. YOLOv4 target detection

YOLOv4 combines many advanced target detection techniques to improve the accuracy and running speed of CNN as it is clarified in Figure 1. It can be run in real-time on a typical GPU, which makes it widely used. Yolo V4 consists of four parts. The main methods and tricks

![Figure 1. The architecture of YOLOv4.](image-url)
utilized in each part are as follows (Bochkovskiy, Wang, and Liao 2020): **Input**: Mosaic data augmentation, cross mini batch normalization (CmBN), and self-adversarial training (SAT). **BackBone**: Cross stage partial connections Darknet53 (CSPDarknet53), mish-activation, and dropblock regularization. **Neck**: SPP, modified feature pyramid network (FPN), path aggregation network (PAN). Prediction: Modified complete IOU (C-IOU) loss, distance IOU (D-IOU) nms. Some of these tricks can obviously improve the network performance. Mosaic data augmentation mixes four training images with clipping and scaling, which significantly reduces the requirement for large mini-batch processing and GPU computing. SAT alters the original SAR image to create the deception of undesired object. These modified images are used for network training in next stage, which could improve the robustness of the network. Cross stage partial connections split the gradient flow propagate into different network paths, which greatly reduces the amount of computation, and improves the inference speed of the network. Other tricks also improve the other details of the networks. For example, SPP, FPN and PAN improve the feature extraction. Modified C-IOU loss and DIOU nms improve the IOU loss to achieve better convergence speed and accuracy of regression problem.

3. The proposed method for SAR target recognition

Inspired by great success of very deep convolutional neural networks (VDCNNs), this paper presents a robust feature extraction method for SAR target classification using very deep features without performing any noise filtering or pre-processing techniques. **Figure 2** shows the main steps of the proposed network: after targets detection from the respective SAR target images, SAR-oriented network (SARON) is used for extracting very deep features from detected data. Then the resulting deep features are then selected and fused by MCCA. Finally, the classification of the targets with respect to its training set is done according to the classification error rates using SVM. In the following sections, each step of the proposed method is described in detail.

![Figure 2. Overall architecture of the proposed method.](image-url)
3.1. **SAR-oriented network**

Using deep neural networks to learn effective feature representations has become popular in SAR target classification (Chen et al. 2016; Srivastava et al. 2014; Wagner 2016; Zhou et al. 2018). It is also employed in the military field as automatic target recognition (ATR) (Mason, Yonel, and Yazici 2017). In contrast to most DCNNs that usually have five or seven layers, the proposed network is based on the very deep VGG net for feature extraction, which has a much deeper architecture (up to 19 weight layers) and hence can provide much informative and descriptive features (Simonyan and Zisserman 2014). To achieve better classification performance, the network is fine-tuned from scratch on the MSTAR database and then the fine-tuned network is treated as a fixed feature descriptor for SAR target images.

Figure 3 shows the architecture of the proposed very deep CNN model, which is a stack of convolutional layers (Conv.) followed by three fully-connected (Fc.) layers, and each stack of Conv. is followed by a max pooling layer (Pool.). Besides, all hidden layers are supplied with ReLU. Fc.1 is regularized using dropout technique, while the last layer acts as a C-class SVM classifier. Figure 4 shows the outputs from the first convolutional layer corresponding to a sample SAR image from the MSTAR dataset.

![Figure 3. The proposed SAR-oriented very deep network for target image classification.](image)

![Figure 4. (Left) Output feature maps and (right) Learned kernels of the first layer.](image)
3.1.1. Configuration

The network used in this paper has 19 weight layers (16 conv. and 3 Fc. layers). The width of the convolutional layers (channels number) is quite small, which begins from 64 in the first layer, then increases by a factor of 2 after each max pooling layer until it reaches 512.

Our network configuration is rather different from the ones used in the previous MSTAR SAR target classification (Chen et al. 2016; Amrani and Jiang 2017; Bochkovskiy, Wang, and Liao 2020; Zhou et al. 2018). Rather than using relatively large receptive fields in the first conv. layers (e.g. 13 \times 13 with stride 4 in Bochkovskiy, Wang, and Liao (2020), 7 \times 7 with stride 1 in Amrani and Jiang (2017), or 5 \times 5 with stride 1 in Chen et al. (2016)), we use very small 3 \times 3 receptive fields throughout the whole net, which are convolved with the input at every pixel (with stride 1). Furthermore, we remove local response normalization (LRN) as it does not improve the performance on our SAR images dataset but leading to increase the computation time and the memory consumption, and we modify the output layers to appropriately match the C classes in our dataset. The configuration of the fully connected layers is the same in all network and all hidden layers are supplied with the rectified linear unit (ReLU) to have more discriminative decision functions and lower computation cost (Simonyan and Zisserman 2014; LeCun, Kavukcuoglu, and Farabet 2010). To achieve better classification accuracy, the linear support vector machine (L2-SVM) is used as a baseline classifier instead of maximum entropy classifier (Mason, Yonel, and Yazici 2017; Tang 2013). Consequently, we come up with expressively more accurate CNN architecture, which achieves the state-of-the-art accuracy on SAR target classification task.

3.2. MCCA based feature level fusion

Our proposed method is based on the fusion of the very deep features learned by our network model. The outputs of some selected layers are used as a feature descriptor of the input target. In particular, we considered the fully connected layers output with 4096 feature channels, which contain more low-level information for the description of texture and edges.

Several feature fusion methods have been proposed to obtain more informative descriptors to describe the image targets. The serial strategy (Liu and Wechsler 2001) simply combines two feature sets into one real union-vector and assumes that x and y are two feature sets with p and q vector dimensions, respectively. Contrary to the serial feature fusion, the parallel strategy (Yang and Yang 2002) combines the two feature sets into a complex feature set z = x + iy where i is the imaginary unit. However, these strategies neglect the class structure among the sample images and attain high dimension of the fused feature sets. Besides, SAR images have their own characteristics (speckle noise, illumination, pose variations, depression angle, corruption, and occlusion), and the choice of features is dictated by the statistical structure of the data (Oliver and Quegan 2004).

Let \( \mathbf{X}_{(p \times n)} \) and \( \mathbf{Y}_{(q \times n)} \) be two matrices that contain \( n \) training feature vectors. In our study, we adopt CCA (Huang et al. 2017; Canty 2014) to find the linear combination \( \mathbf{a}^\top \) and \( \mathbf{b}^\top \) that maximize the pairwise correlations \( \mathbf{X}^* = \mathbf{a}^\top \mathbf{X} \) and \( \mathbf{Y}^* = \mathbf{b}^\top \mathbf{Y} \) across the two feature sets, restrict the correlations to be between classes, and to face noise (Andrew et al. 2013). The top layers of our proposed SAR-OVDN contain more semantic information
and describe the global feature of the images, whereas the intermediate layers describe
the local features, and the bottom layers contain more low-level information for the
description of texture, edges. MCCA is proposed to combine information from different
levels to obtain distinguishing (relevant) features and overcome the curse of dimension-
ality problem. The transformation matrices $a$ and $b$ are diagonalized by solving the
eigenvalue equations as (Krzanowski 2000):

$$
V = \begin{pmatrix}
cov(x) & cov(x, y) \\
cov(y, x) & cov(y)
\end{pmatrix} = \begin{pmatrix}
V_{xx} & V_{xy} \\
V_{yx} & V_{yy}
\end{pmatrix}
$$

(1)

$$
\begin{align*}
V_{xx}^{-1}V_{xy}V_{yy}^{-1}V_{yx}\hat{a} &= \Lambda^2\hat{a} \\
V_{yy}^{-1}V_{yx}V_{xx}^{-1}V_{xy}\hat{b} &= \Lambda^2\hat{b}
\end{align*}
$$

(2)

where $\hat{a}$ and $\hat{b}$ are the eigenvectors and $\Lambda^2$ is the eigenvalues diagonal matrix, the non-
zero eigenvalues number in each equation is $r = \text{rank}(V_{xy}) \leq \min(n, p, q)$ that is organ-
ized in decreasing order, $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_r$, the transformation matrices $a$ and $b$ contain
the consistent eigenvectors to the non-zero eigenvalues, and $X^*, Y^* \in \mathbb{R}^{r \times n}$ are the can-
nical variates. To find the transformed training features sets, the covariance matrix in
Equation 4 will be defined as:

$$
V^* = \begin{pmatrix}
1 & 0 & \cdots & 0 & \gamma_1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 & 0 & \gamma_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & 1 & 0 & 0 & \cdots & \gamma_r \\
\gamma_1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\
0 & \gamma_2 & \cdots & 0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \gamma_r & 0 & 0 & \cdots & 1
\end{pmatrix}
$$

(3)

where the canonical variates have nonzero correlation only on their relative indices. The
upper left and lower right corners in the identity matrices indicate that the canonical
variates are uncorrelated with in training features sets. Figure 5 shows the statistical
transformation of the CCA framework.

In our work, feature-level fusion is then performed by using summation of the trans-
formed feature vectors as:

$$
M = X^* + Y^* = a^TX + b^TY = \begin{pmatrix}
a \\
b
\end{pmatrix}^T \begin{pmatrix}
X \\
Y
\end{pmatrix}
$$

(4)

where $M$ is the Very Deep Canonical Correlation Discriminant Features (VDCCDFs). Multi-
canonical correlation analysis (MCCA) generalizes CCA to be appropriate for more than
two features sets. We suppose that we have $\lambda$ feature sets $F_i \in \mathbb{R}^{p_i \times n}, i = 1, 2, \ldots, \lambda$, which
are arranged based on their rank, $\text{rank}(F_1) \geq \text{rank}(F_2) \geq \cdots \geq \text{rank}(F_\lambda)$. MCCA applies CCA to
two features sets at the same time. In each phase, the two feature sets with the highest
ranks are fused together to keep the maximum possible feature vectors length as shown
in Figure 6.
Figure 5. CCA transformation framework.

Figure 6. Multi-canonical correlation analysis techniques for five sample sets with rank \( F_1 > F_2 > F_3 > F_4 = F_5 \).
3.3. Support Vector Machine (L2-SVM)

Recently, most of the deep learning models utilize multi-class logistic regression for prediction and minimize cross-entropy loss for SAR target classification (Chen et al. 2016; Srivastava et al. 2014; Wagner 2016; Ding et al. 2016). In this paper, for better performance and parameter optimization, L2-SVM (Tang 2013) is adopted to train our SAR-oriented Very deep CNN model, when the learning minimizes a margin-based loss by back propagating the gradients from the top layer linear SVM. Therefore, we differentiate the SVM objective function with respect to the activation of the penultimate layer as follows. Let \( \mathbf{x}_n \in \mathbb{R}^D \) be training feature set and \( \mathbf{y}_n \) its labels, where \( n = 1, N \), and \( t_n \in \{-1, +1\} \). Given the objective \( L(w) \) in Equation 8, and the input \( x \) is replaced with the penultimate activation \( m \) as:

\[
\min_w \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N \max(1 - \mathbf{w}^T \mathbf{x}_n t_n, 0)
\]

(5)

\[
\frac{\partial L(w)}{\partial m_n} = -C t_n \mathbf{w} \mathbb{1}\{1 > \mathbf{w}^T \mathbf{m}_n t_n\}
\]

(6)

where \( \mathbb{1}\{\cdot\} \) is the indicator function and \( C \) is a constant (\( C \geq 0 \)). As before, for the L2-SVM, we have:

\[
\frac{\partial l(w)}{\partial m_n} = -2C t_n \mathbf{w} \max(1 - \mathbf{w}^T \mathbf{m}_n t_n, 0))
\]

(7)

Based on this point, the back-propagation algorithm is just the same as the standard softmax for deep learning networks.

4. Experimental results and analysis

The results of SAR target detection based on the MF dataset, and to validate the effectiveness of the proposed method, we conduct extensive experiments on two MSTAR datasets, which are publically available (SDMS 2016). In the following sections, we first describe the implementation details, and then introduce the used datasets. Finally, the experimental results are presented and analysed.

4.1. Implementation details

The pre-trained AlexNet, CaffeRef, VGGs and Very Deep-16 are fine-tuned on real SAR images from MSTAR database. The proposed network SARON is trained on a 2.7-GHz CPU with 64 GB of memory and a moderate graphics processing unit (GPU) card. All methods have been implemented using Microsoft Windows 10 Pro 64-bit and MATLAB R2016a. We randomly select the SAR image samples for the training and the testing datasets.

Initially, different variants of gradient descent are considered, among which mini-batch gradient descent is the most standard. We have then investigated algorithms that are most commonly used for optimizing Stochastic Gradient Descent (sgd), such as RMSprop and Adam (Ruder 2016). Figure 7 depicts the stepwise change in accuracy and loss across the training process at each iteration step: loss denotes the value of cross entropy. The
training is carried out by optimizing the primal $L^2$-SVM objective to learn lower level features: $L^2$ penalty multiplier and dropout ratio for the first fully connected layer are set to $5 \times 10^{-4}$ and 0.5, respectively. The learning rate is initially set to $10^{-2}$, and then is decreased by a factor of 10 when the accuracy of the validation set stopped improving. In total, the learning rate was decreased two times, and the learning has been stopped after 73 epochs. The classification of the targets with respect to its training set is done according to the classification error rates using the linear SVM. The main advantages of our proposed method as follows: first, using small-size convolution throughout the whole net to reduce the speckle noise. Second, decreases the number of parameters in each layer and therefore reduces computation cost as the CNN goes deeper. Moreover, ReLU layers are integrated to have more discriminative decision functions and lower computation cost. Third, the MCCA algorithm is utilized to fuse the feature vectors from the fully connected layers by summation and concatenation forming new robust and discriminant features. The analysis of the advantages is discussed in the following subsections.

4.2. Datasets

The SAR images used in experiments are from the Moving and Stationary Target Acquisition and Recognition (MSTAR) database. This benchmark data is acquired by the Sandia National Laboratories Twin Otter SAR sensor payload, operating at X-band with a high resolution of 0.3 m, spotlight mode, and HH single 320 polarizations (i.e. single-channel): where the phase content is entirely discarded because it is random and uniformly distributed (El-Darymli et al. 2015; Olivier and Quegan 1998). The first dataset is the MSTAR public mixed target dataset, which includes 10 military vehicle targets: (armoured personnel carrier: BMP-2, BRDM-2, BTR-60, and BTR-70; tank: T-62, T-72; rocket launcher: 2S1; air defence unit: ZSU-234; truck: ZIL-131; bulldozer: D7). The optical images and their relative SAR images are shown in Figure 8.

The second dataset chosen for evaluation is the MSTAR public T-72 Variant dataset, which contains eight T-72 variants: A04, A05, A07, A10, A32, A62, A63, and A64. Optical images and the corresponding SAR images of the eight T-72 targets are shown in Figure 9.
4.3. Results on the MSTAR Public Mixed Target Dataset

The dataset details including depression angles with target signatures of all MSTAR images used in this task are listed in Table 1. BMP-2 and BTR-70 refers to man-made (metal) objects armoured personnel carrier targets, and T-72 refers to main battle tank. The single and overall accuracies are well explained in Tables 2 and 3, respectively. The confusion matrix is clarified in Figure 10.

The noise in SAR images can be multiplicative or additive (Vidal-Pantaleoni and Marti 2004; Choi, Yu, and Jeong 2019). Figure 11 shows the noise simulation scheme, where the value of a randomly selected pixel is replaced by a value from a uniform distribution. The anti-noise performance of the proposed method is compared with the noise simulation
paradigm in Chen et al. (2016) and Dong, Wang, and Kuang (2014) and shows that our proposed method is more robust to noise corruption as clarified in Table 4.

4.4. Results on the MSTAR Public T-72 Variants Dataset

In another scenario, the proposed method is evaluated on the more challenging T-72 eight-target identification problem, as all the targets are almost indistinguishable. The number and the depression angles of the training and the testing sample images are listed in Table 5. The single and overall accuracies are explained in Tables 6 and 7, respectively. The confusion matrix is presented in Figure 12.
Results on the MSTAR with Large Depression Angle Variations

Since SAR images are very sensitive to depression angle variation, the credibility of the proposed method is further evaluated on large depression angles. In these experiments, four targets (2S1, BRDM-2, T-72, and ZSU-234) with 30° depression angle are assessed. The types and the number of the considered SAR images are shown in Table 9.

The single and overall accuracies are presented in Tables 9 and 10, respectively. The confusion matrix is shown in Figure 13.
Table 6. The identification accuracy for each target class on MSTAR public T-72 variants dataset.

| Class | Single Accuracy | Error Single | Total Accuracy | Error Total | Sensitivity | Specificity | Precision | False Positive Rate |
|-------|----------------|-------------|---------------|-------------|-------------|-------------|-----------|---------------------|
| A04   | 0.96715        | 0.032847    | 0.12112       | 0.0041133   | 0.96715     | 0.9953      | 0.96715   | 0.0047022           |
| A05   | 0.9708         | 0.029197    | 0.12157       | 0.0027422   | 0.9708      | 0.99687     | 0.97794   | 0.0031348           |
| A07   | 0.9708         | 0.029197    | 0.12157       | 0.0027422   | 0.9708      | 0.99687     | 0.97794   | 0.0031348           |
| A10   | 0.99262        | 0.0073801   | 0.12294       | 0.00073126  | 0.99262     | 0.99165     | 0.94386   | 0.0083464           |
| A32   | 0.9781         | 0.021898    | 0.12249       | 0.0018282   | 0.9781      | 0.99791     | 0.98529   | 0.0020899           |
| A62   | 0.96715        | 0.032847    | 0.12112       | 0.0054845   | 0.96715     | 0.99373     | 0.95668   | 0.0062696           |
| A63   | 0.96703        | 0.032967    | 0.12066       | 0.0036563   | 0.96703     | 0.99582     | 0.97059   | 0.0041775           |
| A64   | 0.94526        | 0.054745    | 0.11837       | 0.0022852   | 0.94526     | 0.99739     | 0.98106   | 0.0026123           |

Table 7. The overall identification accuracy on MSTAR public T-72 Variants dataset.

| Accuracy | Error | Sensitivity | Specificity | Precision | False Positive Rate |
|----------|-------|-------------|-------------|-----------|---------------------|
| 0.9698   | 0.0302| 0.9699      | 0.9957      | 0.9701    | 0.0043             |

Figure 12. Confusion matrix of the proposed method on T-72 Variants dataset. The rows and columns of the matrix indicate the actual and predicted classes, respectively.

Table 9. The RECOGNITION accuracy for each target class on MSTAR public database with 30° depression angle.

| Class | Single Accuracy | Error Single | Total Accuracy | Error in Total | Sensitivity | Specificity | Precision | False Positive Rate |
|-------|----------------|-------------|---------------|----------------|-------------|-------------|-----------|---------------------|
| 2S1   | 0.98958        | 0.010417    | 0.24761       | 0.0026064      | 0.98958     | 0.99652     | 0.98958   | 0.0034762           |
| BRDM-2| 0.98955        | 0.010453    | 0.24674       | 0.004344       | 0.98955     | 0.99421     | 0.9827    | 0.005787            |
| T-72  | 0.98264        | 0.017361    | 0.24587       | 0.0026064      | 0.98264     | 0.99652     | 0.98951   | 0.0034762           |
| ZSU-234| 0.99306       | 0.0069444   | 0.24848       | 0.0017376      | 0.99306     | 0.99768     | 0.99306   | 0.0023175           |

Table 10. The overall recognition accuracy on MSTAR public database with 30° depression angle.

| Accuracy | Error | Sensitivity | Specificity | Precision | False Positive Rate |
|----------|-------|-------------|-------------|-----------|---------------------|
| 0.9887   | 0.0113| 0.9887      | 0.9962      | 0.9887    | 0.0038             |
4.6. Comparison with Different Fusion Methods

In order to study the sensitivity of feature level fusion on the recognition performance, we have compared different fusion strategies: very deep canonical correlation analysis (VDCCA), very deep multi-canonical correlation analysis (VDMCCA) and SAR-oriented GBVS (Amrani et al. 2018) as shown in Table 11, our proposed method using MCCA achieves the best performance.

4.7. Influence of number of layers and different receptive fields on the results

To study the effect of network depth on its accuracy in SAR image recognition setting, several successful CNN models pre-trained on MSTAR are evaluated in our work, which are the famous baseline model AlexNet, the Caffe reference model (CaffeRef), and the VGG network (Zhou et al. 2017). Due to a large learning capacity, dominant expressive power and hierarchical structure of our SAR-oriented very deep network (19 layers): a high-level, semantic and robust feature representation for each region proposal is obtained as it is illustrated in Table 12.

In addition, using a stack of two $3 \times 3$ convolutional layers (without pooling in between) has a $5 \times 5$ effective receptive field, and using three of $3 \times 3$ layers has a $7 \times 7$ effective receptive field. So the reasons for using, for instance, a stack of three $3 \times 3$ convolutional layers instead of a single $7 \times 7$ layer are: First, we integrate three non-linear rectification layers rather than one, which leads to a more discriminative decision function. Second, we reduce the number of parameters by supposing that the three layer $3 \times 3$ convolution stack has K channels, and the stack is parameterized by $3(3^2K^2) = 27K^2$

Table 11. Comparison with different fusion methods on MSTAR database.

| Method                              | MSTAR public mixed target (%) | MSTAR public T-72 Variants (%) | MSTAR with 30° (%) |
|-------------------------------------|-------------------------------|-------------------------------|-------------------|
| Proposed without fusion             | 98.79                         | 94.50                         | 96.13             |
| VDCCA                               | 99.12                         | 95.31                         | 97.55             |
| VDMCCA                              | 99.70                         | 96.98                         | 98.87             |
| SAR-oriented GBVS (Amrani et al. 2018) | 99.68                         | 96.90                         | 98.70             |

Figure 13. Confusion matrix of the proposed method on MSTAR with large depression angle variations. The rows and columns of the matrix indicate the actual and predicted classes, respectively.
Table 12. The effect of network depth on the performance.

| Models          | Conv1       | Conv2       | Conv3       | Conv4       | Conv5       | Fc1          | Fc2          | Fc3          | MSTAR public mixed | MSTAR T-72 Variants | MSTAR depression 30° |
|-----------------|-------------|-------------|-------------|-------------|-------------|--------------|--------------|--------------|-------------------|---------------------|----------------------|
| AlexNet         | 11 × 11 × 96| 5 × 5 × 256 | 3 × 3 × 384 | 3 × 3 × 384 | 3 × 3 × 256 | 4096 dropout | 4096 dropout | C softmax     | 95.21             | 86.82               | 89.66                |
| CaffeRef        | 11 × 11 × 96| 5 × 5 × 256 | 3 × 3 × 384 | 3 × 3 × 384 | 3 × 3 × 256 | 4096 dropout | 4096 dropout | C softmax     | 95.45             | 85.96               | 88.99                |
| VGGF            | 11 × 11 × 64| 5 × 5 × 256 | 3 × 3 × 256 | 3 × 3 × 256 | 3 × 3 × 256 | 4096 dropout | 4096 dropout | C softmax     | 95.66             | 86.9                | 90.6                 |
| VGGM            | 7 × 7 × 96  | 5 × 5 × 256 | 3 × 3 × 512 | 3 × 3 × 512 | 3 × 3 × 512 | 4096 dropout | 4096 dropout | C softmax     | 96.7              | 88.08               | 91.5                 |
| VGGM-128        | 7 × 7 × 96  | 5 × 5 × 256 | 3 × 3 × 512 | 3 × 3 × 512 | 3 × 3 × 512 | 4096 dropout | 4096 dropout | C softmax     | 96.13             | 89.43               | 90.12                |
| VGGM-1024       | 7 × 7 × 96  | 5 × 5 × 256 | 3 × 3 × 512 | 3 × 3 × 512 | 3 × 3 × 512 | 4096 dropout | 128 dropout  | C softmax     | 97.6              | 90.86               | 91.6                 |
| VGGM-2048       | 7 × 7 × 96  | 5 × 5 × 256 | 3 × 3 × 512 | 3 × 3 × 512 | 3 × 3 × 512 | 4096 dropout | 2048 dropout | C softmax     | 96.74             | 90.12               | 91.55                |
| VGGS            | 7 × 7 × 96  | 5 × 5 × 256 | 3 × 3 × 512 | 3 × 3 × 512 | 3 × 3 × 512 | 4096 dropout | 4096 dropout | C softmax     | 97.95             | 91.66               | 92.88                |
| VGG16           | 16 weight layers including 13 convolutional layers 3 × 3 and 3 fully-connected layers | | | | | | | | | | |
| SAR-OVDN (19 weight layers including 16 convolutional layers and 3 fully-connected layers) | | | | | | | | | | | |
weights; in contrast to a single $7 \times 7$ convolutional layer, which requires $7^2K^2 = 49K^2$, which needs $81\%$ more of parameters.

### 4.8. Comparison with recent representative methods

The recognition performance of the proposed method is compared with the most widely cited approaches and the recent representative paradigms, such as MSRC (Dong, Wang, and Kuang 2014), SVM (Srinivas, Monga, and Raj 2014), Cond Gauss (O’Sullivan et al. 2001), MSS (Dong and Kuang 2015), A-Convnets (Chen et al. 2016), CNN (Srivastava et al. 2014), BCS (Zhang, Qin, and Li 2013), CNN–SVM (Bochkovskiy, Wang, and Liao 2020) and Deep Leaning (Wagner 2016). Because the SVM method (Srinivas, Monga, and Raj 2014) only considers the three-target classification problem, we have run the published online code from Srinivas, Monga, and Raj (2014) to obtain the results on our classification tasks (10 targets, T-72 Variants, and 30° depression angle). All other compared methods do the same classification tasks as ours. Therefore, we directly use the classification accuracies presented in the corresponding papers. As illustrated in Table 12, our proposed method achieves the highest accuracy rates. 

#### Table 13 The performance comparison between the proposed method and the state-of-the-art methods on the MSTAR public database.

| Method                                      | MSTAR public mixed target (%) | MSTAR public T-72 variants (%) | MSTAR with 30° (%) |
|---------------------------------------------|-------------------------------|-------------------------------|-------------------|
| SVM (Srinivas, Monga, and Raj 2014)         | 90                            | 78.5                          | 81                |
| Cond Gauss (O’Sullivan et al. 2001)         | 97                            | 77.32                         | 80                |
| MSRC (Dong, Wang, and Kuang 2014)           | 93.6                          | -                             | 98.4              |
| MSS (Dong and Kuang 2015)                   | 96.6                          | -                             | 98.2              |
| CNN (Srivastava et al. 2014)                | 84.7                          | 81.08                         | 81.56             |
| A-Convnets (Chen et al. 2016)               | 99.13                         | 95.43                         | 96.12             |
| BCS (Zhang, Qin, and Li 2013)               | 92.6                          | 88.76                         | 92.6              |
| Deep Leaning (Wagner 2016)                  | 92.74                         | -                             | -                 |
| CNN–SVM (Bochkovskiy, Wang, and Liao 2020)  | 99.5                          | 95.75                         | 96.6              |
| Proposed with softmax                       | 99.51                         | 96.13                         | 98.34             |
| Proposed with SVM                           | 99.70                         | 96.98                         | 98.87             |

---

Figure 14. MSTAR/IU Mixed Targets recognition results using different MF datasets.
5. Conclusion

This paper developed SARON to automatically learn very deep features from the datasets, select adaptive feature layers, and fuse the learned features. The proposed network increases the depth by using more convolutional layers to lessen the speckle noise effect. A dropout technique is supplied to deal with the over-fitting problem due to limited training datasets. In addition, the MCCA algorithm is introduced to combine the selective adaptive layer’s features and improves the representations with respect to the correlation objective measured on SAR target images, making the proposed method feasible for SAR image processing. Experimental results on the benchmark of MSTAR database demonstrate the effectiveness of the proposed method compared with the state-of-the-art methods.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Moussa Amrani http://orcid.org/0000-0003-3035-2739

References

Amrani, M., S. Chaib, I. Omara, and F. Jiang, “Bag-of-visual-words Based Feature Extraction for Sar Target Classification,” in Ninth International Conference on Digital Image Processing (ICDIP 2017), 10420. Honkong, China: International Society for Optics and Photonics, 2017, p. 104201J.
Amrani, M., and F. Jiang. 2017. “Deep Feature Extraction and Combination for Synthetic Aperture Radar Target Classification.” Journal of Applied Remote Sensing 11 (4): 042616. doi:10.1117/1.JRS.11.042616.
Amrani, M., F. Jiang, Y. Xu, S. Liu, and S. Zhang. 2018. “SAR-oriented Visual Saliency Model and Directed Acyclic Graph Support Vector Metric Based Target Classification.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11 (10): 3794–3810. doi:10.1109/JSTARS.2018.2866684.
Amrani, M., K. Yang, D. Zhao, X. Fan, and F. Jiang. 2017, September. “An Efficient Feature Selection for SAR Target Classification.” In Pacific Rim Conference on Multimedia, 68–78. Cham: Springer.
Andrew, G., R. Arora, J. Bilmes, and K. Livescu, “Deep Canonical Correlation Analysis,” in International Conference on Machine Learning, 2013, pp. 1247–1255.
Bochkovskiy, A., C. Y. Wang, and H. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” arXiv preprint arXiv:2004.10934, 2020.
Canty, M. J. 2014. Image Analysis, Classification and Change Detection in Remote Sensing: With Algorithms for ENVI/IDL and Python. Crc Press.
Chen, S., and H. Wang, “Sar Target Recognition Based on Deep Learning,” in Data Science and Advanced Analytics (DSAA), 2014 International Conference on. IEEE, 2014, pp. 541–547
Chen, S., H. Wang, F. Xu, and Y.-Q. Jin. 2016. “Target Classification Using the Deep Convolutional Networks for Sar Images.” IEEE Transactions on Geoscience and Remote Sensing 54 (8): 4806–4817. doi:10.1109/TGRS.2016.2551720.
Choi, H., S. Yu, and J. Jeong (2019, June). “Speckle Noise Removal Technique in SAR Images Using SRAD and Weighted Least Squares Filter.” In 2019 IEEE 28th International Symposium on Industrial Electronics (ISIE) (pp. 1441–1446). IEEE.
Ciresan, D. C., U. Meier, J. Masci, L. Maria Gambardella, and J. Schmidhuber, “Flexible, High Performance Convolutional Neural Networks for Image Classification,” in IJCAI Proceedings-
International Joint Conference on Artificial Intelligence, vol. 22, no. 1. Barcelona, Spain, 2011, p. 1237.

Ding, J., B. Chen, H. Liu, and M. Huang. 2016. “Convolutional Neural Network with Data Augmentation for Sar Target Recognition.” IEEE Geoscience and Remote Sensing Letters 13 (3): 364–368.

Dong, G., and G. Kuang. 2015. “Classification on the Monogenic Scale Space: Application to Target Recognition in Sar Image.” IEEE Transactions on Image Processing 24 (8): 2527–2539. doi:10.1109/TIP.2015.2421440.

Dong, G., N. Wang, and G. Kuang. 2014. “Sparse Representation of Monogenic Signal: With Application to Target Recognition in Sar Images.” IEEE Signal Processing Letters 21 (8): 952–956.

Du, K., Y. Deng, R. Wang, T. Zhao, and N. Li. 2016. “Sar Atr Based on Displacement-and Rotation-insensitive Cnn.” Remote Sensing Letters 7 (9): 895–904. doi:10.1080/2150704X.2016.1196837.

El-Darymli, K., P. Mcguire, E. W. Gill, D. Power, and C. Moloney. 2015. “Characterization and Statistical Modeling of Phase in Single-channel Synthetic Aperture Radar Imagery.” IEEE Transactions on Aerospace and Electronic Systems 51 (3): 2071–2092. doi:10.1109/TAES.2015.140711.

Furukawa, H., “Deep Learning for Target Classification from Sar Imagery: Data Augmentation and Translation Invariance,” arXiv preprint arXiv:1708.07920, 2017.

Goodfellow, I. J., Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, “Multi-digit Number Recognition from Street View Imagery Using Deep Convolutional Neural Networks,” arXiv preprint arXiv:1312.6082, 2013.

Huang, X., B. Zhang, H. Qiao, and X. Nie. 2017. “Local Discriminant Canonical Correlation Analysis for Supervised Polsar Image Classification.” IEEE Geoscience and Remote Sensing Letters 14 (11): 2102–2106. doi:10.1109/LGRS.2017.2752800.

Krzanowski, W. 2000. “Principles of Multivariate Analysis.” In OUP Oxford.

LeCun, Y., K. Kavukcuoglu, and C. Farabet, “Convolutional Networks and Applications in Vision,” in Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on. IEEE, 2010, pp. 253–256.

Liu, C., and H. Wechsler. 2001. “A Shape-and Texture-based Enhanced Fisher Classifier for Face Recognition.” IEEE Transactions on Image Processing 10 (4): 598–608. doi:10.1109/83.913594.

Mason, E., B. Yonel, and B. Yazici, “Deep Learning for Radar,” in Radar Conference (RadarConf), 2017 IEEE IEEE, 2017, pp. 1703–1708.

O’Sullivan, J. A., M. D. DeVore, V. Kedia, and M. I. Miller. 2001. “Sar Atr Performance Using a Conditionally Gaussian Model.” IEEE Transactions on Aerospace and Electronic Systems 37 (1): 91–108. doi:10.1109/7.913670.

Oliver, C., and S. Quegan. 2004. “Understanding Synthetic Aperture Radar Images.” In SciTech Publishing.

Olivier, C., and S. Quegan. 1998. Understanding Synthetic Aperture Radar Images. Artech House.

Olson, C. F., and D. P. Huttenlocher. 1997. “Automatic Target Recognition by Matching Oriented Edge Pixels.” IEEE Transactions on Image Processing 6 (1): 103–113. doi:10.1109/83.552100.

Ruder, S. “An Overview of Gradient Descent Optimization Algorithms.” arXiv preprint arXiv:1609.04747 (2016).

Schwegmann, C. P., W. Kleynhans, B. P. Salmon, L. W. Mdakane, and R. G. Meyer, “Very Deep Learning for Ship Discrimination in Synthetic Aperture Radar Imagery,” in Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International. IEEE, 2016, pp. 104–107.

SDMS, “Mstar Data,” https://www.sdms.afrl.af.mil/index.php?collection=mstar, last accessed: 2016-11-23.

Simonyan, K., and A. Zisserman, “Very Deep Convolutional Networks for Large-scale Image Recognition,” arXiv preprint arXiv:1409.1556, 2014.

Srinivas, U., V. Monga, and R. G. Raj. 2014. “Sar Automatic Target Recognition Using Discriminative Graphical Models.” IEEE Transactions on Aerospace and Electronic Systems 50 (1): 591–606. doi:10.1109/TAES.2013.120340.
Srivastava, N., G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting.” Journal of Machine Learning Research 15 (1): 1929–1958.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going Deeper with Convolutions, Corr Abs/1409.4842,” http://arxiv.org/abs/1409.4842, 2014.

Tang, Y., “Deep Learning Using Linear Support Vector Machines,” arXiv preprint arXiv:1306.0239, 2013.

Vidal-Pantaleoni, A., and D. Marti. 2004. “Comparison of Different Speckle-reduction Techniques in SAR Images Using Wavelet Transform.” International Journal of Remote Sensing 25 (22): 4915–4932. doi:10.1080/01431160410001688277.

Wagner, S. A. 2016. “Sar Atr by a Combination of Convolutional Neural Network and Support Vector Machines.” IEEE Transactions on Aerospace and Electronic Systems 52 (6): 2861–2872. doi:10.1109/TAES.2016.160061.

Yang, J., and J.-Y. Yang. 2002. “Generalized K–l Transform Based Combined Feature Extraction.” Pattern Recognition 35 (1): 295–297. doi:10.1016/S0031-3203(01)00152-2.

Zhang, X., J. Qin, and G. Li. 2013. “Sar Target Classification Using Bayesian Compressive Sensing with Scattering Centers Features.” Progress In Electromagnetics Research 136: 385–407. doi:10.2528/PIER12120705.

Zhou, F., L. Wang, X. Bai, and Y. Hui. 2018. “SAR ATR of Ground Vehicles Based on LM-BN-CNN.” In IEEE Transactions on Geoscience and Remote Sensing.

Zhou, W., S. Newsam, C. Li, and Z. Shao. 2017. “Learning Low Dimensional Convolutional Neural Networks for High-resolution Remote Sensing Image Retrieval.” Remote Sensing 9 (5): 489. doi:10.3390/rs9050489.