remote sensing

MDPI

Article

Multi-Stream Convolutional Neural Network for SAR Automatic Target Recognition

Pengfei Zhao [1,2,3], Kai Liu [1,2,3], Hao Zou [4] and Xiantong Zhen [1,2,3,*]

[1] School of Electronics and Information Engineering, Beihang University, Beijing 100191, China; pf_zhao@buaa.edu.cn (P.Z.); liuk@buaa.edu.cn (K.L.)
[2] Beijing Key Lab for Network-Based Cooperative ATM, Beijing 100191, China
[3] Beijing Laboratory for General Aviation Technology, Beijing 100191, China
[4] University of Chinese Academy of Sciences, Beijing 100190, China; zouhao15@mails.ucas.ac.cn
[*] Correspondence: zhenxt@buaa.edu.cn; Tel: +86-10-82317846

Received: 23 July 2018; Accepted: 6 September 2018; Published: 14 September 2018

check for updates

Abstract: Despite the fact that automatic target recognition (ATR) in Synthetic aperture radar (SAR) images has been extensively researched due to its practical use in both military and civil applications, it remains an unsolved problem. The major challenges of ATR in SAR stem from severe data scarcity and great variation of SAR images. Recent work started to adopt convolutional neural networks (CNNs), which, however, remain unable to handle the aforementioned challenges due to their high dependency on large quantities of data. In this paper, we propose a novel deep convolutional learning architecture, called Multi-Stream CNN (MS-CNN), for ATR in SAR by leveraging SAR images from multiple views. Specifically, we deploy a multi-input architecture that fuses information from multiple views of the same target in different aspects; therefore, the elaborated multi-view design of MS-CNN enables it to make full use of limited SAR image data to improve recognition performance. We design a Fourier feature fusion framework derived from kernel approximation based on random Fourier features which allows us to unravel the highly nonlinear relationship between images and classes. More importantly, MS-CNN is qualified with the desired characteristic of easy and quick manoeuvrability in real SAR ATR scenarios, because it only needs to acquire real-time GPS information from airborne SAR to calculate aspect differences used for constructing testing samples. The effectiveness and generalization ability of MS-CNN have been demonstrated by extensive experiments under both the Standard Operating Condition (SOC) and Extended Operating Condition (EOC) on the MSTAR dataset. Experimental results have shown that our proposed MS-CNN can achieve high recognition rates and outperform other state-of-the-art ATR methods.

Keywords: CNN; deep learning; multi-view; ATR; SAR; MSTAR

1. Introduction

Thanks to its superior characteristics, including all-weather day-and-night observation, high-resolution imaging capability, and so forth, synthetic aperture radar (SAR) imaging plays an indispensable role in both military and civil applications. Essentially, SAR is an active microwave detection device for remote sensing, which is diversely utilized in geographical surveying, environment and Earth system monitoring, climate change research [1,2], and more. The combination of the electromagnetic scattering mechanism and a coherent imaging system enables SAR images to contain rich features, which provides important information for target recognition [3]. However, such features are contaminated by coherent speckle noise and geometric distortions in the images, accounting for the lower quality of SAR images. This tendency has a negative impact on target detection and recognition [4]. Furthermore, SAR images are highly sensitive to observation depression and aspect angle variations, as well as inevitable imaging deformation, including the perspective scale-variant, shadow, and layover. Even with a limited observation azimuth gap, the shapes of targets in SAR images are almost distinct from each other. All these factors pose severe difficulties in SAR image interpretation and target recognition. In order to overcome these obstacles, the research of SAR image automatic interpretation algorithms has attracted increasing attention; notably, the automatic target recognition (ATR) has been extensively researched.

ATR is used to locate and classify the target in SAR images. The SAR ATR procedure is typically composed of three steps: preprocessing, feature extraction, and classification. The preprocessing provides a region of interest (ROI) cropped from a specific SAR image using a constant false alarm rate (CFAR) detector. The output of this CFAR detector contains not only the ROI target, but also false alarm clutter, like trees, building, and cars. In addition, data augmentation operations, including rotation, flipping, and random cropping, are also deployed in this step. The second step is to extract the effective features from the output with reduced dimensions and eliminate the false clutters in the meantime. In terms of typical feature extraction, the available SAR image features, mainly consisting of scattering features, polarization features, and ROI features, are mostly extracted from single and independent SAR images [5–7]. Finally, a classifier is then applied to specify the category of the target. As the most important step, adopted classification methods in the ATR procedure can be approximately divided into three genres: template-based approaches, model-based approaches, and machine learning approaches [8–10]. The template-based methods rely on template-matching, and therefore, are dependent on the template library. But, once some SAR attributes change, the classification rate will drop sharply. To achieve better robustness, the model-based methods introduce a high-fidelity model with both an offline model construction component and online prediction and recognition component. These methods are adaptive and online adjustable but increase the computation overhead significantly. With the advent of machine learning, classification approaches based on deep learning or a support vector machine (SVM) are proven to be feasible and promising [10]. Accordingly, increasing research efforts have been devoted in the field of SAR image recognition with a deep learning structure, which has thus far reported extraordinary recognition rates [11–13].

Although ATR in SAR has been extensively studied in recent years, it remains an unsolved problem. Its significant challenges arise from a severe lack of raw SAR image data and great variations of SAR images, due to their aspect-sensitive characteristics. Firstly, collecting SAR images can be very burdensome, resulting in the awkward fact that there are not enough training and testing samples to train a deep model with almost perfect performance in the usual manner. Naturally, it is vital to design a network structure that is capable of making full use of the limited available data (such as the Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset). Recent studies regarding ATR began to adopt convolutional neural networks (CNNs); this approach provides a powerful tool for ATR in SAR, with significant progress in the past years [14–16]. However, CNNs remains unable to overcome the aforementioned challenges, partly because of their high dependency on large data for training an excellent model. In addition, many proposed methods attempted to increase the number of training samples by data augmentation for better recognition performance [12]. Indeed, the data augmentation approach shows some benefits, but the inherent connection between these raw SAR images in the MSTAR dataset has not been well explored.

Moreover, owing to the SAR imaging mechanism and SAR parameter settings, the SAR images are highly sensitive to the aspect and depression angle changes. In other words, SAR images with different aspects and depression angles contain largely distinct information about the same ground target. In general, SAR images of different aspect angles of the same target can be obtained by either multiple airborne uninhabited aerial vehicle (UAV) SAR joint observations, or single SAR observations along a circular orbit. Therefore, these images may contain space-varying scattering features, and contain much more information than a single image. SAR ATR algorithms can perhaps consider more in terms of multi-view images of the same target as the network input, in order to build more comprehensive representations [17,18]. This idea may make full use of the inherent connections of a limited raw MSTAR dataset, possibly enhancing the recognition accuracy, which, however, has not yet been explored.

In this paper, we propose a novel deep convolutional learning architecture, called a Multi-Stream CNN (MS-CNN), by using SAR images from multiple views of the same targets to effectively recognize the target classes. MS-CNN handles the aforementioned challenges by disentangling the relationships between images and classes in the learning architecture which is composed of four parts: a multi-stream convolutional layer, a Fourier feature fusion layer, a fully connected layer, and a softmax layer as shown in Figure 1.

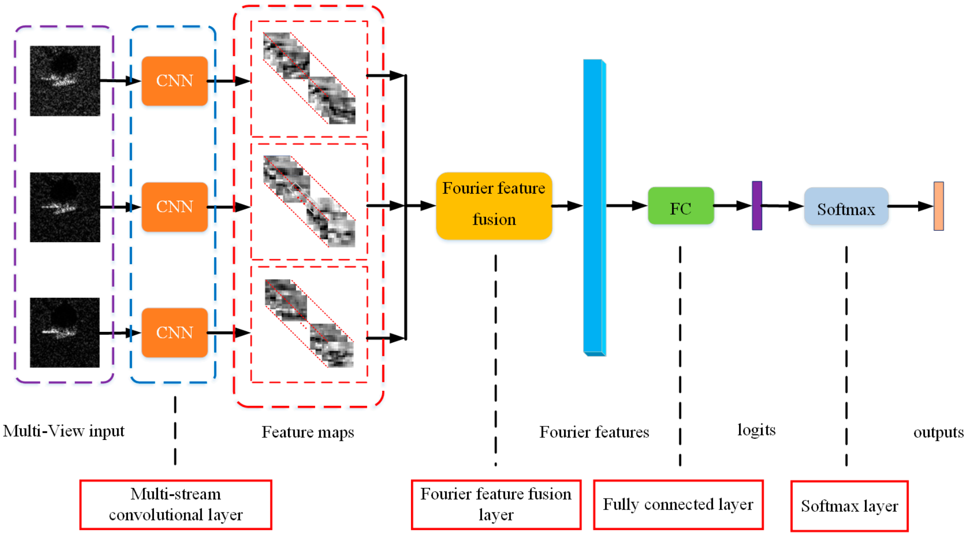


Figure 1. The learning structure of the proposed multi-stream convolutional neural network (MS-CNN); that is, the three-views MS-CNN.

In order to extract not only typical features but also space-varying features induced by multiple views, MS-CNN incorporates a multi-stream convolutional layer, which is more efficient but with fewer parameters. This multiple-input architecture can effectively and efficiently extract features from multiple views of the same targets. Therefore, it can make full use of limited SAR images to improve recognition performance compared to regular CNN, which probably suffers a problem that it can hardly extract effective and interconnected features.

In conjunction with the multi-stream convolutional layer, we introduce a Fourier feature fusion layer into the learning architecture. This part is able to fuse the features of multiple views from upper outputs, and then build strong holistic representations. The Fourier feature fusion is derived from kernel approximation based on random Fourier features, which takes advantage of strength of kernel methods for nonlinear feature extraction and fusion, and it helps unravel the highly nonlinear relationship between images and classes. Furthermore, the Fourier feature fusion turns out to be a nonlinear layer with a cosine activation function, which can make the back-propagation learning process ready to use.

In terms of practical value, our proposed MS-CNN can be easily and quickly operated in real SAR ATR scenarios. This superiority stems from our unique training and testing samples construction approach, which only needs multiple continuous aspect information of multi-view SAR images, rather than requiring the aspects with a fixed interval or larger changeable range mentioned in other methods; these aspects can be calculated by the real-time GPS information of airborne SAR within a tiny time slot.

Our main contributions can be summarized as follows:

We propose a novel convolutional learning architecture, called the multi-stream convolutional neural network (MS-CNN), for ATR in SAR. Our MS-CNN makes full use of discriminating space-varying features, and can largely improve the recognition rates.

We introduce a novel feature extraction structure, the Fourier feature fusion layer, to effectively extract and fuse the features of multi-view SAR images to achieve a strong representation, which in turn establishes the highly nonlinear relationship between SAR images and their associated classes.

We conceive a specific construction approach for corresponding multi-view training and testing samples in our proposed MS-CNN. Its practical value in real SAR ATR scenarios is obvious, simply because it only needs multiple continuous aspect information within a small time slot to construct its testing samples.

The remainder of this paper is organized as follows: The MS-CNN structure and the descriptions of each part are introduced in Section 2. Experimental results and analyses on the MSTAR dataset are given in Section 3. In Section 4, we present some discussions regarding the feasibility and reasonability of the MS-CNN, and future work. Finally, Section 5 concludes this paper.

2. Multi-Stream Convolutional Neural Network

In this part, we first introduce our proposed multi-stream convolutional neural network (MS-CNN), beginning with problem formulation. Next, we will describe each key part of MS-CNN, including the multi-stream convolutional layer and the Fourier feature fusion layer. Then, the learning process of MS-CNN, such as learning rate setting, convolutional kernel updating and so forth, will be given. Finally, we propose an easily accessible approach for training and testing samples construction in real SAR ATR scenarios.

2.1. Preliminaries

SAR ATR is a classification task with the purpose of establishing the mapping between SAR image input and the corresponding classes the targets belong to. The proposed MS-CNN explores and leverages the space-varying information, which means that different views of the same target contain some different features, from multi-view SAR images to enhance the recognition rates, alleviating the problem of the lack of raw SAR images. Specifically, the multi-stream convolutional layer and the Fourier feature fusion layer can effectively and efficiently extract the nonlinear features—both image-based features and space-varying features—which can be used to identify the relationship between images and target categories, and make it possible to improve recognition rates. Therefore, MS-CNN has got a great generalization ability.

2.2. Multi-Stream Convolutional Layer

Image representation is essential for SAR ATR, and CNNs has been identified as an efficient and powerful tool to extract feature in diverse tasks. However, limited by a lack of raw SAR images for training data, traditional CNNs is unable to deeply explore the inherent correlation of limited SAR images, and in turn, cannot adequately dig out effective features in the training process of ATR tasks.

Instead of using regular CNNs, we introduce a multi-stream convolutional layer, which is inspired by inherent connections of the multiple views of the same targets, to make full use of limited raw SAR data, and then extract complementary features from multi-view SAR images for more informative SAR image representations. Moreover, this method can not only adequately extract multi-view features, but also largely reduce the number of parameters and boost the training efficiency, while improving the recognition performance, which fits the SAR ATR tasks well.

As shown in Figure 1, we simply provide multi-view SAR images of the same targets as inputs for the multi-stream convolutional layer. The design of the multi-stream convolutional layer combined with Fourier feature fusion layer enables it to sieve more interacted features from multi-view SAR images, and then help identify their corresponding classes. In addition, it removes the flattening operation by setting rational parameters such as the size of the convolutional kernel and pooling, largely reducing the number of parameters, while making possible further accelerating the training process.

Figure 2 shows the details of one stream of the multi-stream convolutional layer, which consists of four convolutional layers and four pooling layers alternately. The details of each layer and operation will be explained below.

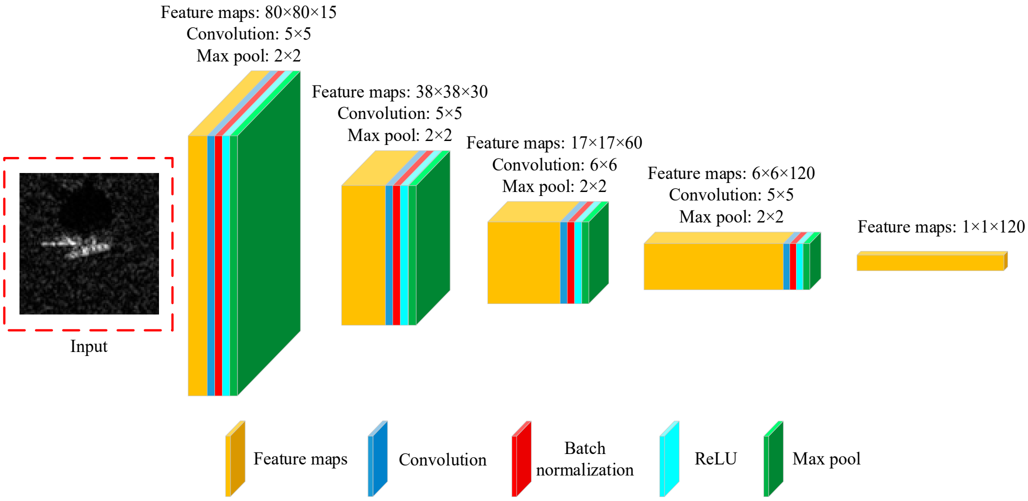


Figure 2. The details of one stream of the multi-stream CNN framework.2.2.1. Convolutional Operation.

2.2.1. Convolutional Operation

The convolutional layer in each stream of the MS-CNN serves to extract one-view features from the multi-view input images, and all streams work parallelly to additionally extract multi-view complementary feature. Compared to standard CNNs, the number of parameters is relatively smaller because of shared convolutional kernels and biases among multiple streams. Generally, the hyperparameters in the convolutional process consist of the number of feature maps, convolutional kernel size, stride and padding. For instance, if the size of previous feature maps is $V_1 \times V_2$, $W_1 \times W_2$ is the convolutional kernel size, and the stride and padding are S and P correspondingly. Then, the size of the feature map's output of each branch is $((V_1 - W_1 + 2P)/S + 1) \times ((V_2 - W_2 + 2P)/S + 1)$. Typically, we remove the flattening operation by rationally setting the size of the convolutional kernel, further reducing the number of parameters and computations.

2.2.2. Batch Normalization

Batch normalization is also required for our multi-stream convolutional layer. Specifically, a batch normalization operation is utilized after the convolutional operation [19]. It is essential to do so, because batch normalization of each stream can solve the instability problem of gradient descent in the process of backpropagation, and ultimately speed up the convergence of the whole network.

Batch normalization can change the distribution of the original data, and render most data to be pulled into the linear part of the activation function. However, in our MS-CNN, we choose the ReLU activation function following batch normalization. This is actually a nonlinear function, so the nonlinear transformation $y_i = \gamma \hat{x}_i + \beta$ is rendered unnecessary.

2.2.3. Nonlinearization

The role of the nonlinear activation function is to increase the nonlinear relationship between layers of the neural network. In this paper, we choose Rectified Linear Units (ReLUs) as the activation function for all streams of multi-stream CNN [20]. ReLU can largely decrease the training time, and achieve a better performance on labeled SAR data without any unsupervised pre-training.

2.2.4. Pooling Operation

There are two types of pooling functions: max pooling and average pooling. In this paper, the max pooling operator is utilized [21], and the relevant operation only reserves the maximum value within the pooling-size region extracted from a certain filter, so this operation is processed in each feature map separately.

2.3. Fourier Feaature Fusion Layer

After the multi-stream CNN process, we have extracted multiple feature vectors correspondingly from multi-view images. Next, we need to generate a high-level and holistic representation by fusing these multiple feature vectors in a proper approach. It would not be optimal to simply sum or put all the vectors together, because there are semantic gaps between those separate outputs. Therefore, we propose a Fourier feature fusion layer to integrate these feature vectors by means of kernel methods, so as to leverage its great strength to fill the semantic gaps [22,23]. In contrast to regular Fourier features, ours are learned from data in an end-to-end way. This enables us to obtain more compact but discriminant features for more accurate recognition. In addition, feature fusion in the kernel level can also acquire nonlinear feature extraction if nonlinear kernels are utilized. The proposed Fourier feature fusion layer is derived from the approximation of shift-invariant kernels, which is supported by the Bochner's theorem.

Theorem 1. (*Bochner* [24]) *A continuous shift-invariant kernel function $k(x,y) = k(x-y)$ on $\mathbb{R}^d$ is positive definite if and only if $k(\delta)$ is the Fourier transform of a non-negative measure on $\mathbb{R}^d$.*

If the kernel $k(\delta)$ is properly scaled, then its Fourier transform $p(w)$ will also be a proper probability distribution. Defining $\xi_w(x) = e^{jw^T x}$, for any $x, y \in \mathbb{R}^d$, we have:

$$k(x-y) = \int_{\mathbb{R}^d} p(w) e^{jw^T(x-y)} dw = E[\xi_w(x)\xi_w(y)^*] \tag{1}$$

where * is the conjugate and $\xi_w(x)\xi_w(y)^*$ is an unbiased estimate of $k(x-y)$ when is w drawn from $p(w)$.

Actually, we focus only on the real part, so the integrand $e^{jw^T(x-y)}$ can be simplified as $\cos w^T(x-y)$. We assume $z_w(x) = \sqrt{2}\cos(w^T x + b)$ that satisfies $E[z_w(x) z_w(y)^*] = k(x,y)$.

We can approximate the kernel $k(x,y)$ by randomly choosing D random samples and calculating the sum of their inner products:

$$k(x,y) \approx \sum_{i=1}^{D} \left\langle \sqrt{\frac{2}{d}} \cos(w_i^T x + b_i), \sqrt{\frac{2}{d}} \cos(w_i^T y + b_i) \right\rangle \tag{2}$$

where w is drawn from $p(w)$ and b obey to the uniform distribution over [0,2π].

Therefore, the corresponding feature maps $\phi_{w,b}(x_i)$ can be simplified as:

$$\phi_{w,b}(x_i) = \sqrt{\frac{2}{d}} [\cos(w_i^T x_i + b_i)]_{1:D} \tag{3}$$

where $\phi_{w,b}(x_i)$ is called the random Fourier feature, and $\cos(\cdot)$ is the cosine function on the element-wise level.

Kernel approximation remains largely underdeveloped. Specifically, the chosen samples are drawn independently from the distributions, and in order to achieve a satisfactory recognition performance, high-dimensional feature maps are always essential for the class prediction. However, on the basis of an approximation operation, it perhaps induces extra computational cost because of the approximate feature maps with high redundancy and low generalization ability. Another problem in sampling is how to select the most suitable kernel configuration parameters. Therefore, approximating the kernel with the random sampling operation probably cannot enhance the recognition performance as expected.

Instead of randomly sampling w from the distribution $p(w)$ and b from [0,2π], we learn these two parameters in a supervised way. Thus, we get a nonlinear layer with the cosine activation function:

$$z_{W,b}(x) = \sqrt{2}\cos(Wx + b) \tag{4}$$

where $W = [w_1^T, w_2^T, \cdot\cdot\cdot, w_D^T] \in \mathbb{R}^{D\times d}$ is the weight matrix and $b = [b_1, b_2, \cdot\cdot\cdot, b_d] \in \mathbb{R}^d$ is the bias vector.

The Fourier feature fusion seamlessly aggregates feature maps from the upper multi-stream convolutional layer and achieves a nonlinear activation in the meantime. Fourier feature fusion can concatenate these features by changing them into the same semantic level, and those space-varying potential features can be identified and utilized in this process to strengthen the discrimination of feature maps. To sum up, the induced Fourier feature fusion can be integrated with the multi-stream CNN to achieve a novel and efficient learning structure, which can be trained through back-propagation readily.

In the end, after a fully connected layer, we utilize the softmax layer as the output layer for classification. It will generate the posterior probability distribution for the inputting feature vector. The final output of the network is a k-dimension probability vector, and each element in this vector represents the probability of identifying as the corresponding class.

2.4. Learning Process of the MS-CNN

2.4.1. Learning Rate

The learning rate indicates the speed at which the parameters reach the optimal value. In the beginning, the initial learning rate should be set as a relatively large value to help the trainable network parameters approach a convergence value faster. However, if it is always trained with a large learning rate through the training process, the parameters may finally fluctuate randomly around the optimal value, instead of reaching the optimum. Therefore, we need to adjust the learning rate according to loss and validation accuracy during the training process. For instance, the learning rate can be decreased by multiplying a factor τ ($0 < \tau < 1$) when we either find the validation accuracy stops improving for a long time, or just change it in fixed epochs d. In this paper, we utilized the second updating approach, and its initial value is set to α, where

$$\begin{cases} \alpha \leftarrow \tau\alpha & if \ \mathrm{mod}(i,d) = 0 \\ \alpha \leftarrow \alpha & \qquad others \end{cases} \tag{5}$$

where i ($I \geq 1$) is denoted as the training epochs, α takes 0.001, d takes 1, and τ takes 0.96.

2.4.2. Cost Function with $L2$ Regularization and Backpropagation

The original cost function in this paper uses the cross-entropy cost function, so the formula of the cost function can be written as:

$$L(w,b) = -\frac{1}{C}\sum_{i=1}^{C} y_i \log \rho(y_i|Z^{(L)}; w, b) \tag{6}$$

After the operation of $L2$ regularization, the cost function $L^R(w,b)$ can be rewritten as:

$$L^R(w,b) = -\frac{1}{C}\sum_{i=1}^{C} y_i \log \rho(y_i|Z^{(L)}; w, b) + \frac{\lambda}{2C}(\|w\|_2)^2 \tag{7}$$

where $\|w\|_2 = \sqrt{\sum_{i=1}^{N} w_i^2}$ is the Euclid norm, and λ is the weight decay value and is set to 0.00001.

The aim of backpropagation is to train and update the trainable network parameters to achieve minimum loss, and the parameters w and b are updated through the formula below:

$$w \leftarrow w - \alpha \frac{\partial L^R}{\partial w} = w - \alpha \frac{\partial L}{\partial w} - \frac{\alpha\lambda}{C} w = (1 - \frac{\alpha\lambda}{C}) w - \alpha \frac{\partial L}{\partial w} \tag{8}$$

$$b \leftarrow b - \alpha \frac{\partial L^R}{\partial b} = b - \alpha \frac{\partial L}{\partial b} \tag{9}$$

where the value of α is the learning rate. It is obvious that the $L2$ regularization operation only influences the update of convolutional kernel w without any impacts on bias b.

In the backpropagation algorithm, these two partial derivatives, $\frac{\partial L}{\partial w}$ and $\frac{\partial L}{\partial b}$, can be calculated by resorting to the error term $\delta_j^{(l)}$ of each layer. Thus, the formulas of $\frac{\partial L}{\partial w}$ and $\frac{\partial L}{\partial b}$ can be written as:

$$\frac{\partial L}{\partial w_{ij}^{(l)}} = \delta_j^{(l)} * (a_{ij}^{(l-1)})^T \tag{10}$$

$$\frac{\partial L}{\partial b_j^{(l)}} = \sum_{x,y} \delta_j^{(l)}(x, y) \tag{11}$$

where the error term of each layer $\delta_j^{(l)}$ can be calculated by $\delta_j^{(l)} = \sigma'(z_j^{(l)}) \odot \sum_i \delta_i^{(l+1)} * (w_{ij}^{(l+1)})^T$ and the error term of the output layer is $\delta_j^{(L)} = -(y_i - \rho(y_i|Z^{(L)}))$.

2.5. Training and Testing Sample Construction

Generally, multi-view SAR images can be acquired by either multiple airborne/UAV SAR joint observations from same depression angle and different aspect angles, or single airborne SAR observations along a circular orbit. We assume that the depression angle is known in advance, so only varying aspect angles are demanded here. As shown in Figure 3, the airborne SAR sensors within the plane, moving along a circular orbit for a given target, can produce continuous SAR images with different aspects. On the basis of this, we can make our own multi-view SAR image samples for training and testing.

We assume that the raw SAR image sequence for a specific experiment is defined as $X^r = \{X_1, X_2, X_3, \ldots, X_C\}$, where $X_i = \{x_1, x_2, x_3, \ldots, x_{n_i},\}$ is the image set for a specific class y_i, and their relevant aspect angles are $\theta(x_{n_i})$. The set $\{y_i \in [1, 2, 3, \ldots, C]\}$ indicates the class labels, and C is the sequence number of classes. For a given view number k, the k-view SAR image combinations for each class can be gained by regrouping the current SAR images. Specifically, we first sort these SAR images by azimuth angles in ascending order for each class of each target type. In other words, each image set $X_i = \{x_1, x_2, x_3, \ldots, x_{n_i},\}$ is put in order according to their aspect angles, such as $\theta(x_{s_1}) < \theta(x_{s_2}) < \theta(x_{s_3}) < \ldots < \theta(x_{s_{n_i}})$. Then, as shown in Figure 4, we combine these sorted images according to the view number k to generate multi-view training and testing samples whose size equals the original, such as $\{x_{s_1}, x_{s_2}, x_{s_3}, \ldots, x_{s_k}\}$, $\{x_{s_2}, x_{s_3}, x_{s_4}, \ldots, x_{s_{k+1}}\}$, $\ldots$,$\{x_{s_{n_i-(k-2)}}, x_{s_{n_i-(k-1)}}, x_{s_{n_i-k}}, \ldots, x_{s_1}\}$.

For a typical class y_i, let $X_k^i = \left\{X_{s_1}^i, X_{s_2}^i, X_{s_3}^i, \ldots, X_{s_{n_i}}^i\right\}$ be the set of n_i sizes of sorted k-view images, where $X_{s_j}^i = \{\{x_{s_1}, x_{s_2}, x_{s_3}, \ldots, x_{s_k}\}, \ldots, \{x_{s_{n_i-(k-2)}}, x_{s_{n_i-(k-1)}}, x_{s_{n_i-k}}, \ldots, x_{s_1}\}, \ldots, \{x_{s_{n_i}}, x_{s_1}, x_{s_2}, \ldots, x_{s_{k-1}}\}\}$ and $j = \{1, 2, 3, \ldots, \mathrm{n}_i\}$ is one k-view SAR image combination. Thus, the sorted k-view SAR image dataset is $X_k^s = \left\{X_k^1, X_k^2, X_k^3, \ldots, X_k^C\right\}$.

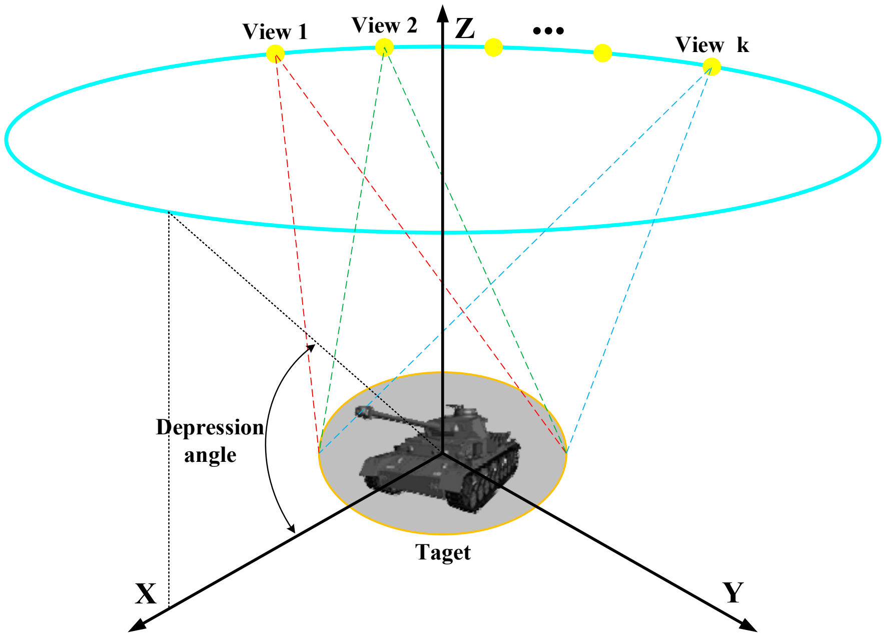


Figure 3. Illustration of multi-view SAR image acquisition.

As we mentioned before, the SAR images are very sensitive to the aspect angles. Therefore, before we start the training process, some pre-processing concerning the aspect angles is required. As such, in the second step, the rotation operation will be carried out for all training and testing data to make them stay with the same orientation. For each specific aspect angle of each SAR image, the MSTAR dataset provides us with its precise value for every aspect while we can calculate these aspects by the real-time GPS information of airborne SAR in real ATR scenarios.

Suppose that the multi-view value k is 3, reflecting three SAR images in the sorted 3-view SAR image dataset X_3^s. These three SAR images are $A_{\theta_1}, A_{\theta_2}, A_{\theta_3}$ with the relevant aspect angles $\theta_1, \theta_2, \theta_3$, respectively. After the rotation operation for these three images, they can be depicted as follows:

$$\begin{cases} R = \text{rotate}(A_{\theta_1}, \theta_2 - \theta_1) \\ G = \text{rotate}(A_{\theta_2}, 0) \\ B = \text{rotate}(A_{\theta_3}, \theta_2 - \theta_3) \end{cases} \tag{12}$$

where $\text{rotate}(X, \varphi)$ represents that the image X is rotated by φ degrees counterclockwise.

Actually, we are not likely to obtain the specific orientation of the target in advance in the real scene, and thus, we cannot directly attain accurate information of the ground target, such as the three aspect angles $\theta_1, \theta_2, \theta_3$. However, the airborne SAR is capable of acquiring the angle difference between the two adjacent observation angles in the process of SAR images acquisition with the help of GPS information, which means that the value of the angle difference, like $\theta_2 - \theta_1, \theta_2 - \theta_3$, can be obtained. Therefore, the flight platform only needs to acquire the diverse SAR images of the ground target at multiple continuous azimuth angles to meet the requirements of the inputs of MS-CNN designed in this paper. Our proposed MS-CNN outperforms other state-of-the-art methods due to its easy and quick maneuverability in this regard. More specifically, this method only needs multiple continuous aspect information, which can be obtained from the real-time GPS information within a small time slot, to create testing samples for ATR recognition tasks.

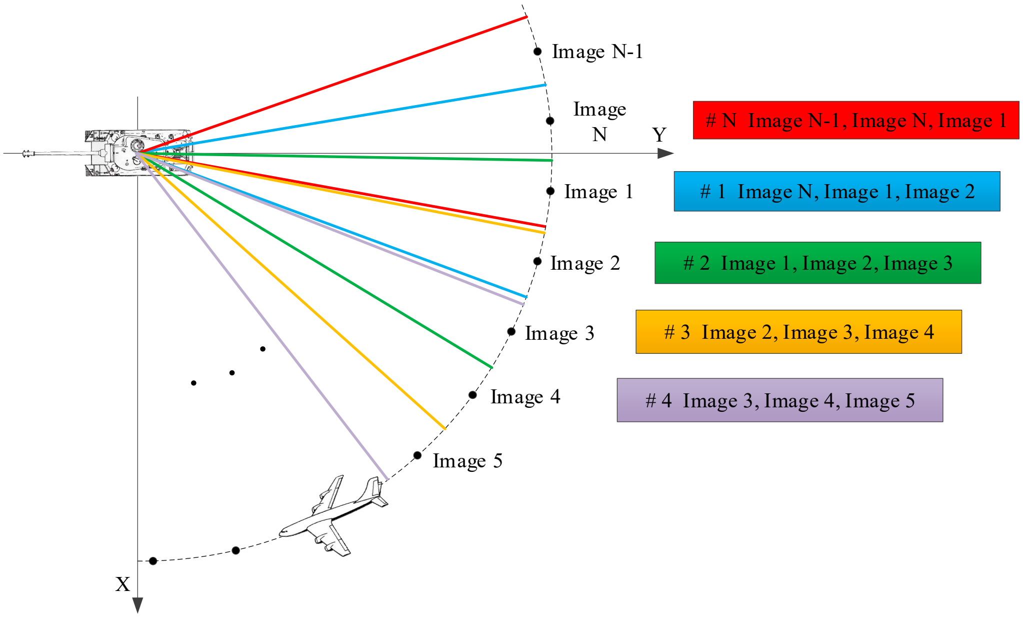


Image N-1
Image N
Y
Image 1
Image 2
Image 3
Image 4
Image 5
X
# N Image N-1, Image N, Image 1
# 1 Image N, Image 1, Image 2
# 2 Image 1, Image 2, Image 3
# 3 Image 2, Image 3, Image 4
# 4 Image 3, Image 4, Image 5

Figure 4. Example of three-view SAR image combinations.

3. Results

3.1. Implementation Details

In the experiments, we deploy three network instances with two, three, and four-view inputs to comprehensively assess the recognition performance of MS-CNN. As mentioned before, the MS-CNN is composed of one multi-stream convolutional layer, one Fourier feature fusion layer, and one softmax layer. To tell the details of these three instances, the size of multi-view SAR image inputs is 80 × 80, the kernel sizes of four convolutional layers are 5 × 5, 5 × 5, 6 × 6, and 5 × 5, respectively, the strides of convolutional layers and pooling layers are 1 × 1 and 2 × 2, respectively, and the dropout ratio is set as 0.5. As shown in Figure 2, the multi-stream convolutional layer removes the flattening operation by setting aforementioned parameters, largely reducing the number of parameters. Of course, the setting of hyperparameters depends on the specific experimental methods.

For our proposed MS-CNN, the framework Tensorflow 1.2 is applied to implement our design. As for hardware supports, a server with four Nvidia TITAN XP GPU is employed for training and testing our proposed network. The parameter for weight decay is 0.00001, and we choose the stochastic optimization algorithm Adam with the cross-entropy loss function to learn the parameters of MS-CNN. The learning rate begins with 0.001 and with 0.96 exponential decay every epoch, and the mini-batch size is set to 24. The epochs of training process vary from 20 to 30 with a constant interval of 5 epochs.

3.2. Dataset

We use the MSTAR dataset provided by Sandia National Laboratory, and in this dataset, all images have a resolution of 0.3 m × 0.3 m, and each target covers each azimuth from 0° to 360°, covering military targets of different categories, different models, different azimuth angles, and different depression angles. However, only a small proportion are publicly available. The publicly released datasets consist of ten different categories of ground targets (BMP-2, BRDM-2, BTR-60, BTR-70, T-62, T-72, 2S1, ZSU-23/4, ZIL-131, and D7), and this available MSTAR benchmark dataset is widely used to evaluate and verify the recognition performance of SAR ATR methods.

On the basis of public SAR data, we have undertaken extensive experiments on this dataset under both the Standard Operating Condition (SOC) and Extended Operating Condition (EOC). Specifically, SOC assumes that the training and testing sets hold the same serial number and target configurations while there are some variations under EOC between training and testing sets, including depression angle variants, target configuration and version variants. Finally, we conduct a comprehensive performance comparison with other state-of-the-art methods mainly from the recognition rates and the number of network parameters. Moreover, our proposed MS-CNN consistently gains high recognition rates and outperforms other previous methods.

3.3. Experiments under SOC

The experiment under SOC is the classic experiment of 10 class ground target recognition; the SAR image dataset consists of T62, T72, BMP2, BRDM2, BTR60, BTR70, D7, ZIL131, ZSU23/4, and 2S1. The optical images and relevant SAR images of the same orientation are shown in Figure 5. It can be seen that the optical images of different targets vary greatly, and their corresponding SAR images also have discernable differences observable by human eyes. Table 1 shows the class types and the number of training samples and test samples used in the experiment. Among them, SAR images acquired at a 17° depression angle were used for training, and SAR images acquired at a 15° depression angle were used for testing.

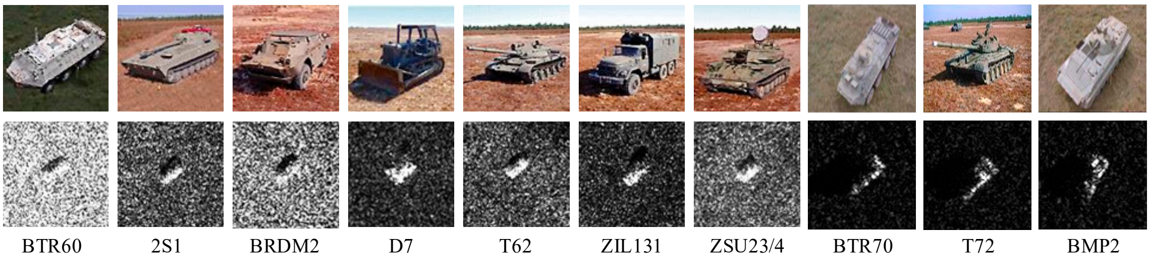


Figure 5. Optical images and SAR images of targets under SOC.

Table 1. Dataset of Experiment under SOC.

Class	2S1	BMP2	BRDM2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU23/4
Training samples	299	233	298	256	233	299	299	232	299	299
Testing samples	274	195	274	195	196	274	273	196	274	274

Tables 2–4 show the recognition accuracy confusion matrix of a two, three, and four-view MS-CNN, respectively. The confusion matrix is widely used for performance illustration; each row in the confusion matrix represents the real category to which the target belongs, and each column represents the prediction result of the network. We found that the recognition rates increase with the change of the number of views under SOC in Tables 2–4, reaching 99.84%, 99.88%, and 99.92%, respectively. From this increase among these three instances, we can conclude that our proposed MS-CNN is able to identify and extract more features from multiple views to improve the recognition performance along with the increasing views, while the recognition rates of the four-view instance are nearly all correct. In Tables 3 and 4, we can see that the testing targets of nine classes have been completely identified (except for BTR60 partly because these types of tanks look similar in terms of the appearance and seem hard to classify in certain aspects), and the overall recognition rate reaches 99.88% and 99.92%, which reaffirms that the proposed MS-CNN in this paper can effectively identify the SAR targets.

Table 2. Confusion matrix of a two-view MS-CNN under SOC.

Class	2S1	BMP2	BRDM2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU234	P_{CC} (%)
2S1	272	0	0	0	1	0	1	0	0	0	99.27
BMP2	0	195	0	0	0	0	0	0	0	0	100.00
BRDM2	0	0	274	0	0	0	0	0	0	0	100.00
BTR60	0	0	2	193	0	0	0	0	0	0	98.97
BTR70	0	0	0	0	196	0	0	0	0	0	100.00
D7	0	0	0	0	0	274	0	0	0	0	100.00
T62	0	0	0	0	0	0	273	0	0	0	100.00
T72	0	0	0	0	0	0	0	196	0	0	100.00
ZIL131	0	0	0	0	0	0	0	0	274	0	100.00
ZSU234	0	0	0	0	0	0	0	0	0	274	100.00
Total											99.84

Table 3. Confusion matrix of a three-view MS-CNN under SOC.

Class	2S1	BMP2	BRDM2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU234	P_{CC} (%)
2S1	274	0	0	0	0	0	0	0	0	0	100.00
BMP2	0	195	0	0	0	0	0	0	0	0	100.00
BRDM2	0	0	274	0	0	0	0	0	0	0	100.00
BTR60	0	0	0	192	1	0	0	0	2	0	98.46
BTR70	0	0	0	0	196	0	0	0	0	0	100.00
D7	0	0	0	0	0	274	0	0	0	0	100.00
T62	0	0	0	0	0	0	273	0	0	0	100.00
T72	0	0	0	0	0	0	0	196	0	0	100.00
ZIL131	0	0	0	0	0	0	0	0	274	0	100.00
ZSU234	0	0	0	0	0	0	0	0	0	274	100.00
Total											99.88

Table 4. Confusion matrix of a four-view MS-CNN under SOC.

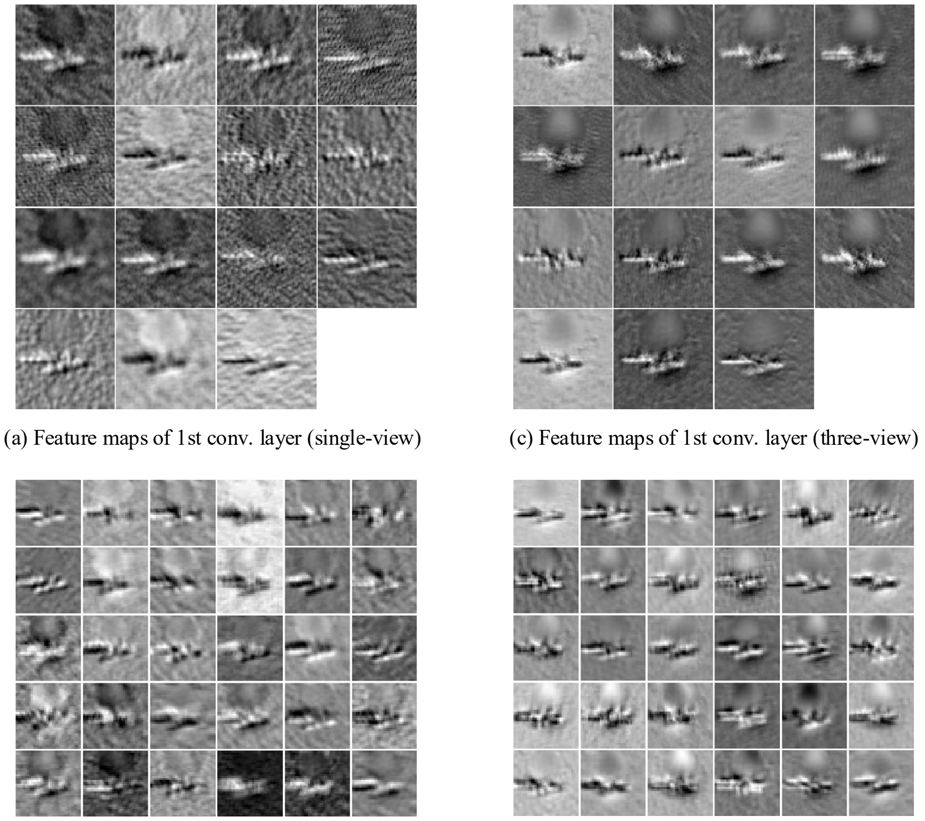
Class	2S1	BMP2	BRDM2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU234	P_{CC} (%)
2S1	274	0	0	0	0	0	0	0	0	0	100.00
BMP2	0	195	0	0	0	0	0	0	0	0	100.00
BRDM2	0	0	274	0	0	0	0	0	0	0	100.00
BTR60	0	0	2	193	0	0	0	0	0	0	98.97
BTR70	0	0	0	0	196	0	0	0	0	0	100.00
D7	0	0	0	0	0	274	0	0	0	0	100.00
T62	0	0	0	0	0	0	273	0	0	0	100.00
T72	0	0	0	0	0	0	0	196	0	0	100.00
ZIL131	0	0	0	0	0	0	0	0	274	0	100.00
ZSU234	0	0	0	0	0	0	0	0	0	274	100.00
Total											99.92

Table 5 shows the comparison of our MS-CNN with other methods from the perspective of FLOPs, number of parameters, and recognition accuracy. It can be seen that the recognition rate is loosely related to the number of parameters of the network. In other words, the recognition rates increase along with the quantity of parameters, indicating that too few parameters are not sufficient to extract enough effective features from different categories of targets, and result in lower recognition rates. In order to achieve high recognition rates, Furukawa's ResNet-18 mentioned in [25] uses millions of parameters and the FLOPs inevitably attains the order of magnitude of ten billion, which demands heavy computing resources and more computation time when training and testing the network, causing the low efficiency. Among these three multi-view methods, our MS-CNN obtains the highest recognition rates with the least number of parameters and FLOPs, benefiting from both parameters sharing of multi-stream convolutional layer and rational parameters setting of MS-CNN. Compared with our proposed MS-CNN, Pei et al.'s network mentioned in [26] contains more parameters, which resulted from the strategy of fusing multiple layers progressively, leading into low training efficiency. Moreover, the lack of further feature representations, like Fourier random features and Gabor features, accounts for lower recognition rates. However, the MA-BLSTM [27] encodes the Gabor features with TPLBP operator, achieving relatively high recognition rates. All in all, the comparison results of recognition rates and the quantity of parameters clearly validate the superiority of our proposed MS-CNN in the SOC scenario.

Table 5. Comparison of the number of parameters.

Input	Method	FLoating-Point Operations (FLOPs)	Number of Parameters	P_{CC} (%)
Typical	Morgan's [13]	2.514×10^7	8.8×10^4	92.30
	A-ConvNets [11]	3.761×10^7	3.03×10^5	99.13
	Furukawa's [25]	1.244×10^{10}	2.75×10^6	99.56
Multi-view	MA-BLSTM [27]	7.562×10^8	9.58×10^6	99.90
	2-VDCNN [26]	1.667×10^8	2.22×10^6	97.81
	3-VDCNN [26]	2.235×10^8	2.38×10^6	98.17
	4-VDCNN [26]	2.506×10^8	2.87×10^6	98.52
	MS-CNN (2-view)	5.044×10^7	2.59×10^5	99.84
	MS-CNN (3-view)	7.566×10^7	2.60×10^5	99.88
	MS-CNN (4-view)	1.008×10^8	2.61×10^5	99.92

Figure 6 shows the comparison of recognition degree of feature maps to verify the robustness of MS-CNN, mainly including two feature maps for both single-view inputs and three-view inputs, respectively. As mentioned before, the initial SAR image input is 80 × 80, the outputs of the first convolutional layer are 76 × 76 × 15 feature maps, and then it outputs 34 × 34 × 30 feature maps after the second convolutional layer. Figure 6a,c show 15 feature maps acquired at the first convolutional layer. Apparently, the coherent speckle noise in the raw SAR image has a strong impact on the feature maps in Figure 6a, so the targets and shadows are not obvious at all, and therefore, hard to distinguish, because the ambient noise around the target is amplified. However, the targets in the feature maps of Figure 6c are clearly visible, including both the outline of the target and the shadow, without much influence on recognition by speckle noise. Therefore, we can conclude that the three-view SAR image input, containing more information of the targets, is more robust, and can alleviate the effect of speckle noise. Figure 6b,d are feature maps acquired from the second convolutional layer, from which we can see that the features in Figure 6b becomes turbulent, but the feature map in Figure 6d is still very clear—the features extracted in Figure 6d are much better than those in Figure 6b, which means higher recognition rates.



(a) Feature maps of 1st conv. layer (single-view)

(c) Feature maps of 1st conv. layer (three-view)

(b) Feature maps of 2nd conv. layer (single-view)

(d) Feature maps of 2nd conv. layer (three-view)

Figure 6. Comparison of recognition degree of feature maps.

3.4. Experiments under EOC

According to the EOC experiment settings of A-ConvNets [11], we first evaluated the EOC performance with respect to a large depression angle (30°). This big change in depression angle from the 15° of SOC to 30° seems to damage the testing performance in most of the current existing methods, because of the sensitivity characteristic of SAR ATR for depression and aspect angle variance. The MSTAR dataset only contains four types of target samples for testing (2S1, BRDM-2, T72, and ZSU-23/4), which are observed at a depression angle of 30°, while the four corresponding training samples are at the depression angle of 17°. As shown in Figure 7, the optical images of the four different types of tanks used in this experiment correspond the following SAR images with the same direction. Therefore, we validate the EOC performance of a large depression angle change, called EOC-1, on these four samples, as listed in Table 6. The corresponding recognition confusion matrix is shown in Table 7.

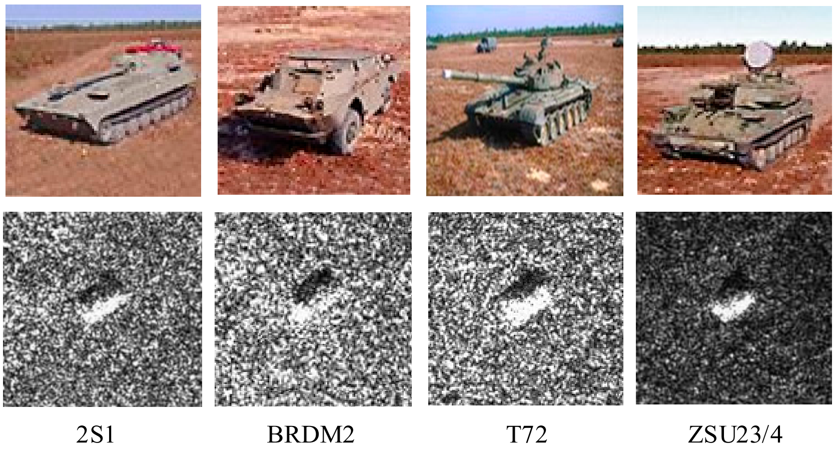


Figure 7. Optical images and SAR images of four different types of tanks at a depression angle of 30°.

Table 6. Dataset of EOC-1.

Class	2S1	BRDM2	T72	ZSU234
Training samples (17°)	299	298	232	299
Testing samples (30°)	288	287	288	288

Table 7 shows that the proposed MS-CNN with two, three, and four views achieves great recognition performance in this EOC experiment, reaching 96.96%, 97.48%, and 98.61%, respectively. Specifically, in the four-view MS-CNN experiment, the recognition rates of 2S1, BRDM2, and ZSU23/4 are more than 98%, while for type T72 they are still 96.88%. This lower performance of T72 is caused by the difference in both the depression angle and serial number variation of training and testing samples, since other tank types change only in depression angle. We can draw the conclusion that the proposed MS-CNN is robust and resilient to the sensitive depression angle variation.

Table 7. Confusion matrix of the MS-CNN under EOC-1 (large depression angle).

Views	Class	2S1	BRDM2	T72	ZSU234	P_{CC} (%)	Total
2-views	2S1	271	14	3	0	94.10	96.96
	BRDM2	3	284	0	0	98.95	
	T72	0	3	275	10	95.49	
	ZSU23/4	2	0	0	286	99.31	
3-views	2S1	272	6	10	0	94.44	97.48
	BRDM2	3	284	0	0	98.95	
	T72	0	6	278	4	96.53	
	ZSU23/4	0	0	0	288	100.00	
4-views	2S1	283	3	2	0	98.26	98.61
	BRDM2	1	286	0	0	99.65	
	T72	0	3	279	6	96.88	
	ZSU23/4	0	1	0	287	99.65	

As for the target configuration variants and version variants, another two EOC experiments were carried out to evaluate the performance of the MS-CNN with respect to EOC. Configuration variants are different from version variants. According to their definition, version variants are built to different blueprints, while configuration variants are built to the same blueprints but have had different post-production equipment added. Specifically, the MSTAR data includes version variants of the T-72 and the BTR-70. Version variants occur when the chassis of the original version has been adapted for an alternate function, such as Personnel Carrier, Ambulance, Command Post, Reconnaissance, and so forth. On the other hand, configuration variants involve the addition or removal of objects, not due to damage. Examples of how configurations may vary include fuel drums on the back of a T72, crewmembers on the vehicle, and mine excavation equipment on the front of vehicle. Moreover, configuration variants also involve the rotation or repositioning of objects, including turret rotations, opening of doors and hatches, and repositioning of tow cables. In these two EOC experiments, we selected four categories of targets (BMP-2, BRDM-2, BTR-70, and T-72) as training samples with 17° depression angles from Table 1, while the testing samples—acquired at both 17° and 15° depression angles—consisted of two-version variants of BMP-2 and ten-version variants of T-72, as listed in Tables 8 and 9. These two relevant recognition confusion matrixes are shown in Tables 10 and 11.

Table 8. Dataset of EOC-2 (version variants).

Class	T72 (15° & 17°)				
Serial No.	S7	A32	A62	A63	A64
Testing samples	419	572	573	573	573

Table 9. Dataset of EOC-2 (configuration variants).

Class	BMP2 (15° & 17°)			T72 (15° & 17°)			
Serial No.	9566	c21	812	A04	A05	A07	A10
Testing samples	428	429	426	573	573	573	567

Table 10 reveals the MS-CNN performance under EOC-2 (version variants) in three types of situations. It shows that there were only nine images which are incorrectly classified into other types of tanks in the two-view instance, and its relevant recognition rate reaches 99.67%. Remarkably, the recognition accuracy of the three-view and four-view instances are all 100%, showing that the proposed MS-CNN is superior in discerning the targets with version variations.

Table 11, which represents the recognition performance of EOC-2 (configuration variants), shows excellent recognition ability in discriminating the BMP2 and T72 targets with configuration differences in training samples. We can see that these three instances reach the recognition accuracy of 98.71%, 99.08%, and 99.58%, respectively. It is obvious that the recognition rate rises with the increasing number of multiple views.

Table 10. Confusion Matrix of the MS-CNN under EOC-2 (version variants).

Views	Class	Serial No.	BMP2	BRDM2	BTR70	T72	P_{CC} (%)	Total
2-views	T72	A32	0	0	0	572	100.00	99.67
		A62	0	0	0	573	100.00	
		A63	0	0	0	573	100.00	
		A64	0	0	0	573	100.00	
		S7	5	0	4	410	97.85	
3-views	T72	A32	0	0	0	572	100.00	100.00
		A62	0	0	0	573	100.00	
		A63	0	0	0	573	100.00	
		A64	0	0	0	573	100.00	
		S7	0	0	0	419	100.00	
4-views	T72	A32	0	0	0	572	100.00	100.00
		A62	0	0	0	573	100.00	
		A63	0	0	0	573	100.00	
		A64	0	0	0	573	100.00	
		S7	0	0	0	419	100.00	

Table 11. Confusion matrix of the MS-CNN under EOC-2 (configuration variants).

Views	Class	Serial No.	BMP2	BRDM2	BTR70	T72	P_{CC} (%)	Total
2-views	BMP2	9566	411	3	0	14	96.03	98.71
		c21	403	4	0	22	93.94	
	T72	812	2	0	0	424	99.53	
		A04	1	0	0	572	99.83	
		A05	0	0	0	573	100.00	
		A07	0	0	0	573	100.00	
		A10	0	0	0	567	100.00	
3-views	BMP2	9566	410	3	8	7	95.79	99.08
		c21	417	4	0	8	97.20	
	T72	812	0	0	1	425	99.77	
		A04	1	0	0	572	99.83	
		A05	0	0	0	573	100.00	
		A07	1	0	0	572	99.83	
		A10	0	0	0	567	100.00	
4-views	BMP2	9566	425	2	0	1	99.30	99.58
		c21	423	4	0	2	98.60	
	T72	812	5	0	0	421	98.83	
		A04	1	0	0	572	99.83	
		A05	0	0	0	573	100.00	
		A07	0	0	0	573	100.00	
		A10	0	0	0	567	100.00	

To sum up, our proposed MS-CNN shows high performance under EOC, including with a large depression angle (EOC-1), configuration variants (EOC-2), and version variants (EOC-2), which demonstrate the significate value of MS-CNN in SAR ATR tasks.

3.5. Recognition Performance Comparison

In this section, we undertake a performance comparison between our proposed MS-CNN and ten other SAR ATR methods, including the extended maximum average correlation height filter (EMACH) [28], support vector machine (SVM) [28], adaptive boosting (AdaBoost) [28], iterative graph thickening (IGT) [28], sparse representation-based representation of Monogenic Signal (MSRC) [29], monogenic scale-space (MSS) [30], modified polar mapping classifier (M-PMC) [31], all-convolutional networks (A-ConvNets) [11], combined discrimination trees (CDT) [32], multi-aspect-aware bidirectional LSTM recurrent neural networks (MA-BLSTM) [27], and multi-view DCNNs (MVDCNNs) [26]. All these aforementioned ATR methods hold state-of-the-art performance, and therefore, we choose them for comparisons.

Although all these aforementioned SAR ATR methods are based on the MSTAR dataset, they might utilize different training dataset and implement distinct principles. In addition, the quantity of SAR images inputs varies from single-view methods to multi-view. All those factors lead to the difficulty of recognition performance comparison. We can simply compare all these methods by recognition rates and the number of training samples inputs, and assume that an ATR method with higher recognition rates but fewer inputs holds a better recognition performance. On the basis of this assumption, we select the recognition rates under both SOC and EOC and the number of network inputs as criterions to compare the recognition performance.

Table 12 shows comparison with other state-of-the-art methods, including EMACH, SVM, A-ConvNets, MA-BLSTM, and so forth. All these ATR methods are based on the MSTAR dataset, so the results cited from corresponding papers for the recognition rate comparison are reliable. As listed in Table 12, it is obvious that the deep learning approaches outperform the methods of conventional machine learning, like SVM, AdaBoost, and so forth, in the field of SAR ATR in both SOC and EOC scenarios. Moreover, the performance of the multi-view methods, MVDCNNs, MA-BLSTM, and MS-CNN, is better than that of the other deep learning methods that use single-view input, partly because the extra space-varying information extracted from interconnected multi-view images can improve the recognition performance.

Table 12. Recognition accuracy comparison of the MS-CNN and other methods.

Method	SOC	Inputs	EOC-1	Inputs	EOC-2 (Vrsion Variants)	Inputs
EMACH [28]	88	3670	77	1129	68	1593
SVM [28]	90	3670	81	1129	75	1593
AdaBoost [28]	92	3670	82	1129	78	1593
IGT [28]	95	3670	85	1129	80	1593
MSRC [29]	93.6	2747	98.4	896	-	-
MSS [30]	96.6	2747	98.2	896	-	-
M-PMC [31]	98.81	3671	-	-	97.31	996
A-ConvNets [11]	99.13	2747	96.12	698	98.93	698
CDT [32]	99.30	3681	97.50	1370	96.9	997
MA-BLSTM [27]	99.90	2320	-	-	99.59	928
2-VDCNN [26]	97.81	2754	93.29	1130	93.75	998
3-VDCNN [26]	98.17	2760	94.34	1134	95.08	1002
4-VDCNN [26]	98.52	2760	94.61	1132	95.46	1004
MS-CNN (2-view)	99.84	2747	96.96	1128	99.67	996
MS-CNN (3-view)	99.88	2747	97.48	1128	100.00	996
MS-CNN (4-view)	**99.92**	**2747**	**98.61**	**1128**	**100.00**	**996**

Due to the similar mechanisms by which MVDCNNs [26] and our proposed MS-CNN are based on the multi-view concept and CNNs, it is necessary to conduct a more detailed comparison to show the superiority of MS-CNN.

- Training and testing sample construction. Our proposed training and testing samples construction approach makes full use of MSTAR dataset to produce equivalent amounts of multi-view SAR images, while the multi-view SAR data formation approach mentioned in [26] merely leverages part of raw images of MSTAR dataset to multiply its training and testing samples by many times. In other words, MS-CNN can be better trained simply because it has more raw images from the training samples to learn, compared with the MVDCNNs. In addition, as shown in Table 12, due to multiplying the training samples, the quantity of MVDNNs inputs is slightly larger than MS-CNN. Therefore, we can conduct a conclusion that our proposed multi-view training samples construction method is more effective.

- Network architecture. Since we incorporate the Fourier feature fusion layer into MS-CNN to achieve high-level and holistic representation, MVDCNNs only rely on the great strength of CNNs, so naturally, some limitations to further improving the recognition rates exist.
- Time complexity. In MS-CNN, we remove the flattening operation by setting rational parameters such as the size of the convolutional kernel and pooling, largely reducing the number of parameters, and then decreasing time complexity. Moreover, we parallelly conceive a multi-stream convolutional layer to extract features of multi-view SAR images, instead of fusing feature maps from inputs to last layer progressively in the network topology described in MVDCNNs. This design makes possible to share parameters among multi-view inputs, further reducing the quantity of parameters and accelerating the training process.

All these experiments carried out in this paper reveal that the proposed MS-CNN has a better generalization and recognition ability than other state-of-the-art methods, and naturally achieves a superior recognition performance.

4. Discussion

We have conducted extensive experiments on the MSTAR benchmark dataset, and the aforementioned experimental results have also verified the superiority of our proposed MS-CNN under both SOC and EOC, compared with ten state-of-the-art SAR ATR methods. In this section, we will mainly discuss the feasibility and reasonability, and future work regarding the MS-CNN.

4.1. Feasibility and Reasonability Discussion

At first, in terms of the feasibility and reasonability of multiple views implemented with the deep learning approaches in SAR ATR, we have undertaken comprehensive investigations from existing literature regarding deep learning, SAR ATR with single inputs, and ATR with multi-view inputs.

- Deep Learning. Deep learning has attained significant development in the fields of natural language processing, speech recognition, target detection, image classification, human-machine games, and autopilot. Naturally, many novel deep learning algorithms and systems have been proposed, including convolutional neural networks, deep belief networks (DBNs) [33], and recurrent neural networks (RNNs) [34]. Most of them, especially CNN, have been widely used in the field of computer vision, such as in target detection and target recognition. In 2012, Krizhevsky et al. designed an AlexNet deep learning network with an eight-layer network structure; he won the championship 2012 ILSVRC, with a Top5 error rate of 15.3%, which is much lower than the previous 26% [35]. Szegedy et al. designed a Google Inception network with a 22-layers network structure in 2014 which largely reduced the number of parameters and calculations and won the championship with a 6.67% Top5 error rate [36]. In 2015, Kaiming et al. continued to deepen the network hierarchy, and proposed a 152-layers Residual Network (ResNet), reducing the error rate to 3.57%, which exceeded the manual error rate 5% [37].
- SAR ATR with Single Inputs. Many target recognition algorithms proposed for optical images have been widely applied to SAR images with a high probability of correct cognition (PCC). In 2014, Chen et al. designed a convolutional neural network with a single hidden layer to identify SAR image targets, and achieved a recognition accuracy of 84.7% on a 10 class military target dataset [38]. The same year, Chen et al. designed a novel convolutional neural network, with five convolutional layers and three pooling layers, for SAR image target recognition. In order to refine the objective conditions of limited raw SAR image data and sensitive observation conditions, they replaced the fully connected layers with convolutional layers, and achieved 99.13% recognition accuracy [11]. However, they augmented the training data set by means of

randomly cropping and flipping, which increased the scale of image training samples 10 times compared to the original, and the cost was an increase in training time. In 2017, Furukawa et al. designed a network structure with 18 convolutional layers for SAR image target recognition by imitating the idea of the residual network and achieved extremely high recognition accuracy through data augmentation methods such as random cropping [25]. However, the parameters of this network were up to a million levels, which would take up a huge amount of computing resources, and consume much time when training the network.

- SAR ATR with Multi-View Inputs. Most of the SAR target recognition approaches only use a single view of the observation target as the input of a network, without considering the acquisition characteristics of the SAR images. Recently, researchers have studied ATR problems with multi-view images from multiple aspects, and have reached an agreement that multi-view inputs could enhance the recognition rates [18]. In 2011, Zhang et al. introduced a joint sparse representation based multi-view ATR method and achieved 94.69% recognition rate on a 10 class problem with three consecutive views on the MSTAR database [39]. In 2017, Zhang et al. proposed a multi-aspect aware bidirectional LSTM recurrent neural network with a Gabor filter and Three-Patch Local Binary Pattern (TPLBP) extracting the spatial features, followed by a fully-connected Multi-Layer Perceptron (MLP) network reducing the feature dimensionality. Although this novel idea attained 99.90% recognition accuracy on 10 class problem using an MSTAR dataset and showed good anti-noise and anti-confusion performance [27], it is probably troublesome to apply in real ATR situations, because it needs a 50 images sequence as the input, which will take too much time to be suitable for real-time scenarios. Moreover, the structure of BA-LSTM is very complex, and the cost of this network is much higher due to longer training time and bigger storage. In 2018, Pei et al. proposed a deep convolutional neural network framework with multi-view images, containing a parallel topology in which the learned features from the distinct views can be fused progressively. In addition, this literature adopted pre-processing of data augmentation and finally achieved 98.17% on the 10 class problem using an MSTAR dataset [26]. However, these SAR image input sequences with a specific aspect interval may not perform well if the testing image falls out of default interval.

From these investigations, we can reasonably conclude that the multi-view method combined with the deep learning algorithm has great strength to improve the recognition rates in SAR ATR compared with single inputs. And our proposed MS-CNN not only leverages the strength of multiple views to extract potential space-varying features of multi-view SAR images, but also introduces the Fourier feature fusion framework into the multi-view architecture to fuse these features in a kernel level to achieve more holistic representations. Therefore, our proposed MS-CNN is both feasible and reasonable and can effectively and efficiently tackle the SAR ATR tasks in real time.

In addition, the reasonability of experiments for compressively evaluating the proposed MS-CNN is discussed as follows.

- Experiments under SOC. SOC refers to the same serial numbers and target configuration for both training and testing samples, but with only a 2° depression angle difference. This typical experimental setting aims to evaluate whether the proposed network has the ability to classify targets in general situations. Our proposed MS-CNN achieves 99.84%, 99.88%, and 99.92% in the two, three, and four-view instances, respectively, which can demonstrate its superior recognition ability; it outperformed other state-of-the art methods by large margin.
- Experiments under EOC. EOC, which represents the extreme circumstances in ATR tasks, mainly consists of the large depression angle gap, target configuration variants, and version variants. Those methods which can effectively recognize corresponding categories of targets with high recognition rates have the characteristic of robustness for changeable attributes, and can be better served in real SAR ATR tasks. Compared with other ten excellent methods, our proposed MS-CNN again shows its great superiority for recognition in the EOC scenarios. Remarkably,

the experiment of EOC-2 (version variants) attains the recognition rates of 100%, which reaffirms its excellent capability in ATR tasks.

4.2. Future Work

In order to further improve the recognition performance of ATR in future work, we will make some attempts to explore the ATR algorithm with a mathematical approach, like introducing the orthogonal low-rank loss into our proposed MS-CNN, to see whether it can further improve the recognition performance. In addition, we will attempt to apply the MS-CNN to other fields of computer version, especially inspired by handcrafted features extraction and classification task mentioned in [40]. We could perhaps exploit output of MS-CNN as a feature vector to feed an SVM to obtain better classification performance.

5. Conclusions

In this paper, we have presented a novel convolutional learning architecture, i.e., multi-stream convolutional neural networks (MS-CNN) for multi-view SAR ATR. The MS-CNN is composed of a multi-stream convolutional layer, a Fourier feature fusion layer, a fully connected layer, and a softmax layer. Specifically, the multi-view SAR image features can be efficiently extracted by the multi-stream convolutional layer, and then combined by the Fourier feature fusion layer, and finally successively fed into the fully connected layer and softmax layer for classification. These layers jointly establish the nonlinear relationships between raw SAR images and corresponding classes, making full use of the discriminating space-varying features of limited raw SAR images to enhance the classification performance and robustness. In addition, our proposed MS-CNN is qualified with the desired characteristic of easy and quick maneuverability in real SAR ATR scenarios, because it only needs to acquire real-time GPS information of airborne SAR to calculate multiple aspects. Experimental results on the MSTAR dataset have shown that the recognition performance of our MS-CNN surpasses other state-of -the-art methods under both SOC and EOC. Thus, our proposed MS-CNN is an effective method for SAR target recognition and offers promise for wider SAR ATR applications.

Author Contributions: P.Z. and H.Z. conceived and designed the experiments; P.Z. performed the experiments; X.Z. assisted to ameliorate the idea and the experiments; K.L. and X.Z. provided materials and computing resources; All authors contributed in writing the paper.

Funding: This research was funded by the National Key R & D Program of China under grant number. 2016YFB1200100, NSFC under grant numbers: 61871016, 61271195, and 61231013, National Science Foundation for Innovative Research Groups of China under grant number 61521091.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Moreira, A.; Prats-Iraola, P.; Younis, M.; Krieger, G.; Hajnsek, I.; Papathanassiou, K.P. A tutorial on synthetic aperture radar. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–43. [CrossRef]
2. Wang, P.; Liu, W.; Chen, J.; Niu, M.; Yang, W. A high-order imaging algorithm for high-resolution spaceborne SAR based on a modified equivalent squint range model. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 1225–1235. [CrossRef]
3. Blacknell, D.; Griffiths, H. Radar Automatic Target Recognition (ATR) and Non-Cooperative Target Recognition (NCTR). *IET Digit. Libr.* **2013**, *296*, 5–35.
4. El-Darymli, K.; Gill, E.W.; Mcguire, P.; Power, D.; Moloney, C. Automatic target recognition in synthetic aperture radar imagery: A state-of-the-art review. *IEEE Access* **2017**, *4*, 6014–6058. [CrossRef]
5. Potter, L.C.; Moses, R.L. Attributed scattering centers for SAR ATR. *IEEE Trans. Image Process.* **1997**, *6*, 79–91. [CrossRef] [PubMed]
6. Novak, L.M.; Halversen, S.D.; Owirka, G.; Hiett, M. Effects of polarization and resolution on SAR ATR. *IEEE Trans. Aerosp. Electron. Syst.* **1997**, *33*, 102–116. [CrossRef]
7. Dudgeon, D.E.; Lacoss, R.T. An overview of automatic target recognition. *Linc. Lab. J.* **1993**, *6*, 3–10.

8. Sun, Y.; Liu, Z.; Todorovic, S.; Li, J. Adaptive boosting for SAR automatic target recognition. *IEEE Trans. Aerosp. Electron. Syst.* **2007**, *43*, 112–125. [CrossRef]
9. Hummel, R. Model-based ATR using synthetic aperture radar. In Proceedings of the Record of the IEEE 2000 International Radar Conference, Alexandria, VA, USA, 12 May 2000; pp. 856–861.
10. Zhao, Q.; Principe, J.C. Support vector machines for SAR automatic target recognition. *IEEE Trans. Aerosp. Electron. Syst.* **2001**, *37*, 643–654. [CrossRef]
11. Chen, S.; Wang, H.; Xu, F.; Jin, Y. Target classification using the deep convolutional networks for SAR images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4806–4817. [CrossRef]
12. Ding, J.; Chen, B.; Liu, H.; Huang, M. Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 364–368. [CrossRef]
13. Morgan, D.A. Deep convolutional neural networks for ATR from SAR imagery. In Proceedings of the Algorithms for Synthetic Aperture Radar Imagery XXII, Baltimore, MD, USA, 13 May 2015; pp. 1–13.
14. Wilmanski, M.; Kreucher, C.; Lauer, J. Modern approaches in deep learning for SAR ATR. *Int. Soc. Opt. Photonics* **2016**, *9843*, 1–10.
15. Shin, H.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.J.; Summers, R.M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Image* **2016**, *35*, 1285–1298. [CrossRef] [PubMed]
16. Li, X.; Li, C.S.; Wang, P.B.; Men, Z.R.; Xu, H.P. SAR ATR based on dividing CNN into CAE and SNN. In Proceedings of the 5th Asia-Pacific Conference on Synthetic Aperture Radar, 1–4 September 2015; pp. 676–679.
17. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3D shape recognition. In Proceedings of the International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 945–953.
18. Johns, E.; Leutenegger, S.; Davison, A.J. Pairwise decomposition of image sequences for active multi-view recognition. *Comput. Vis. Pattern Recognit.* **2016**, 3813–3822.
19. Loffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
20. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the International Conference on International Conference on Machine Learning, Haifa, Isreal, 21–24 Jnue 2010; pp. 807–814.
21. Lecun, Y.; Kavukcuoglu, K.; Farabet, C. Convolutional networks and applications in vision. *IEEE Int. Symp. Circuits Syst.* **2011**, *14*, 253–256.
22. Li, F.; Sminchisescu, C. Fourier kernel learning. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2012; Springer: Heidelberg, Germany, 2012; pp. 459–473.
23. Cho, Y.; Saul, K.L. Kernel Methods for Deep Learning. In *Advances in Neural Information, Proceedings of Neural Information Processing Systems 2009, Vancouver, BC, Canada, 7-10 December 2009*; Curran Associates, Inc.: Red HooK, NY, USA, 2009; pp. 342–350.
24. Bochner, S. *Lectures on Fourier Integrals*; Princeton University Press: Princeton, NJ, USA, 1959.
25. Srinivas, U.; Monga, V.; Raj, R.G. SAR automatic target recognition using discriminative graphical models. *IEEE Trans. Aerosp. Electron. Syst.* **2014**, *50*, 591–606. [CrossRef]
26. Dong, G.; Wang, N.; Kuang, G. Sparse representation of monogenic signal: With application to target recognition in SAR Images. *IEEE Signal. Process. Lett.* **2014**, *21*, 952–956.
27. Dong, G.; Kuang, G. Classification on the monogenic scale space: application to target recognition in SAR image. *IEEE Trans. Image Process.* **2015**, *24*, 2527–2539. [CrossRef] [PubMed]
28. Park, J.I.; Kim, K.T. Modified polar mapping classifier for SAR automatic target recognition. *IEEE Trans. Aerosp. Electron. Syst.* **2014**, *50*, 1092–1107. [CrossRef]
29. Zhao, X.; Jiang, Y.; Stathaki, T. Automatic target recognition strategy for synthetic aperture radar images based on combined discrimination Trees. *Comput. Intell. Neurosci.* **2017**, *2017*, 7186120. [CrossRef] [PubMed]
30. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
31. Williams, R.J.; Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* **1989**, *1*, 270–280. [CrossRef]

32. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. In imageNet classification with deep convolutional neural networks. In Proceedings of the International Conference on Neural Information Processing Systems, Nevada, NV, USA, 3–6 October 2012; pp. 1097–1105.
33. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erthan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imageNet classification. In Proceedings of the IEEE International Conference on Computer Vision, Washington DC, WA, USA, 9 October 2015; pp. 1026–1034.
35. Chen, S.; Wang, H. SAR target recognition based on deep learning. In Proceedings of the International Conference on Data Science and Advanced Analytics, Shanghai, China, 30 October–2 November 2015; pp. 541–547.
36. Furukawa, H. Deep learning for target classification from SAR imagery: Data augmentation and translation invariance. *IEICE Tech. Rep.* **2017**, *117*, 11–17.
37. Zhang, H.; Nasrabadi, N.M.; Zhang, Y.; Huang, T.S. Multi-view automatic target recognition using joint sparse representation. *IEEE Trans. Aerosp. Electron. Syst.* **2012**, *48*, 2481–2497. [CrossRef]
38. Zhang, F.; Hu, C.; Yin, Q.; Li, W.; Li, H.; Hong, W. SAR target recognition using the multi-aspect-aware bidirectional LSTM recurrent neural networks. *arXiv*, 2017; arXiv:1707.09875.
39. Pei, J.; Huang, Y.; Huo, W.; Zhang, Y.; Yang, J.; Yeo, T. SAR automatic target recognition based on multiview deep learning framework. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2196–2210. [CrossRef]
40. Nanni, L.; Ghidoni, S.; Brahnam, S. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognit.* **2017**, *71*, 158–172. [CrossRef]