National College of Ireland

Military and Non-Military Vehicle Detection by Faster R-CNN and SSD300 Models using Transfer Leaning

MSc Research Project

MSc in Data Analytics

Venkata Devaraju Nandimandalam

Student ID: x18181422

School of Computing

National College of Ireland

Supervisor: Hicham Rifai

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Venkata Devaraju Nandimadalam
Student ID:	x18181422
Programme:	MSc Data Analytics
Year:	2019-2020
Module:	Research Project
Supervisor:	Hicham Rifai
Submission Due Date:	28/09/2020
Project Title:	Military and Non-Military Vehicle Detection by Faster R-CNN and SSD300 Models using Transfer Leaning
Word Count:	7510
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	N.V. Devaraju
Date:	24th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	Q
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	Q
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	Q

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Military and Non-Military Vehicle Detection by Faster R-CNN and SSD300 Models using Transfer Leaning

Venkata Devaraju Nandimandalam

X18181422

Abstract

The need for using deep learning in military applications is increasing every day and this research is helpful for surveillance, target detections, and taking precautionary measures using Unmanned Aerial Vehicles (UAV) and strategic warning systems by detecting military vehicles. Detecting small military vehicles from Aerial images and segregating military and non-military vehicles is a challenging task. Solving image detection problems involve deep learning algorithms like Faster R-CNN and SSD300 that are proposed and implemented with the help of transfer learning. For evaluating the model's performance twelve different metrics are calculated. As opposed to the performance of SSD300 model Faster R-CNN with FPN detected objects with high mAP value around 82 percent. The customized dataset is used for training the models and it contains five military and two non-military vehicle classes.

Keywords: Faster R-CNN, SSD300, transfer learning, military and non-military vehicle detection, aerial images.

1. Introduction

Vehicle perception plays a crucial role in administering the law and has found its application in defense areas. Distinguishing between military and non-military vehicles is vital for image detection applications. The most addressed problem in vehicle detection is filtering the objects from the framework. The outcome of vehicle detection has a significant impact on the further processing of the entire system. The object detection domain has seen vast growth in the development of techniques that aids in meeting the demands of robotic technology and associated counterparts. Automatic target detection (ATD) is primarily essential for monitoring military operations. The military operations involve sensors being placed on the grounds. These sensors are mounted on unmanned aerial vehicles and unmanned ground vehicles to extract sensory data. The collected data will be utilized by the ATD algorithms to locate the targets through bounding boxes (Liu and Liu, 2017). In military research conducted battlefields, the primary focus has always been on detecting the armored vehicles to become proficient in military techniques such as recognition, precision striking, etc., Detecting armored vehicles in the battlefield environment comes with various challenges with regards to the complex background such as changing background, shelter and few other interferences optic illusions, changing weather conditions. Varied models of armored vehicles and the complexity in their shapes have also impacted the detection process due to factors like different shooting angles, scale changes, etc.,

Recognizing vehicles in aerial images has been given greater stress in recent times. Although it comes with certain drawbacks like vehicles appearing smaller in large aerial images making it hard to locate the regions with vehicles. Also, classification of appropriate regions has become a

challenging task due to the varied directions of vehicles and the presence of more number structures that resemble vehicles (Tianyu *et al.*, 2017).

This research is contributing to military defense by identifying the objects precisely during combat operations. Many countries conducting researches by using machine learning and artificial intelligence in military operations in project Maven pentagon using deep learning to detect targets quickly from the information collected by the drones. Defense Advanced Research Projects Agency (DARPA) is a military research organization working on autonomous warplanes is controlled from another aircraft and warships to detect submarines and it operates using machine learning and artificial intelligence without requiring any human intervention. This research uses Faster R-CNN using FPN with Resnet 101 backbone and SSD300 with VGG16 backbone to detect the vehicles from images and the transfer learning approach is performed to improve the model accuracy.

1.1. Research question

To what extent deep learning algorithms like Faster R-CNN and SSD models can improve the efficiency in detecting military and non-military vehicles from aerial and normal view images using transfer learning.

1.2. Research Objectives

The objectives of this research contain

Detecting the military and non-military vehicles from aerial and normal images using deep learning techniques.

Performing Faster R-CNN model with FPN feature extractor and SSD300 model with VGG16 feature extractor based on transfer learning approach.

Tuning both the models during training phase to accomplish best results and comparing the results of both the developed models.

This document is arranged as follows: Section 1 contains Introduction, section 2 displays the related works in detecting the military vehicles and objects in aerial and satellite images, section 3 displays the methodology, section 4 illustrates the design specification of the models, section 5 demonstrates the model implementation, section 6 presents the evaluation section 7 contains conclusion and future work.

2. Related Work

Previous studies conducted in detecting and classifying the military vehicles and objects from aerial and satellite images using different deep learning models and transfer learning approaches are described below. These studies are supportive and helpful to conduct the research in the right direction.

2.1. Detecting Vehicles using Faster R-CNN and SSD Techniques

Recognition of military vehicles was performed in this research using deep learning algorithms like SSD, faster R-CNN, and R-FCN by using pre-trained networks as feature extractors. The Inception v2, Inception Resent v2, and Resnet 101 are used for producing the feature maps. Data was prepared using aerial images, real and toy images gathered from IMAGENET, VEDAI, and segregated into three types of categories like military, Non-military, and Non-vehicle by using VOTT tool they annotated the data. For all the three models 300 proposals are generated and sent to the classifier. For R-FCN and SSD the batch size used was 4 and for Faster R-CNN it was 2 all the three models are estimated by using the Average Precision and IOU. For Hemet and tank classes output Average precision score was high compared to other classes due to more training data. The SSD architecture model trained with 800k iterations provided a better precision score compared to R-FCN and Faster R-CNN and when state of art architecture is used Faster R-CNN performed better (Kamran *et al.*, 2019).

In this research, the optical cable and vehicles are detected using Faster R-CNN model VEDAI dataset aerial images used for model training. For extracting the features in the proposed model ZFNET and VGG16 both are conducted and using the Stochastic Gradient Descent method in both VGG16 extracted best features. The RPN generates the proposals, other different models, HOG+LBP+SVM, DPM performed on the same dataset. The precision Recall curves, and loss graphs of the proposed model are represented these are evaluated by utilizing Mean Average Precision. The Faster R-CNN model detected cables and vehicles with high mAP value than other models (Zhang *et al.*, 2018).

Identifying tanks and artillery in battlefield environments Faster R-CNN model was implemented using ZF Net for extracting the features from images the trained data has two categories with the help of RPN they produced proposals. Only tanks and wheel chariot are two classes used in this research The ZF Net designed using Alex Net as base model for categorizing the SVM classifier was used at the end the Caffe framework was used for model implementation and IOU threshold value for positive samples set to 0.7 only the Faster R-CNN with ZF Net used and recognized the vehicles with good efficiency (Xie and He, 2017).

SSD Multi-Scale detection model was proposed to solve the issue in detecting the small objects the proposed system consist of Area proposal Network and MSN. From the raw images, areas are generated using dividing image strategy the area proposals are generated using APN network and VGG16 used for extracting these in this model they adding the clipped areas to the classifier using RPN the proposals are generated Based on IOU value the areas are labeled as positive it contains objects and negative if not. Backpropagation was performed to reduce the loss during training for selecting boxes Non-Max Suppressor was used the SSD MSN model compared with other methods evaluated using a map it overcomes all other models (Chen *et al.*, 2019).

Aircrafts are detected using the proposed model Aircraft Targets Region Proposal Network R-CNN which was based on faster R-CNN model and used Resnet 50 pre-trained model for feature map generation these maps fed as input to ATRPN box generation network and these are given to

Pooling layer and at last, calculating the position and confidence scores the ATRPN consist of proposal layer and RPN and boxes are selected with a high confidence score. For clustering aircraft sizes K means method was used the designed model was evaluated using precision and recall rate. With this model target aircrafts in remote sensing images are detected accurately (Wang *et al.*, 2019).

DF-SSD network proposed to enhance the detection of small objects it is SSD algorithm based on feature fusion and dense network they VGG16 network changed with Dense Net S-32-I. For combining low level and high-level semantic features fusion mechanism was introduced. The network structure built on residual prediction and feature fusion the own feature network was developed using Dense Net stem structure was designed by getting motivation from inception v4 the model implemented from scratch gave the best output when conducted experiment using pascal and MS COCO dataset changing the feature extraction increased the accuracy (Zhai *et al.*, 2020).

Geospatial detection was carried out using remote sensing data to detect the objects multi-scale feature fusion was implemented in the SSD framework the images are labeled by the HBB based labeling. The developed framework for this study was based on YOLO and SSD and extracting futures using the Darknet-53 for predicting the bounding boxes the anchor's design are taken using Faster R-CNN. Concatenating the predicting outputs before that the feature maps are fused by the up sampling. Using the same data proposed method and other models are trained and validated the method which was proposed gave 4.8 percent more AP than other models and using Soft NMS technique the performance of detecting objects was improved (Zhuang *et al.*, 2019).

From satellite images, the airplanes are identified using deep learning techniques YOLO, SSD, and Faster R-CNN the data augmentation was conducted high-resolution images were used for training the model. For SSD model Inception v2 feature extractor was used and at the end, the NMS suppressor was used to reduce the occurrences on object regions at the output the YOLO model was developed based on Dark Net layers and Faster R-CNN models were implemented and evaluated by the COCO metrics out of three models Faster R-CNN predicted well SSD gave good result in localization but not in object detection, tuning parameters and transfer learning gave best results in detection (Alganci *et al.*, 2020).

2.2. Detecting Vehicles and Objects using Convolutional Neural Networks by Hybrid, Ensemble and Cascade Approaches

From the satellite images detecting vehicles is a hard task to resolve this problem the authors proposed Deep neural network model because it achieved best results on image processing in previous studies it extracts the same scale features only and it can’t allow objects with large scale variance to solve this they used Hybrid Deep Neural Network and maps are divided in pooling and convolution layer of DNN and multi-scale features can be extracted with help of HDNN and Back propagated the trained HDNN model. They used a technique called the sliding window to locate the objects different methods are compared in this article DNN, LBP+SVM, Adaboost, and HOG+SVM and evaluated by recall rate for training they used San Francisco city vehicle data gathered from google earth. HDNN effective in extracting multiscale characteristics in which DNN can’t support HDNN accomplished superior results compared to DNN (Chen *et al.*, 2014).

Convolutional Neural network with state of art showing tremendous performance in detection of objects over a decade to overcome the detection problems in spectrum images they implemented unsupervised fused multichannel CNN. The proposed methodology uses the backbone from Fast R-CNN and contains an image fusion approach. The framework contains image fusion, feature extraction, proposal generation, and classification different fusion architectures are present they used pixel-level fusion. To find the effectiveness of the developed fusion model they took six different types of images like Motion, MWIR, Visible, three channels, Visible-MWIR the Accuracy, Precision, and IOU values are evaluated the implemented unsupervised image fusion method three channels image accomplished high precision and accuracy values (Liu and Liu, 2017).

Cascade Convolutional neural networks have proposed for target detection in the military sector the proposed framework consists of three deep CNN levels for detecting objects from aerial images input images are scaled and for CNN input passed as image pyramid. Generating proposals with help Fully Convolution Network in the L1 net and predicting class confidence and the bounding box and in L2 net rejecting the false cases and is a deep network contain Fully connected layers and depth of network was increased in L3 net and provide the results used Caffe framework for developing the model and used Sophisticated Gradient Decent technique for setting the parameters and model was trained on VIVID datasets the Cascade framework model predicted objects from aerial images accurately (Zhang *et al.*, 2019).

Detecting small objects from aerial images is a difficult task to deal with this they proposed a Region-based CNN and framework divided into Feature generation network, Vehicle proposal Network and the Vehicle Classification Network the collected data was augmented and proposal network based on ZF model and generated 300 proposals using pre-trained ZF net model in classification network SoftMax was used for classification. The implemented model was compared with Faster R-CNN and ACF detector these are evaluated using F1 score, precision, and recall the Region-based CNN gave the best results and as future work tried to implement the model on satellite images (Tianyu *et al.*, 2017).

For identifying objects Ensemble approach was followed in this methodology combines and classifies the proposal generated from different Convolution neural network models and these are used for extracting features after combining the region proposals from different models these are classified using box classifier the main objective in performing ensemble method is to achieve the best accuracy. Resnet 50, inception v2, Resent 101 and Resent 152 are used and using the box voting and model selecting based on the mAP values by considering the class and size of the object the best models are combined the Pascal VOC data was used by approaching this technique assisted in boosting the accuracy and made the model robust (Lee *et al.*, 2018).

Locating vehicles and non-vehicles from satellite images Deep Convolutional Neural Network was proposed in this research and showed best results in earlier studies on recognition and classification of images for extracting the patches from images the super-pixel segmentation was introduced and fed input to the DNN network and for dealing with backpropagation SGD was implemented and for detecting vehicles in satellite images the approached segmentation method increased the model effectiveness by decreasing the sliding windows the DNN with classifier outputs the best

performance as a future work the developed model to train across distinct resolutions (Wang *et al.*, 2016).

For identifying objects using CNN from aerial images is inefficient because of this model struggle in precise localization and in detecting a small object to solve the problem the researchers used Cascade Convolutional neural network. The two CNN networks are trained separately the VPN used for generating the regions and these are given as input to the VDN. In the VPN network the VGG16 was used and the network last layers are modified and giving output to connected layers for predicting score and bounding box. The VPN network also used VGG16 and used the SGD technique the model was trained on Munich and VEDAI data. The implemented framework predicted with good accuracy in less time (Zhong *et al.*, 2017).

2.3. Detecting with help of Transfer Learning

Detecting military vehicles two deep learning models are implemented the author developed a Convolutional neural network model from scratch and a pre-trained Resent50 model. The used data was gathered using social media and for training, it's very less, so they used Augmentation techniques for the CNN model and they changed the dense layers in the Resnet model and SoftMax classifier is used for identifying best parameters random search was carried and both models used 10 fold cv. The experiments conducted on Keras and cross-entropy performed to reduce loss both the models evaluated using accuracy. The transfer learning Resnet50 model gave the best performance and overcome the Convolutional neural network model (Hiippala, 2017).

When having small training data to train a model the transfer learning plays a helpful role to overcome the issue for identifying military objects they used transfer learning approach conducted experiments till will layer should be transferred and retrained the mixed layer scheme was used and the proposed convolutional model with transfer learning is differentiated with the models like SVM, RESNET and inception the proposed CNN with transfer learning achieved the best result in recognizing military objects (Yang *et al.*, 2019).

Detecting targets in synthetic aperture radar images is hard due to lack of insufficient labeled data in this research they approached two methods for solving the lack of data they are data augmentation and transfer learning and used the single-shot detector model using VGGNet as feature extractor and applied NMS algorithm at the end of the network they added the auxiliary structure models are evaluated the proposed method detected targets in SAR images with high precision and F1 score and less recall score compared to other approaches using SSD (Wang *et al.*, 2019).

2.4. Object Detection using FPN and PCA

The framework was created by using deep FPN and Gabor filtering was introduced to effectively detecting the small objects these Gabor filters in the FRPN helps in generating quality proposals and reducing the proposals by boosting the output and reducing the processing time and developed highly utilized FPN and processed both the networks parallel to lower the time. Compared the time

and average recall between the RPN, Selective search, Edge box, and created FRPN using the same MOD VOC MSTAR dataset. The implemented model with HU-FPN and deep FPN improved the results in locating objects when compared with YOLOv2, Faster R-CNN, DSOD300, DSSD513 models (Hu *et al.*, 2019).

For identifying the objects in the military, the researcher in this paper used Hyperspectral Imagery. Multiple details were drawn out for detecting objects from unknown images using HIS. With the help of a constrained energy minimization method, it only generates one spectrum from images and can't able to detect the big objects accurately it can identify only small objects. The strategy proposed creates super pixel principal component analysis and K-means. For unknown pixel average energy and max value calculated using CME. The framework contains three main steps generating super pixel, similarity detection using CME, and shape extraction. Image features are pulled out using the Histogram of Oriented Gradients. For validating the model atmospheric and geometric corrections made on collected San Diego hyperspectral imagery to produced super pixels PCA and k-means with different values are tested out of them at k=3 generated smaller pixels and saves time. The CME and proposed results are compared CME detected background pixels as objects, the proposed method with shape matching was efficient and gave accurate results in detecting airplanes (Ke, 2017).

3. Methodology

To detect the military objects, present in Aerial images and Normal images the KDD approach was followed during the implementation it was represented in Figure 1. The data was gathered and preprocessed using the tools then applied the deep learning techniques the Faster R-CNN and SSD300 are performed using Transfer Learning to detect the military and non-military vehicles and evaluated both the models using precision, Recall, Accuracy, mAP to find out the best model between them.

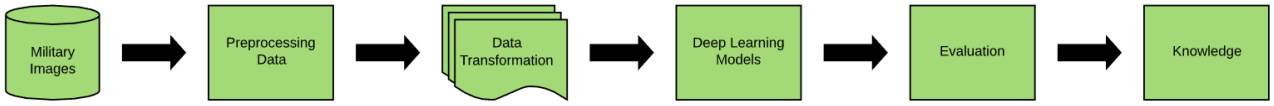

Figure 1: KDD Methodology

3.1. Data Collection

Images are gathered from two different sources using ImageNet and VIVID tracking Evaluation Website. There are no proper military data sets available due to confidentiality and security reasons. From ImageNet the military images and non-military vehicles are downloaded using the URLs ImageNet contains 14 million images belong to different classes and its publicly available for researches to use the data. In the VIVID tracking EgTest03 data was downloaded it contains Aerial view military images.

3.2. Data Preparation

The total 1400 images are collected, and the images are divided for training, test, and validation. By using a tool called Labellmg image annotation tool the images are annotated and saved in XML files in PASCAL VOC format. 1100 images are used for training, 100 images for validation, and 200 images are used for testing. The prepared dataset contains Military and Non-Military Vehicles with 8 different classes like Warplane, Plane, Tank, Warship, Ship, Military Truck, Military Jeep, and Person. During labeling the labels are assigned with the respective class names the generated XML file contains the class names and x, y box coordinate min, and max values which are present in the image for each image one XML file will be generated.

3.3. Creating Binary Mask Images and Generating COCO JSON File

Using the XML files and images which are used for creating XML's files both are used to generate the binary mask images separating the object from the background contain black color and foreground contain the white box and for example if an image having 3 objects in it based on the objects three separate binary mask images instances are generated. For generating the binary mask and COCO Json File python code was written and pycoccreator tools have used the images are renamed and with the help of created files COCO Json file was created this file is helpful to train the models. The created COCO Json file contains Image ID, Width, Height, Filename, annotations like boundary box, and category labels. Figure 2, shows the binary mask image of single image 0 having for objects in it at different locations and Figure 3, represents the visualization of ground truth boxes with class names for Aerial images and Normal images.

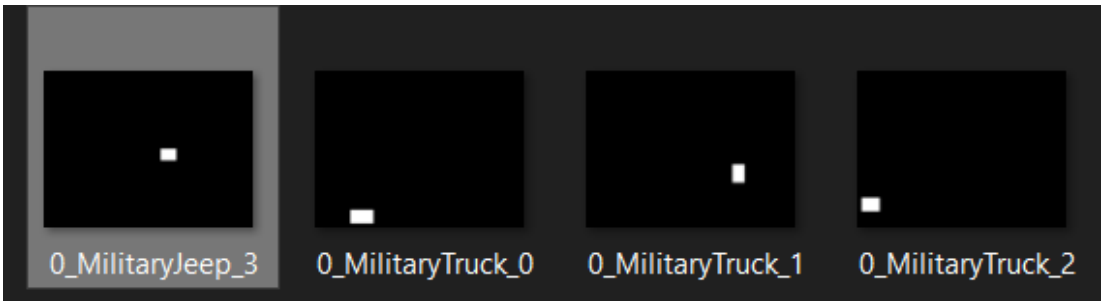

Figure 2: Created Binary Mask Images

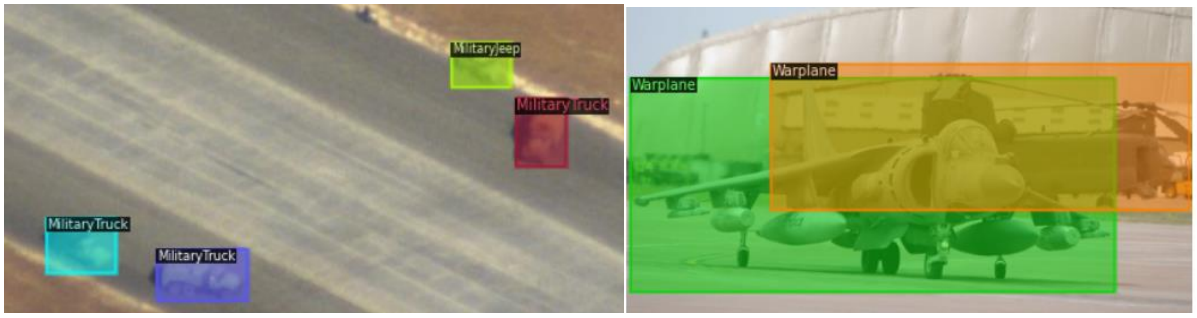

Figure 3: Ground Truth Boxes with Class Names

4. Design Specification

Two deep learning models are implemented to solve the problem of detecting military vehicles from Aerial Images and Normal Images. Faster R-CNN model with FPN using Resnet 101 model and Single Shot Detector300 with VGG16 Backbone networks are implemented. The Faster R-CNN model developed using the Detectron2 library and SSD model was developed using the MM Detection library.

Detectron2 is a next-generation object detection library that was developed by the Facebook Research Team it contains State of the art object detection algorithms. It is an open-source, extensible, and flexible library based on PyTorch. Detectron2 was an upgraded version of Detectron with help of this can build deep learning models. MM Detection is an object detection open-source toolbox helpful to build object detection models using MMCV.

4.1. Faster R-CNN with FPN using Resnet 101 Backbone

In Faster R-CNN the FPN is used as a feature extractor for generating the feature maps the FPN network made up of Bottom-up and top-down pathway. As shown in Figure 4, FPN uses Resnet 101 in the Bottom-up path is the feature extraction network Resent consist of a basic stem and res2, res3, res4, res5 Bottleneck blocks contain convolutional layers. In the basic stem, the input image is passed and down sampling it twice by 7x7 convolution with Batch norm layer and ReLu activation function and the bottom-up network goes up by stride =2 down sampling the network and output from the Bottom-up convolution layers used in the top-down network. FPN is made up of ResNet, up samplers Lateral and output Conv layers, and last level max pool layer.

The lateral Conv layers take Resent generated Features with different channel numbers from res2, res3, res4 and res 5 and return the 256 channel output maps and these are fed to the 3x3 convolutional output layer these result the output p5, p4, p3, p2, and p6 is generated by adding max pool layer to Resnet final block these are with different scales the FPN generates the multiscale Features maps are fed as input to the RPN network. The main use of the Multiscale network is to recognize large and small objects p2 and p3 helps in identifying small objects and p4 to p6 helps in identifying large objects.

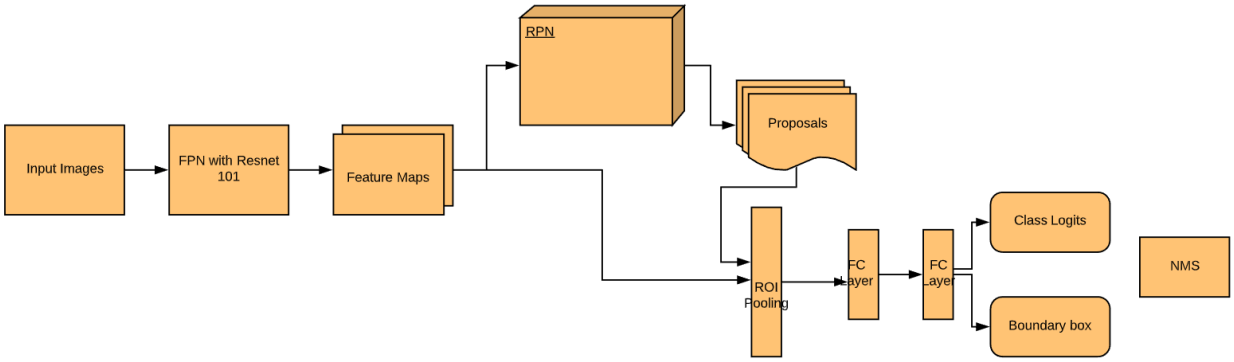

Figure 4: Faster R-CNN Using FPN with Resnet 101

Resent Output features are res2, res3, res 4, res5.

FPN Input features are res2, res3, res 4, res5.

FPN Output features are p2, p3, p4, p5, p6.

RPN Input Features are p2, p3, p4, p5, p6 these are multiscale feature maps generated by FPN. The RPN network consists of three convolutional layers these maps are fed to network one after another. For combining the generated ground truth with objectness and anchor deltas anchor boxes are used. Ground truth boxes values are taken from the generated COCO JSON file. The different anchor sizes and aspect ratios are used to generate anchor boxes, five different sizes and three aspect ratios are used and for 5 feature maps, 15 cell anchors are generated. By placing cell anchors on grid generates anchor boxes and with the help of the Intersection over union finding the anchor boxes which are close to the two ground truth boxes by setting the threshold value to 0.7 if its greater than that are labeled as foreground and less than 0.3 labeled as background and otherwise ignored. For generating the bounding boxes near to ground-truth the anchor deltas are calculated and resampling the boxes. Two functions are used for calculating the loss localization loss and objectness loss at the end from the RPN p2 to p6 top score boxes 1000 region proposals are taken by applying Non-Max Suppression.

In the ROI pooling the proposals from the RPN are used on feature maps to create ROIs by cropping the feature maps by following the assigned feature level rule the appropriate feature maps allocate to proposal box and before that the proposals are sampled are given to box head these consist of two fully convolutional layers these are flattened and in data there are eight classes. So, for each proposal eight different scores for eight classes and one background score is calculated, bounding box regressor predict boxes for all the classes. In the box head class scores and bounding, box deltas are calculated. The class loss was calculated using SoftMax cross-entropy and box regression using localization loss. At last, the predicted boxes with low scores are filtered and NMS is used to eliminate the overlapping boxes.

4.2. SSD300 Using Vgg16 Backbone

Single Shot Multi-Box Detector as shown in Figure 5, is implemented using Feature extractor as VGG16 which is a pre-trained network the SSD network is based on generating feature maps and for detecting objects convolutional filters are applied. From VGG16 it used conv4_3 layer and generates feature maps of size 38x38x512 and it uses till fully connected layer 7 and generates feature maps of size 19x19x1024 after that the convolutional layers are used to generate the feature maps the size of the generated feature maps gradually decreases across the layers. Multiple feature maps are produced at a different stage of the network with different sizes. The feature maps which are produced by the conv4_3 layer are raw maps from the input with the help of that model can able to detect the small objects and large objects are detected by the network using the low-resolution feature maps which are produced at the end by convolutional layers after VGG16.

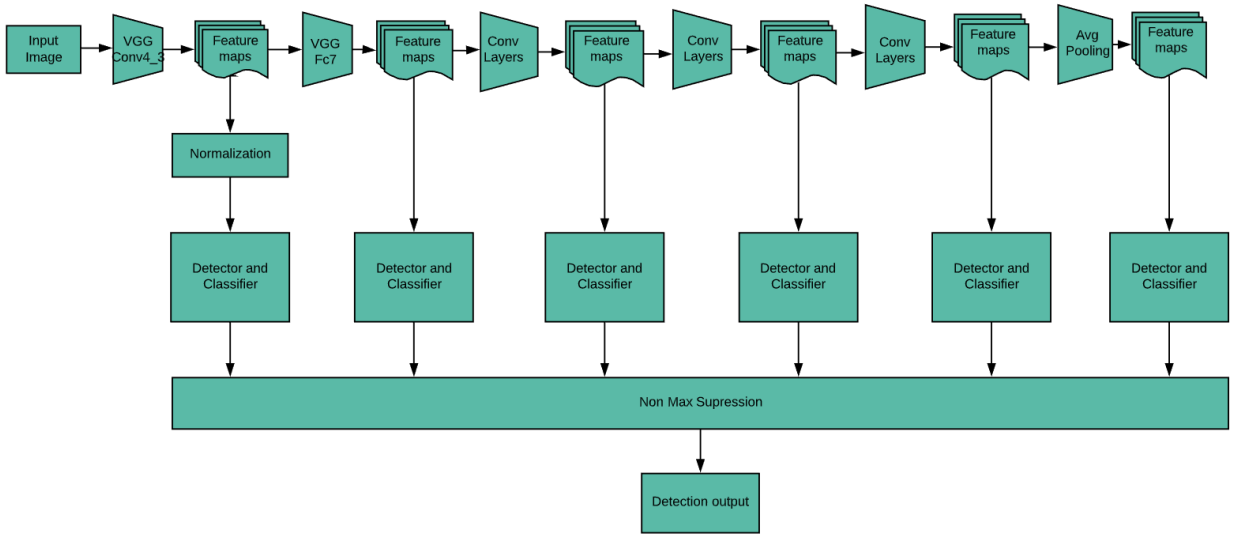

Figure 5: SSD300 Using VGG16

The Detector and classifier stage consist of generating the default boundary boxes SSD does not contain RPN network to generate the boundary boxes using anchors, it uses multi-box technique the SSD select these default boundary boxes based on defining the scale values from 0.2 to 0.9 to each feature layer and combining these with aspect ratios calculating the height and width of the boxes the generated default boxes are compared with the ground truth images if IOU value greater than 0.5 considered as positive else as negative these are initial guess by the network and taking the positive boxes as starting point it helps to predict shapes close to ground truth and applied small 3x3 convolutional filters on feature maps. for each location, the bounding boxes are produced with different size and aspect ratio the class scores are computed for each boundary boxes. All the feature maps with different input sizes generated and go through 3x3 convolution. The generated boundary boxes loss is calculated and defined by L1 loss it is mismatch between predicted and ground truth and predicted boxes. The SoftMax is used to calculate the confidence score of respective classes at the end Non-Max Suppression is used to discard the boxes which are having less confidence scores this removes the overlapping boxes.

5. Implementation

This section contains information about the implementation of the Faster R-CNN model and SSD models which are executed to detect the military vehicles and non-military vehicle classes in Aerial View images and Normal Images. To train deep learning models from scratch requires an enormous amount of data available due to the lack of Military images data transfer Learning approach is used for training both the models and helps to improve the performance.

5.1. Environment

The model implementation is carried using Google Collaboratory with 12.72 GB RAM and 68.40 GB Disk Space the data was loaded by mounting google drive and Google collab has runtime changing options both models are performed using GPU Runtime. For labelling the images labellmg tool was used. Detectron2, MM Detection, PyTorch, CUDA toolkit, Tensor Board, and other python libraries like mmcv, pycocotools are installed and python programming language is used for model's creation and to generate COCO JSON files.

5.2. Training Faster R-CNN using Transfer Learning

The Faster R-CNN model was trained using transfer learning by taking the weights from the pre-trained model zoo (Erickson *et al.*, 2017) and applied to the model. In Detectron2 the model training was iteration based the input images are resized to 800x800 and the training data contains 1100 images and these are validated by 120 images during training batch size is set to 2, the initial learning rate is 0.01 and trained with different iterations 5000, 10,000 and 20,000 and for validated Eval Period was defined according to iterations. For finding the best optimal values Stochastic Gradient Descent was used as an optimizer. After 20,000 iterations the model started to overfit, so the training was stopped before the model overfits. The model is evaluated using COCO Evaluation metrics by using the training predictions the training data was evaluated it contains 200 images during testing the threshold score was set to 0.8 these value can be adjusted it will just show the boxes with a score greater than that if the value is less than that it won't show the bounding box in output.

5.3. Training SSD300 using Transfer Learning

The Single Shot Multi-Box Detector was trained using transfer learning by using pre-trained model vgg16_caffe weights. The input images are resized to 300x300 and the input channels used are 512, 1024, 512, 256, 256, 256. The same data used for training Faster R-CNN was used to train SSD. In MM Detection the model training was epoch based. Stochastic Gradient Descent optimizer is used to discover the finest optimal values initially the learning rate is 0.001 and momentum is 0.9 and the bbox metric is used for training model evaluation. The model is trained with different epochs 10, 15, 20 and 30 at starting when trained with 10 epochs the accuracy of the predictions was less so started training with different epochs the model performance is improved after 30 epochs model is not learning and mAP value also not increasing the training predictions are applied on the test images.

6. Evaluation

For evaluating the implemented Faster R-CNN and SSD deep learning models the metrics used in the COCO detection challenge are implemented here it gives the model performance using twelve metrics at different levels. The model mean average precision, each class separates average precision value, average precision at different IOU thresholds, and at different object sizes are calculated. Average Recall for the number of detections per image and different scales are calculated. The Tensor Board Visualization tool is used to visualize the model and the tensor board can only visualize training or validation metric each at a time it can't able to plot both in a single graph.

6.1. Evaluation of Faster R-CNN model

The Table 1 represents the results of the Faster R-CNN model during training it consists of twelve different metrics the first average precision represents mean average precision calculated over ten thresholds incremented by 0.5. The mAP value is 82 it represents how well the objects detected are localized by bounding boxes and IOU at 0.50 and 0.75 are high. The model Mean average precision and mean average recall in identifying the small, medium, and large objects and average recall based on max detections per image are calculated. The dataset contains eight different classes the precision value for each is calculated and represented in Table 2, the Ship class has high value and person class having low precision value.

Table 1: Precision and Recall Values of Faster R-CNN Model on Training data

Model	mAP	AP at IOU=0.5	AP at IOU=0.75	AP for small	AP fro Medium	AP for large	AR for small	AR for Medium	AR for large	AR dets per image max =1	AR dets per image max =10	AR dets per image max =100
Faster R-CNN	0.82	0.99	0.98	0.76	0.75	0.84	0.78	0.79	0.88	0.67	0.85	0.85

Table 2: Each Class Average Precision Value on Training data

Classes	AP
Warplane	75.06
Plane	87.55
Warship	86.23
Ship	88.88
MilitaryTruck	81.95
MilitaryJeep	77.14
Tank	86.58
Person	72.40

The following Figure 6 shows the mean average precision and the Total loss of the Faster R-CNN the mAP value is increased over the iterations and the loss is decreased when iterations are increasing after 18k iterations loss again increased and decreased over 20k iteration.

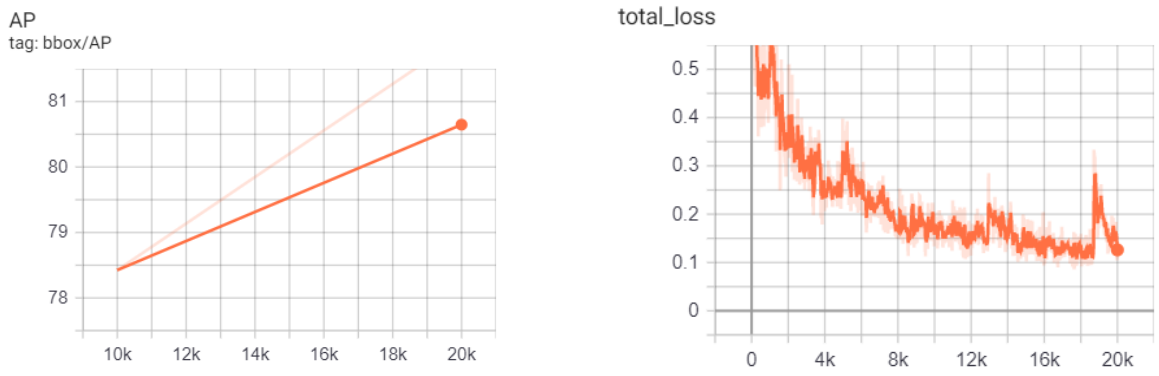

Figure 6: Mean Average Precision and Total Loss of Faster R-CNN Model

The classification and localization loss are visualized in the Figure 7, the loss_box_reg is the boundary loss when compared between the model predicted boxes with ground truth boxes and loss_cls is the loss that shows how well the model classified the predicted box is background or foreground it is SoftMax cross-entropy loss.

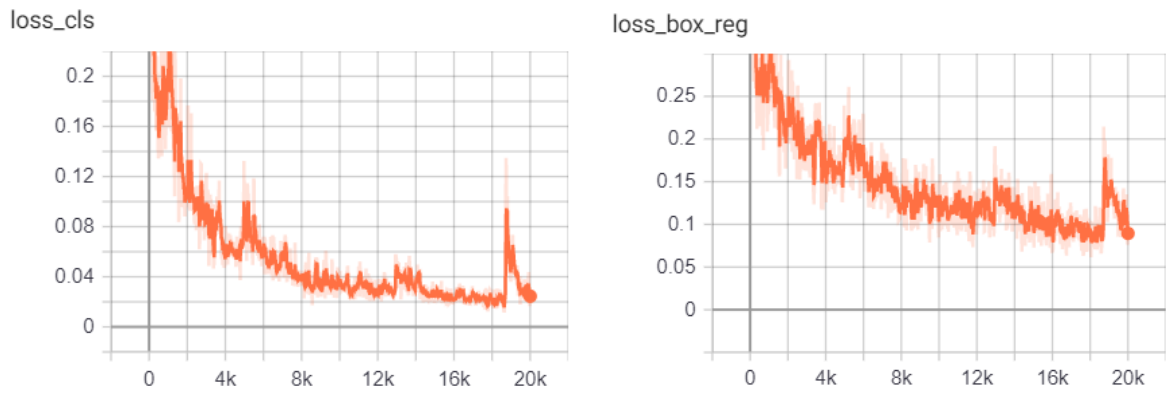

Figure 7: Classification and Localization Loss of Faster R-CNN model

These are the some of the predictions in below Figure 8, made by the Faster R-CNN model this model detected the warplanes and the persons with good accuracy.

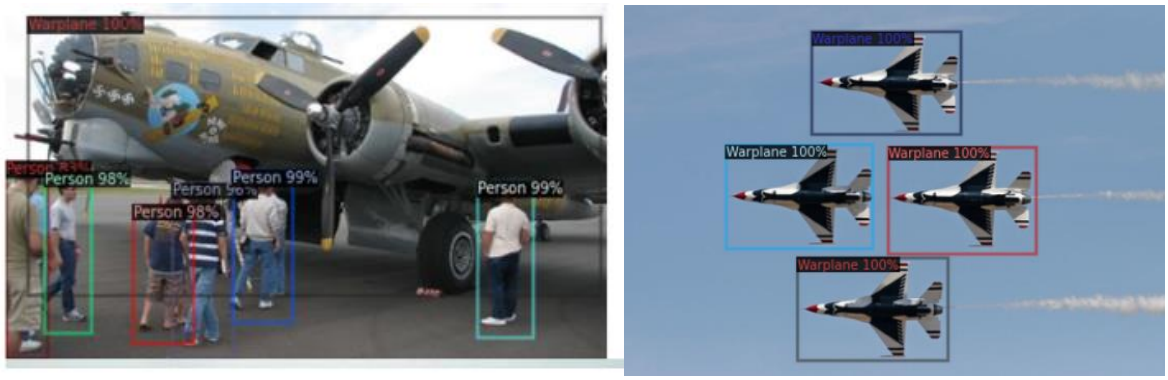

Figure 8: Faster R-CNN Model Predictions

The generated model is tested using the test data it contains 200 images the Table 3 represent the model mAP value it is 0.67 for test data and in Table 4 shows the model precision score for all the classes are calculated all the classes have decent values expect the person class because in training data the images with persons are very less so model didn't predict well for person class.

Table 3: Precision and Recall Values of Faster R-CNN Model on Test data

Model	mAP	AP at IOU=0.5	AP at IOU=0.75	AP for small	AP fro Medium	AP for large	AR for small	AR for Medium	AR for large	AR dets per image max =1	AR dets per image max =10	AR dets per image max =100
Faster R-CNN	0.67	0.85	0.79	0.17	0.35	0.73	0.17	0.38	0.76	0.60	0.71	0.71

Table 4: Each Class Average Precision Value on Test data

Classes	AP
Warplane	62.29
Plane	87.46
Warship	77.55
Ship	78.31
MilitaryTruck	63.71
MilitaryJeep	77.53
Tank	63.36
Person	27.20

6.2. Evaluation of SSD300 Model

Single Shot Multi-Box Detector was trained on 1100 images and evaluated at different metrics the Table 5 below describes the mAP value at different IOU thresholds is 0.68 and the model performed better in detecting the small objects when compared to medium and large scale objects the average recall and average precision values are high.

Table 5: Precision and Recall Values of SSD300 on Training data

Model	mAP	AP at IOU=0.5	AP at IOU=0.75	AP for small	AP fro Medium	AP for large	AR for small	AR for Medium	AR for large	AR dets per image max =100	AR dets per image max =300	AR dets per image max =1000
Faster R-CNN	0.68	0.92	0.78	0.70	0.51	0.61	0.73	0.58	0.68	0.73	0.73	0.73

The Figure 9 represents SSD model mAP and loss values mAP value increased suddenly from 2k to 6k iterations and after, that it gradually increased till 16k iterations the training on SSD using MM Detection is epochs based here in the graph it is representing in iterations. The training loss of the model reduced when the iterations increases.

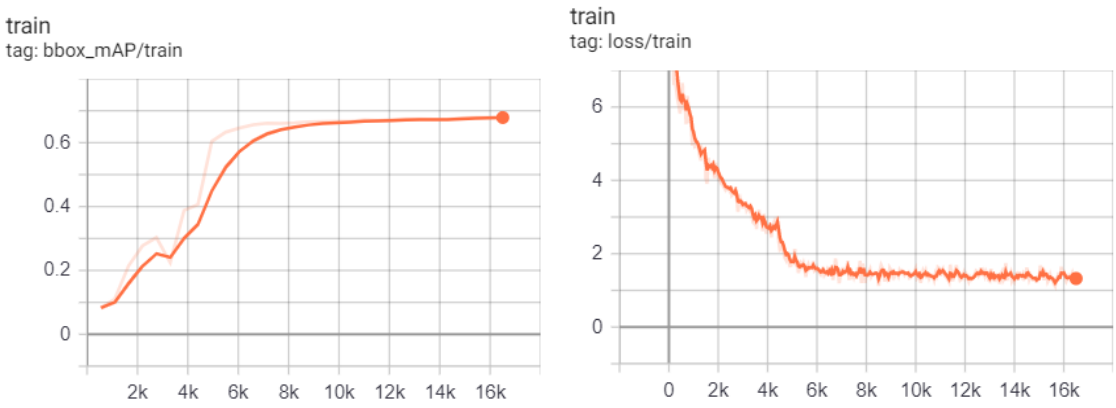

Figure 9: mAP and Total Loss of SSD300 Model

The below Figure 10 represents the loss_bbox and loss_cls for SSD model

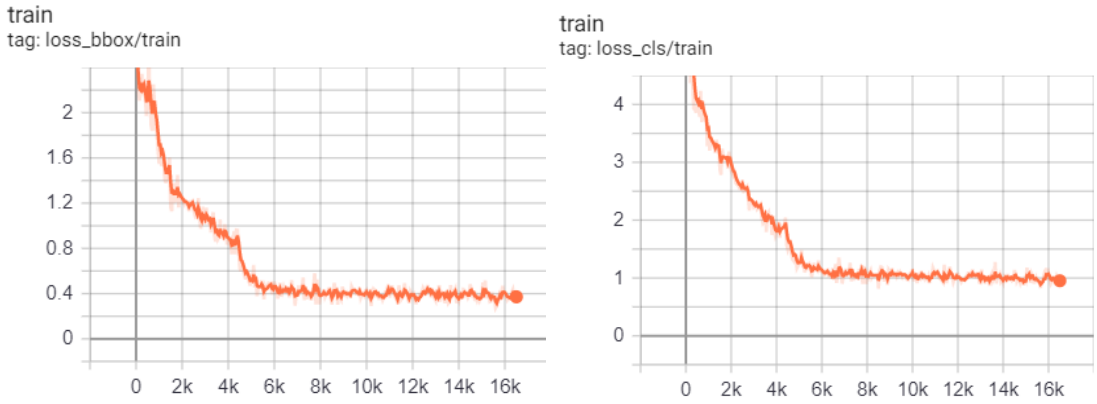

Figure 10: Classification and Localization Loss of SSD300

The below Figure 11 shows predicted outputs are made by the SSD model when testing the threshold score is set to 0.6 it will show only the box score is greater than that in left side image all the warplanes are detected the score was greater than 0.6 and so in right side image the model predicted the ships with high confidence score at the front but the model predicted ships at back with less score than 0.6.

Figure 11: SSD300 Model Predicted Output

6.3. Discussion and Comparison of Two Models

Both the developed SSD300 and Faster R-CNN models are compared. For detecting military vehicles and non-military vehicles the Faster R-CNN model achieved better results compared to the SSD300. The table 6 below shows the model, backbone, and mAP values. The mAP value is high for Faster R-CNN and the above image on left was tested on both the models. Faster R-CNN predicted with good accuracy compared to SSD300. In this research, to improve the precision and recall values of the Faster R-CNN and SSD300 model transfer learning approach and hyper parameters tuning are performed. Faster R-CNN model gave 82 percent mAP precision and SSD300 gave 68 percent mAP. SSD300 model has been implemented with the VGG16 backbone and also been predicted with less average precision value compared to SSD using Inceptionv2 pre-trained network reported in (Kamran *et al.*, 2019), the reason may be due to feature extractor and fine-tuning the network.

However, the Faster R-CNN that was implemented using FPN with ResNet 101 backbone is found to have predicted a high mAP value when compared to Faster R-CNN that was implemented using Inception-ResNet v2 network as described in (Kamran *et al.*, 2019) as FPN using ResNet 101 achieved better results in extracting features.

Table 6: Comparison of Faster R-CNN and SSD300

Model	Backbone	mAP	Iterations/Epochs
Faster R-CNN using FPN	ResNet 101	0.82	20000 Iterations
SSD300	VGG16	0.68	30 Epochs

7. Conclusion and Future Work

In this research, two deep learning techniques are applied to detect and classify the military and non-military vehicles. The dataset is prepared by aerial and normal images that are collected from ImageNet and VIVID vehicle tracking website. To deal with an inadequate amount of data, the transfer learning approach is implemented for Faster R-CNN and SSD300 models. In contrast to SSD300, the Faster R-CNN model predicted better results as it involves the usage of FPN with resnet101 for extracting features. On the other hand, the SSD300 model also detected small objects precisely compared to medium and large objects. As future work, my target is to detect military and non-military objects in videos using deep learning.

8. Acknowledgment

As a token of gratitude, I would like to thank my professor Hicham Rifai for being immensely supportive from the scratch of my research and providing timely guidance which helped me a lot in accomplishing the goals of the research. I also extend my thanks to the National College of Ireland, Dublin which has been a backbone throughout my master's program and helping me learn new technologies by giving innovative projects which have also played a vital part in shaping my technical background. Finally, I would like to thank my parents and friends who have been constantly encouraging me from the beginning to work hard despite all the odds and boosting up my confidence level to face any type of challenge in the work undertaken by me.

References

Alganci, U., Soydas, M. and Sertel, E. (2020) 'Comparative research on deep learning approaches for airplane detection from very high-resolution satellite images', *Remote Sensing*, 12(3). doi: 10.3390/rs12030458.

Chen, X., Xiang, S., Liu, C. L. and Pan, C. H. (2014) 'Vehicle detection in satellite images by hybrid deep convolutional neural networks', *IEEE Geoscience and Remote Sensing Letters*. IEEE, 11(10), pp. 1797–1801. doi: 10.1109/LGRS.2014.2309695.

Chen, Z., Wu, K., Li, Y., Wang, M. and Li, W. (2019) 'SSD-MSN: An Improved Multi-Scale Object Detection Network Based on SSD', *IEEE Access*. IEEE, 7, pp. 80622–80632. doi: 10.1109/ACCESS.2019.2923016.

Erickson, B. J., Korfiatis, P., Akkus, Z., Kline, T. and Philbrick, K. (2017) 'Toolkits and Libraries for Deep Learning', *Journal of Digital Imaging*. Journal of Digital Imaging, 30(4), pp. 400–405. doi: 10.1007/s10278-017-9965-6.

Hiippala, T. (2017) 'Recognizing military vehicles in social media images using deep learning', *2017 IEEE International Conference on Intelligence and Security Informatics: Security and Big*

Data, ISI 2017, pp. 60–65. doi: 10.1109/ISI.2017.8004875.

Hu, X., Zhang, P. and Xiao, Y. (2019) 'Military object detection based on optimal gabor filtering and deep feature pyramid network', *ACM International Conference Proceeding Series*, pp. 524–530. doi: 10.1145/3349341.3349462.

Kamran, F., Shahzad, M. and Shafait, F. (2019) 'Automated Military Vehicle Detection from Low-Altitude Aerial Images', *2018 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2018*. IEEE, pp. 1–8. doi: 10.1109/DICTA.2018.8615865.

Ke, C. (2017) 'Military object detection using multiple information extracted from hyperspectral imagery', *Proceedings of 2017 International Conference on Progress in Informatics and Computing, PIC 2017*, pp. 124–128. doi: 10.1109/PIC.2017.8359527.

Lee, J., Lee, S. K. and Yang, S. Il (2018) 'An Ensemble Method of CNN Models for Object Detection', *9th International Conference on Information and Communication Technology Convergence: ICT Convergence Powered by Smart Intelligence, ICTC 2018*. IEEE, pp. 898–901. doi: 10.1109/ICTC.2018.8539396.

Liu, S. and Liu, Z. (2017) 'Multi-Channel CNN-based Object Detection for Enhanced Situation Awareness', *IEEE Access*, pp. 1–9. Available at: http://arxiv.org/abs/1712.00075.

Tianyu, T., Shilin, Z., Zhipeng, D., Lin, L. and Huanxin, Z. (2017) 'Fast Multidirectional Vehicle Detection on Aerial Images Using Region Based Convolutional Neural Networks', *IEEE Access*, 5, pp. 5–8.

Wang, B., Zhou, Y., Zhang, H. and Wang, N. (2019) 'An aircraft target detection method based on regional convolutional neural network for remote sensing images', *ICEIEC 2019 - Proceedings of 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication*, pp. 474–478. doi: 10.1109/ICEIEC.2019.8784637.

Wang, C., Jiang, Q., Cheng, M., Li, J. and Cao, L. (2016) 'DEEP NEURAL NETWORKS-BASED VEHICLE DETECTION IN SATELLITE IMAGES Fujian Key Laboratory of Sensing and Computing for Smart City School of Information Science and Engineering , Xiamen University Xiamen , China', *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*.

Wang, Z., Du, L., Mao, J., Liu, B. and Yang, D. (2019) 'SAR target detection based on SSD with data augmentation and transfer learning', *IEEE Geoscience and Remote Sensing Letters*. IEEE, 16(1), pp. 150–154. doi: 10.1109/LGRS.2018.2867242.

Xie, X. and He, C. (2017) 'Object detection of armored vehicles based on deep learning in battlefield environment', *Proceedings - 2017 4th International Conference on Information Science and Control Engineering, ICISCE 2017*, pp. 1568–1570. doi: 10.1109/ICISCE.2017.327.

Yang, Z., Yu, W., Liang, P., Guo, H., Xia, L., Zhang, F., Ma, Y. and Ma, J. (2019) 'Deep transfer learning for military object recognition under small training set condition', *Neural Computing and Applications*. Springer London, 31(10), pp. 6469–6478. doi: 10.1007/s00521-018-3468-3.

Zhai, S., Shang, D., Wang, S. and Dong, S. (2020) 'DF-SSD: An Improved SSD Object Detection Algorithm Based on DenseNet and Feature Fusion', *IEEE Access*, 8, pp. 24344–

24357. doi: 10.1109/ACCESS.2020.2971026.

Zhang, M., Li, H., Xia, G., Zhao, W., Ren, S. and Wang, C. (2018) 'Research on the Application of Deep Learning Target Detection of Engineering Vehicles in the Patrol and Inspection for Military Optical Cable Lines by UAV', *Proceedings - 2018 11th International Symposium on Computational Intelligence and Design, ISCID 2018*. IEEE, 1, pp. 97–101. doi: 10.1109/ISCID.2018.00029.

Zhang, W., Li, J. and Qi, S. (2019) 'Object detection in aerial images based on cascaded CNN', *Proceedings - 2018 International Conference on Sensor Networks and Signal Processing, SNSP 2018*. IEEE, pp. 434–439. doi: 10.1109/SNSP.2018.00088.

Zhong, J., Lei, T. and Yao, G. (2017) 'Robust vehicle detection in aerial images based on cascaded convolutional neural networks', *Sensors (Switzerland)*, 17(12). doi: 10.3390/s17122720.

Zhuang, S., Wang, P., Jiang, B., Wang, G. and Wang, C. (2019) 'A single shot framework with multi-scale feature fusion for geospatial object detection', *Remote Sensing*, 11(5). doi: 10.3390/rs11050594.