

A Comparison of Fixed Threshold CFAR and CNN Ship Detection Methods for S-band NovaSAR Images

Tobias Carman, Avyaya Kolhatkar
Defence Science and Technology Laboratory
Dstl Portsmouth West, Portsmouth Hill Road, Fareham PO17 6AD, United Kingdom
tcarman@dstl.gov.uk, akolhatkar@dstl.gov.uk

ABSTRACT

NovaSAR is a commercial S-band Synthetic Aperture Radar (SAR) small satellite, built and operated by SSTL in the UK. One of its primary mission objectives is to carry out maritime surveillance and monitoring for security and defence applications. An investigation was carried out into comparing and contrasting conventional and new methods to perform automated ship detection in NovaSAR images. The outcome of this investigation could show the potential effectiveness of ship detection using spaceborne S-band SAR for Maritime Domain Awareness (MDA).

The conventional approach is to apply a suitable distribution model to characterise sea surface clutter, followed by the implementation of a fixed threshold, Constant False Alarm Rate (CFAR) detection algorithm. In comparison, a RetinaNet-based convolutional neural network (CNN) solution was developed and trained on an open-source C-band dataset in order to determine the validity of applying non-native training data to S-band imagery. The detection performance was then compared with the CFAR technique, finding that for two selected test acquisitions a CNN-based ship detection algorithm was able to outperform a fixed threshold, CFAR-based method in the absence of native training data. CNN ship detection performance was further improved by applying transfer learning to a native S-band NovaSAR image dataset.

INTRODUCTION

NovaSAR Mission

NovaSAR is a small (430kg) commercial S-band Synthetic Aperture Radar (SAR) satellite, built by SSTL in the UK and launched in September 2018. It is capable of acquiring images with up to 6m resolution in Stripmap mode, and also features a Maritime mode with a 400km swath. In addition, the satellite hosts an Automatic Identification System (AIS) receiver to aid ship identification. The main focus of the mission is to serve as a demonstrator of low cost space-based SAR. One of the primary objectives is to demonstrate Maritime Domain Awareness (MDA) for security applications, including the prevention of illegal fishing. The global economic impact of illegal and unreported fishing losses has previously been estimated at between \$10-23.5 billion annually¹. British maritime protected areas are distributed across the globe, and are therefore difficult to monitor without space-based Earth Observation (EO) assets. Other objectives for the UK government in this domain that space-based EO may be able to contribute to could include:

- Deterring arms and narcotics smuggling
- Countering terrorism and counter-piracy operations
- Monitoring movement of refugees and preventing people trafficking
- Protecting vital maritime trade, including energy transportation routes
- Protecting the integrity of UK and British Overseas Territories marine areas
- Marine pollution detection and attribution

- Sea ice monitoring and shallow bathymetry to aid safe transit
- Supporting overseas evacuation operations of British citizens
- Search and rescue

The contemporaneous collection of both SAR images and AIS signals over maritime areas provides two complementary streams of geospatial intelligence that can be applied to the above problems. AIS information is not considered reliable enough on its own for a number of reasons, including:

- AIS transponders can be switched off
- Information broadcast such as location, vessel name or unique identifier can be fabricated
- Low probability of detection by satellite receivers over congested areas²

Ship Detection

Ships present a highly reflective cross-section to radar, with multiple opportunities for double-bounce backscattering. They therefore tend to appear bright in SAR images in comparison to the relatively dark sea background, and are theoretically easy to detect. However, in ports or rough sea conditions there can be a lot of clutter present in the images, making this more difficult. Conventional automated detection techniques have operated on the basis of masking out the land and modelling the sea surface clutter according to one of a number of statistical distributions, with a Constant False Alarm Rate (CFAR) detection algorithm³. In recent years, methods including the Generalised Likelihood Ratio Test (GLRT)⁴ as well as deep learning/computer

vision techniques including Convolutional Neural Networks (CNN) have demonstrated improved detection performance over CFAR.

Previous studies in this area have, however, utilised either Sentinel-1 (C-band), Gaofen-3 (C-band) or TerraSAR-X (X-band) SAR images, and the application of S-band data to this problem is believed to be a new area of research. It is unknown whether or not a CNN-based methodology outperforms a CFAR-based one for S-band images. Additionally, the impact of applying training datasets made up of imagery of different band/resolution to the testing dataset has not previously been investigated in depth. This investigation was designed to determine, for S-band SAR imagery:

- i. Whether a CNN-based ship detection methodology could outperform a CFAR-based one
- ii. The impact on detection performance of training this CNN on C-band imagery, compared with training on a native S-band dataset.

Performance Metrics and Terminology

In the object detection field for CNNs, success is measured in terms of *Intersection over union (IoU)*, *precision*, *mean average precision (mAP)* and *recall*. CFAR methodology uses *probability of false alarm (P_{fa})* and *probability of detection (P_d)*.

Intersection over Union (IoU), also known as the Jaccard index J , measures the overlap between the true bounding box A of an object in an image and the predicted bounding box B , as shown in Figure 1.

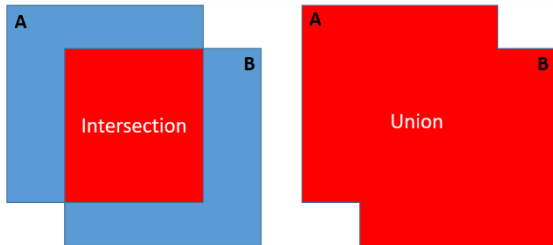


Figure 1: Intersection (overlapping red area on the left) and Union (combined red area on the right) of two bounding boxes A and B.

IoU is given by the equation:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

The intersection $|A \cap B|$ is the overlapping region, and the union $|A \cup B|$ is the total area of the combined region formed.

Predictions can be described as True Positives (TP), False Negatives (FN) or False Positives (FP), determined by their IoU value. If the IoU of a predicted bounding box is above the threshold that has been set, the prediction is a true positive. If the IoU is below this threshold then the prediction is a false positive; there is not sufficient overlap between the prediction that has been made and the ground truth. This may occur when the object is present, but has not been bounded correctly, or when there is no object present. A false negative occurs when the object is present but no prediction is made.

Precision is defined as the number of true positives out of the total number of positive predictions:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Qualitatively, this may be thought of as the proportion of predictions made that were correct.

Recall is defined as the number of true positives out of the total number of true positives and false negatives, equivalent to the total number of ground truths:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Qualitatively, this may be thought of as the proportion of objects which were detected.

The F_1 score is often used to combine precision and recall scores into a single metric, defined as the harmonic mean of the precision and recall:

$$F_1 = \left(\frac{Recall^{-1} + Precision^{-1}}{2} \right)^{-1} \quad (4)$$

This simplifies to:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5)$$

Average precision (AP) is the precision averaged across all recall values. Mean average precision (mAP) takes all AP values for the classes and IoU thresholds considered and finds the mean of these. For a simple ship detection (rather than classification) system, there is only one class to consider (ship) and therefore the mAP for a given IoU is simply the average precision across all test images.

Reducing the IoU threshold required for a detection, or in the context of a CFAR detector, raising the false alarm rate, would be expected to lead to an increased number of both true and false positives. This will in general have the effect of increasing the recall whilst lowering the precision, and vice versa if the IoU threshold or false alarm rate is raised.

CNN Training and Challenges

Full-size SAR images will often contain more than 10000 pixels. It is usual to segment the image into smaller sub-image tiles for training and detection purposes.

Once the neural network has been trained, typically beginning from a set of pre-trained weights, the resulting model may be used for inference. In the wider object detection field, training datasets can range into the millions of images for problems involving multiple classes of objects. However, for ship detection, thousands of image tiles can be sufficient to obtain high levels of detection performance if classification between types of ships is not required.

This still presents a problem for new systems during their first months or years of operational life, since a training dataset must first be accumulated through hundreds of acquisitions. These acquisitions should ideally feature globally distributed locations in a variety of sea states in order to maximise the robustness of the network and ensure its geo-generalisability.

The images must then be individually hand-labelled by an analyst before a neural network can be trained in order to start to make predictions with a useful degree of accuracy. However, if ground truth data in the form of either accompanying optical imagery or AIS data is not available, this process can be challenging since many objects that backscatter brightly can appear similar to ships.

The training process itself is also time-consuming, with models taking days or even weeks to be fully trained dependent on hardware, size of the training dataset and number of epochs (number of times the network sees the entire training dataset). Any changes in configuration of the network require retraining in full before they can be tested, which drastically lengthens the timescale necessary to find the optimal configuration.

LITERATURE REVIEW

Land Masking

It can be difficult to find ships in littoral regions of an image due to the highly reflective coastal and land regions that can make the surrounding areas quite noisy and sometimes obscure maritime regions due to specular reflections. It is therefore critical to mask these regions in order to detect vessels or offshore objects accurately using a fixed threshold CFAR based method. Ensuring all the land is correctly masked also ensures that there are no false alarms generated from reflective surfaces on land.

Several methods have been used to land mask SAR images, the simplest of which is to simply overlay a shoreline shape file or DEM model over the GeoTIFF image. This requires the geolocation accuracy of the sensor to be relatively accurate and therefore does not work for TIFF SAR images that have not been accurately georeferenced. Another quick method proposed by Kefeng⁵ is to down sample the image until the largest vessels occupy a single pixel. Then apply a median filter to eliminate ships from the low-resolution image. Then a 2-threshold histogram-based segmentation method is used to remove bright regions. This method only works well for images with relatively calm sea state as it works on the assumption that the land regions are always brighter.

Martin-de-Nicolas⁶ provides a comparison of several segmentation based techniques for land masking including Canny edge detection, wavelet-transform based edge detection, mean shift algorithm and clustering based segmentation techniques. Edge detection methods measure the intensity gradient across pixels to identify land sea boundaries and edge orientation. The Canny edge detection method⁷ developed by John Canny convolves the image pixel gradient with a two dimensional Gaussian first derivative (G_n) distribution model to identify the peak intensity and peak gradient as a smoothed step would demonstrate a low edge strength in-line with the edge and a strong gradient normal to the edge. The directional magnitude can be described by:

$$|G_n * I| = |\nabla(G * I)| \quad (6)$$

where I is the image intensity. When selecting the edges that correctly define the boundary between land and sea, it is critical to apply the appropriate threshold values. A double threshold is required for this method as a single threshold does not reflect the variation of coastline contours, which will have areas of softer edges that would subsequently cause several break points in the detection. A range of acceptable thresholds enables the boundary to be defined as a solid line but also risks marking noise edges if the range is too large.

Clutter Modelling

The next critical step in determining the presence of vessels in the maritime environment is to model the sea state accurately. This is an incredibly complex problem and does not have a single solution. The sea clutter can be modelled by analysing the histogram of the land masked image. Rough sea states tend to produce 'spikey' tail features in the histogram that can be difficult to model. Several papers use a number of distributions to attempt to model sea states. Sea clutter tends to display an underlying mean intensity with a modulating speckle

component⁸. The K distribution is the most widely accepted model for SAR imagery⁹. The K distribution probability density function (PDF) is very similar in shape to the Weibull distribution. It is the compound formulation of the K distribution which is important⁸. Jian Sun¹⁰ uses a Gamma, Weibull, Nakagami, Log-Normal, Rayleigh and K distribution across a number of wavelengths and found that a K distribution provided the best parameters to fit the test data. Sebastien Angelliaume used K + noise (KN), Pareto + noise (PN), K + Rayleigh (KR) and trimodal discrete (3MD) distributions¹¹. His results showed that the KR and 3MD model provided the better ‘goodness of fit’ metric to the S band NetRAD dataset. 3MD had the best performance at the cost of a greater number of parameters.

The probability density function (PDF) for the lognormal distribution is defined as:

$$PDF_{lognormal}(x; \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \quad (7)$$

Where σ is the scale parameter and μ is the shape parameter. The K distribution better captures the long spikey tail of the image distribution. It usually includes a gamma functions Γ and fast fluctuating component that uses a modified Bessel function of the second kind K_α . The three parameter PDF is given by:

$$PDF_k(x; \mu, \nu, L) = \frac{2\xi^{(\beta+1)/2} x^{(\beta-1)/2}}{\Gamma(\nu)\Gamma(L)} * K_\alpha(2 * \sqrt{\xi x}) \quad (8)$$

Where μ is the calculated mean of the image data, L is the number of looks and ν is the shape parameter¹². The gamma distribution is given by:

$$PDF_{gamma}(x; h\nu) = \frac{h^\nu}{\Gamma(\nu)} x^{\nu-1} \exp(-hx) \quad (9)$$

Where h is the scale parameter and ν is the shape parameter. Measuring the ‘goodness of fit’ can be accomplished in a number of ways, two of which are by using the Maximum likelihood estimation (MLE) or threshold error¹¹. The threshold error is usually calculated using the cumulative distribution function (CDF). In this context it is also referred to as the probability of false alarm and acts as a useful metric to describe how far over or under estimate a CFAR threshold would be set.

Probability of False Alarm and Probability of Detection

Clutter can be described in terms of its amplitude distribution with the probability of detection P_d and probability of false alarm P_{fa} given for a fixed threshold that does not vary spatially. To get a more dynamic threshold the mean amplitude across over local spatial variations can be taken to provide a more accurate

threshold in regions of the image with higher or lower average intensities.¹³ The P_{fa} for an ideal threshold is given by:

$$P_{fa} = \int_t^\infty P(x)dx \quad (10)$$

where the threshold varies along the distribution. This can be particularly useful for large images with non-uniform backscatter properties. For a uniform backscatter, a single threshold can be calculated by setting the false alarm to a value, usually 10^{-4} to 10^{-6} ¹¹. The PDF of the cell-averaged threshold $P(t)$ is taken as the sum of M independent Rayleigh distributed samples. $P(t)$ is given by¹³:

$$P(t) = \left(\frac{M}{\alpha}\right)^{MN} \frac{t^{MN-1}}{\Gamma(MN)} \exp\left(-\frac{Mt}{\alpha}\right) \quad (11)$$

Then the average \bar{P}_d is calculated using:

$$\bar{P}_d = \int_0^\infty \left(\int_t^\infty P(x)dx\right) P(t)dt \quad (12)$$

It is worth noting that the P_{fa} being set dynamically allows the P_d to be evaluated in a range of sea states due to some SAR images such as strip map mode, covering large distances in azimuth. An acceptable false alarm rate can be determined based on the situation. A trade off must be made between a high false alarm rate with high probability of detections and a low false alarm rate with the risk of missing many detections. A receive operating characteristic (ROC) curve is useful for characterising the performance of the model using these metrics.

CNN Ship Detection

Several CNN-based ship detection and classification techniques have been proposed in the last 3-4 years. Some^{14, 15} have used CFAR in conjunction with a CNN in order to reduce false-alarm rate compared to a pure CFAR solution. Several¹⁵⁻¹⁷ have even had success classifying different types of ships and other marine objects such as wind turbines and oil platforms using high resolution TerraSAR-X and Gaofen-3 imagery.

Pure CNN-based methods applied to both optical and SAR imagery have predominantly used either two-stage R-CNN derivatives^{18, 19} (Fast R-CNN²⁰, Faster R-CNN²¹) which are dependent on region proposals, or one-stage regression-based detectors SSD²²⁻²⁴ (Single Shot Detector) or YOLOv2^{25, 26} (You Only Look Once).

YOLOv2 showed²⁵ improved performance (90.05% mAP) when compared to Faster R-CNN with an order of magnitude reduction in detection execution time. YOLOv3²⁷ introduced improvements in bounding box and class prediction, as well as feature extraction, which increased detection performance for small objects in

comparison to YOLOv2 and SSD. The backbone network developed for use with YOLOv3 is named Darknet53, since it contains 53 convolutional layers. There have not yet been any published studies evaluating the use of YOLOv3 for ship detection.

Recently, RetinaNet²⁸ has also been applied^{23, 29, 30} to ship detection in SAR images, demonstrating³⁰ the highest precision seen for any CNN with up to 97.56% mAP. RetinaNet introduces two key advances in one-stage object detection: feature pyramid networks (FPN) for feature extraction³¹ and focal loss for dense sampling²⁸.

FPNs³¹ feed feature maps representing the input image at different scales into an object detector, allowing for more accurate detections since objects may occupy a range of different scales. Crucially, FPNs allow all of these scales to be evaluated as part of the neural network's inherent structure with increased resolution but without significant impact on processing time.

Focal loss aims to rectify the class imbalance introduced between easy and hard examples during training. In object detection, far more negative samples are evaluated since the majority of candidate locations are in empty background regions, and detectors therefore focus the majority of their efforts on learning to classify easy background areas rather than the more difficult to detect objects of interest.

Typical cross-entropy (CE) loss measures the performance of a binary classification model, penalising predictions that are wrong with a high loss value. CE loss takes the following form²⁸:

$$CE(p_t) = -\log(p_t) \quad (13)$$

where \log here denotes the natural logarithm and p_t is essentially the correctness of the prediction, formally:

$$p_t = \begin{cases} p, & y = 1 \\ 1 - p, & \text{otherwise} \end{cases} \quad (14)$$

where p is the predicted confidence of the class being present, and y is the class label, equal to 1 if the class is present or -1 if not.

Therefore if the classifier predicts the probability of the class being present is 0.9, and the class is present, $p_t = 0.9$ and $CE = 0.105$ (to 3 s.f.). If the class was not in fact present, $p_t = 0.1$ and $CE = 2.30$ (to 3 s.f.). The further the prediction diverges from reality, the higher the loss incurred. However even when negative examples are correctly classified (i.e. a low probability is predicted), the total loss incurred is still significant since there are so many of them. Focal loss addresses this problem by

introducing a focusing parameter $\gamma \geq 0$, defining focal loss (FL) as²⁸:

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t) \quad (15)$$

If p_t is near 0, i.e. the example is misclassified, the $(1 - p_t)^\gamma$ factor is close to 1 and $FL \approx CE$. However as p_t tends towards 1, i.e. the example is classified correctly with high confidence, this factor tends towards 0. For $\gamma = 2$ and $p_t = 0.9$ as before, $FL = 0.00105$; 100 times smaller than the CE loss, whereas for $p_t = 0.1$, $FL = 1.87$; only 1.23 times smaller than the CE loss. This has the effect of down weighting the loss contribution from easily classified examples, leading training to be focused towards the more difficult examples in order to reduce the overall loss.

For ship detection in satellite imagery, it is expected that focal loss will be highly applicable, since there is a large amount of background in comparison to the relatively small objects to be detected. Ships in harbours or close to other ships may also be more easily distinguished by RetinaNet compared to other networks since these harder examples will be focused on more during training than the easier examples single, bright ships in open water.

METHODOLOGY

This section seeks to detail the unforeseen but necessary steps involved in ensuring a high detection precision is achieved. In order to mask the land regions in the image a Canny edge detector was implemented, however its performance was poor due to the high number of noisy edges detected in the original image. Some pre-processing algorithms were used to improve the performance. A Gaussian filter was used initially, as it reduced the speckle noise from the image and by moving a kernel over the image one pixel at a time, creating a smoothing effect.

Land Mask

The image was down-sampled using the average cell value within the kernel. In addition to a reduction in false edges being detected, this reduced the overall size of the image and therefore improved processing speed.

A Canny edge detector upper and lower threshold were set manually to optimise its performance in detecting land edges and ignoring softer edges detected in the ocean. In order to make sure any breakages in the edge detection were properly connected the edges were dilated, to close the gaps in the image. The land regions that touched the ends of the image also needed to be closed off in order to fill the gaps. Once the gaps were filled, the land regions in the image were counted; with key features such as region centroids and areas extracted. In order to make sure any vessels were not mistaken for

small land masses, a percentage of the mean area of all the land masses was taken. This required a manual percentage allocation for each image.

Finally, the land masked regions were converted to a binary array and scaled up to match the original image size. This created a small offset as it did not always scale to an integer number of pixels. This offset was rectified using a dilation function once again. This loss in shoreline details was seen as acceptable due to the model being aimed at open water vessel detection rather than littoral regions. Upon revisit, it was found that the offset problem when scaling up the mask was eliminated if the image was not down sampled using cell averaging. This in turn increased the processing time however proved more effective for single acquisitions.

Distribution Fitting

The PDF of a lognormal, Gamma and K distributions were calculated for the image dataset. Based on these distributions the \log_{10} CDF was calculated to find the distribution that fitted the empirical data the best. The false alarm threshold was set to 10^{-6} initially and the increased to measure the effect on detector performance. In order to get an accurate distribution for the dataset all zero pixel values occurring due to the land mask were removed, as this would have heavily skewed the distribution.

In some cases, the SAR image was heavily saturated and caused the image to appear bright. In order to reduce the effect of this the pixel intensity was capped at a maximum value, which allowed the intensity distribution to be stretched for better contrast between ship and sea surface. The stretched image however was not used for the thresholding in order to preserve information about the brighter pixels.

Thresholding

For the strip map images a single ideal threshold was used as this proved to perform well. The ScanSAR images would require more adaptive thresholding methods using cell averaging as described in the previous section. Then the mean power could be calculated to find the false alarm threshold across averaged cells¹³.

Once the false alarm was set, the distribution that had the lowest error to image data was used to calculate the ideal threshold. The error was measured in dB and converted to an 8-bit value for the threshold. The image is then converted to binary and regions detected above and below the pre-defined size range of vessels to be detected, are removed. This limits the minimum detectable ship length but also removes any non- vessel objects that may be highly reflective on the sea surface.

Probability of Detection

For a fixed threshold model, the probability of detection can be measured as a function of pixel power in dB. The aim of this report is to compare the performance of a fixed threshold, CFAR detector against the CNN approach, therefore precision and recall were calculated by cross-referencing detected regions with the labelled NovaSAR images for the acquisition.

Bounding boxes were drawn around regions that were thought to be detections. These were pixel positions rather than Cartesian coordinates in order to compare with labelled images and calculate the IoU. The detections could be converted to georeferenced coordinates for comparison with other sensor data e.g. AIS, however this test has not been taken further in this report.

Model Sensitivity

To have a truly robust tool the subtle and not so subtle variances in different types of SAR imagery must be considered. As mentioned in the land masking section of the methodology, high-resolution imagery with a small swath will perform differently to lower resolution imagery with a wider swath. This is due to a greater range of sea states that may be captured in the larger image, making a single threshold less effective. Many studies have been carried out to show that polarisation and incidence angle also have a large impact on reflectivity of the ocean surface.

CNN Configuration

Two CNN-based object detectors were chosen for initial investigation: the AlexeyAB fork³² of YOLOv3²⁷ and the Fizyr keras-retinanet implementation³³ of RetinaNet²⁸. YOLOv3 was chosen since YOLOv2 previously demonstrated strong performance in ship detection²⁵, and YOLOv3 was shown to have further improved performance in object detection²⁷. Both were trained on the open source SAR Ship Detection Dataset²³ (SSDD), which consists of 43,819 ship tiles, each of resolution 256×256 pixels with 50% overlap between them. The tiles are cropped from a total of 210 images captured by Gaofen-3 and Sentinel-1, both C-band SAR satellites, and are provided with the coordinates of bounding boxes for the locations of ships in accompanying label files.

The dataset was split randomly into 70% training, 20% validation and 10% test portions for both YOLOv3 and RetinaNet. Due to the differing formats and file configurations between the two networks, they were each trained on a different random split, however the results are still expected to be broadly comparable.

YOLOv3 was trained using the Darknet53 backbone from the darknet53.conv.74 starting weights, with a batch size of 64, 32 subdivisions, input image size of 512×512 and a learning rate of 0.001 for 12,000 batches. Batch and subdivision sizes of 1 were used for testing. The network was trained once without any image augmentation, and once with augmentation on the same data split of up to 5 degrees in image rotation and up to 1.5 in exposure magnitude to investigate the applicability of traditional augmentation techniques to SAR imagery. Hue and saturation colour augmentations were not applied since the images are single channel i.e. greyscale.

The validation mAP appeared to plateau during training after 9000 batches, so training was stopped after 12000 batches to avoid overfitting. Image augmentation appeared only to decrease stability and contribute a requirement for longer training times without improvement in precision or recall. It was therefore concluded that these classical image augmentation techniques did not provide benefit to detection performance in SAR imagery and so were not applied when training RetinaNet.

RetinaNet was trained on the SSDD using the ResNet-50 backbone from the resnet50_coco_best_v2.1.0 starting weights, with a batch size of 2, a step size of 15337 (no. images in training set divided by batch size), an input image size of 800×800 and a learning rate of 1×10^{-5} for 12 epochs. Anchor optimization for RetinaNet³⁴ was used to generate optimal anchors.

The anchor configurations control the sizes and scales of candidate bounding boxes, and may result in some objects being omitted from training in the event that there is no candidate with greater than 0.5 IoU. Due to the small sizes of some of the ships, the optimal scales were found to be much smaller than the default.

The anchor configuration for training RetinaNet on the SSDD was:

Sizes: 32, 64, 128, 256, 512
 Strides: 8, 16, 32, 64, 128
 Ratios: 0.440, 1.000, 2.274
 Scales: 0.488, 0.775, 1.221

NovaSAR Dataset

The NovaSAR dataset is made up of 35 multilook detected ground range acquisitions; 24 in Stripmap mode (6m resolution) and 11 in ScanSAR mode (8 at 20m and 3 at 30m resolution). In total, they contained 616 ships; 424 in Stripmap and 192 in ScanSAR. Two were acquired in VV polarisation, with the rest in HH. A 0.1% contrast stretch was applied to all of the images in the dataset to improve visibility.

Each full size acquisition in the NovaSAR dataset was first labelled manually using LabelImg³⁵. Coincident AIS data was used to verify the labelling was correct in two of the acquisitions, however this data was not available for the vast majority of the dataset. Whilst every effort was made to label all ships present in the images and avoid mistakes, there may have been a small number of ships that were omitted due to uncertainty or objects that closely resembled ships that were mistakenly labelled as such.

Two NovaSAR acquisitions, with ID 6102 (20m ScanSAR HH) and 8498 (6m Stripmap HH) were selected to form the test set for comparison with CFAR, which will be referred to as NovaSAR Test Set B. This was because the fixed threshold CFAR technique is applied on whole images, and an acquisition-level comparison is a better example of an operational use case for a ship detection technique. Acquisition 6102 contains 30 labelled ships, and 8498 contains 52 labelled ships. Together they account for 13.3% of the ships in the NovaSAR dataset.

Table 1: NovaSAR Test Set B acquisition properties.

ID	Mode	GRD (m)	Swath (km)	Pol	No. of looks
8498	Stripmap	6	20	HH	1 (range) 4 (azimuth)
6102	ScanSAR	20	~100	HH	2 (range) 2 (azimuth)

The remaining 33 acquisitions, containing 534 ships, were divided into tiles, since the resolution of the full-size images was too high to be used as input to a CNN without significant downscaling resulting in information loss.

The tiles were generated using a sliding window approach, with 128 pixels of vertical and horizontal overlap between tiles in order to ensure that any ships that would otherwise have been split between tiles by the edge of the window were fully captured in at least one of the tiles. This overlap has the effect of artificially inflating the number of ships in the dataset through duplication, and is similar to the approach taken in constructing the SSDD. The final tiles in each row and column contained an additional, variable amount of overlap with the previous tile to account for the fact that the tile sizes were not generally perfect factors of the full size image dimensions.

The tiles were only saved and incorporated into the dataset if the label files indicated that they contained ships. It was not seen as helpful to include a large number of negative examples, i.e. tiles that did not contain ships, since these may overwhelm the training dataset and drastically increase training times.

An example of a labelled NovaSAR image tile is shown in Figure 2.

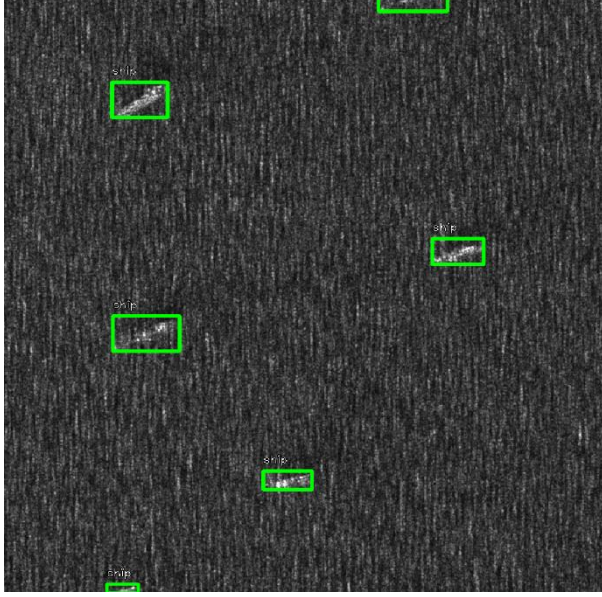


Figure 2: A labelled portion of a 6m resolution NovaSAR Stripmap mode Ground Range Detected (GRD) HH image. The bounding box coordinates reside in a separate annotation file and are displayed for demonstration purposes; they are not part of the image itself. Image Copyright SSTL.

This dataset of tiles was further split randomly into 70% training, 20% validation and 10% test. This test portion will be referred to as NovaSAR Test Set A.

The resolution of the NovaSAR images was generally higher than that of the acquisitions used to generate the SSDD, and therefore the apparent sizes of ships would have varied from the SSDD if the same 256×256 tile size was used, reducing the applicability of the SSDD learning to the NovaSAR dataset. RetinaNet was therefore tested directly on the NovaSAR dataset, using the weights generated by training on the SSDD, to determine the ideal tile sizes for both Stripmap and ScanSAR images. This step was necessary in order to ensure maximum transferability from the SSDD learning to a model trained on the NovaSAR dataset. If the NovaSAR dataset were sufficiently large, detection performance and speed may be improved by using a larger tile size.

The optimal square tile size (by F1-score) for the Stripmap images was found to be 480 pixels, as shown in Figure 3, while for ScanSAR 448 pixels was found to be optimal as shown in Figure 4. The results are dimensionless quantities with values between 0 and 1.

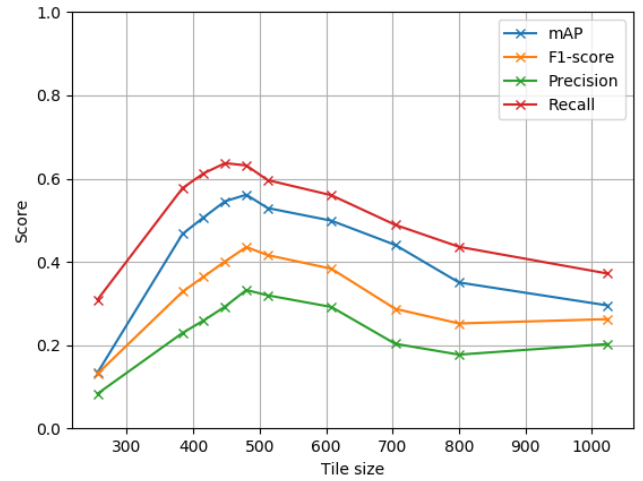


Figure 3: Performance metrics for the SSDD model for a range of tile sizes when applied to NovaSAR Stripmap images after 23 training batches.

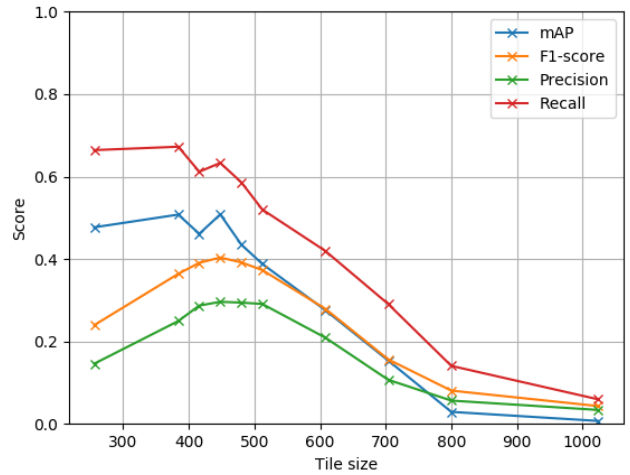


Figure 4: Performance metrics for the SSDD model for a range of tile sizes when applied to NovaSAR ScanSAR images after 23 training batches.

The NovaSAR dataset of image tiles, which excluded acquisitions 6102 and 8498, was therefore generated using these tile sizes, and is described in Table 2. The total number of ships in these tiles more than doubled in comparison to the true number of ships in the full size images, since even in the larger Stripmap tiles, the majority of each tile is made up of overlapping regions.

Table 2: NovaSAR image tile dataset.

	No. of tiles	No. of ships
Training	602	846
Validation	172	248
Test	87	127

Transfer Learning Approach

The model weights which gave the highest performance on the SSDD were used a starting point from which to train RetinaNet on the NovaSAR dataset. RetinaNet was trained until the validation mAP plateaued and the weights that gave the highest mAP were used for testing.

The anchor configuration for training RetinaNet on the NovaSAR dataset was:

Sizes: 32, 64, 128, 256, 512
 Strides: 8, 16, 32, 64, 128
 Ratios: 0.432, 1.00, 2.312
 Scales: 0.400, 0.504, 0.640

The optimal anchor scales for NovaSAR were found to be significantly smaller than the optimal SSDD anchor scales, since the ships in the NovaSAR images were generally smaller as a proportion of the image than the ships in the SSDD.

Prediction Combination

In order to achieve a final set of detections for an entire acquisition and compare these directly with the CFAR based method, the coordinates of the detections in each tile of Test Set B had to be translated back into the original image space by accounting for the coordinates of each tile. Additionally, duplicate ships may have been correctly detected in multiple tiles, resulting in several overlapping bounding boxes that have detected the same ships. This was accounted for by comparing overlap regions and discarding all but the highest confidence counterparts for those boxes that were predicted in multiple tiles, as illustrated by Figure 5. This preserved predictions that overlap within the same tile, as in the case of ships that are close together, as well as retaining predictions made in one tile but not in any others. This will have the effect of increasing the probability of both true and false positives, which in turn increases recall whilst lowering precision. For an operational use case, it is expected that recall is likely to be valued over precision, since the consequences of missing a detection are potentially greater than a false alarm being reported.

Testing

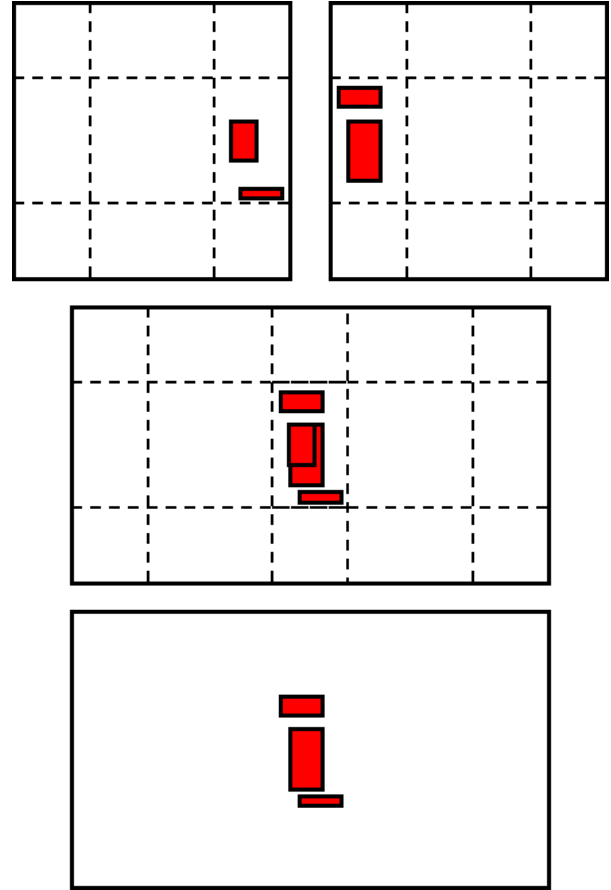


Figure 5: Bounding box prediction combination process, showing two tiles and their overlapping regions (dashed lines, not drawn to scale). Top: two side-by-side tiles in which different predictions (red boxes) have been made in the overlapping region. Middle: The predictions from both tiles are overlaid in the original coordinate space. Bottom: Duplicate predictions are discarded while all unique predictions are retained. This process is repeated for all overlapping regions.

The final results given for NovaSAR Test Set B were computed using the predictions resulting from this process, whereas the results for Test Set A were derived directly from the individual tiles. The two sets of results are not directly comparable since, for the comparison acquisitions in Test Set B, the entire image was divided into tiles and input to RetinaNet for prediction, whereas only tiles which were known to contain ships were included in the shuffled Test Set A. Additionally, some ships may have been divided into fragments at tile borders, causing the performance of the detector to be reduced if it failed to detect the fragments as ships in Test Set A. For Test Set B, the performance test would not have penalised this behaviour since the ship would have been fully present in an adjacent tile, and that prediction

would have been carried into the final set of predictions used to measure overall detection performance.

Test Set B gave a comparison with the CFAR method and demonstrated the process which would be applied to an image for which the presence and locations of ships was unknown. Test Set A allowed average performance across a range of acquisitions to be determined and compared with the detection performance for the C-band imagery in the SSDD.

RESULTS

SSDD - RetinaNet & YOLOv3

Both YOLOv3 and RetinaNet were tested on their respective 10% test portions of the SAR Ship Detection Dataset (SSDD), each using a confidence threshold of 0.25 to allow for a direct comparison. RetinaNet defaults to a 0.1 confidence threshold which does result in a higher mean average precision (mAP) of 95.4%, however false positives (FP) overwhelm the true positives (TP), making the detections considerably less useful. The model produced after 23 training batches was found to perform the best on the validation dataset, so this model was used for testing on the SSDD test set. An IoU threshold of 0.5 for a positive detection was required throughout testing for all neural networks and models.

It can be seen from the results in Table 3 that the mean average precision (mAP), F1-score and recall of RetinaNet are excellent, far exceeding the performance of YOLO. RetinaNet predicted a higher number of true positives, a lower number of false negatives and only marginally more false positives, resulting in increased precision in addition.

Table 3: Results of testing both CNN object detectors on the SAR Ship Detection Dataset (SSDD) at 0.25 confidence.

	YOLOv3	RetinaNet
mAP	0.774	0.928
F1-score	0.75	0.90
Precision	0.83	0.85
Recall	0.69	0.95
TP	4195	5709
FP	884	1004
FN	1870	324

Based on these results, RetinaNet was selected for testing on the NovaSAR dataset due to its high detection performance.

NovaSAR Test Set A - RetinaNet

Testing the trained SSDD models directly on the validation portion of the NovaSAR dataset revealed that the model produced after 10 batches performed best, likely because models produced beyond this point in training were overfitted to the SSDD images. Performance on the validation set during NovaSAR model training is shown in Figure 6. All metrics plateaued completely after ~75 batches, indicating that any learning to be gained from the relatively small dataset had been exhausted.

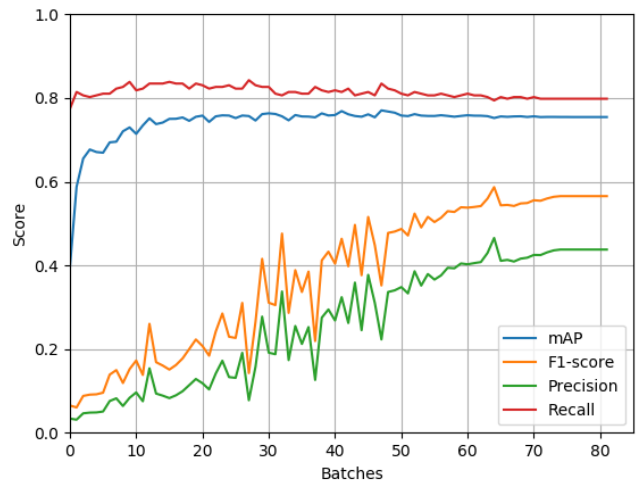


Figure 6: NovaSAR model validation performance over the course of training on the NovaSAR dataset.

The test results for confidence thresholds between 0.1 and 0.9 for the SSDD model and NovaSAR model are shown in Figure 7 and Figure 8 respectively.

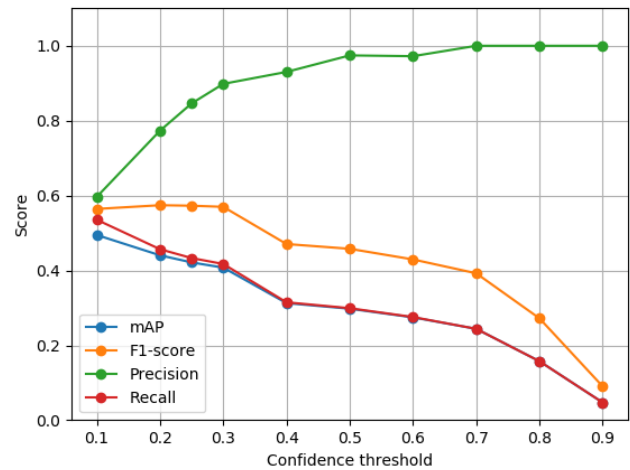


Figure 7: Performance of the SSDD model at a range of confidence thresholds for NovaSAR Test Set A.

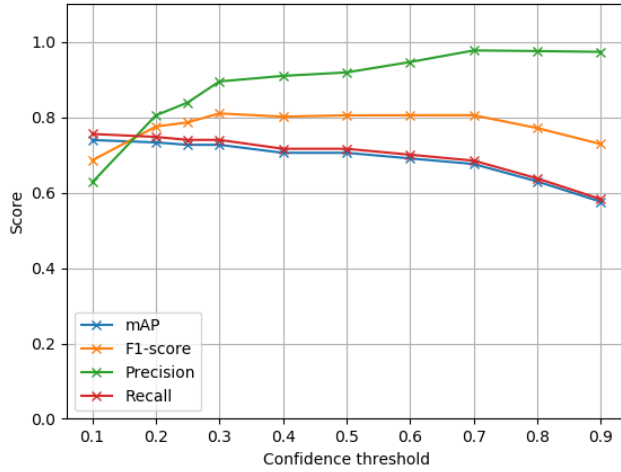


Figure 8: Performance of the NovaSAR model at a range of confidence thresholds for NovaSAR Test Set A.

Lower thresholds provided higher performance for the SSDD model since it had not been trained directly on NovaSAR data and therefore predictions were generally low confidence. The NovaSAR model was able to perform well at high thresholds since the predictions made were generally high confidence.

The NovaSAR transfer-learned model outperformed the SSDD model across all confidence thresholds, demonstrating higher mAP, F1-score and recall. The SSDD model appears to outperform the NovaSAR model at very high confidence thresholds in terms of precision, but this is only due to the extremely low recall at this level.

The results for the highest performing (by F1-score) confidence thresholds for each model on NovaSAR Test Set A are shown in Table 4.

Table 4: Performance of RetinaNet SSDD and NovaSAR models for NovaSAR Test Set A. The SSDD model was evaluated at a 0.2 confidence threshold, while the NovaSAR model was evaluated at a 0.3 confidence threshold.

	SSDD model	NovaSAR model
mAP	0.440	0.727
F1	0.574	0.810
Precision	0.773	0.895
Recall	0.457	0.740
TP	58	94
FP	17	11
FN	69	33

While the NovaSAR model clearly provided the best detection performance, the SSDD model was able to identify nearly half of the ships in the images with relatively few false positives, despite having been trained on SAR images of different band and resolution to the test set.

NovaSAR Test Set B - CFAR

The land masking for acquisition 8498 can be seen below. The high land mask performed better as it was able to identify ships in the littoral regions as shown in Figure 9.

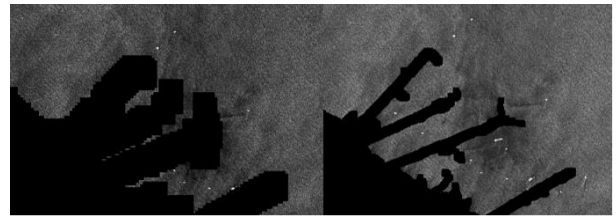


Figure 9: Land mask with cell averaged image (left) and with full resolution image (right). Image Copyright SSTL.

Once the land mask was applied, the Gamma, Lognormal and K distributions were plotted against pixel intensity in dB. This can be seen in Figure 10.

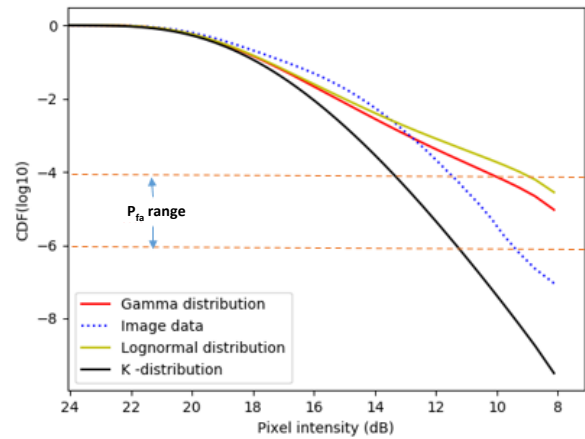


Figure 10: logCFD distributions plotted against pixel intensity in y (dB).

With a range of false alarms from 10^{-4} to 10^{-6} the Gamma distribution performed the best at a low P_{fa} but the k-distribution had the lowest error at a high P_{fa} , as can be seen in Table 5. The table also shows the CDF error associated with the P_{fa} for Both imaging modes. A minimum region size of 24 pixels was set in order to eliminate small, highly reflective surfaces at the bottom left of the image. The results can be seen in Figure 11.

Table 5: Sensitivity of false alarm values against number of detections in image and CDF error.

False alarm	No. of detections		CDF error (dB)	
	6102	8498	6102	8498
10^{-4}	10190	70	0.51	1.40
10^{-5}	9395	76	1.75	1.82
10^{-6}	8606	65	2.4	1.16



Figure 11: Vessels detected in Stripmap image. Image Copyright SSTL.

The bright regions of the detected vessels were in some cases captured as independent vessels. These centroids were clustered to produce a new location and corresponding bounding box for the resulting images. The results of this can be seen below in Figure 12.

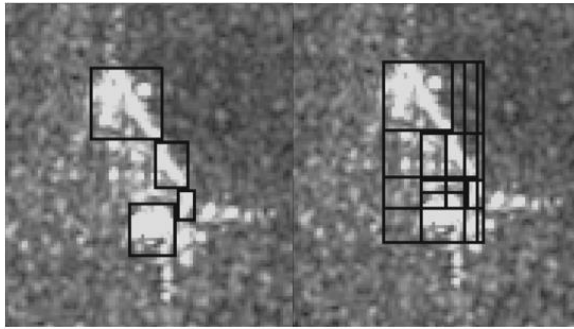


Figure 12: Bright regions of a ship as separate detections (left) and clustering of bounding boxes to find more accurate ship area (right). Image Copyright SSTL.

The ScanSAR image 6102 proved much more difficult to land mask as the image was originally saturated. The image was capped at a max intensity to stretch the dynamic range of the image. This improved the contrast in the image as can be seen in Figure 13 below. The reflectivity can be seen to vary in range, resulting in bright sea regions (bottom) and darker regions towards the top. The darker regions created softer edge gradients resulting in poorer land mask performance.

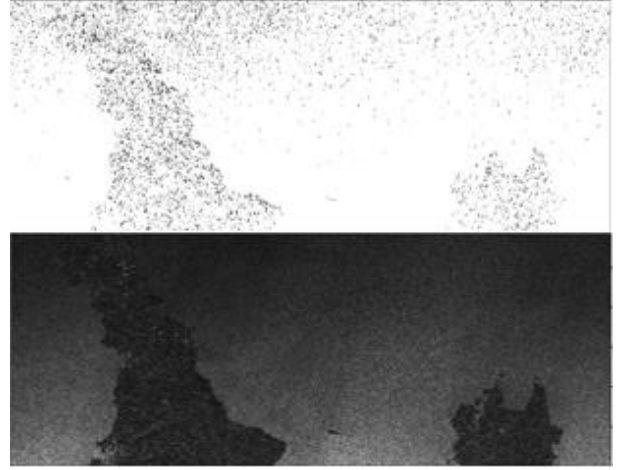


Figure 13: Visualisation of ScanSAR image before (top) and after distribution is stretched. Image Copyright SSTL.

The distributions were fitted to the land masked image as shown in Figure 14. The CDF divergence is more uniform due to the image being stretched across a smaller dynamic range. The K- distribution produced the lowest error.

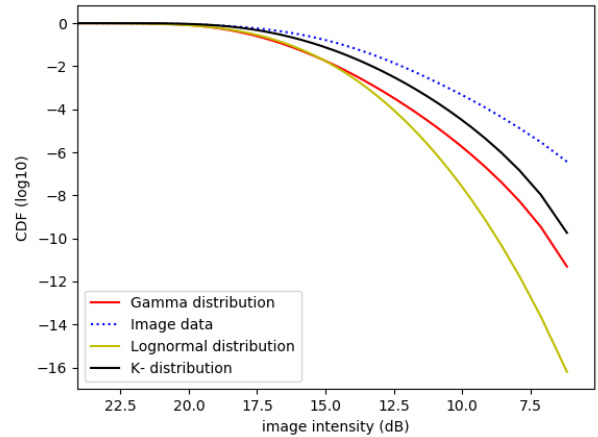


Figure 14: logCDF of distributions in ScanSAR image against pixel intensity (dB).

Due to the varying sea state in range a great deal of highly reflective surfaces in the bright regions were tagged as vessels, increasing the number of false positives dramatically. The detected image can be seen in Figure 15.

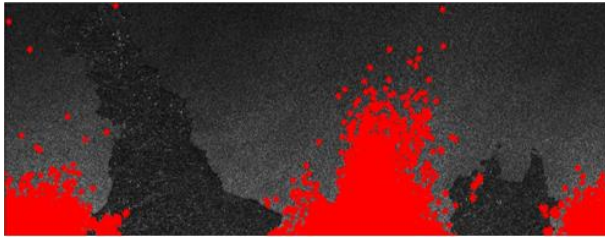


Figure 15: Ship detection performance in ScanSAR image. There are a large number of detections in the bright region of the image. Image Copyright SSTL.

NovaSAR Test Set B - RetinaNet

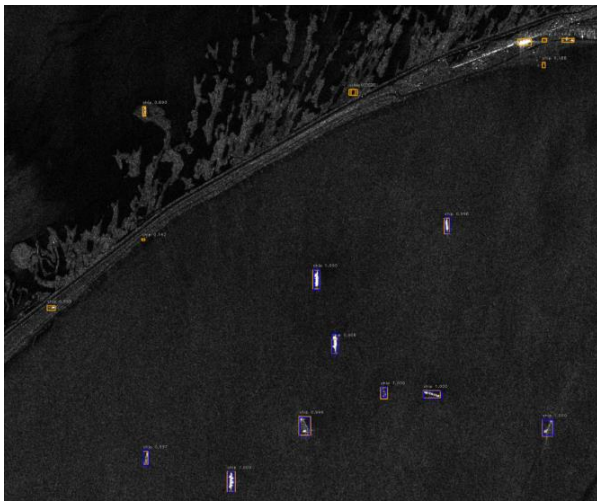


Figure 16: A region of the 8498 acquisition after detection by RetinaNet (NovaSAR model), showing detections (orange boxes) and labels (blue boxes). Detections are accompanied by a class and confidence label. Many of the detections are difficult to see due to the near-perfect IoU with the labels, however there are some false positives on the small strip of land which could not be masked out due to the NovaSAR image geolocation error. Image Copyright SSTL.

Figure 16 shows a portion of Stripmap image 8498, demonstrating good detection performance; IoU for the correctly detected ships is nearly 1.0, and all 9 ships which are clearly visible are correctly detected. There are, however, 8 false positives shown in the area of land that have high prediction confidence.

The majority of the false positives occurred over regions of land, as can be seen in Figure 17. Land masking using a shape file was applied to the detections, however due to the geolocation error in the NovaSAR images, this was offset and therefore unable to fully mask out the false positives in land regions. The time taken to perform detections could have been drastically reduced if the land mask had been applied prior to detection, however this method allowed for comparison and evaluation of the

need for application of a land mask using a CNN-based object detector.

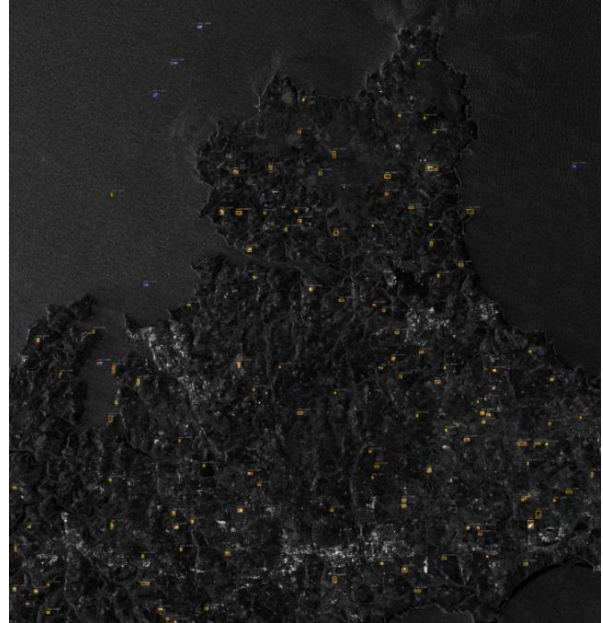


Figure 17: A region of the 6102 acquisition after detection by RetinaNet (SSDD model), showing detections (orange boxes) and labels (blue boxes). Some correct detections with high IoU can be seen in the top-left and top-right of the image, however it is obvious that an overwhelming number of false positives were produced over land in regions of bright backscatter. Image Copyright SSTL.

RetinaNet was also able to identify some ships by their wakes, which are clearly visible in Figure 18 even though the ships themselves are difficult to see. The bounding boxes for these ships, however, were erroneously predicted as being much too large, leading to these detections being counted as false positives since their intersection over union with the labels was lower than the required value of 0.5.

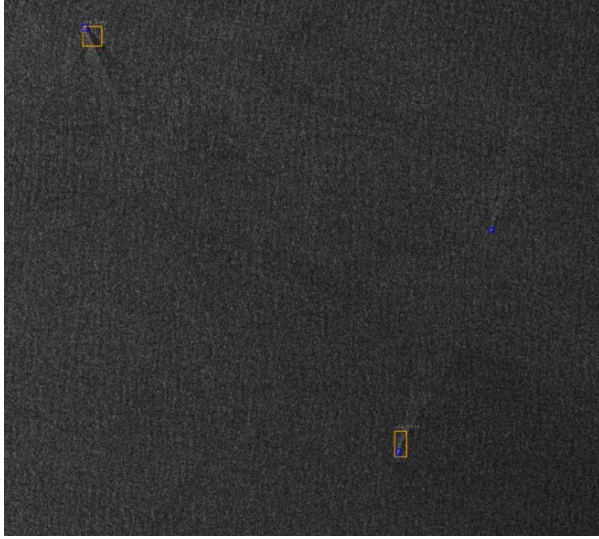


Figure 18: RetinaNet (NovaSAR model) detected (orange boxes) two ships by their wake in the 6102 ScanSAR image, though it predicted bounding boxes that were too large. A third, smaller ship (blue box, middle-right) was not detected. Image Copyright SSTL.

The performance of RetinaNet using both SSDD and NovaSAR trained models on NovaSAR Test Set B is shown in Figure 19. For both models, it can be seen that the best results were obtained with higher confidence thresholds than for Test Set A before a land mask was applied.

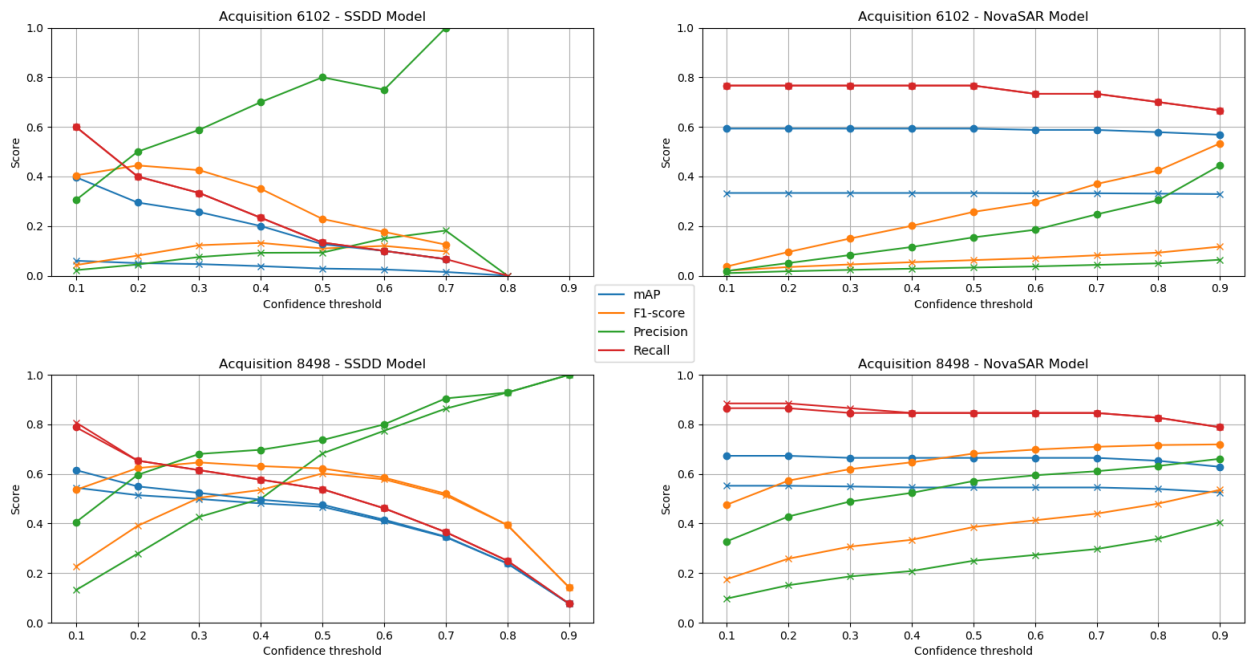


Figure 19: Ship detection performance by RetinaNet for NovaSAR Test Set B at a range of prediction confidence thresholds. Cross markers denote the results before land masking, while circular markers denote the results after land masking.

Lower thresholds resulted in an overwhelming number of false positives, leading to extremely low levels of precision.

The highest F1-scores for the SSDD model were at 0.4 confidence before land masking, and 0.2 confidence after land masking in the 6102 acquisition. The highest F1-scores in the 8498 acquisition were at 0.5 confidence before land masking and 0.3 confidence after land masking. For the NovaSAR model F1-scores were highest in both images, before and after land masking, at 0.9 confidence. Each of these thresholds was therefore applied to yield the results in the performance comparison with CFAR, in order to give a best-case scenario for each method.

CNN & CFAR Comparison

The performance comparison for all methods is shown in Figure 20 for acquisition 8498 and in Figure 21 for acquisition 6102. The bounding boxes produced by the CFAR method were compared to the labels at an IoU threshold of 0.5, allowing mean average precision, F1-score, precision and recall to be calculated as with the CNN-based method. Any duplicate boxes were counted as false positives as with the RetinaNet results. Metrics were not calculated for the CFAR results for the 6102 image due to the large number of false alarms produced; precision was effectively zero.

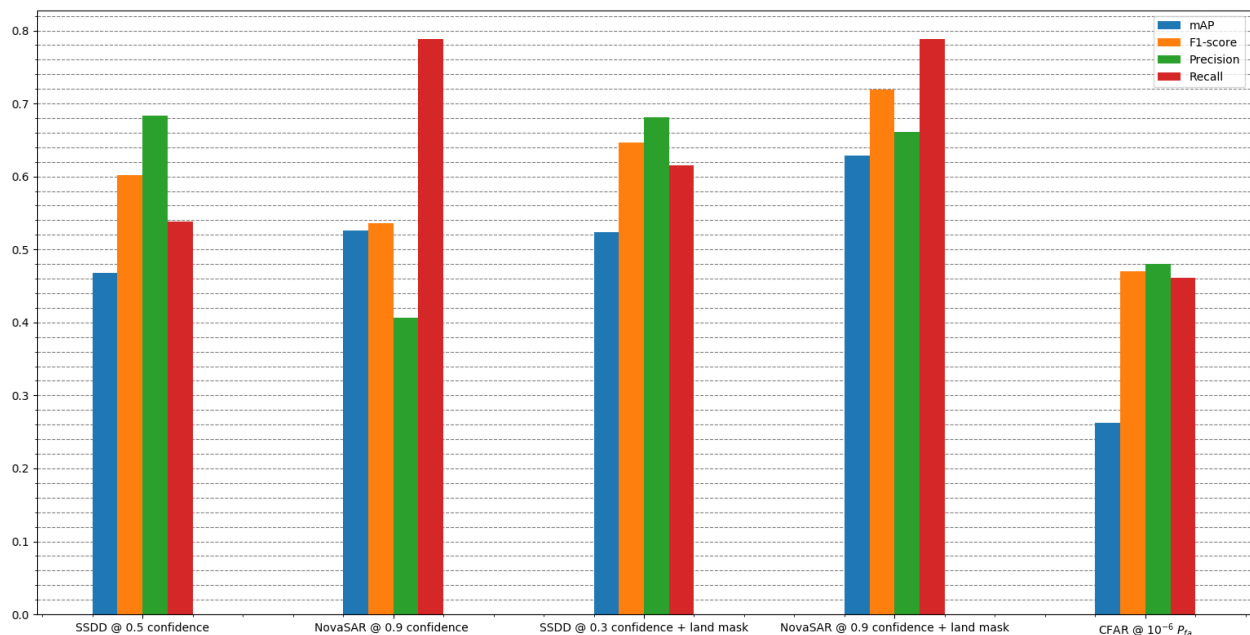


Figure 20: NovaSAR acquisition 8498 performance comparison between RetinaNet and CFAR based methods.

Both methods were able to identify ships in the 8498 image, with all RetinaNet models outperforming the CFAR method on nearly all metrics both with and without land masking. The NovaSAR RetinaNet model outperformed the SSDD model after land masking, however prior to land masking the NovaSAR model produced more false alarms resulting in reduced precision. Land masking was necessary for each of the RetinaNet models to increase precision in the 6102 image, where the NovaSAR model again outperformed

the SSDD model as expected, except in terms of precision. The NovaSAR model was capable of detecting more than 66% of the ships, while the SSDD model was able to detect 40% of the ships present with slightly fewer false positives. The CFAR detector produced more than 8000 false positives after applying a land mask. Before land masking, the NovaSAR model (at 0.9 confidence) produced 292 false positives, and the SSDD model (at 0.4 confidence) produced 69.

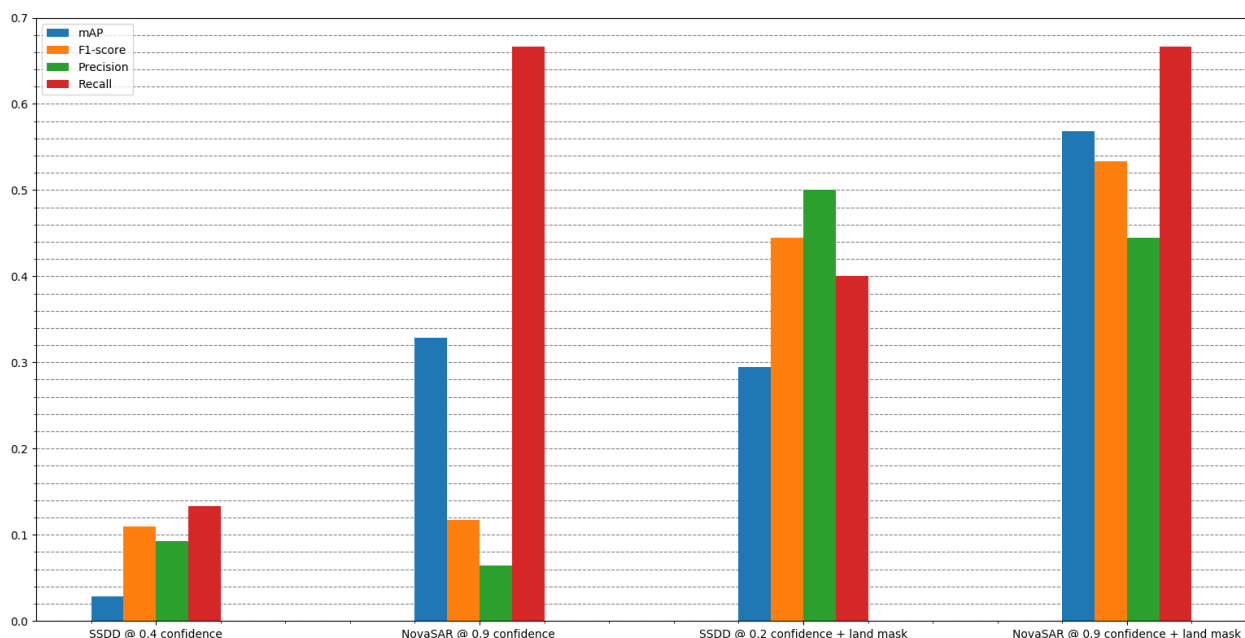


Figure 21: NovaSAR acquisition 6102 performance comparison between RetinaNet methods.

ANALYSIS

CFAR Results

The Stripmap image provided good results in the detection of vessels. The original image had many land features such as harbours and bridges connecting islands that were not masked at all with a land boundary shape file. The ScanSAR land mask was less accurate than the higher resolution Stripmap image due to the low intensity gradient between sea and land. However, it proved to be effective in masking out smaller land masses, such as islands. Setting a suitable region size above which any detected objects are considered land is a difficult and this can cause false alarms in regions with islands that may have a similar size and similar reflectivity to a large vessel. Awareness of the size of a vessel that the user may be interested in could greatly improve the distinction. The ScanSAR image also shows a multitude of false detections that encroach on land regions due to the edge boundary not being enclosed. With most existing systems a simple DEM or shape file could be applied, as the geolocation accuracy is good enough to mask land accurately. In the event of a GPS malfunction an effective land masking solution is critical to effective detection of offshore objects.

Each image showed a different outcome with an increase in P_{fa} . The Stripmap image demonstrated better performance with a Gamma distribution at a highest false alarm. At the lowest false alarm rate, the K distribution had a lower error, as the other two distributions started to diverge from the original. The ideal threshold that was derived from this error worked very well for the Stripmap image as the sea state variation was much less than that of the ScanSAR image; this is partly due to the larger area covered by ScanSAR exposing it to more range in sea surface roughness and reflectivity. To properly threshold the ScanSAR image, an adaptive threshold would need to be used. This could be achieved by taking the cell averaged mean scatter of the land masked image.

The overlap in bounding boxes shows a clear requirement for the tool to cluster the centroids of detected regions to eliminate duplications and to ensure a more accurate representation of vessel size. The duplication resulted in a perceived loss of performance of the detector as they were counted as false negatives although this is not reflective of its true performance.

Image Labelling

The labelling process imposes an artificial limit on the performance of the CNN-based object detectors since they will not have been trained to detect objects that a human cannot recognise. They may have otherwise been

capable of exceeding human detection performance if it were possible to label the dataset with perfect accuracy. They may also be more prone to making mistakes since they may have been erroneously trained to recognise objects that are not ships.

Labelling is improved with user domain experience and access to ground truth. It will therefore introduce bias into the system - smaller ships, those that are located near the coastline, those that may be miscategorised by a novice operator and those that appear in geographical areas where access to ground truth is limited are more likely to be mislabelled. It is possible that some of the detections that have been interpreted as false positives were in fact ships that were not labelled.

CNN Results

The majority of literature focuses solely on the mean average precision (mAP), which as shown in the results of this study does not fully describe the performance of a ship detection system. F1-score provides a more useful measurement of the utility of each ship detection technique, since both high recall and high precision are important in this domain.

RetinaNet was found to outperform YOLOv3 for ship detection, likely due to the inclusion of the aforementioned feature pyramid networks and focal loss. This was consistent with findings in literature^{17, 30} which demonstrated the highest levels of performance with RetinaNet in comparison to a variety of other networks. The SSDD results were similar to those found in previous studies, with mAP of 95.4% at 0.1 confidence. This was higher than the 91.4%²³ achieved by the SSDD authors, but slightly lower mAP compared to the 96.6%³⁰ using a dataset solely made up of Gaofen-3 images.

The 6102 image was more difficult for a number of reasons - the resolution was lower at 20m compared to 6m for the Stripmap image and there was a much greater region of land contributing to false alarms. Additionally, the majority of the NovaSAR training dataset was made up of 6m Stripmap images, which meant that the NovaSAR model had only received limited training for this type of image.

Applying a land mask to the RetinaNet detections had very little impact on recall, which should be expected since true positives ought to lie in the sea. However a small number of correct detections were masked out, either due to the image geolocation error or the extent of the land mask itself. Precision, on the other hand, was greatly improved by applying a land mask due to the reduction in false positives over land, despite the fact that

land masking is not typically utilised for CNN-based ship detection.

The reason for the slightly lower precision and combined need for higher confidence thresholds with the NovaSAR model is unknown. It may have been due to overtraining, or simply a limitation of the small dataset.

As the resolution of the NovaSAR images was generally comparable to the Gaofen-3 and Sentinel-1 images in the SSDD, with a large enough training dataset it would be expected that similar mAP could be reached. However, as shown in Table 3, the mAP plateaued more than 20% below the peak for the SSDD. This supports the hypothesis that the small number of images available in the NovaSAR training dataset was a limiting factor in its performance.

CONCLUSIONS

The results in this study should be treated with caution - it cannot be conclusively determined from a dataset of this size that any one method has better performance in all circumstances.

It is likely that, particularly with such a small dataset, certain acquisitions will perform better than others. For example, the training set may be predominantly made up of calm water conditions, or water of a certain depth, and the two acquisitions chosen in NovaSAR Test Set B may not reflect this. Therefore the performance results shown are only indicative and may have been improved with additional testing.

Conversely, the NovaSAR dataset contains only a small number of unique locations, some of which are featured in multiple acquisitions. It may therefore also be possible that the CNN results for both NovaSAR Test Sets would not have been as favourable if the locations featured had not previously been imaged and included in the training dataset.

However, the results do indicate that a CNN object detector can outperform a CFAR methodology for ship detection in S-band SAR imagery, even in the absence of native training data. This is an important finding as it could potentially allow a new satellite such as NovaSAR to incorporate a ship detection capability either on the ground or on-board, which would provide utility from the start of operations. This would avoid the need to amass an extensive training dataset - wasting a significant portion of the satellite's operational lifetime - before automated image exploitation could become fully operational.

Additionally, performance was found to improve with the application of transfer learning to a small native

dataset. Though performance would almost certainly have been further improved with a larger dataset, these two findings combined indicate that a satellite mission like NovaSAR could initially use an open-source training dataset, and gradually train on a native dataset as images are captured, improving detection performance throughout its operational lifetime.

Whilst the CNN and CFAR methods both demonstrated utility to an image analyst, neither proved that detection performance could match or exceed human levels, and therefore would not yet be suitable as a complete replacement for defence and security purposes where high levels of accuracy are required.

However, with enough tolerance for false alarms and missed detections, it is completely possible to automate the process of detecting ships in SAR images. The RetinaNet methodology is fully automated and produces detections in approximately 3 minutes for a standard Stripmap image and approximately 10 minutes for a standard ScanSAR image on an Nvidia Quadro P3200 using the tile sizes specified. This time could be reduced significantly by dividing the images into larger tiles and could theoretically be incorporated onto a satellite system for tipping and cueing of an accompanying optical Earth Observation satellite.

FUTURE WORK

CFAR Ship Detection

Through the development of the CFAR ship detector, there were several parameters that required adjustment to optimise the performance on individual images. A sensitivity analysis could be run as an independent study to measure the effects and adaptively set the following parameters:

- Upper and lower hysteresis threshold for Canny edge detection
- Standard deviation on Gaussian filter
- Cell averaging scale factor. This is dependent on the speed requirements and processing capability of the system. It also creates an offset in the land mask when scaling up to original size.
- Minimum and maximum region detection size to remove smaller land masses and sea surface specular reflections.
- Sample size for distribution fitting.
- False alarm value.

Setting of the above parameters can be made easier with a toggle interface. Since CFAR in its nature is an optimisation, it is difficult for the process to be truly automated. These parameters could be adaptively set by measuring parameters such as edge gradient for the

hysteresis threshold and mean scatter for the localised adaptive threshold, however this would incur a heavy penalty on processing speed.

CNN Ship Detection

YOLOv4³⁶ was released after testing with YOLOv3 was conducted on the SSDD. YOLOv4 promises improved performance over YOLOv3 and may rival RetinaNet, however it is unknown how applicable these improvements are likely to be to SAR ship detection.

Application of traditional image augmentation techniques including angle and exposure were investigated and found not to be applicable to SAR images. Future investigation could involve SAR-specific image augmentation techniques e.g. speckle filtering, multilooking and variation in ground range projection. Intensity plots and varying contrast thresholds may be found to improve detection performance. Simulated data could also be produced to determine whether its inclusion in the dataset improves model performance or generalisability.

In this paper, transfer learning was used to apply the learning from the large SSDD to the small NovaSAR dataset. Future work could investigate the benefit of training on a large, NovaSAR-exclusive dataset once enough maritime acquisitions from the satellite have been collected.

Negative examples, i.e. image tiles that did not contain ships, were not included in the NovaSAR dataset for training. Including these in future may reduce the number of false positives, especially in images containing large regions of land. If this were successful, masking out areas of land in the images may not be necessary to derive useful detections.

One tri-polar image containing ships was available for the NovaSAR dataset, though only the HH polarisation was used. Preliminary results based on this image indicate that ships appeared more clearly in HV polarisation than in HH or VV. It would be useful to acquire further HV polarisation images to investigate in more detail which one yields the best results. Additionally, while all of the work in this study was carried out on single channel images, tri-polar images could be combined into three channels for training and detection like standard RGB images, and the different modes of backscatter in each polarisation channel may improve detection performance.

Assisted labelling is suggested as a method of making the labelling process faster and more efficient. The AIS sensor on-board NovaSAR could be used to aid with this, by confirming detections made by a neural network, and

saving them as labels, if there is a nearby AIS signal that was transmitted close to the time of imaging. Additionally, AIS messages contain detailed information about the ships that they were transmitted by. While AIS is not completely reliable, classification into different types of vessels could be achieved even if the difference between e.g. a cargo ship and a tanker is not visible to the human eye.

SAR images are susceptible to interference by active deception jamming techniques³⁷. Additionally, neural networks have been shown³⁸ to be deceived by adversarial attacks with a change of a single pixel alone. Therefore, if a CNN ship detector were to be used for defence and security purposes, it would need to be robust against both of these types of potential attacks.

ACKNOWLEDGEMENTS

The authors would like to thank Surrey Satellite Technology Ltd. (SSTL) for their help and technical guidance while producing this paper.

© All images copyright Surrey Satellite Technology Ltd.

© Crown copyright (2020), Dstl. This material is licensed under the terms of the Open Government Licence except where otherwise stated. To view this licence, visit <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3> or write to the Information Policy Team, The National Archives, Kew, London TW9 4DU, or email: psi@nationalarchives.gsi.gov.uk

REFERENCES

1. Agnew, D.J., et al., *Estimating the Worldwide Extent of Illegal Fishing*. PLOS ONE, 2009. **4**(2): p. 1-8.
2. Skauen, A., *Quantifying the Tracking Capability of Space-Based AIS Systems*. Advances in Space Research, 2015. **57**.
3. Xing, X.W., et al. *A fast algorithm based on two-stage CFAR for detecting ships in SAR images*. in *2009 2nd Asian-Pacific Conference on Synthetic Aperture Radar*. 2009.
4. Iervolino, P. and R. Guida, *A Novel Ship Detector Based on the Generalized-Likelihood Ratio Test for SAR Imagery*. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017. **PP**: p. 1-15.
5. Ji, K., *A landmasking algorithm for ship detection in SAR images*, in *IGARSS*. 2016, National University of Defense Technology. p. 3.
6. Martín-de-Nicolás, J., *Segmentation techniques for land mask estimation in SAR imagery*, in *Fifth International Conference on Computational Intelligence*. 2013, Communication Systems and Networks.
7. Canny, J., *A Computational Approach to Edge Detection*. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, November 1986. **VOL. PAMI-8**(NO. 6).
8. Watts, S., *Sea clutter presentation*. February 2020, Simon watts: SW research consultancy.
9. Ward, K., R.J.A. Tough, and S. Watts, *Sea Clutter: Scattering, the K Distribution and Radar Performance*. 2007: The Institution of Engineering Technology. p. 233-234.
10. Sun, J., *The Dependence of Sea SAR Image Distribution Parameters on Surface Wave Characteristics*. Remote Sens. 2018(1843).
11. Angelliaume, S., *Modeling the Amplitude Distribution of Radar Sea Clutter*. Remote Sensing, 2019. **11**: p. 319.
12. Redding. *k-distribution*. 2020 [cited April 2020; 3 parameter solution to the K-distribution].
13. Watts, S., *Detection in sea clutter* February 2020: SW research consultancy.
14. Kang, M., et al. *A modified faster R-CNN based on CFAR algorithm for SAR ship detection*. in *2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP)*. 2017.
15. Bentes, C., D. Velotto, and B. Tings, *Ship Classification in TerraSAR-X Images With Convolutional Neural Networks*. IEEE Journal of Oceanic Engineering, 2018. **43**(1): p. 258-266.
16. Lin, H., S. Song, and J. Yang, *Ship Classification Based on MSHOG Feature and Task-Driven Dictionary Learning with Structured Incoherent Constraints in SAR Images*. Remote Sensing, 2018. **10**: p. 190.
17. Ma, M., et al., *Ship classification and detection based on CNN using GF-3 SAR images*. Remote Sensing, 2018. **10**: p. 2043.
18. Lin, Z., et al., *Squeeze and Excitation Rank Faster R-CNN for Ship Detection in SAR Images*. IEEE Geoscience and Remote Sensing Letters, 2019. **16**(5): p. 751-755.
19. Zhang, S., et al., *R-CNN-Based Ship Detection from High Resolution Remote Sensing Imagery*. Remote Sensing, 2019. **11**: p. 631.
20. Girshick, R., *Fast R-CNN*. 2015.
21. Ren, S., et al., *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2015.
22. Wang, Y., et al. *Combining Single Shot Multibox Detector with transfer learning for ship detection using Chinese Gaofen-3 images*. in *2017 Progress in Electromagnetics Research Symposium - Fall (PIERS - FALL)*. 2017.
23. Wang, Y., et al., *A SAR Dataset of Ship Detection for Deep Learning under Complex Backgrounds*. Remote Sensing, 2019. **11**(7): p. 765.
24. Liu, W., et al., *SSD: Single Shot MultiBox Detector*. Lecture Notes in Computer Science, 2016: p. 21-37.
25. Chang, Y.-L., et al., *Ship detection based on YOLOv2 for SAR imagery*. Remote Sensing, 2019. **11**: p. 786.
26. Redmon, J. and A. Farhadi, *YOLO9000: Better, Faster, Stronger*. 2016.
27. Redmon, J. and A. Farhadi, *YOLOv3: An Incremental Improvement*. arXiv, 2018.
28. Lin, T.-Y., et al., *Focal Loss for Dense Object Detection*. 2017.
29. Su, H., et al. *Ship Detection Based on RetinaNet-Plus for High-Resolution SAR*.

- Imagery. in *2019 6th Asia-Pacific Conference on Synthetic Aperture Radar (APSAR)*. 2019.
30. Wang, Y., et al., *Automatic Ship Detection Based on RetinaNet Using Multi-Resolution Gaofen-3 Imagery*. *Remote Sensing*, 2019. **11**: p. 531.
 31. Lin, T.-Y., et al., *Feature Pyramid Networks for Object Detection*. 2016.
 32. Bochkovskiy, A. *AlexeyAB/darknet*. 2020 29/04/2020]; Available from: <https://github.com/AlexeyAB/darknet>.
 33. Gaiser, H., et al. *fizyr/keras-retinanet*. 2020 29/04/2020]; Available from: <https://github.com/fizyr/keras-retinanet>.
 34. Zlocha, M.a.D., Qi and Glocker, Ben, *Improving RetinaNet for CT Lesion Detection with Dense Masks from Weak RECIST Labels*. arXiv preprint arXiv:1906.02283, 2019.
 35. tzutalin. *LabelImg*. 2020 29/04/2020]; Available from: <https://github.com/tzutalin/labelImg>.
 36. Bochkovskiy, A., C.-Y. Wang, and H.-Y.M. Liao, *YOLOv4: Optimal Speed and Accuracy of Object Detection*. 2020.
 37. Yuang, G., S.-b. Li, and X.-h. Cao, *SAR counter deception jamming based on radiometric calibration*, in *IET International Radar Conference*. 2009.
 38. Su, J., D.V. Vargas, and K. Sakurai, *One pixel attack for fooling deep neural networks*. *IEEE Transactions on Evolutionary Computation*, 2019.