UOW-Vessel: A Benchmark Dataset of High-Resolution Optical Satellite Images for Vessel Detection and Segmentation

Ly Bui[1], Son Lam Phung[1,*], Yang Di[1], Hoang Thanh Le[1], Tran Thanh Phong Nguyen[1],
Sandy Burden[1], Abdesselam Bouzerdoum[1,2]
[1]University of Wollongong, [2]Hamad Bin Khalifa University

Abstract

In this paper, we introduce UOW-Vessel, a benchmark dataset of high-resolution optical satellite images for vessel detection and segmentation. Our dataset consists of 3,500 images, collected from 14 countries across 4 continents. With a total of 35,598 instances in 10 vessel categories, UOW-Vessel is to date the largest satellite image dataset for vessel recognition. Furthermore, compared to the existing public datasets that only provide bounding box ground-truth, our new dataset offers more accurate polygon annotations of vessel objects. This dataset is expected to support instance segmentation-based approaches, which is a less investigated area in vessel surveillance. We also report extensive evaluations of the recent algorithms for instance segmentation on the new benchmark dataset.

1. Introduction

Vessel recognition from remote sensing images has attracted significant interest in maritime surveillance. It has applications in many domains, including defence, border control, law enforcement, fisheries management, maritime search-and-rescue, vessel traffic management, and marine environmental control [11]. Among several sensing modalities, optical satellite images are increasingly utilized for vessel surveillance. High-resolution optical satellite images offer valuable visual information about vessel features (e.g., size, shape, and structure), which greatly benefits the fine-grained vessel recognition task.

Several optical satellite image datasets for vessel recognition have been proposed, including HRSC2016 [16], ShipRSImageNet [28] and VHRShips [12]. These datasets comprise a large number of images and vessel categories, but they only provide the common horizontal or rotated bounding box as the ground-truth [16, 28]. Other datasets, such as FGSCR-42 [4] and FGSC-23 [26], only provide cropped images of the vessel objects, which do not capture the complexity of real-life scenes involving vessels.

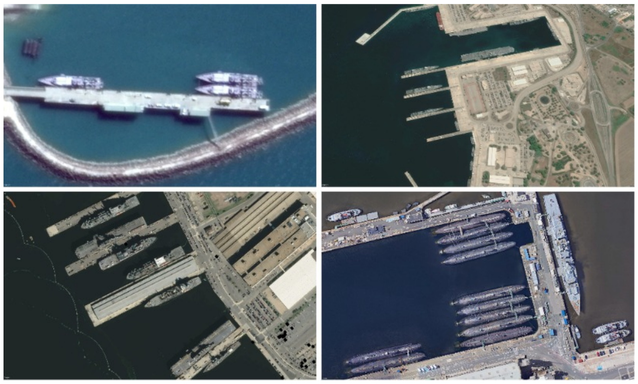


Figure 1. Sample images in the UOW-Vessel dataset.

To address this gap, we introduce the UOW-Vessel dataset, consisting of 3,500 high-resolution satellite images and 35,598 instances in 10 vessel categories. Figure 1 shows the sample images in the dataset. Moreover, our dataset offers more precise polygon annotations of the vessel objects. This annotation type can capture intricate details of the vessels and better account for occlusions, overlapping instances, and arbitrary orientations.

With the increasing availability of large-scale datasets, deep learning methods for vessel recognition have received significant attention in the research community. Because the public datasets provide bounding box and classification labels, existing methods mostly adopted the object detection and image classification approaches for vessel recognition [8, 13, 15, 27]. The segmentation-based methods for vessel recognition in satellite images remain a less investigated area in the literature. Therefore, in this paper, we conduct extensive experiments to verify the effectiveness of the recent instance segmentation methods on our proposed dataset. The prediction accuracy, model sizes and inference speeds are provided as the performance baselines.

The main contributions of the paper are two-fold. Firstly, we introduce a novel benchmark dataset for vessel detection and segmentation, called UOW-Vessel. Our dataset

[*]Corresponding author: phung@uow.edu.au

Dataset	Year	Sources	# Images	Image size	# Classes	# Objects	Annotation
HRSC2016 [16]	2016	Google Earth	1,070	300 × 300 ∼ 1500 × 1500	25	2,976	H-Box R-Box
Airbus [10]	2018	SPOT 6/7	192,556	768 × 768	1	213,723	H-Box
FGSC-23 [26]	2020	Google Earth Gaofen-1	4,080	40 × 40 ∼ 800 × 800	23	4,080	Classification
FGSCR-42 [4]	2021	Multi-sources	9,320	50 × 50 ∼ 1500 × 1500	42	9,320	Classification
ShipRSImageNet [28]	2021	Multi-sources	3,435	930 × 930 ∼ 1400 × 1400	50	17,113	H-Box R-Box
S2-Ships [3]	2021	Sentinel-2	16	1783 × 938	1	1,053	Polygon
DOTA-v2.0 [5]	2022	Multi-sources	11,268	800 × 800 ∼20K × 20K	1	251,883	R-Box
VHRShips [12]	2022	Google Earth	5,312	280 × 720	35	11,179	H-Box
UOW-Vessel (ours)	2023	Google Earth	3,500	8192 × 4320 ∼ 8192 × 6881	10	35,598	Polygon

Table 1. Comparison between existing public datasets on optical remote sensing ship recognition and the proposed UOW-Vessel dataset.

contains 3,500 high-resolution optical satellite images and 35,598 instances in 10 vessel categories[1]. To the extent of our knowledge, UOW-Vessel is the largest satellite image dataset with fine-grained ground-truth for vessel recognition to date. Secondly, we evaluate the recent instance segmentation methods for the vessel detection and segmentation tasks, and provide the baseline performances on our benchmark dataset.

2. Related work

In this section, we analyze the existing satellite image datasets for vessel recognition and compare them with our proposed dataset. We then review the existing deep learning methods for the vessel recognition task.

2.1. Benchmark datasets

Over the past decade, several Earth observation datasets have been created for vessel recognition from the aerial view. Table 1 summarizes the comparison between the existing datasets and the UOW-Vessel dataset. DOTA-v2.0 [5] is a large-scale dataset of high-resolution optical images for general object detection from aerial view. The dataset consists of 18 categories and 251,883 instances belong to the vessel class. Airbus [10] is a large-scale ship detection dataset with 192,556 satellite images and 213,723 vessel instances. Although there is a large number of vessel objects in these datasets, only one class is labeled as ‘Ship’ or ‘Vessel’, which is not applicable to the fine-grained vessel recognition task. Compared to these datasets, our UOW-Vessel dataset provides 10 sub-categories of the vessels, including aircraft carrier, destroyer and frigate.

FGSC-23 [26] and FGSCR-42 [4] datasets label each image with one category for the vessel classification task. FGSC-23 [26] consists of 4,080 images and 23 vessel categories in total. FGSCR-42 [4] comprises 9,320 images and 42 categories, which is an improved dataset compared to the FGSC-23 dataset. These datasets only provide cropped vessel images, which do not capture the complexity of real-life scenes involving vessels. Compared to these datasets, the images in our dataset encompass a wide range of vessel scenes, with complex backgrounds and in challenging weather conditions. Moreover, our dataset captures large variations in vessel sizes, which is an inherent challenge in practical applications.

HRSC2016 [16] was released in 2016 as the first large-scale optical satellite image dataset for vessel recognition. It features 1,070 images, 25 vessel categories, and 2,976 vessel instances. VHRShips [12] has a total of 5,312 images and 11,179 instances in 35 vessel categories. ShipRSImageNet [28] is currently the largest dataset for fine-grained vessel recognition, with 3,435 images and 17,113 instances in 50 vessel categories. Compared to these datasets, UOW-Vessel provides more precise polygon annotations of the vessel objects, thereby capturing the exact boundary of the vessels. With the polygon annotations, our dataset supports multiple tasks for vessel recognition, including object detection, semantic segmentation, and instance segmentation. Additionally, our dataset contains a larger amount of annotated vessels, with double the instance count of the ShipRSImageNet dataset. While S2-Ships [3] provides segmentation masks for vessel objects, the dataset only con-

[1] URL: https://documents.uow.edu.au/∼phung/UOW-Vessel.html

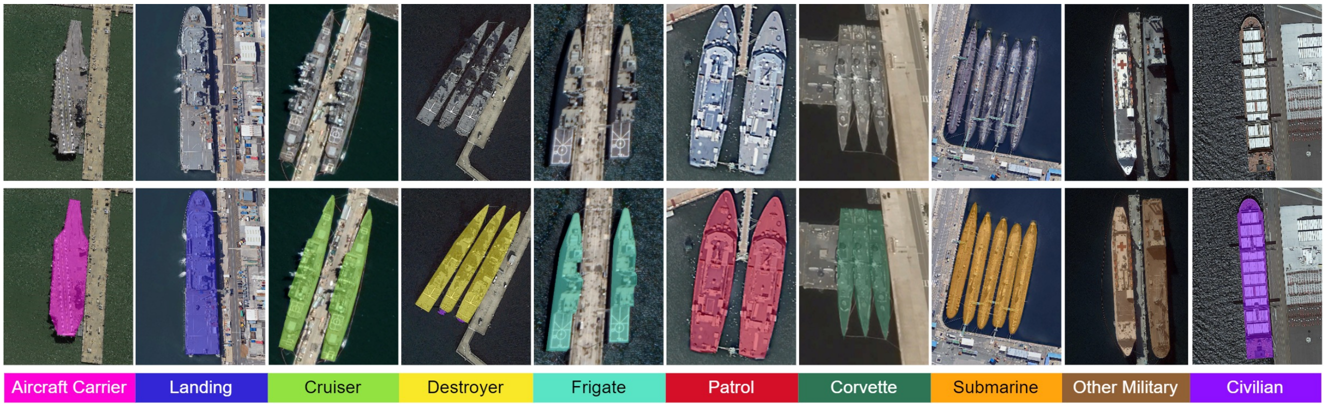


Figure 2. Visualization of the UOW-Vessel dataset. The UOW-Vessel dataset contains 10 vessel categories. The first row shows the color images of the vessels, and the second row shows the corresponding ground-truth annotations for the instance segmentation task.

tains 16 images from Sentinel-2 and one vessel category. Compared to S2-Ships, UOW-Vessel is more suitable for fine-grained vessel recognition tasks. In summary, these distinctions make the UOW-Vessel dataset a valuable resource for vessel surveillance research.

2.2. Algorithms

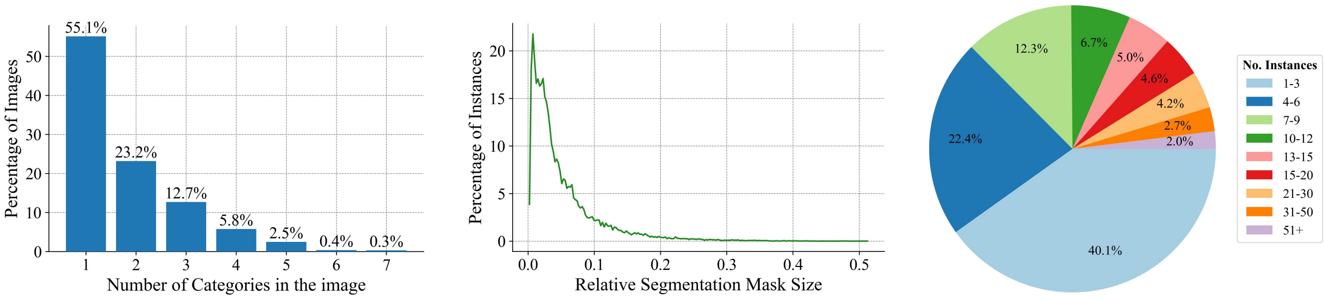
Traditional vessel recognition methods generally adhere to the following workflow: sea-land separation, removal of environmental effects, vessel candidate detection, removal of false alarm detection, and classification [11]. Within this sequence, the detection and classification of vessel candidates are the areas of significant research interest. Traditional vessel detection methods include threshold-based methods [21], salient-based methods [25], and methods based on shape and texture features of the vessels [14, 29]. These methods are generally fast, however, there still exist some limitations such as high false alarm rates and poor generalization to complex backgrounds [11].

With the growing number of optical satellite image datasets, deep learning-based methods have emerged as the leading approach for vessel detection. Detecting vessels via satellite imaging is a challenging task due to the objects' varied aspect ratios, arbitrary orientations and sometimes dense distributions. For example, the horizontal bounding boxes for arbitrary-oriented vessels that are closely docked to each other include redundant background. Hence, detection methods that rely on horizontal bounding boxes are not sufficiently precise for practical needs. Some rotated detectors have been proposed to solve this issue. An anchor-free rotation vessel detector is introduced in [27], which utilizes a Gaussian mask branch to model the vessels more accurately based on their geometry characteristics. In [13], a dual-branch regression is designed to generate rotated region proposals. It consists of two independent regression branches: orientation-agnostic regression branch to predict the bounding box variables, and orientation regression branch to predict the ship orientation. In [15], the rotated bounding box is determined by the regression of the long and short sides of the vessels. This approach transforms the angle regression task into the side regression task, which avoids the problems of predicting the angle directly.

For vessel classification, the aim is to categorize the detected vessels into a specific class, e.g. destroyer, landing, or frigate. Inter-class similarity and intra-class difference are challenging problems in vessel classification, i.e. vessels from the same class may have different visual appearances while vessels from different classes may appear similar. Deep learning methods can tackle these problems due to their capability to extract discriminative features automatically from large labeled data. In [26], a multilevel enhanced visual feature representation is designed to fuse the re-weighted regional features, thereby focusing on the silent region and suppressing other regions. Additionally, an attribute-guided feature extraction branch is designed to generate complementary attribute features (i.e., scale and aspect ratio), which are combined with enhanced features for classification. An attention mechanism is applied in several studies to differentiate the vessels more effectively in dense and complex scenes. In [18], an attention mechanism is employed to improve the accuracy of the model by focusing on high-response channels. In [8], a dual-mask attention module is used to suppress clutter and enhance the distinction between closely docked vessels.

3. UOW-Vessel dataset

In this section, we describe the data collection and annotation process of satellite images for identifying vessel targets. We then present the statistics and the organization of our dataset.



(a) Distribution of the number of unique vessel categories per image.

(b) Distribution of the relative size of instance masks with respect to the image size.

(c) Distribution of the number of vessel instances per image.

Figure 3. Statistics of the UOW-Vessel dataset. Best viewed digitally.

3.1. Data collection

We collected satellite images from various naval bases and maritime routes worldwide. The collection tool allows us to retrieve high-resolution satellite images at various locations on Earth. In total, we collected 3,500 images from 35 naval bases, across 14 countries and 4 continents. To enhance the diversity of the dataset, we systematically collected the images with various timestamps from July 2007 to May 2023. We employed map rotation, tilt and zoom functionalities to acquire diverse viewpoints of the maritime scenes. Furthermore, we intentionally captured images in challenging scenarios (e.g., with sunglint effect, poor illumination, or cloudy/hazy conditions) to account for a wide range of real-world scenarios. For each image, we also saved the metadata, including the GPS coordinates, date, and time. This information can facilitate image data analysis, especially in locating specific vessel features.

In this dataset, the vessels of interest include Aircraft Carrier, Landing, Cruiser, Destroyer, Frigate, Patrol, Corvette, Submarine, Other Military, and Civilian. The class “Other Military” contains auxiliary vessels such as training ships, survey ships, medical ships, and oil replenishment ships. Our dataset covers a wide range of military vessels from various navies around the world. Furthermore, our dataset is the only one that provides annotations for the Corvette class, which is a significant type of warships not available in the existing datasets.

3.2. Data annotation

For this dataset, we use polygons to annotate the vessel objects. With polygon annotations, objects with irregular shapes or multiple components can be more precisely annotated compared to other annotation types, e.g., horizontal or oriented bounding boxes. This is important for vessel detection tasks because vessels can have arbitrary orientations and complex shapes, with multiple components such as the hull, deck, and superstructure. This level of detail can help deep learning models differentiate vessels from other objects, even when the vessels are partially occluded or closely docked in a complex background.

Table 2. The number of instances per ship category in each set and the category distribution in the UOW-Vessel dataset.

Class	Train	Val	Test	Total	Dist.
Aircraft Carrier	127	23	30	180	0.51%
Landing	723	117	238	1,078	3.03%
Cruiser	160	22	33	215	0.60%
Destroyer	512	80	154	746	2.10%
Frigate	666	117	180	963	2.71%
Patrol	1,037	120	280	1,437	4.04%
Corvette	444	47	131	622	1.75%
Submarine	584	95	166	845	2.37%
Other Military	3,602	502	1,001	5,105	14.34%
Civilian	16,599	3,277	4,531	24,407	68.56%
Total	24,454	4,400	6,744	35,598	100%

The annotation process consists of three rounds. Initially, our team manually categorized vessel objects into 10 classes. Subsequently, a team member reviewed the annotations for accuracy and adjusted the polygon boundaries as needed. Finally, the ground-truth data was processed into suitable formats for network training. The annotation ground-truth is available in the JSON format, which contains the polygon coordinates for all images. Figure 2 shows the visualization of annotated vessels in our dataset.

3.3. Dataset statistics

The UOW-Vessel dataset consists of 3,500 total images with the image resolution ranging from $8,192 \times 4,320$ to $8,192 \times 6,881$ pixels. The dataset is partitioned into the training, validation and test sets with the proportions of 70%, 10% and 20%, respectively.

Category statistics. There are 10 vessel categories present in our dataset. Table 2 shows the distribution of the number of instances per category in each set. The distribution indicates that there is an inherent class imbalance in the dataset, as 68.6% of the instances belong to the Civilian category. The class imbalance reflects the real-life scenarios, where civilian vessels tend to outnumber other classes, and certain classes like Aircraft Carrier and Cruiser have fewer instances. Additionally, Figure 3a presents the distribution of the unique class counts per image. Overall, 55.11% of the dataset has one unique class present in an image. The largest number of unique classes an image has is 7 classes, and these images account for only 0.37% of the dataset.

Instance statistics. Our dataset consists of 35,598 instances in total. On average, each image is annotated with 10.2 instances. The most instances an image has in our dataset is 682 of mostly civilian vessel objects. Figure 3c shows the distribution of the number of vessel instances per image. Furthermore, most of the vessel instances in our dataset have a relatively small size compared to the image size. Figure 3b illustrates the distribution of the relative sizes of instance masks with respect to the image size, which is calculated as the square-root of [mask area divided by image area]. The distribution shows that approximately 65.4% of the instances have a relative area smaller than 5% of the image area.

3.4. Dataset organization

We provide the ground-truth annotations at three levels in our dataset to cater for various purposes. The labels are organized as follows:

- **Level 0:** All vessel instances are labeled as “Vessel”.
- **Level 1:** The vessel instances are labeled into two categories, “Military” and “Civilian”. All vessel categories except ”Civilian” are grouped into the ”Military” class.
- **Level 2:** The vessel instances are labeled into ten sub-categories: Aircraft Carrier, Landing, Cruiser, Destroyer, Frigate, Patrol, Corvette, Submarine, Other Military, and Civilian.

4. Methodology

In this section, we present the recent instance segmentation methods that are employed for vessel detection and segmentation on our dataset. Compared to object detection, instance segmentation methods not only detect objects in an image but also predict the pixel-level masks for the objects. We investigate three categories of instance segmentation methods: two-stage, multi-stage, and single-stage methods.

Two-stage methods. The two-stage approach for instance segmentation first generates candidate object proposals using an object detection algorithm. It then assigns a class label and generates a binary mask for each object proposal. Mask R-CNN [9] is a commonly used method for object detection and instance segmentation. It is built upon Faster R-CNN [20], by adding an extra FCN mask branch to produce the object mask. The mask prediction branch is added in parallel with the existing classification and bounding box regression branches.

Multi-stage methods. Cascade Mask R-CNN [2] extends Cascade R-CNN [1] for instance segmentation. Because Cascade R-CNN consists of multiple detection and classification branches, Cascade Mask R-CNN inserts the segmentation branch in three different ways. The first and second strategies add a single mask prediction branch at either the first or the last stage. The third strategy adds a mask prediction branch to every cascade stage.

Single-stage methods. The single-stage approach for instance segmentation performs both detection and segmentation in a single architecture and therefore reduces the inference time. SOLOv2 [24] decouples the mask learning process into convolution kernel learning and feature learning. The method directly maps the input image to the desired object classes and object masks without bounding box detection. CondInst [22] uses instance-aware mask heads to predict the masks, which are dynamically generated and conditioned on each target instance. This eliminates the need for ROI cropping and feature alignment in Mask R-CNN [9].

YOLO (You Only Look Once) is an effective one-stage object detection method, first introduced in [19]. The YOLO network typically has the following structure: Backbone for feature extraction, Neck for feature aggregation, and Head for result regression. YOLOv5 [6] extends its predecessors by introducing a Cross-Stage Spatial connection in the convolutional layers, and a Spatial Pyramid Pooling module to capture features at multiple scales. Based on YOLOv5, YOLOv7 [23] further enhances the detection performance by using a new backbone called Extended Efficient Layer Aggregation Network (E-ELAN) and applying bag-of-freebies to optimize the training process. YOLOv8 [7] integrates multiple state-of-the-art modules into the network architecture. The YOLO models support instance segmentation tasks by adding a segmentation head to generate the mask for each object.

5. Experiments

In this section, we first describe the implementation details and the evaluation metrics for training and testing. We then report the quantitative and qualitative results of recent instance segmentation methods on our benchmark dataset, and present the analysis.

5.1. Implementation details

For Mask R-CNN and Cascade Mask R-CNN, we used the default implementations provided by MMDetection framework [17]. The batch size was set as 4, and the number of epochs was set as 12. We used the Stochastic Gradient Descent (SGD) as the optimizer for training, with a learning rate of 0.02, a momentum rate of 0.9, and a weight decay of 0.0001.

Method	**Backbone**	**mAP$^b_{50}$**	**mAP^b**	**mAP$^m_{50}$**	**mAP^m**	**Params**	**GFLOPs**	**FPS**
Level 0								
Mask R-CNN [9]	R-50-FPN	64.5	46.5	63.1	44.0	43.9	113.0	9.9
Cascade Mask R-CNN [2]	R-50-FPN	66.4	49.8	64.6	45.5	77.0	1,632.0	9.2
SOLOv2 [24]	R-50-FPN	-	-	51.6	30.4	46.2	113.9	4.8
CondInst [22]	R-50-FPN	64.9	45.2	59.1	33.4	33.9	133.9	-
YOLOv5-L [6]	CSPDarknet	87.9	69.4	79.9	52.3	47.4	147.0	50.7
YOLOv7 [23]	CSPDarknet	**88.6**	70.3	80.4	52.2	37.8	142.6	50.1
YOLOv8-L [7]	CSPDarknet	88.0	**70.7**	**83.5**	**55.8**	45.9	220.8	35.5
Level 1								
Mask R-CNN [9]	R-50-FPN	66.0	47.2	64.2	44.1	43.9	113.0	9.9
Cascade Mask R-CNN [2]	R-50-FPN	67.2	50.8	65.8	45.8	77.0	1,632.0	9.0
SOLOv2 [24]	R-50-FPN	-	-	53.2	32.2	46.2	113.9	5.1
CondInst [22]	R-50-FPN	63.9	45.1	58.7	34.2	33.9	133.9	-
YOLOv5-L [6]	CSPDarknet	85.2	68.0	78.2	51.9	47.5	147.0	50.8
YOLOv7 [23]	CSPDarknet	**85.3**	68.3	78.5	52.0	37.8	142.7	50.2
YOLOv8-L [7]	CSPDarknet	84.5	**69.0**	**80.3**	**54.8**	45.9	220.8	39.5
Level 2								
Mask R-CNN [9]	R-50-FPN	55.3	41.1	53.4	37.7	44.0	113.0	9.7
Cascade Mask R-CNN [2]	R-50-FPN	59.3	48.2	57.6	41.7	77.1	1,637.0	9.0
SOLOv2 [24]	R-50-FPN	-	-	53.7	36.7	46.2	114.1	4.8
CondInst [22]	R-50-FPN	56.8	43.2	54.0	35.5	34.0	133.9	-
YOLOv5-L [6]	CSPDarknet	**82.4**	70.6	78.9	57.3	47.5	147.2	51.5
YOLOv7 [23]	CSPDarknet	82.1	70.5	78.7	57.7	37.9	142.8	50.5
YOLOv8-L [7]	CSPDarknet	81.6	**72.5**	**80.0**	**61.7**	45.9	220.8	35.8

Table 3. Performance of instance segmentation methods on the UOW-Vessel data set. Top-1 performance of each level is shown in **bold**.

For SOLOv2 and CondInst, we trained the models using $1\times$ learning schedule as implemented in the *detectron2* framework. The batch size was set as 8. The models were trained for 90K iterations with an initial learning rate of 0.01. The learning rate was reduced by a factor of 10 at iteration 60K and 80K. The weight decay and momentum were set as 0.0001 and 0.9, respectively.

For YOLOv5, YOLOv7 and YOLOv8, we adopted the official GitHub repository for training. The batch size was set as 16 for all models. We trained the models for 200 epochs with early stopping, i.e., if the validation performance did not improve after 50 epochs. The SGD was also used as the optimizer for training, with a learning rate of 0.01, a momentum rate of 0.937, and a weight decay of 0.0005. All models were initialized with the pretrained weights on the MS-COCO dataset. We trained and evaluated all models on an NVIDIA GeForce RTX 3090 GPU with 24GB memory.

5.2. Evaluation metrics

We used three metrics to evaluate the performances of the instance segmentation models: Mean Average Precision (mAP), giga floating-point operations (GFLOPs), and frames per second (FPS).

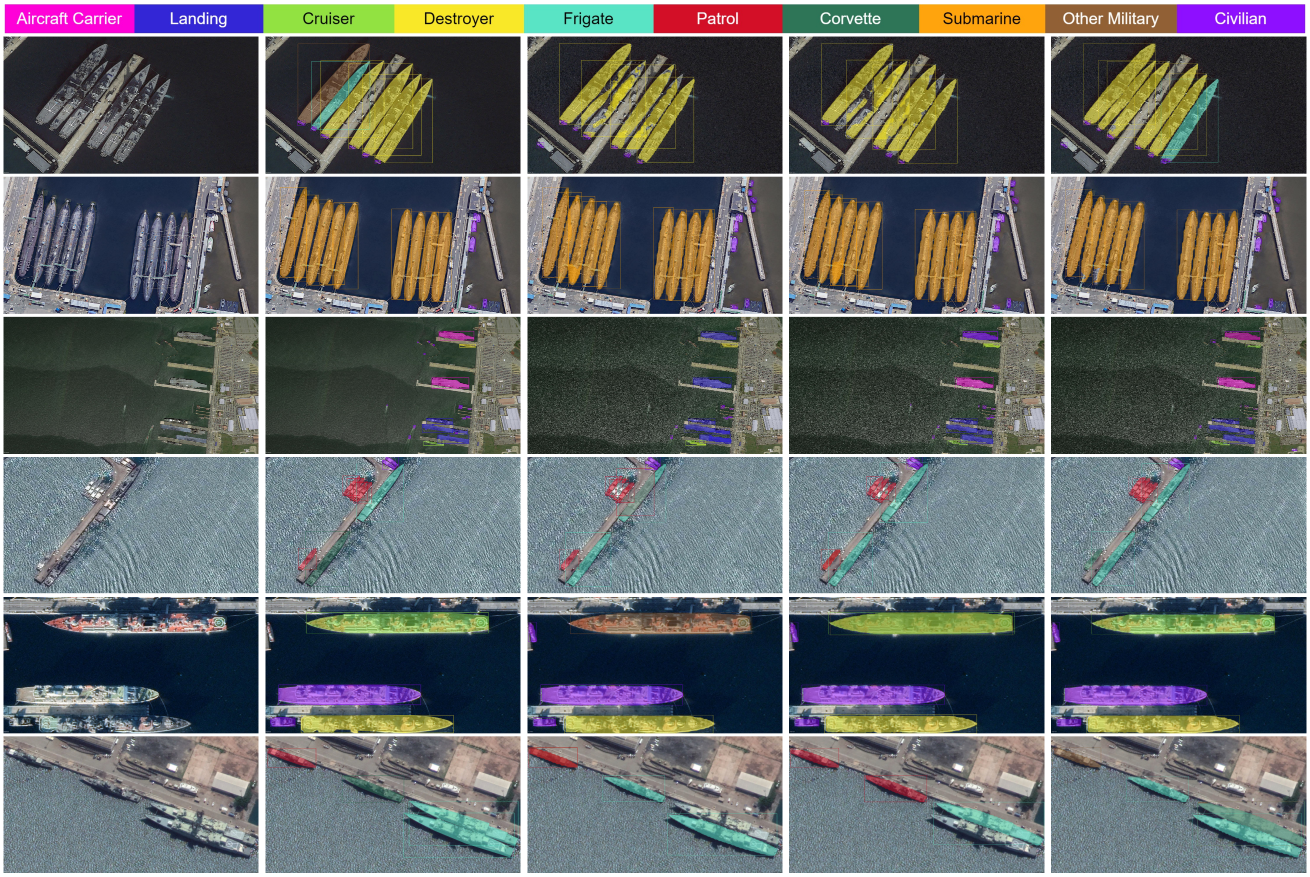
Mean Average Precision (mAP) assesses the model performance to simultaneously detect and segment objects within an image. We compute the precision and recall values as follows. Let A and B denote the ground-truth and a predicted bounding box, respectively. The Intersection-over-Union (IoU) is the ratio between the areas of the intersection versus the union of A and B:

$$\mathrm{IoU} = \mathrm{J}(A, B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid}. \qquad (1)$$

A predicted bounding box is considered a true positive (TP) if the IoU is higher than the threshold. The AP value for each IoU threshold is computed by interpolating the precision values at various recall levels. Finally, we take the mean of AP values across all IoU thresholds to obtain the mAP. We report the mAP_{50} and mAP for the detection and segmentation predictions of the baseline methods. mAP_{50} is calculated at the IoU threshold of 0.5, and mAP is calculated across the IoU thresholds from 0.5 to 0.95.

Giga floating-points operations (GFLOPs) is the number of billions of floating-point operations required to process one image.

Frames per second (FPS) is the number of images that a model can process per second.



(a) Image (b) Ground-truth (c) YOLOv5 [6] (d) YOLOv7 [23] (e) YOLOv8 [7]

Figure 4. Visualization of the vessel detection and segmentation on the UOW-Vessel dataset.

5.3. Baseline performances

Quantitative results. Table 3 summarizes the performance of the recent instance segmentation methods on the UOW-Vessel dataset. The performance of each method is evaluated on three levels of annotation complexity, as detailed in Section 3.4.

Firstly, the experimental results show that the detection and segmentation performance generally decreases as the task becomes more fine-grained (from Level 0 to Level 2). For example, the performances of Mask R-CNN in Level 2 are lower than those in Level 0 by 5.4% to 9.7%.

Secondly, the results show a relatively consistent trend in performances across three different levels. Cascade Mask R-CNN performs better than Mask R-CNN across all mAP metrics. However, the model complexity (GFLOPs) of Cascade Mask R-CNN is approximately 14× higher than that of Mask R-CNN. SOLOv2 and CondInst have a slightly inferior performance compared to Mask R-CNN and Cascade Mask R-CNN. Specifically, the mask mAP of Mask R-CNN in Level 2 is 1.0% and 2.2% higher than SOLOv2 and CondInst, respectively. Note that the detection accuracy for SOLOv2 is not reported as the method does not predict the bounding box for the target instances.

Moreover, it can be observed that the YOLO methods achieve better performance than other methods in terms of prediction accuracy and inference speed. In terms of detection accuracy, even though YOLOv8 has lower mAP_{50} than YOLOv5 and YOLOv7, it still has the highest mAP overall. In terms of segmentation accuracy, YOLOv8 also achieves the best performances compared to YOLOv5 and YOLOv7. The inference speed of YOLOv8 is slower compared to those of YOLOv5 and YOLOv7, however, it still meets the real-time requirement for image processing.

Qualitative results. Considering the superior performances of the YOLO algorithms, we show the visualization of the detection and segmentation predictions of YOLOv5, YOLOv7 and YOLOv8. The visualization results along with the ground-truth are illustrated in Figure 4. It can be observed that YOLOv8 can generate finer segmentation masks compared to YOLOv5 and YOLOv7 (See Row 1). Moreover, YOLOv8 can detect and segment objects in clutter scenes, which are sometimes missed by YOLOv5 and

YOLOv7 (See Row 5 and 6). However, it can be seen that all methods still wrongly classify the vessel categories (See Row 1, 5 and 6). The misclassification is likely due to the inter-class similarity problem in vessel recognition, where vessels in different classes may have similar appearances.

5.4. Further analysis

Challenging conditions. We investigate the effects of the various challenging conditions on the performance of the model. We divided the test set (701 images) into four different scenes: normal, sunglint, cloud/haze, and low illumination. We selected YOLOv8-Large as the baseline model to evaluate the performance. The results are reported in Table 4. Overall, the results show that the model performs well in normal and sunglint conditions. However, the performance decreases in poor illumination and cloud/haze settings. The drop in performance is possibly because the visual features of vessel objects in these settings are obstructed by poor weather and lighting conditions.

Table 4. Performance comparison on multiple weather conditions.

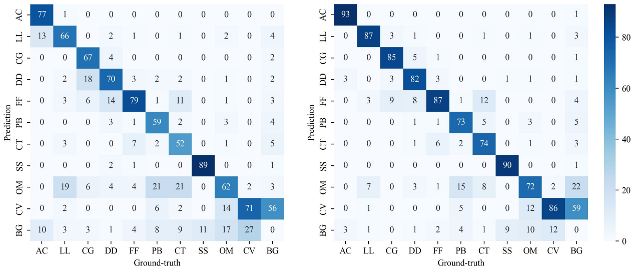
Scene	# Images	mAP$^b_{50}$	mAP^b	mAP$^m_{50}$	mAP^m
Normal	473 (67.4%)	84.1	75.3	82.9	65.7
Sunglint	130 (18.5%)	80.8	72.0	80.4	63.1
Cloud/Haze	79 (11.2%)	74.8	62.2	69.3	46.5
Poor illumination	19 (2.7%)	60.4	54.7	59.5	46.4

Input image resolutions. We investigate the effects of the input image resolutions on the performance of the model. In this study, the image resolution is varied from 480 × 480 to 1440 × 1440 pixels. We selected YOLOv8-Large as the baseline model to evaluate the performance. All models are trained from scratch. The results are reported in Table 5.

Table 5. Performance comparison on multiple image sizes.

Image size	mAP$^b_{50}$	mAP^b	mAP$^m_{50}$	mAP^m
480 × 480	73.0	60.6	70.6	49.7
640 × 640	76.7	65.2	75.5	56.2
800 × 800	78.0	66.9	77.0	60.1
960 × 960	85.6	76.9	84.5	68.8
1120 × 1120	81.5	71.8	81.0	66.2
1280 × 1280	79.9	70.2	79.4	64.8
1440 × 1440	87.8	79.4	87.1	72.9

The results show that the detection and segmentation accuracy improves as the image size increases. The detection and segmentation mAP enhance by 6.3% and 19.1%, respectively, when the image size increases from 480 × 480 to 960 × 960 pixels. The performance also improves when trained on the image size of 1440 × 1440 pixels. In particular, the performance is enhanced by 2.5% and 4.1% in detection and segmentation mAP, respectively, compared to the model trained on 960 × 960 pixels.



(a) 640 × 640 (b) 1440 × 1440

Figure 5. Confusion matrices of models trained on input image sizes of 640 × 640 and 1440 × 1440 pixels. Best viewed digitally.

The confusion matrices of the models trained on 640 × 640 and 1440 × 1440 pixels are shown in Figure 5. The predictions of all categories improve when increasing the image size, especially for lower-performing categories such as Patrol, Corvette and Landing. The improvement is likely due to the model's ability to capture finer details of the vessel objects when the sizes of the vessel objects increase, which is crucial in fine-grained vessel recognition tasks. However, the trade-off between prediction accuracy and inference speed should be considered when the model is applied to practical applications.

6. Conclusion

We introduced a large-scale dataset for fine-grained vessel recognition in optical satellite images called UOW-Vessel. The dataset contains the polygon annotation of the vessel targets, which supports multiple approaches for vessel recognition, including object detection, semantic segmentation, and instance segmentation. We conducted extensive evaluations of recent instance segmentation methods on the new benchmark dataset. The experimental results showed that YOLOv8 achieves the highest performance for detection and segmentation. We further examined how the model performed with challenging scenarios and different input image sizes. While recent methods have shown notable performances, there is still significant potential for improvement. We expect that UOW-Vessel dataset will facilitate the development of new algorithms for vessel detection and segmentation.

Acknowledgements

This research was supported by the NSW Space Research Network and NSW State Government through an SRN Research Pilot Project (Grant ID: SRN-PILOT-2022-03). The research was also supported by the Australian Geospatial-Intelligence Organisation, FrontierSI, and Intelligent System Design Pty Ltd.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6154–6162, 2018. 5

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(5):1483–1498, 2021. 5, 6

[3] Alina Ciocarlan and Andrei Stoian. Ship detection in sentinel 2 multi-spectral images with self-supervised learning. *Remote Sensing*, 13(21), 2021. 2

[4] Yanghua Di, Zhiguo Jiang, and Haopeng Zhang. A public dataset for fine-grained ship classification in optical remote sensing images. *Remote Sensing*, 13(4), 2021. 1, 2

[5] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7778–7796, 2022. 2

[6] Jocher Glenn. YOLOv5, 2020. 5, 6, 7

[7] Jocher Glenn, Chaurasia Ayush, and Qiu Jing. YOLOv8, 2023. 5, 6, 7

[8] Yaqi Han, Xinyi Yang, Tian Pu, and Zhenming Peng. Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 60:1–18, 2022. 1, 3

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Int. Conf. Comput. Vis.*, pages 2980–2988, 2017. 5, 6

[10] Kaggle. Airbus ship detection challenge, 2019. 2

[11] Urška Kanjir, Harm Greidanus, and Krištof Oštir. Vessel detection and classification from spaceborne optical images: A literature survey. *Remote Sensing of Environment*, 207:1–26, 2018. 1, 3

[12] Serdar Kızılkaya, Ugur Alganci, and Elif Sertel. Vhrships: An extensive benchmark dataset for scalable deep learning-based ship detection applications. *ISPRS Int. J. Geo-Information*, 11(8), 2022. 1, 2

[13] Linhao Li, Zhiqiang Zhou, Bo Wang, Lingjuan Miao, and Hua Zong. A novel cnn-based method for accurate ship detection in hr optical remote sensing images via rotated bounding box. *IEEE Trans. Geosci. Remote Sens.*, 59(1):686–699, 2021. 1, 3

[14] Ge Liu, Yasen Zhang, Xinwei Zheng, Xian Sun, Kun Fu, and Hongqi Wang. A new method on inshore ship detection in high-resolution satellite images using shape and context information. *IEEE Geosci. Remote Sens. Letters*, 11(3):617–621, 2014. 3

[15] Qiangwei Liu, Xiuqiao Xiang, Zhou Yang, Yu Hu, and Yuming Hong. Arbitrary direction ship detection in remote-sensing images based on multitask learning and multiregion feature fusion. *IEEE Trans. Geosci. Remote Sens.*, 59(2):1553–1564, 2021. 1, 3

[16] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *Int. Conf. Pattern Recog. Appl. Methods*, pages 324–331, 2017. 1, 2

[17] MMDetection Contributors. OpenMMLab Detection Toolbox and Benchmark, 2018. 5

[18] Peng Qin, Yulin Cai, Jia Liu, Puran Fan, and Menghao Sun. Multilayer feature extraction network for military ship detection from high-resolution optical remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 14:11058–11069, 2021. 3

[19] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, real-time object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 779–788, 2016. 5

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 5

[21] Tong Shuai, Kang Sun, Xiangnan Wu, Xia Zhang, and Benhui Shi. A ship target automatic detection method for high-resolution remote sensing. In *IEEE Int. Geosci. Remote Sens. Symposium*, pages 1258–1261, 2016. 3

[22] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Eur. Conf. Comput. Vis.*, 2020. 5, 6

[23] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7464–7475, 2023. 5, 6, 7

[24] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. SOLOv2: Dynamic and fast instance segmentation. In *Adv. Neural Inform. Process. Syst.*, volume 33, pages 17721–17732, 2020. 5, 6

[25] Feng Yang, Qizhi Xu, and Bo Li. Ship detection from optical satellite images based on saliency segmentation and structure-LBP feature. *IEEE Geosci. Remote Sens. Letters*, 14(5):602–606, 2017. 3

[26] Xiaohan Zhang, Yafei Lv, Libo Yao, Wei Xiong, and Chunlong Fu. A new benchmark and an attribute-guided multilevel feature representation network for fine-grained ship classification in optical remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 13:1271–1285, 2020. 1, 2, 3

[27] Xiangrong Zhang, Guanchun Wang, Peng Zhu, Tianyang Zhang, Chen Li, and Licheng Jiao. GRS-Det: An anchor-free rotation ship detector based on gaussian-mask in remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 59(4):3518–3531, 2021. 1, 3

[28] Zhengning Zhang, Lin Zhang, Yue Wang, Pengming Feng, and Ran He. ShipRSImageNet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 14:8458–8472, 2021. 1, 2

[29] Changren Zhu, Hui Zhou, Runsheng Wang, and Jun Guo. A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features. *IEEE Trans. Geosci. Remote Sens.*, 48(9):3446–3456, 2010. 3