Improve Illegal Ship Detection Using Pixel-Wise Masks

Phillip Hale, Microsoft, Southern Methodist University

***Abstract*—This report describes the techniques and experiments for improving automatic ship detection from synthetic aperture radar (SAR) satellite imagery as a participant in the xView3 Dark Vessels Challenge 2021. The xView3 Challenge provides a large multi-dimensional dataset of SAR satellite views to benchmark new approaches to automatically detect illegal fishing activities at a global scale. Computer vision methods and Azure Machine Learning services are utilized in this challenge aiming to advance research contributions in extracting accurate ships mask and dimensions to enable performance improvements in ship detection, ship classification and estimating the length of detected ships. The initial technique has been tested and evaluated on the xView3 Challenge public dataset for benchmarking the performance of the trained models were the proposed method ranked 45 on the leaderboard. While areas of improvement are reflected, the detector and classifier do not outperform the xView3 reference model for this challenge however the proposed method for estimating the length of ships provided positive results. Visual comparisons of the proposed method for delineating the vessel outline in SAR images using mask segmentation indicated more better ground truths from those provided by experts using manual analysis, however manual expert review still may be needed for verifying the classification of ships. Source code and additional analysis is available at** https://github.com/naivelogic/xview3_ship_detection.

***Index Terms*—computer vision, azure machine learning, satellite remote sense imagery, xView3 Dark Vessels Challenge**

I. Introduction

The xView3 Dark Vessels Challenge was announced in August 2021 with the goal of benchmarking challenge participant solutions to detect dark vessel ships using computer vision and global Synthetic Aperture Radar (SAR) satellite imagery. The problem statement for the competition is to address Illegal, unreported, and unregulated (IUU) fishing activities that has negative effect to the stability of human food supply, marine ecosystem health and geopolitical systems [1]. This challenge seeks to provide actionable data to counter economic disparities to developing countries where sustainable fisheries in national waters are impacted by IUU fishing activities and connected with trans-national crimes. The outcome from participating in the xView3 Challenge is to continue research advancements in applying computer vision techniques to aid Global Fishing Watch (and the Defense Innovation Unit (DIU) with tools to discover IUU fishing activities in scalable fashion.

Remote sensing using satellite and aerial are a topic of much research as the field is harnessing the power of computer vision to automatically measure, track and process signals of objects on earth transmitted from satellites. Aerial imagery like SAR continue to face challenges for tools to utilize this information to identify small objects on Earth, appropriately classify them as image resolution, sensor angle, wind and wave conditions and lack of ground-truth data continue to be advanced in research [1]. To effectively monitor ships in multi-sensor SAR images advanced machine learning techniques have become commonplace [2].

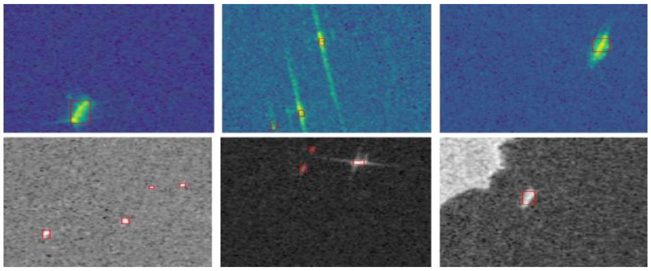

Figure 1 Samples from xView3 vh channel chips

Advances is computer vision and cloud computing has significantly accelerated the ability to process image datasets at scale at low cost and on-demand infrastructure buildout. Microsoft Azure platform offers cloud-based services that enable training computer vision models utilizing low-cost, general purpose cloud computational devices viz. virtual machines and storage devices viz. Azure Blob storage. The many of the traditional burden on hardware and software requirements the volume of images and the state-of-the-art networks dependency is transferred to Azure service offering that enable an on-demand research experience. This is needed when advancing research into remote sensing area due to the significant size of aerial images and SAR context for effective solutions.

The main contribution in this report is enhancing the SAR training data labels by extracting accurate ship pixel-wise mask to create segmentation labels to improve the accuracy of ground truth ships. The pixel-wise mask is then implemented in the deep neural network (DNN) utilizing Yolact segmentation algorithms [3] utilizing in Azure ML services to train on a cluster of GPUs. The goal of these proposed enhancements is to outperform the xView3 reference model by detecting and classifying and estimating the length of ships in the xView3 public test set. The key idea is to improve upon the naïve data processing techniques that was provided in the original challenge helper scripts to provide more accurate bounding box annotation as well as providing mask that would enable additional training methods such as segmentation computer vision models to be utilized for this challenge. The hypothesis of improving the ground truth annotations provides could increase the quality of the features during model training that would positively benefit the model's ability to predict dark vessels which produces a higher leaderboard score, indicating the proposed method is capable of generalizing performance in new environments and/or adapting to difficult scenes.

The remainder of the report is organized as follows. To start an overview of the xView3 Challenge dataset and the proposed approach for utilizing synthetic aperture radar (SAR) satellite imagery. Following that, training approach and experiment results are described and analyzed. Finally, the report highlights key discussion points on improvement areas and an overall conclusion.

II. xView3 Challenge and Dataset Enhancements

A. *xView3 Challenge Dataset*

The xView3 challenge presents participants an opportunity to be a part of the solution by building automatic detection and classification tools that counters illegal fishing activities. DIU, the host of the xView3 challenge, has provided a public dataset that include large high and medium resolution SAR satellite imagery along with contextual data to benchmark top methods for detecting and classifying objects. The provided dataset is based on pre-processed Sentinel-1 SAR images with ship detections ground truth obtained from GPS coordinates broadcasted by the ship. To our knowledge this is the largest satellite-based SAR dataset publicly available for identifying and characterizing vessels. For the xView3 Challenge, the objective can be broken into three parts: (1) identify the maritime objects in each scene, (2) for each detected object classify the ship as a ship or non-ship, additionally for each classified ship another binary classification is required to determine if it is a fishing ship or non-fishing ship. Lastly, (3) for each detected object estimate the length of the object.

The provided xView3 Challenge dataset images were broken into a training, validation, public leaderboard for benchmarking and a private set used to verify the reperformance of the leaderboard requiring the provided solution to be implemented in a container that is constrained to a set limit of time and memory. In total, the challenge dataset contains over 1,000 SAR Sentinel-1 scenes where each scene contains five specific SAR files that are in a GeoTIFF format each around 1GB in size. As detailed in following sections, there were unexpected project cost when storing and processing these raw SAR image files. The SAR scene files provides consist five image files that make up the entire scene, however, each image file in the scene provide different satellite imagery information that can be utilized as a multi-dimensional input image for training a model to account for feature such as wind or land segmentation. For our approach two of the five SAR scene images were primary use in creating the training dataset. Two different polarization band-SAR image referred to as VH and VV that were primary utilized for the ship detection. While VH (vertical-horizontal) polarization is usually better suited for ship detection task as it shows better separation between vessels and the sea clutter, the VV (vertical-vertical) polarization is better at identifying sea surface features (e.g., waves, wind, oil and sediments), which are often useful in the context of ship detection and characterization [1].

In addition to the two forementioned polarization images, xView3 also included three other large files to incorporate additional SAR satellite related features including bathymetry, wind speed and direction, wind quality, and land/ice masks. While the cost was incurred utilizing these ancillary image files early in the challenge execution, the approach only utilized the VH polarized as the initial attempt in solving the challenge.

B. *Data Pipeline Proposed Enhancements*

The xView3 dataset collected from the Sentinel-1 SAR images and processed for the challenge participants contain around 1,000 scenes at a 10-meter resolution, including both VH and VV polarization images that have an image size around 30000 x 30000 pixels. After downloading the scenes to the Azure Blob container, used for low-cost cloud storage, the next stage is converting the raw SAR images into a dataset format to train a DNN known as image processing. The main objective in image processing is to create a usable dataset from the SAR images to train a DNN to detect the maritime objects, classify them appropriately and estimate the length of the ships.

Image Cutting: All images are cropped in the patches of 200 x 200 pixels with a 20% overlap. The overlap helps to ensure the training dataset does not have objects from being truncated in the batch where an object part is distributed in multiple images and therefore reduces accuracy ground truth in the dataset. The xView3 training dataset utilized for training utilized 44,383 SAR chipped images with size of each image at 200 x 200. xView3 dataset also included labels for the training and validation dataset that gives information about the ship length, the scene id, the scene pixel row and column containing the detected object, binary value for if the detection is a ship and if the detection is a fishing ship. The detect scene pixel row and column during the image cutting process were re-computed to the new image and saved the pixel row and column coordinates for each detection in the new annotation file that would be used for training.

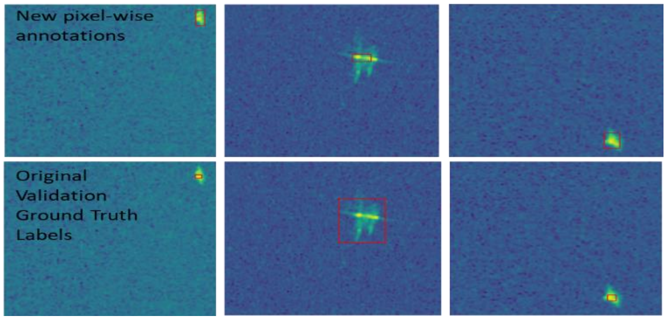
New pixel-wise annotations
Original Validation Ground Truth Labels

Figure 2 Dataset Label Enhancements from fixed bound box size to pixel-wise segmentation mask

Fixed Bounding Box Approach: The image annotation process is shown in Figure 2. The annotation technique used to improve the original ground truth is discussed next to enhancing the annotations and creating segmentation ground truth to train a DNN for image segmentation. After visual inspection of the new annotations, it was identified that modifications were needed to accurately train the DNN that the dataset labels did not include the height (h) and width (w) of the ship ground truth which are needed to create a region prediction (e.g., a bounding box) that are standard labels for state-of-the-art DNNs. By not having height and width ground truths bounding boxes based on the provide labels would be inaccurate as fixed sizes based on the center of the bounding box would have to be used. This fixed region bounding box was included in one of the earlier experiments and determined this method was not an effective solution for this challenge as one of the primary goals is to estimate the length of detected ships. Therefore, we looked at a new approach to apply enhancements to this annotation process by creating ground truth labels based on evaluating the ship pixel-wise segmentation mask to compute more accurate bounding boxes.

Pixel-wise Mask Segmentation: Figure 1 illustrates how bright the ship pixels are represented in the SAR images where the darker background indicates the ocean and water features and the ship (as well as other non-ship) contrasted clearly. In effort to improve the provided annotations in the xView3 dataset were only the centroid of the ship is provided, enhancements in more accurate bounding boxes and ship mask can be obtain that would lead to more accurate training data and less false positives. The Constant False Alarm Rate (CFAR) [3] approach and Otsu's algorithm is utilized in the preprocessing stage of this challenge in effort to provide a training dataset that provides additional features such as segmentation to improve ship detection, ship length estimation and classification of detected ships.

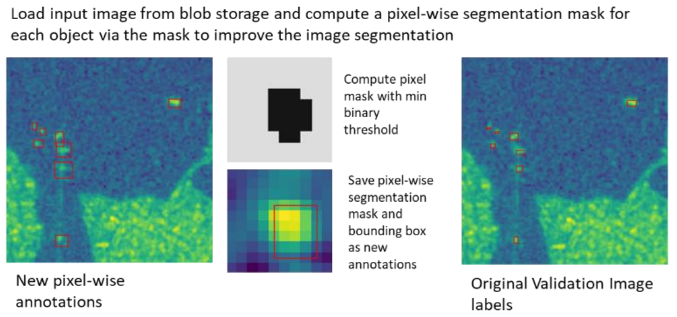
Load input image from blob storage and compute a pixel-wise segmentation mask for each object via the mask to improve the image segmentation
Compute pixel mask with min binary threshold
Save pixel-wise segmentation mask and bounding box as new annotations
New pixel-wise annotations
Original Validation Image labels

Figure 4 Data loader improved annotation ground truth SAR

The process implemented for enhancing the dataset was later identified to be like a CFAR approach described in the remote sensing letter paper [4]. One key difference in our proposed approach is the creation of segmentation annotations for a training dataset. This process is visualized in Figure 3 where a SAR image of a vessel is cropped using the provided centroid coordinate and an Otsu threshold is computed where all pixels are measured as histogram with the goal of clustering the brightest pixels representing the ship as one value and setting the ocean clutter to another value. The result is a threshold value that is set that targets the ship signatures while reducing the likelihood that pixels representing the ocean are constrained. The threshold is applied to create a binary mask of only the ships that is in filtered so individual ship rectangular shape, segmentation and ship length can be accurately estimated relative to the original image resolution and dimension.

III. Methodology and Approach

In this section the approach used in participating in the xView3 challenge is detailed including the azure cloud resources used in the training process, the model selection and hyperparameters, as well analysis in the results of the experiments and methods used in the submission. To evaluate the performance of the ship detector, a set of five scenes from the xView3 validation set were held out of the training dataset used to evaluate the performance of the model during training to for selecting the best model to use on the leaderboard test dataset for the challenge benchmarking. Model performance evaluation was implemented as required by the xView3 challenge, but at training time, the holdout validation set was only evaluated using the Average Precision (AP) metric. The AP metric [5] averaged over Intersection-over-Union (IoU) thresholds using the instance masks and tested different scores to rank predictions. After training was completed, each of the models were then evaluated on the aggregate evaluation metrics required in the submission on the same validation set.

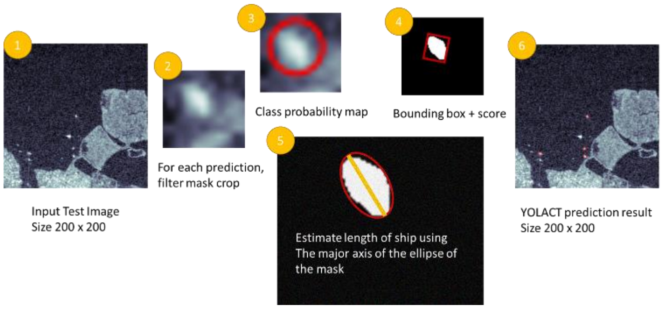

Figure 3 Estimating Ship Length using mask and ellipse

xView3 Challenge had several submission requirements in solving the maritime object detection and classification in global SAR scenes. Figure 4 provides a visualization for how our proposal approached the solution on both the validation set and the public leaderboard test set. Given a scene as input (1) each of the image splits were batched into the model for an input size of 200 x 200 pixel. For each of the image scenes, the Yolact model cropped each detected maritime object (2) within the scene and determined the (3) class probability map, (4) computed the bounding box and prediction confidence score, and (5) estimated the length of the ship using the major axis of the ellipse of the mask. Finally, (6) if the confidence score was greater than 0.4, the Yolact predictions would store the predictions in a json file that would map back the local image pixel coordinates and length estimations to the original scene dimensions to be submitted for the challenge. The pixel coordinates are defined as the vertical axis and the horizontal x axis for each detection.

The next section described the training process and network implementation for this challenge using Azure cloud resources.

A. *Microsoft Azure Cloud Resources*

Microsoft's Azure Machine Learning (AzureML) platform was utilized as the cloud computing infrastructure to develop and test our solutions in participation in this challenge. The cloud architecture for this vision system utilizing virtual machines for code development, AzureML API calls for configuring training deployments and Azure Blob storage containers. Together these services provided benefits in serverless computing to minimize code development and maximizing scalability through training on GPU clusters. We train all the detections methods on AzureML where a cluster of 2 NVIDA Telsa K80 GPUs were provisioned as the processing unit on an Ubuntu 18.04 machine with CUDA11.0. The benefits of AzureML were realized as we were able to reserve 32 nodes that could be allocated across the cluster for multiple training experiments running in parallel. All that was required once the AzureML GPU clusters were reserved was the deployment of the model container environment that would host and run the PyTorch python training code.

For our submissions to be qualify for the challenge prize additional challenge constraints had to be addressed. xView3 Challenge participates are required to verify their submission by also submitting the model and the inference code in a container to the xView3 submission portal. The submission verification constraints required the created container to process each scene for a maximum of 15 minutes utilizing only 60GB of RAM on a V100 GPU per scene. Additionally, containers do not have network access, so the submitted container must be packaged with everything required to run independently without dynamic downloads [1]. To address these constrains, a separate VM was provisioned with V100 GPUs however, after the first day, the compute cost outweighed the total cost allocated for the project was an alternative method was proposed. In attempts to match the verification with a smaller VM GPUs were provisioned using only 2 K80 GPUs on a local azure VM.

B. *Yolact Instance Segmentation Experiments*

The DNN implemented for this challenge utilized the You Only Look At CoefficienTs (Yolact) [6] that run on PyTorch. This was selected because of the efficient results shown by training large datasets on a small number of GPUs. Additionally, Yolact also has been shown to be a model fitting for this proposed approach by its ability to break instance segmentation into two parallel subtasks: (1) generating a set of prototype masks and (2) predicting per instance mask coefficients [6]. This was an ideal selection based on the enhancements made in the training datasets by provide accurate segmentation annotations where the Yolact model would be able to process the mask coefficients effectively so at inference time our inference algorithm would be able to consume the cropped detected objects to accurate estimate the length of the objects without having to rely on computing the length using the bounding boxes.

Pre-trained Resnet50 weights were utilized for the Yolact models based on transfer learning best practices indicated in the paper that were trained with ImageNet. Similar hyperparameters were implemented based off the Yolact paper [6] that utilizing stochastic gradient decent for 200K iterations and divide the learning rate four times by 70K, 150K, 176K and 186K using the initial learning rate of 0.001, a weight decay of 0.0005, and a momentum of 0.9. After the Yolact container was created training experiments were launched in AzureML workspace using the Tesla K80 GPU clusters. The average training time took about 96 hours (~4 days).

C. *Training Results*

Table 1 summarizes the average precision scores that can be read as AP@0.5 / AP@0.5:0.95 where the validation dataset was five validation scenes that were held out from the training process. While the mask results are disappointing, the detected bounding box showed somewhat reasonable results on the initial attempts. By systematically improving the ground truths and adding backgrounds for minimized the false detection rate, the final enhance dataset showed improvement from prior training results. As described in the discussion section, areas of improvements have been defined where continual improvements in an interactive process with additional time could continue seeing higher performance results based current experiment improvements indicate.

	all	**0.50**	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
Box	28.18	56.63	52.95	47.65	41.02	53.71	23.82	15.88	8.21	2.58	0.38
Mask	7.03	24.46	19.89	13.76	7.96	3.27	0.71	0.16	0.05	0.00	0.00

Table 1 Bounding Box Inference Results Bounding box results of inference with different detection methods that can be read as AP@0.5 / AP@0.5:0.95

D. *Submission Results*

Table 2 summarizes the final submission results along with the xView3 reference model results on the xView3 public test set for the competitive leaderboard challenge. The table breaks down each of the scoring metrics that are then combined for an aggregate score to determine the challenge overall metrics score that is used for raking submissions to the challenge. The aggregate score is ranges between 0 low and 1 as high. The best Yolact model came in the top 50 in the overall challenge, rank 45, qualifying for prize and rewards, however, the personal challenge targets were to outperform the xView3 reference model which neither model did. Our method to utilize the mask to compute the length based on pixels outperform other length estimation methods that relies on the detected bounding box from a length score of 34% (fixed) to 52% (pixel).

Model	**Length Est version**	**loc fscore**	**l-fscore score**	**vessel fscore**	**fishing fscore**	**length acc**	**aggregate**
xView3	Fixed	**0.190**	**0.426**	**0.121**	0.712	**0.398**	**0.000**
yolact	Pixel	0.163	0.243	0.166	0.921	**0.752**	0.516
yolact	Fixed	0.155	0.249	0.166	0.921	**0.752**	0.341

Table 2 xView3 Challenge Submission Leaderboard Results The View3 challenge scores shown with the proposed Yolact model and the benchmark xView3 reference model results at the end of the challenge. The scoring metrics is an aggregate of five scores measuring the performance in overall (F1) maritime object detection, F1 score for close-to-shore objects, F1 score for vessel classification and F1 score for fishing classification and the aggregate F1 percentage error for object length [1].

E. *xView3 Submission Inference Speed vs Accuracy Tradeoff*

The tradeoff of speed and accuracy was particularly challenging for our approach as it was unclear if significant preparing would be required to meet the challenge constraints. By the time we were satisfied with the data loader and improving it upon the provided naïve methods, the cost to re-chip the training data to incorporate image size of 800 or 1024 would make us this project go overbudget in the current Azure Blob storage credits provided.

For experiments and analysis purposes only, a smaller dataset was created to compare the performance with different image sizes. Using the same Yolact training parameters, training experiments were quickly performed however formal metrics are not provided as we view this as quick use case study requiring further analysis. However, comparing the result of the input size for Yolact 1024-pixel vs 200 pixels, the 1024 accuracy decreased however the speed was significantly higher about twice as fast when compared to the Yolact 200-pixel model.

While these results would need to be evaluated further, utilizing a larger input size would indicate positive results in the verification stage of the results as the Yolact 200-pixel model took about 20 mins to evaluate per scene where the Yolact 1024 took about 9 mins. xView3 challenge required the verification container to evaluate each scene within 15 mins therefore, our hypothesis would consider Yolact 1024 to be considered.

IV. Discussion Points

This section highlights provides analysis and context on some of the challenges and areas of improvement have been identified related to the proposed methods and as reflecting on some of the approaches utilized as a participant in the xView3 Dark Vessel 2021 Challenge.

Azure Cloud Resources Challenged: There were many shortcomings in reflecting on the proposed solution for this challenge that can be broken into two categories: cloud computing challenges and project execution. From the cloud computing challenges, loading the entire dataset containing approximately 1000 scenes from maritime regions of interest was unexpected high. From unzipping the raw challenge dataset to creating individual image chips appropriate for training a DNN was a significant price for a project using Azure credits. Additionally, given the large number of scenes, multiple Azure VMs requiring a K80 GPU to support the memory during data processing was required to unpack and split the images in parallel. Moreover, to complete the project using Azure credits, we were restricted to using only K80 GPUs for training the DNN which significantly added additional training time if compared to using the preferred V100 GPUs.

Project Execution Challenges: This project contained a six-week execution timeline where the idea was to iterate as quickly as possible, however, reflecting, most of the time was spent on improving the ground truth annotations for the detector and vessel length predictions to be accurate. While reasonable initial results were obtained in these proposed methods, many shortcomings were identified in the final weeks of the challenge. Notably, the initial chipped images for the training dataset should have been set to 1024x1024 or 800x800 to take advantage of the selected models utilized as the training utilized transfer learning techniques on the models that could process input image sizes greater than initial 200x200. The selected 200x200 image size was initially used for debugging the initial data loader and was an oversight when chipping the remainder of the challenge dataset. However, in the final weeks, re-chipping the training data and test dataset (public leaderboard) would have consumed the remaining Azure credits required for model tuning, model testing and creating the submission inferences for the challenge leaderboard.

Another improvement identified in this project execution was time was not allocated efficiently for iterating on model selection and tuning hyperparameters. Performance gains could be realized through other methods of incorporating multi-scaling and model ensemble techniques. Finally, one of the most significant short comings was the high knowledge required to participate effectively in this challenge. This proposed method did not effectively utilize all the co-registered SAR images (e.g., ‘vh’ and ‘vv’) and the ancillary images that provides additional information such as wind speed, sea condition etc.

Personal Note: As this is the first major ICCV related challenge that we’ve been able to complete end-to-end within a six-week project timeline, reaching the completion of the challenge is satisfying, as noted, there are many areas to improve upon. Looking forward to being completed ICCV related challenged in 2022.

V. CONCLUSION

This report presented a pixel-wise segmentation method to improve the ship detection, classification, and ship length estimation in participation to the xView3 Dark Vessels 2021 Challenge. The proposed method used Yolact one-stage instance segmentation algorithm uses the enhanced pixel-wise ground truth to compare the performance in the xView3 Reference model on the public test challenge dataset. On the leaderboard our model ranked 45 in the top 50 participants. While the xView3 reference model outperforms our proposed Yolact model in the aggregate challenge scoring metrics, our method outperforms the reference model in estimating the length of detected ships and F1 ship classification.

VI. Acknowledgement

The xView3: Dark Vessel challenge datasets [1] was sponsored by the Defense Innovation Unit (DIU) under the Creative Commons Attribution Non-Commercial Share alike 4.0 International (CC BY-NC-SA 4.0) license. We are highly thankful for the research opportunity.

VII. References

[1] Global Fishing Watch, Defense Innovation Unit, and Cambr.io, "xView3 Prize Challenge: Developing Machine Learning Algorithms to Combat Illegal, Unreported, and Unregulated Fishing," 2021.

[2] W. K. B. P. S. L. W. M. a. R. G. M. C. P. Schwegmann, "Synthetic Aperture Radar Ship Detection Using Capsule Networks," *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium,* vol. 2018, pp. 725-728, 2018.

[3] K. Eldhuset, "An Automatic Ship and Ship Wake Detection System for Spaceborne SAR Images in Coastal Regions.," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 34, no. 4, pp. 1010-1019, 1996.

[4] M. Stasolla and H. Greidanus, "The exploitation of Sentinel-1 images for vessel size estimation,," *Remote Sensing Letters,* vol. 7, no. 12, pp. 1219-1228, 2016.

[5] L. V. G. C. K. W. J. W. a. A. Z. M. Everingham, "The pascal visual object classes (voc) challenge," *International journal of computer vision,* vol. 88, no. 2, pp. 303-338, 2010.

[6] D. Bolya, C. Zhou, F. Xiao and Y. Jae Lee, "YOLACT Real-time Instance Segmentation," arXiv, 2019.