Received: 15 September 2022 | Revised: 24 October 2022 | Accepted: 1 November 2022 | Published online: 2 November 2022

RESEARCH ARTICLE

Artificial Intelligence and Applications
2022, Vol. 00(00) 1–11
DOI: 10.47852/bonviewAIA2202412

BON VIEW PUBLISHING

Multi-Stream Fast Fourier Convolutional Neural Network for Automatic Target Recognition of Ground Military Vehicle

Olalekan J Awujoola[1,*], P.O. Odion[1], A.E. Evwiekpaefe[1] and G.N. Obunadike[1,2]

[1]*Department of Computer Science, Nigerian Defence Academy, Nigeria*

[2]*Department of Computer Science, Federal University Dutsin-Ma Katsina, Nigeria*

Abstract: The synthetic aperture radar (SAR) is very useful in both military and civilian applications due to its 24/7, all-weather, and high-resolution capabilities, as well as its ability to recognize camouflage and penetrating cover. In the field of SAR image interpretation, target recognition is an important research challenge for researchers all over the world. With the application of high-resolution SAR, the imaging area has been expanding, and different imaging modes have appeared one after another. There are many difficulties with the conventional understanding of human interpretation. There are issues like slow movement, a lot of labor, and poor judgment. Technology for intelligent interpretation needs to be developed immediately. Although deep convolutional neural networks (CNNs) have proven extremely efficient in image recognition, one of the major drawbacks is that they require more parameters as their layers increase. The cost of convolution operation for all convolutional layers is therefore high, and learning lag results from the inevitable rise in computation as the size of the image kernel grows. This study proposes a three ways input of SAR images into multi-stream fast Fourier convolutional neural network (MS-FFCNN). The technique elaborates on the transformation of rudimentary multi-stream CNN into MS-FFCNN. By utilizing the fast Fourier transformation instead of the standard convolution, it lowers the cost of image convolution in CNNs, which lowers the overall computational cost. The multiple streams of FFCNN overcome the problem of insufficient sample size and further improve the long training time while also improving the recognition accuracy. The proposed method yielded good recognition accuracy of 99.92%.

Keywords: automatic target recognition, convolutional neural network, fast Fourier transform, ground military vehicle, multi-stream, synthetic aperture radar

1. Introduction

The synthetic aperture radar (SAR) imaging technology has gradually matured with the development of large-scale integrated circuits and high-performance electronic devices. Due to its active sensor nature, SAR imaging is less affected by the effects of weather, illumination, and other conditions [1]. Military and civilian applications both benefit from its all-weather, high-resolution capabilities, camouflage recognition, and piercing cover capabilities, among other attributes. Academic researchers all over the world are grappling with the challenge of SAR target recognition. This makes SAR image interpretation extremely important. With the advancement of SAR imaging technologies and the increasing resolution accuracy of SAR images, understanding and using these images have become a pressing issue. There are many difficulties with the conventional understanding of human interpretation. There are issues like slow movement, a lot of effort, and poor judgment. The design of intelligent interpretation technology is essential [2].

Deep learning has advanced dramatically in recent years, finding considerable success across a wide range of industries. Deep learning-based approaches, in contrast to traditional algorithms, frequently use hierarchical designs, such as deep neural networks (DNNs), to extract feature representations of raw data for a variety of applications [3]. Through stacks of convolutional and pooling layers, convolutional neural networks (CNNs) can extract low- and high-level features from raw images and then use these features to perform computer vision tasks such as large-scale image recognition, semantic segmentation, and object detection [4, 5].

However, one of the issues with CNN training is that it is quite expensive to operate all of the convolutional layers. Particularly, an increase in the image or kernel size unavoidably results in an increase in computation, which causes a learning lag [6]. Using the Fourier transform to change the domain and build a CNN in the frequency domain can solve this problem, as convolution in the spatial domain is the same as pointwise multiplication in the Fourier domain. Convolution is more expensive to compute than

*Corresponding author: Olalekan J. Awujoola, Department of Computer Science, Nigerian Defence Academy, Nigeria. Email: ojawujoola@nda.edu.ng

point-by-point multiplication in general. Prior strategies have concentrated on increasing processing speed to address the time cost problem and improving recognition [7, 8].

This research suggests a multi-stream fast Fourier convolutional neural network (MS-FFCNN) to further enhance the recognition accuracy, overfitting, training time, and computational complexity of CNN for SAR target detection. In order to extract and combine multiple levels of features, this work introduced fast Fourier convolutions (FFCs) to replace the standard classical convolution in three simultaneous streams of networks. The network's computation is then further reduced by using Dropout and max-poolings. Experiments demonstrate how successful the network model achieved.

Therefore, this work significantly enhanced Zhao et al. [9] and Pei et al. [10, 14] works by substituting all of the convolutions in the three streams of the network for FFCs and converting the input data into a fast Fourier spectrum.

To the best of our knowledge, this is the first application of deep integrated fast Fourier transformed convolutions in multiple stream CNNs with multiple inputs for automatic target recognition (ATR) of SAR images, notably in classification of military ground vehicles. This approach resolves the issue of a long training time, improves recognition accuracy, and lowers the computing cost of convolution.

The rest of this research work is divided into the following section: Section 2 recaps the review of related literature and Section 3 describes the research methodology adopted in this study and brief description of the dataset used. Section 4 outlines the discussion of results obtained from the experiment and the classification algorithms, which are compared with some other recent state-of-the-art models. Finally, the conclusion and future work are expressed in Section 5.

2. Related Works

The use of numerous feature-based SAR ATR algorithms is a result of the development of machine learning. Feature extraction and categorization are often the first two stages of feature-based techniques [11]. Deep learning, a new machine learning branch, is what is driving the study of artificial intelligence [12]. Recently, it has undergone significant development and seen a growth in applications across numerous businesses and academic disciplines. To put it simply, deep learning is a neural network structure with a large number of hidden layers. The technique of deep learning involves extracting target features automatically through unsupervised learning and avoids issues caused by manual selection of features, in contrast to CNNs [13]. Pei et al. [10] suggested a typical multi-view deep CNN (DCNN) with many inputs (i.e., SAR images from different perspectives) and a low requirement for raw SAR data. The final layer of the network gradually fuses the features discovered from various points of view, resulting in classification rates for standard operating condition (SOC) and experimental operating condition (EOC) of 98% and 93%, respectively. A parallel DCNN, with many inputs, was introduced by Pei et al. [14]. An innovative processing framework was developed for a multi-view SAR ATR pattern, where SAR images were processed in a different way in each view. A multi-view SAR image can include features from both the inter- and intra-view, which are key classification features and have been completely learned by the multi-view DNN. In a problem involving ten classes, its recognition rates with three and four views were 99.30% and 99.62%, respectively, under SOC. Furukawa [15] suggested a CNN made of an encoder and a decoder called the verification support network (VersNet). One key characteristic of the network is that the input SAR picture may be of any size and comprise numerous targets belonging to various classes.

In addition, Zou et al. [16] integrated three continuous azimuth images of the same target as a pseudo-color image input, which is input into CNN. This was done because SAR images are particularly sensitive to azimuth angle. A multi-view CNN and long short-term memory (LSTM) network were created by Wang et al. [17] to extract and merge the data from various nearby azimuth angles. To increase recognition rate, Zhang et al. [13] combined CNN with CBAM, an attention mechanism. The deep semantic knowledge of the target can be extracted using the deep learning technique. It has a higher recognition rate for SAR targets than the model-based method since it eliminates the requirement for human feature extraction. Pei et al. [18] used 2-D principal component analysis-based 2-D neighborhood virtual point's discriminant embedding for SAR ATR to extract SAR image features. Features and models must, however, be retaught when fresh samples are received. Low universality and time requirements characterize this approach. Dang et al. [19] employed the incremental non-negative matrix decomposition method to examine the features online in order to solve this issue and increase the model's computational effectiveness and universality. Different classifiers can be created to categorize targets in SAR images after feature extraction. Furthermore, He et al. [20] achieved a final recognition rate of 99.47% by using CNN to categorize SAR images; nevertheless, only seven categories of targets in the moving and stationary target acquisition and recognition (MSTAR) dataset were recognized. More and more parameters need to be learned for CNN as the number of layers rises. In the meanwhile, overfitting happened frequently, which prevents the network from converging or from converging to the global optimum. Chen et al. [21] suggested a SAR image target recognition approach based on A-ConvNets, which eliminated all the fully connected (FC) layers and only contained sparse connection layers in order to reduce the number of the network parameters. At the network's end, a softmax activation function was used to provide the final classification. The MSTAR dataset was used to verify this method, and it had a 99% recognition rate, which was higher than the previous method.

Two models were developed by Bentes et al. [22], with the classification model being the second. The first model was DNNs with denoising autoencoder (DNN-DAE). This research employs DNN-Conv, also known as DNN-DAE-Conv, and uses convolutional layers (DNN with denoising autoencoder and convolution). Incorporating a DAE will allow for the learning of higher-level feature representation. The DNN-unsupervised DAE's block receives the noisy SAR input image after it has first been processed for object detection using a constant false alarm rate, after which the discovered regions have been removed, normalized, and used as input. The output of the unsupervised block then enters the supervised block, where training was carried out using labeled data from a database that contains targets that are manually recognized as well as labels produced by automated identification system. By using only pertinent characteristics, the DNN-DAE model could then provide labeled data, which it could then use for classification in the DNN-Conv model. In conclusion, the stacked autoencoder layers in the unsupervised block allowed the DNN-DAE-Conv to learn the higher-level representation of features. The authors proposed that denoising the input image before recognizing targets may be useful in subsequent research. Therefore, combining the DNN-DAE with many deeper CNNs to create a hybridized model can be worked on experimentally and may enhance the performance of the network as a whole.

Similar to this, Chen et al. [21]'s work suggests an all-CNN called A-ConvNet that focuses on addressing the overfitting and limited dataset issues when using CNN to classify MSTAR targets. The A-ConvNet architecture's originality lies in the absence of completely linked layers, which restricts the model's degrees of freedom, as well as the architecture's hyper-parameter settings. Additionally, data augmentation was carried out to increase the dataset for the A-ConvNet training, and the classification results of experiments tested under SOC and EOC revealed improvements but few misclassifications. When applied to images with just 1% noise, the model's performance decreased by about 7%. This demonstrates how poorly A-ConvNet handles noise. However, an end-to-end experiment was also carried out in which MSTAR targets detected in a congested environment were also taken into account before categorization. A-ConvNet was employed in two steps, with the first stage consisting of a binary classifier to categorize target and clutter and the second stage being the A-ConvNet itself. In images with little noise, results indicate 98% accuracy with few false alarms and incorrect recognitions. In order to enhance performance and make A-ConvNet adaptable to noisy images, the researchers recommended that it can be trained further for noisy images or that a preprocessing stage may be added to its model. Similar to this, Wang et al. [23] propose a CNN that excludes the conventional FC layers and consists only of sparsely connected levels (convolutional and pooling layers). By lowering the algorithm's free parameters, this technique prevents overfitting brought on by a lack of training images. Unquestionably, their all-convolutional networks (A-ConvNets) are among the best SAR ATR methods.

For the purpose of improving speech, Shchekotov et al. [24] designed two neural network architecture. Fast Fourier convolutional autoencoder is the first one, and FFC layers are used in U-Net architecture for the second one. It was found that in terms of improving speech quality, phase estimation, and parameter efficiency, neural network designs based on FFC greatly outperform conventional convolution-based architectures. The suggested designs are much smaller than the baselines and produce state-of-the-art results on speech denoising benchmarks. Similarly, Sinha et al. [25] proposed an architecture that learns both local and global elements and combines them to produce high-quality images. The architecture subnetwork widens its receptive area and learns long-range relationships using nonlocal attention-aided fast Fourier convolutions. Furthermore, the results demonstrate that these Fourier features implicitly offer faster convergence on low frequency components but require prior knowledge for high frequency components that are not detected. The model adapts effectively to various datasets and further maximizes the performance improvement in the ablation investigation, and the author also investigates the effect of nonlocal attention and the ratio of local and global characteristics.

According to Lu et al. [26], a novel convolutional operator known as FFC has the main features of nonlocal receptive fields and cross-scale fusion. In this model, nonlocal receptive fields are achieved by harnessing Fourier spectral theory. In the proposed operator, three types of computations are encapsulated in a single operation unit to implement cross-scale fusion: a local branch that performs ordinary small kernel convolution, a semiglobal branch that process spectrally stacked image patches, and a global branch that manipulates image-level spectrum. A consistent improvement in performance has been observed in three comprehensive vision benchmarks (ImageNet for image recognition, Kinetics for video action recognition, and MSCOCO for human key point detection) that are clearly attributed to FFC. It consistently improves accuracy in all of the aforementioned tasks by sizable margins.

2.1. Convolutional neural networks

CNNs are specialized multilayer perceptron neural networks designed to recognize two-dimensional shapes with a high level of invariance to skewing, translation, scaling, and other sorts of deformation [27]. Although CNNs are a part of deep learning, they have the advantage of being trained using standard backpropagation techniques. In order to improve performance, CNNs are frequently utilized in pattern recognition systems [28]. The subject of deep learning, in general, became well-liked due to the employment of graphics processing units for calculation and the capacity to initialize the networks via a layer-by-layer pre-training [29]. Max-pooling, batch normalization, and rectified linear unit (ReLU) are additional layers present in CNNs. Convolution and fully linked layers are referred to as weighted layers in CNNs (ReLU). Convolution layers need a lot of computations because of the sliding window and a lot of multiplications, whereas fully linked layers need a lot of memory [30].

The visual system of the brain serves as the driving force behind this type of network, which is a branch of deep learning research [31]. For a long time, CNNs were the only DNN type that could be trained efficiently using the weight-sharing approach. The term "receptive field" (the filter defining field) first appears in the works of Hubel and Wiesel [32]. The max-pooling layer was first introduced by Fukushima [33]. In addition to introducing the standard structure of CNN, LeCun et al. [7] also coined the phrase "convolutional layer" for the first time. Ciresan et al. [34] recently showed the structure of DNNs and achieved the best performance on six benchmark image classification databases, including the MNIST (handwritten digits), NIST SD-19, handwritten Chinese characters, traffic signs, CIFAR10, and NORB. Traffic signs and MNIST findings outperform individuals in every way.

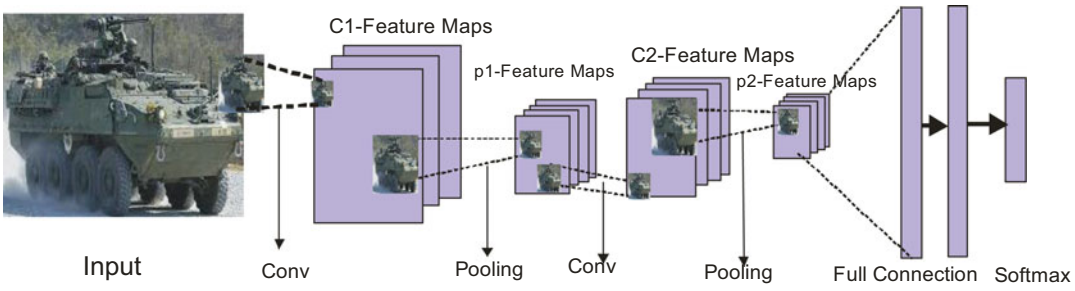
Additionally, CNNs can be thought of as supervised feed-forward networks that performed remarkably well in large-scale object categorization tasks. The primary visual cortex of the human brain, which controls how visual information is processed, stimulates the fundamental structure of CNNs [35]. When doing image classification, CNNs can automatically extract the learnable visual features from the large-scale dataset input photos from the classes, as opposed to the old handmade features extraction approaches. The fact that CNNs employ representations of the features and the classifier in the same network to break down their dependencies is one of its key advantages over traditional classification techniques. Convolution layers, pooling layers, and linking layers make up the majority of the CNNs' usual architecture, which is briefly explored below and seen in Figure 1.

2.2. CNN in Fourier domain

The Fourier convolutional neural network (FCNN) is a technique where just the Fourier domain is used for training. The effect is a significant reduction in training time while maintaining efficacy. The Fourier domain is used to analyze and represent images, and a convolution mechanism is utilized in a manner similar to that used in more traditional CNN algorithms [36].

The most fundamental neural network for computer vision issues, such as classification, segmentation, and denoising, is

Figure 1
Typical convolutional neural network architecture



called a CNN. Additionally, it has been used to extract and learn visual properties like classification, detection, etc. from DNNs [6]. Additionally, DNNs have been widely created and successfully used to a variety of tasks employing big datasets, such as ImageNet, including AlexNet, VGG, DeseNet, and ResNet [20, 37]. However, one of the issues with CNN training is that it is quite expensive to operate all of the convolutional layers. In particular, an increase in the size of the image or kernel unavoidably results in an increase in computation, which causes a learning lag [8]. Two solutions have been offered to address this issue: shifting the domain through Fourier transform and creating a CNN in the frequency domain. This is because the convolution operation in the spatial domain is the same as the pointwise multiplication in the Fourier domain. Convolution is more computationally expensive to compute than point-by-point multiplication in general. Prior strategies for handling the time cost problem concentrated on increasing computing performance [8, 38].

2.3. Fourier transform

A potent mathematical tool that dates to the middle of the 1800s is the Fourier transform. Its premise states that, regardless of the complexity of the periodic function, any periodic function can be written as the sum of sines and cosines of various frequencies, each one weighted by a different coefficient. If such a function is not periodic, the decomposition into sines and cosines can still be used, but from the integral beneath the curves instead. The fast Fourier transform (FFT) was created with the advent of computing and technical advancements [39]. There is a Fourier transform for discrete domains and a Fourier transform for continuous domains; the latter allows for quick computations and efficient approximations. The FFT enables a number of manipulations of digital signals, primarily in pictures, including the application of several frequency-domain filters [39]. In essence, the discrete Fourier transform (DFT) is used to generate a complex array with the same dimension as the original D-dimensional array in addition to the magnitude (or module) and phase angle. The DFT, for instance, is defined mathematically as follows for two dimensions:

$$\begin{aligned} F(u,v) &= R(u,v) + jI(u,v) \\ &= |F(u,v)|e^{j\varnothing(u,v)} \end{aligned} \tag{1}$$

where u and v stand for the two-dimensional matrix coordinates, R and I stand for the real and imaginary components, respectively, and $\varnothing$ is the phase angle. Equations (2) and (3) specify the input signal's amplitude and phase angle, respectively.

$$|F(u,v)| = [R^2(u,v) + I^2(u,v)]^{\frac{1}{2}} \tag{2}$$

$$\varnothing(u,v) = arctan\left[\frac{I(u,v)}{R(u,v)}\right] \tag{3}$$

2.3.1. Utilization of FFCNN

FCNN approach has the advantage of reducing complexity, especially in the context of larger images, and thus increasing network efficiency significantly. The convolution theorem's basic notion claims that for two functions k and u, we have

$$F(k * u) = F(k) \odot F(u) \tag{4}$$

where $F(.)$ stands for Fourier transform, * represents convolution, and the Hadamard pointwise product is shown by $\odot$. Convolution calculations can now be performed more quickly, thanks to FFTs. Given the effectiveness of the Fourier transform and the fact that convolution is equivalent to the Hadamard product in the Fourier domain, this method is substantially quicker and uses a lot fewer computer processes than the sliding kernel spatial method [36].

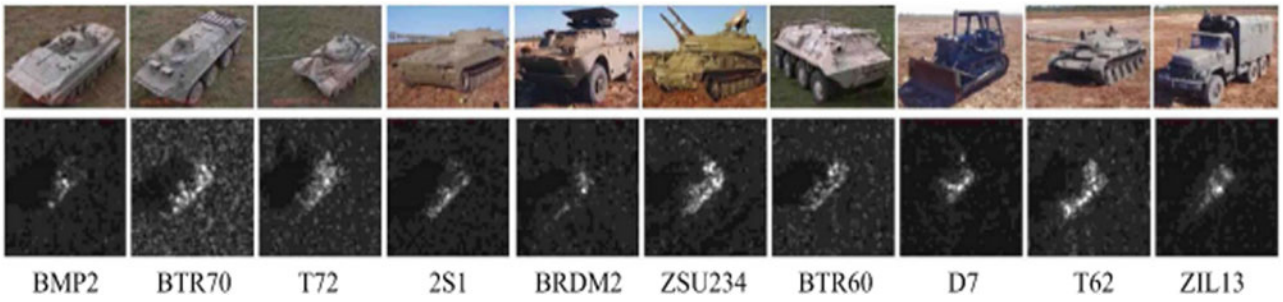
3. Methodology

This section describes the research methodology adopted in this work based on existing scholarly frameworks, provides guidelines for implementing MS-FFCNN, describes some of the principles and structures used in the model, and provides a brief description of the dataset used. The experiment used SAR images to categorize military ground vehicles. However, the MS-FFCNN model was developed and tested using the open-source Python deep learning framework with a tensor flow backend. A typical PC with a 20 GB RAM HP Proliant DL380p Gen8 server served as the platform for all studies. This work therefore built on the work of Zhao et al. [9] and combined various aspects of the methodologies employed in similar works.

3.1. Description of the used dataset

In contrast to the rapid growth of optical image recognition research, it is highly challenging to obtain enough publically accessible datasets in the field of SAR ATR due to the complexity of target detection methods. One of the few datasets that can identify ground vehicle targets among them is the MSTAR, which

Figure 2
Ten kinds of items from the MSTAR dataset are represented by optical images and associated SAR images
Source [17]



is publicly accessible in the United States. The Defense Advanced Research Projects Agency started MSTAR in the middle of the 1990s [40]. In the former Soviet Union, high-resolution focused SAR is utilized to gather SAR images of different military vehicles. To create a reasonably comprehensive and organized field test database, MSTAR intends to undertake SAR field testing on ground targets under various scaling situations, such as target occlusion, camouflage, configuration modifications, and target obscuration. Up until now, this dataset has mostly served as the foundation for international research on SAR ATR. Targets from the following ten categories are included in the MSTAR dataset: ZSU234, ZIL131, T72, T62, D7, BTR70, BTR60, BRDM2, BMP2, and 2S1 as shown in Figure 2.

3.2. Multi-stream fast fourier CNN for SAR ground military vehicle

MS-FFCNN comprises three network streams where each stream composed of input layer block and subnetwork layer block. This work implements the multi-stream CNN as found in Zhao et al [9] and Pei et al. [14] but with FFT-based NN. In comparison to the spatial domain, the distribution of the image data is different in the Fourier domain.

This enables us to retain more information while reducing the data size by the same amount as in the spatial domain. The convolution between the image and the kernel is implemented by the neural network using the FFT. Figure 3 shows the proposed architecture. The input image is first transformed into Fourier spectrum using fast Fourier transformation algorithm.

In this work, a CNN architecture with three streams was developed, where each CNN stream received the Fourier spectrum data image as input. A concatenated fusion operation is further implemented for the three streams. Specifically, fast Fourier convolutional layers learn the three Fourier spectrum input features. Features extracted from each stream were flattened separately before fusion. The choice of concatenation operation is adopted and not multiplication, max, nor sum because it is more flexible and allows the streams structure modification. Therefore, it allows the streams of convolutional layers to have different structures. Then, the output tensors of the three streams concatenated into vector, and this vector will be learned by the FC layer.

4. Result and Discussion

This section deals with the implementation process of MS-FFCNN based on the methodology presented in section 3. The experimentation is based on the development of multiple streams of FFCNN in a frequency domain for classification of fast Fourier transformed SAR image.

4.1. Experimental results analysis

In this work, two experiments were performed in order to comprehensively assess the recognition performance of the

Figure 3
Proposed architecture

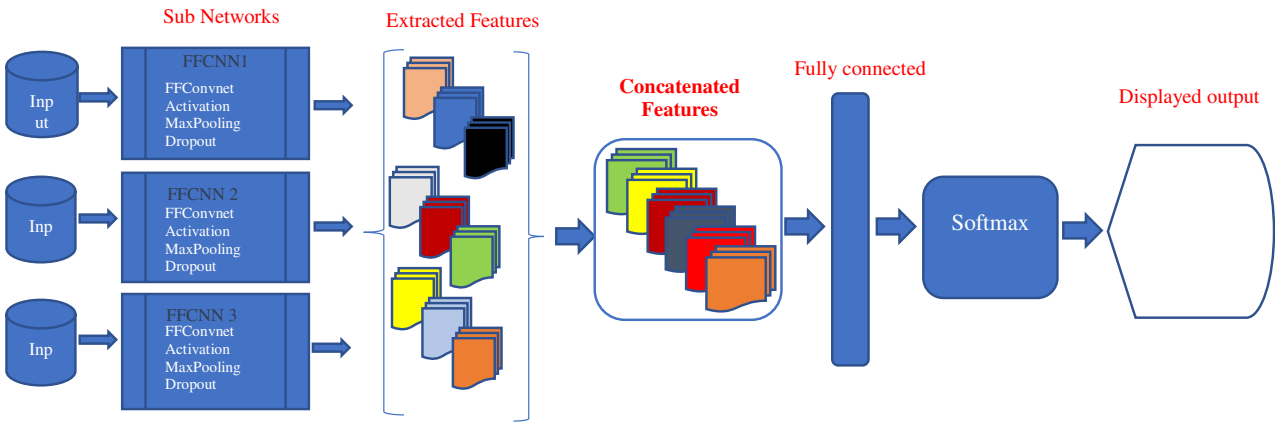


Table 1
Training and test target class number and variation

S/No	Target class	Training set		Test set					
		No. of images	Depression	No. of images	Depression	No. of images	Depression	No. of images	Depression
1	2S1	299	17°	274	15°	288	30°	288	45°
2	BMP2	233	17°	195	15°				
3	BRDM2	298	17°	274	15°	287	30°	287	45°
4	BTR60	256	17°	195	15°				
5	BTR70	233	17°	196	15°				
6	D7	299	17°	274	15°				
7	T62	298	17°	273	15°				
8	T72	232	17°	179	15°				
9	ZIL131	299	17°	274	15°				
10	ZSU234	299	17°	274	15°	288	30°	303	45°
	Total	2746	17°	2425	15°				

proposed model. The experiments were performed on MSTAR dataset that comprises ten different types of military ground vehicles. The first experiment is the traditional multi-stream CNN, where images were convolved in conventional convolutions domain and raw SAR images are used as input data, while in the second experiment, all the images were transformed into a fast Fourier spectrum and fed as input into the FFCs. That is a convolution in the frequency domain.

4.1.1. Experiment protocol

Table 1 shows the total number of images used for training and testing, as well as the number of images for each class of military vehicle.

4.1.2. Experiment with raw SAR image and normal CNN

The result obtained from the first experiment is shown in Table 1 while Figures 4, 5, and 6 show validation versus training error, validation versus training accuracy, and confusion matrix, respectively.

Table 2 shows that the recognition accuracy is 99.38% with time step between 143 and 150 ms/step in 32 min and 45 s training time. The experimental mean square error is 0.205 while the mean square log error is 0.016. The model floating point operation (FLOPs) is 5×10^5 and the model parameter is 2.61×10^5.

In order to make sure the model was not overfitted, Figures 4 and 5 compare the validation error against the training error and the validation accuracy against the training accuracy, respectively. However, the validation accuracy and training accuracy can be reasonably inferred from Figure 5.

The proximity suggests that the model was not overfit. All the correctly and incorrectly classified occurrences are displayed in Figure 6.

4.1.3. Experiment with transformed SAR image and FFCNN

In this experiment, the SAR images were transformed into fast Fourier spectrum before input into the three streams of convolutions in frequency domain. Table 2 shows the result obtained and Table 3 revealed the classification report of the experiment. Figures 7, 8, and 9 show the validation versus training error, validation versus training accuracy, and confusion matrix, respectively.

Table 3 shows that the recognition accuracy of 99.92% obtained in the second experiment is far higher than the accuracy obtained in Table 2. Furthermore, there is slight difference in training time obtained compared to training time in Table 2. Therefore, it is very clear that multi-stream CNN in frequency domain is better than the traditional multi-stream CNN. It improves the recognition accuracy as well as model training time.

Figure 4
Validation versus training error

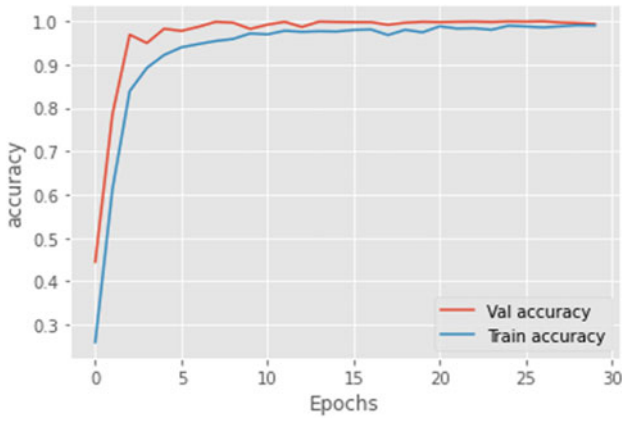


Figure 5
Validation versus training accuracy

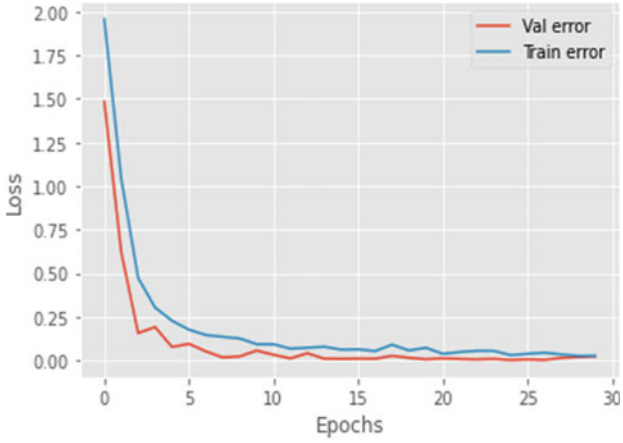


Figure 6
Confusion matrix

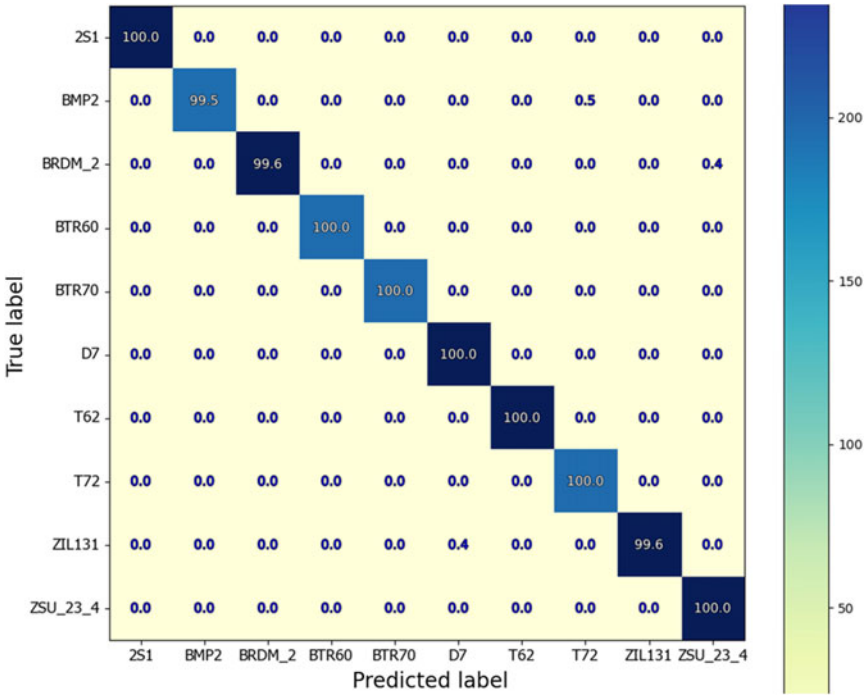


Table 2
MS-CNN experimental results

MCMS-CNN	
Recognition accuracy	99.38%
Training time	00:19:35
Time step	150–143 ms/step
Validation accuracy	99.26%
Mean squared error (MSE)	0.205
Mean squared log error (MSLE)	0.016
Floating point operations (FLOPs)	5×10^{-2} GFLOPS $= 5 \times 10^{7}$
Parameters	$261{,}702 = 2.61 \times 10^{5}$

Table 3
MS-FFCNN experimental results

MCMS-FFCNN	
Recognition accuracy	99.92%
Training time	00:19:21
Time step	220–236 ms/step
Validation accuracy	98.42%
Mean squared error (MSE)	0.00412
Mean squared log error (MSLE)	0.000797
Floating point operations (FLOPs)	0.0449GFLOPS $= 4.49 \times 10^{7}$
Parameters	$261{,}702 = 2.61 \times 10^{5}$

Classification report is a performance evaluation metric in machine learning that shows the precision, recall, F1 score, and support score of the trained classification model. Table 4 shows that the average precision for all the ten vehicle images recognized is 99.93%. Its only 2S1 and ZSU_23_4 have precision accuracy of 99.64%, while all other vehicles have precision accuracy of 100%.

Figure 7
Validation versus training error

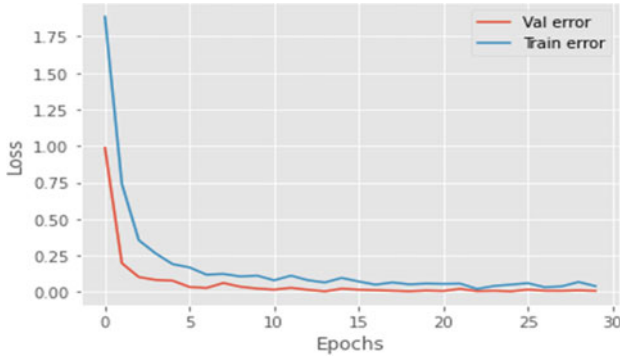


Figure 8
Validation versus training accuracy

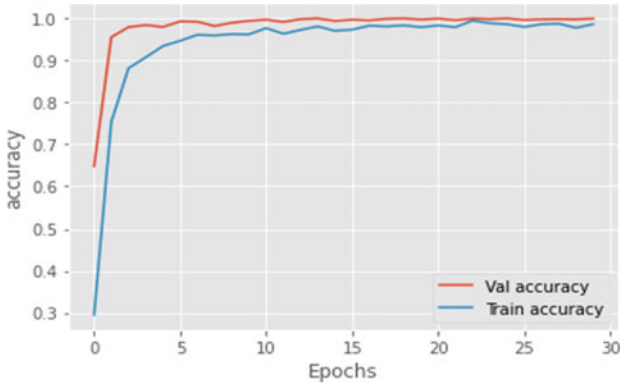


Figure 9
Proposed model confusion matrix

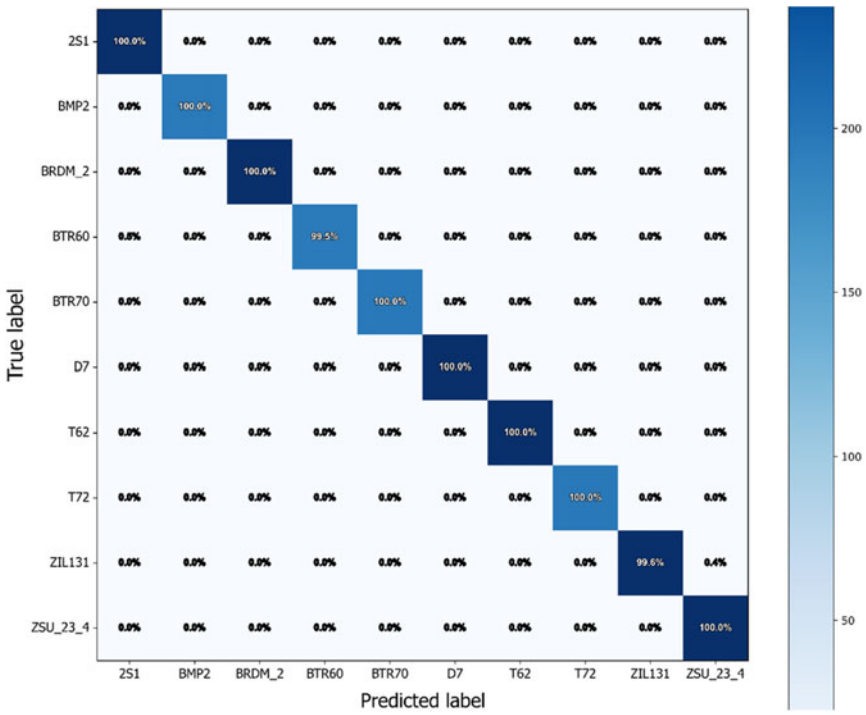
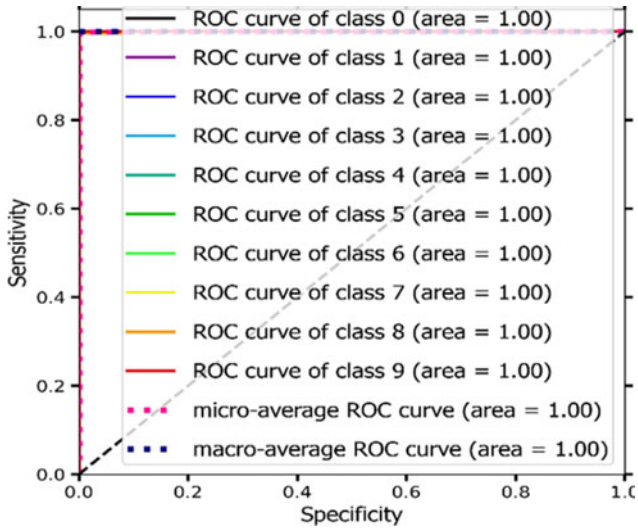


Table 4
Classification report

	Precision	Recall	F1 score	Support
2S1	0.9964	1.0000	0.9982	274
BMP2	1.0000	1.0000	1.0000	195
BRDM_2	1.0000	1.0000	1.0000	274
BTR60	1.0000	0.9949	0.9974	195
BTR70	1.0000	1.0000	1.0000	196
D7	1.0000	1.0000	1.0000	274
T62	1.0000	1.0000	1.0000	273
T72	1.0000	1.0000	1.0000	196
ZIL131	1.0000	0.9964	0.9982	274
ZSU_23_4	0.9964	1.0000	0.9982	274
Accuracy			0.9992	2425
Macro avg.	0.9993	0.9991	0.9992	2425
Weighted avg.	0.9992	0.9992	0.9992	2425

Figure 10
Model ROC curve

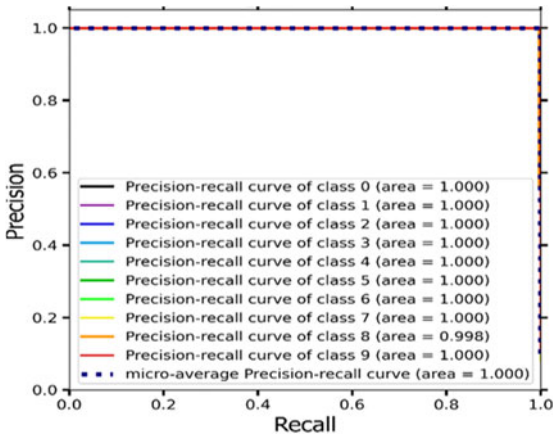


Figures 7 and 8 show that the model did not overfit nor underfit; therefore, the model trained and learned well. Figure 10 represents the validation receiver operating characteristic curve, which is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters, such as true positive rate (sensitivity, recall, or probability of detection) and false positive rate (probability of false alarm and can be calculated as 1 – specificity). Also Figure 11 represents the precision recall curve that shows the tradeoff between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

This MS-FFCNN with three streams produced a 99.92% accuracy rate, which is a significant improvement over the MS-CNN proposed by Zhao et al. [9], which had an overall accuracy of 99.88%. Based on Tables 5 and 6, we can see the differences between our model and state-of-the-art models in terms of model accuracy, number of parameters, and computational complexity.

Based on FLOPs, parameters, and recognition accuracy, Table 6 compares our MS-FFCNN with other methods. The recognition rate is loosely correlated with the number of parameters in the network. A lower recognition rate can be attributed to a lack of effective features from different categories of targets extracted from too few parameters, resulting in insufficient features and lower recognition rates. Furukawa [48] ResNet-18 uses millions of parameters and

Figure 11
Model precision-recall curve



FLOPs that reach ten billion in order to obtain high recognition rates; however, this results in low efficiency because it needs a lot of computational power and more time to train and test the network.

In comparison to the other three multi-view approaches, our MS-FFCNN obtains the highest recognition rates with the least number of parameters and FLOPs, thanks to a multi-stream convolutional layer and sensible parameter setting. Based on the comparison results of recognition rates, FLOPPs, and the quantity of parameters, it is clear that our proposed MS-FFCNN is superior in the SOC scenario.

5. Conclusion and Future Work

The MS-FFCNN is a supervised learning-based image features fusion method that uses the MSTAR dataset and achieves a high recognition rate at a reduced training time with minimal network parameters. It can be concluded that the proposed MS-FFCNN is able to identify and extract more features from multiple streams of

Table 5
Result comparison with other state-of-the-art algorithms

Author	Title	Dataset	Methodology	Results
Pei et al. [14]	Multiview deep feature learning network for SAR automatic target recognition	MSTAR	Multi-view CNN	99.30% – 3 views 99.62% – 4 views
Guo [41]	SAR image classification based on multi-feature fusion decision convolutional neural network	MSTAR	FCNN	99.30%
Ma [42]	Improving SAR target recognition performance using multiple preprocessing technique	MSTAR	CNN	99.02%
Yu et al. [43]	Combination of joint representation and adaptive weighting for multiple features with application to SAR target recognition	MSTAR	Joint sparse representation and adaptive weighting	99.38%
Zhao et al. [44]	Multi-aspect SAR target recognition based on prototypical network with a small number of training samples	MSTAR	CNN	ResNet18 – 99.84% VggNet11 – 99.13% AlexNet – 96.17%
Zhang et al, [45]	Image target recognition model of multi-channel structure convolutional neural network training automatic encoder	MSTAR	DCNN	98.5%
Pei et al. [46]	FEF-Net: A deep learning approach to multiview SAR image target recognition	MSTAR	CNN	2 – views – 98.42% 3 – views – 99.31% 4 – views – 99.34%
Zhao et al. [9]	Multi-stream convolutional neural network for SAR automatic target recognition	MSTAR	MS-CNN	3-views – 99.88% 4- views – 99.92
Proposed	Multi-stream fast Fourier convolutional neural network for automatic target recognition of ground military vehicle	MSTAR	MS-FFCNN	3-streams – 99.92%

Table 6
Comparison of the floating-point operations and number of parameters

	Author	Method	Number of parameters	Floating point operations (FLOPs)	Accuracy
1	Zhao et al. [9]	MS-CNN (2-view)	2.59×10^5	5.044×10^7	99.84%
2		MS-CNN (3-view)	2.60×10^5	7.566×10^7	99.88%
3		MS-CNN (4-view)	2.61×10^5	1.008×10^8	99.92%
4	Dong et al. [47]	VDCNN (2-view)	2.22×10^6	1.667×10^8	97.81%
5		VDCNN (3-view)	2.38×10^6	2.235×10^8	98.17%
		VDCNN (4-view)	2.87×10^6	2.506×10^8	98.52%
6	Furukawa [48]		2.75×10^6	1.244×10^{10}	99.56%
	Proposed	MS-FFCNN–3-streams	2.60×10^5	4.49×10^7	99.92%

the networks and able to achieve 99.93% recognition accuracy. The results of this study show that operating convolutions in frequency domain reduces training time and improves recognition accuracy. Moreover, multi-stream techniques significantly reduced overfitting, which occurs when training data are limited.

For further research on this domain, it may be beneficial to integrate LSTM into multi-stream to further reduce training times.

Conflict of Interest

The authors declare that they have no conflicts of interest to this work.

References

[1] Jia, Z., Guangchang, D., Feng, C., Xiaodan, X., Chengming, Q., & Lin, L. (2019). A deep learning fusion recognition method based on SAR image data. *Procedia Computer Science*, *147*, 533–541. https://doi.org/10.1016/j.procs.2019.01.229

[2] Xinyan, F., & Weigang, Z. (2019). Research on SAR image target recognition based on Convolutional neural network. *Journal of Physics: Conference Series*, *1213*, 042019. https://doi.org/10.1088/1742-6596/1213/4/042019

[3] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

[4] Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, *7*, 87–93. https://doi.org/10.1007/s13735-017-0141-z

[5] Zhao Z.-Q., Zheng P., Xu S.-T., & Wu, X. (2019). Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 30, 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865

[6] Han, Y., & Hong, B. (2021). Deep learning based on Fourier convolutional neural network incorporating random Kernels. *Electronics*, *10*, 2004. https://doi.org/10.3390/electronics10162004

[7] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 2278–2324. https://doi.org/10.1109/5.726791

[8] Mathieu, M., Henaff, M., & Lecun, Y. (2013). Fast Training of Convolutional Networks through FFTs. arXiv 2013, arXiv: 1312.585.

[9] Zhao, P., Liu, K., Zou, H., & Zhen, X. (2018). Multi-stream convolutional neural network for SAR automatic target recognition. *Remote Sensing*, *10*, 1473. https://doi.org/10.3390/rs10091473

[10] Pei, J., Huang, W., Huo, Y., Zhang, Y., Yang, J., & Yeo, T.-S. (2018). SAR automatic target recognition based on multiview deep learning framework. IEEE Transactions on Geoscience and Remote Sensing, 56, 2196–2210. https://doi.org/10.1109/TGRS.2017.2776357

[11] Tian, Z., Wang, L., Zhan, R., Hu, J., & Zhang, J. (2018). Classification via weighted kernel CNN: Application to SAR target recognition. *International Journal of Remote Sensing*, *39*, 9249–9268. https://doi.org/10.1080/01431161.2018.1531317

[12] Zhang, Z. (2018). Joint classification of multiresolution representations with discrimination analysis for SAR ATR. *Journal of Electronic Imaging*, *27*, 1. https://doi.org/10.1117/1.jei.27.4.043030.

[13] Zhang, M., An, J., Yu, D., Yang, L., & Lv, X. (2020). Convolutional neural network with attention mechanism for SAR automatic target recognition. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/LGRS.2020.3031593

[14] Pei, J., Huo, W., Wang, C., Huang, Y., Zhang, Y., Wu, J., & Yang, J. (2021). Multiview deep feature learning network for SAR automatic target recognition. *Remote Sensing*, *13*, 1455. https://doi.org/10.3390/rs13081455

[15] Furukawa, H. (2018). Deep learning for end-to-end automatic target recognition from synthetic aperture radar imagery. *IEICE Technical Report*, *117*, 35–40. https://doi.org/10.48550/arXiv.1801.08558

[16] Zou, H., Lin, Y., & Hong, W. (2018). Research on multi-aspect SAR images target recognition using deep learning. *Journal of Signal Processing*, *34*, 513–522.

[17] Wang, C., Pei, J., Wang, Z., Huang, Y., & Yang, J. (2020). Multi-view CNN-LSTM neural network for SAR automatic target recognition. In Proceedings of the IEEE Geoscience and Remote Sensing Society, Waikoloa Village, HI, USA (pp. 1755–1758).

[18] Pei, J., Huang, Y., Huo, W., Wu, J., Yang, J., & Yang, H. (2016). SAR imagery feature extraction using 2DPCA-based two-dimensional neighborhood virtual points discriminant embedding. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*, 2206–2214. https://doi.org/10.1109/JSTARS.2016.2555938

[19] Dang, S., Cui, Z., Cao, Z., & Liu, N. (2018). SAR target recognition via incremental nonnegative matrix factorization. *Remote Sensing*, *10*, 374. https://doi.org/10.3390/rs10030374

[20] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition, advances in neural information processing systems*, Barcelona, Spain, 5–10, (CVPR) (pp. 770–778). https://doi.org/10.1109/cvpr.2016.90

[21] Chen S., Wang H., Xu F., & Jin Y.-Q. (2016). Target classification using the deep convolutional networks for SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(8), 4806–4817. https://doi.org/10.1109/TGRS.2016.2551720

[22] Bentes, C., Velotto, D., & Lehner S. (2015). Target classification in oceanographic SAR images with deep neural networks: Architecture and initial results. In *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)* (pp. 3703–3706).

[23] Wang, H., Chen, S., Xu, F., & Jin, Q. (2015). Application of deep-learning algorithms to MSTAR data. In *2015 IEEE International geoscience and remote sensing symposium* (IGARSS) (pp. 3743–3745).

[24] Shchekotov, I., Andreev, P. K., Ivanov, O., Alanov, A., & Vetrov, D. (2022). FFC-SE: Fast Fourier convolution for speech enhancement. In *Interspeech 2022.* https://doi.org/10.21437/interspeech.2022-603

[25] Sinha, A. K., Manthira Moorthi, S., & Dhar, D. (2022). NL-FFC: Non-local fast Fourier convolution for image super resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* https://doi.org/10.1109/cvprw56347.2022.00062

[26] Lu C, Borui J, & Yadong M (2020). Fast Fourier convolution. *Advances in Neural Information Processing Systems*, *33*, 4479–4488. https://papers.nips.cc/paper/2020/hash/2fd5d41ec6cfab47e32164d5624269b1-Abstract.html

[27] 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*, 2861–2872. https://doi.org/10.1109/taes.2016.160061

[28] Chen, J., Du, L., He, H., & Guo, Y. (2019). Convolutional factor analysis model with application to radar automatic target recognition. *Pattern Recognition*, *87*, 140–156. https://doi.org/10.1016/j.patcog.2018.10.014

[29] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35,* 1798–1828. https://doi.org/10.1109/tpami.2013.50

[30] Abtahi, T., Shea, C., Kulkarni, A., & Mohsenin, T. (2018). Accelerating convolutional neural network with FFT on embedded hardware. *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems*, *26*, 1737–1749. https://doi.org/10.1109/tvlsi.2018.2825145

[31] Xueyun, C., Shiming, X., Cheng-Lin, L., & Chun-Hong, P. (2014). Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters, 11,* 1797–1801. https://doi.org/10.1109/lgrs.2014.2309695

[32] Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology 148*, 574–591. http://jp.physoc.org/content/148/3/574.full.pdf+html

[33] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–202. https://doi.org/10.1007/BF00344251

[34] Ciresan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3642–3649). Red Hook, NY: Curran.

[35] Bautista, M. A., Sanakoyeu, A., Tikhoncheva, E., & Ommer, B. (2016). CliqueCNN: Deep unsupervised exemplar learning. *Advances in Neural Information Processing Systems*, *29*. https://arxiv.org/abs/1608.08792

[36] Lendave, V. (2021, December 28). *How fast Fourier convolution can replace the Convolutional layer of CNN?* Analytics India Magazine. https://analyticsindiamag.com/how-fast-fourier-convolution-can-replace-the-convolutional-layer-of-cnn/

[37] Liu, Y., & Lu, F. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

[38] Heckbert, P. (1995). Fourier transforms and the Fast Fourier Transform (FFT) algorithm. *Computers & Graphics*, *2*, 15–463. https://www.cs.cmu.edu/afs/andrew/scs/cs/15-463/2001/pub/www/notes/fourier/fourier.pdf

[39] Gonzalez, R. C., & Woods, R. E. (2000). *Processamento de imagens digitais*. Editora Blucher.

[40] Karine, A., Toumi, A., Khenchaf, A., & El Hassouni, M. (2018). Radar target recognition using salient keypoint descriptors and multitask sparse representation. *Remote Sensing, 10,* 843.

[41] Guo, L. (2022). SAR image classification based on multi-feature fusion decision convolutional neural network. *IET Image Processing*, *16*, 1–10. https://doi.org/10.1049/ipr2.12323

[42] Ma, Q. (2021). Improving SAR target recognition performance using multiple preprocessing techniques. Computational Intelligence and Neuroscience, 2021.

[43] Yu, L., Wang, L., & Xu, Y. (2021). Combination of joint representation and adaptive weighting for multiple features with application to SAR target recognition. Scientific Programming, 2021.

[44] Zhao, P., Huang, L., Xin, Y., Guo, J., & Pan, Z. (2021). Multi-aspect SAR target recognition based on prototypical network with a small number of training samples. *Sensors*, *21*, 4333. https://doi.org/10.3390/s21134333

[45] Zhang, S., Cheng, Q., Chen, D., & Zhang, H. (2020b). Image target recognition model of multi-channel structure convolutional neural network training automatic encoder. *IEEE Access*, *8*, 113090–113103. https://doi.org/10.1109/ACCESS.2020.3003059

[46] Pei, J., Wang, Z., Sun, X., Huo, W., Zhang, Y., Huang, Y., . . . & Yang, J. (2021b). FEF-Net: A deep learning approach to multiview SAR image target recognition. *Remote Sensing*, *13*, 3493.

[47] Dong, G., Gangyao, K., Linjun, Z., Jun, L., & Min, L. (2014). Joint sparse representation of monogenic components with application to automatic target recognition in SAR imagery. 2014 *IEEE Geoscience and Remote Sensing Symposium*. https://doi.org/10.1109/igarss.2014.6946481

[48] Furukawa, H. (2017). Deep learning for target classification from SAR imagery: Data augmentation and translation invariance. *IEICE Technical Report*, *117*, 11–17. https://doi.org/10.48550/arXiv.1708.07920

How to Cite: Awujoola, O. J., Odion, P. O., Evwiekpaefe, A. E., & Obunadike, G. N. (2022). Multi-Stream Fast Fourier Convolutional Neural Network for Automatic Target Recognition of Ground Military Vehicle. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA2202412