This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

A Benchmark for Deep Learning Based Object Detection in Maritime Environments

Sebastian Moosbauer[1,2], Daniel König[1], Jens Jäkel[2], and Michael Teutsch[1]

[1] Hensoldt Optronics GmbH, Oberkochen, Germany
{sebastian.moosbauer, daniel.koenig, michael.teutsch}@hensoldt.net
[2] HTWK Leipzig, Leipzig, Germany
jens.jaekel@htwk-leipzig.de

Abstract

Object detection in maritime environments is a rather unpopular topic in the field of computer vision. In contrast to object detection for automotive applications, no sufficiently comprehensive public benchmark exists. In this paper, we propose a benchmark that is based on the Singapore Maritime Dataset (SMD). This dataset provides Visual-Optical (VIS) and Near Infrared (NIR) videos along with annotations for object detection and tracking. We analyze the utilization of deep learning techniques and therefore evaluate two state-of-the-art object detection approaches for their applicability in the maritime domain: Faster R-CNN and Mask R-CNN. To train the Mask R-CNN including the instance segmentation branch, a novel algorithm for automated generation of instance segmentation labels is introduced. The obtained results show that the SMD is sufficient to be used for domain adaptation. The highest f-score is achieved with a fine-tuned Mask R-CNN. This is a benchmark that encourages reproducibility and comparability for object detection in maritime environments.

1. Introduction

Visual object detection in maritime environments belongs to the research topics that gain rather little attention in the field of computer vision. Several applications exist such as harbor surveillance [11, 40] or collision avoidance for autonomously operating vessels [9, 10, 48]. However, in contrast to generic object detection [13, 36, 47], pedestrian detection [12, 25, 58], or face detection [55], there is no sufficiently comprehensive public benchmark available. Hence, object detection results recently reported in this field [26, 32, 37, 50] are neither reproducible nor comparable. Prasad *et al*. [43] introduced the Singapore Maritime Dataset (SMD). This dataset is one of the few publicly available that is specifically dedicated to object detection in maritime environments, but it lacks of representative benchmarking results. There are some other maritime datasets tackling challenges such as boat traffic monitoring [4], piracy detection [1], vessel classification [19], or obstacle detection for unmanned surface vehicles [31]. But none of them is either sufficiently large to train a Deep Convolutional Neural Network (DCNN) or specifically designed for object detection. However, we feel that the SMD can be utilized for both deep learning and representative benchmarking. In addition, it is beneficial that the SMD includes Visual-Optical (VIS) and Near Infrared (NIR) videos since several authors proved the advantages of using multiple spectra for object detection [7, 25, 30]. However, the annotations lack in some aspects as for example there is no split proposed to separate training and test data for machine learning based detection approaches. Furthermore, the annotations are inconsistent regarding class assignments and there are only few samples for some of the ten classes contained in the annotations.

In this paper, the SMD is analyzed in depth and utilized for deep learning. State-of-the-art object detection approaches are evaluated for their suitability to be applied in maritime environments. Since semantic instance segmentation turned out to be a promising addition to multi-task learning for object detection [2, 8, 21], an instance segmentation algorithm is introduced to enrich the SMD data annotation, and an approach for weakly supervised recursive training [28] is evaluated. The overall result of this paper is a benchmark for object detection in maritime environments. Since the SMD provides only few samples for some of the ten contained classes, we cannot sufficiently train our DCNNs and thus consider only two classes: *object* and *background*. The contributions of this paper are (1) the analysis of the SMD in preparation for deep learning and proposing a split into training, validation, and test data, (2) the introduction of a simple algorithm for automated generation

of instance segmentation labels in maritime environments, (3) the introduction of a maritime object detection benchmark using the two state-of-the-art object detection DCNNs Faster R-CNN [45] and Mask R-CNN [22], and (4) the examination of weakly supervised recursive training [28] to improve the results of Mask R-CNN using the generated instance segmentation labels. The evaluation scripts, the annotations including the generated instance segmentation labels, and the necessary python scripts and configuration files for the Detectron [17] framework are available [1].

The remainder of this paper is structured as follows: related work is presented in Section 2. The SMD is analyzed in Section 3. Training strategies, experiments, and results are described in Section 4. We conclude in Section 5.

2. Related Work

Maritime Datasets: The SMD provides 31,653 frames with a Ground Truth (GT) of 240,842 annotated bounding boxes for ten different object classes in total. Mainly due to its large extent compared to other available datasets we chose it as data basis for our benchmark. Several public maritime datasets exist but none of them is promising for deep learning based object detection. Fefilatyev *et al*. [14] introduced a dataset captured from a buoy for horizon detection. As the annotations do not include bounding boxes for object detection, this dataset cannot be utilized. The Large-Scale Image Dataset for Maritime Vessels (MARVEL) [19] and the Maritime Imagery in the VIS and IR spectrums (VAIS) [57] do not provide bounding boxes either since they are datasets for object classification. Kristan *et al*. [31] introduced the Marine Obstacle Detection Dataset (MODD) captured from an unmanned surface vehicle. The aim of this dataset is to provide videos to train and evaluate obstacle detection approaches in maritime environments. Bovco *et al*. [6] improved the MODD by adding segmentation labels for sky, sea, and shore and thus prepared the dataset to be usable for anomaly detection using auto-encoders. However, as this dataset contains only twelve videos and ignores objects above the horizon, it is not applicable for the benchmark we are aiming at. The IPATCH dataset [1] was published to tackle the challenge of piracy detection. Its rather small extent with only fourteen videos and scenes makes us discard this dataset for further consideration. The Maritime Detection, Classification, and Tracking Database (MarDCT) [4] was acquired for surveillance applications. Object detection is one of the challenges, the dataset was created and annotated for. 1,739 bounding boxes are provided in 8,115 frames. Unfortunately, the annotations are fragmentary and the data extent is definitely not sufficient to train DCNNs. Finally, many authors [29, 50, 51] use private datasets that cannot be utilized for benchmarking.

[1] https://github.com/smoosbau/SMD-Benchmark

Object Detection: In maritime environments horizon detection is often used as a preprocessing step. This is not a stringent requirement for this task, but can have a positive effect on the detection robustness. It is either used to align consecutive frames [42, 43] to add spatio-temporal information or to set up search areas for subsequent detection algorithm(s) [51]. Common techniques for horizon detection use edge maps [3, 54, 49] or region based horizon detection [52, 5, 53]. Unfortunately, horizon detection is error prone if the horizon is occluded by objects or fog. The detected horizon is then used to learn a background model and perform background subtraction. Foreground regions are then assumed to contain objects. Prasad *et al*. [43] analyzed various algorithms such as single image statistics [44, 5] or feature based classifiers [59]. However, background models that assume a stationary background usually perform poorly as illumination changes, waves, and foam are highly dynamic and cannot be represented well. An alternative can be Gaussian Mixture Models (GMM) [15, 56], relatively stationary pixels [52], or kernel density estimation [38].

Recent state-of-the-art results in multiple computer vision challenges are achieved using DCNNs and deep learning techniques. In the maritime domain, horizon detection [26], object classification [20, 33], and visual anomaly detection [32], but also object detection are promising tasks for the application of DCNNs: Kim *et al*. [29] propose a spatio-temporal approach. They use Faster R-CNN fine-tuned on a custom dataset to detect objects and then apply short-term tracking. Marie *et al*. [37] utilize statistical machine learning methods to extract Regions of Interest (ROIs), which are then further analyzed using Fast R-CNN [16]. Testing on the SMD their best result achieves an f-score of 0.78. However, the results are not reproducible as the data split for training and testing remains unknown. Tangstad [50] uses Faster R-CNN for obstacle avoidance in maritime environments. Here, object detection is considered as a two-class problem (object vs. background). Just like other authors [4], images showing boats and vessels are borrowed from benchmark datasets for generic object detection [47, 39]. Together with a custom maritime dataset, domain adaptation can be performed for the Faster R-CNN leading to convincing object detection results. However, those results are not reproducible either.

3. Singapore Maritime Dataset

The SMD [43] is probably the most promising, currently available public dataset for object detection in maritime environments. There are 240,842 object labels with ten different classes in 81 videos in total. Table 1 shows some of its properties. Videos in the VIS and the NIR spectrum are provided that were acquired on-shore and on-board from a small boat. Furthermore, different illumination conditions such as hazy, daylight, and dark/twilight are covered.

Table 1. Properties of the SMD.

Subdataset	Videos (Annotated)	Labeled Frames	Number of Labels
NIR	30 (23)	11,286	83,174
VIS on-board	11 (4)	2,400	3,173
VIS on-shore	40 (36)	17,967	154,495
Total	81 (63)	31,653	240,842

However, we can identify some issues that need to be discussed: first, there is no data split suggested to separate training and test data. Hence, we propose such a split in Section 3.1 considering some of the dataset's constraints. Second, the bounding boxes are inconsistently labeled. For occluded objects, either only the visible part or the estimated entire object is annotated. Furthermore, some objects contain rather large background areas, which can be a drawback for training well-generalizing DCNN models [58]. We handle this issue by choosing a smaller Intersection-over-Union (IoU) threshold (0.3 instead of 0.5) to determine True Positive (TP) detections within our experiments in Section 4. Third, the labels are inconsistent in class assignment, which can be an issue when training a DCNN for multi-class object detection. In order to check the class assignment consistency, the GT for object tracking is used that contains a unique ID for every object. With this ID it is possible to verify whether the class label of an object is consistent across the entire sequence or not. Unfortunately, about 9 % of the individual tracks contain at least one switch in the annotated object's class assignment. In addition, the training data for the ten classes contained in the SMD is strongly imbalanced and there are only few samples for some of the ten classes. We consider this issue by evaluating object detection using the SMD not as a ten-class but as a two-class (object vs. background) problem within our experiments in Section 4. Finally, the number of individual objects is rather small. According to the annotations for object tracking, there are not more than 534 individual objects contained in the SMD considering both the VIS and the NIR spectrum. This can increase the danger of overfitting during training. We handle this issue by making sure that an object that is contained in multiple videos belongs either to the training or to the test dataset during data split in Section 3.1. We also consider that it is not beneficial for object detection to utilize each frame of a video for training a DCNN but every second [24].

3.1. Training and Test Data

To utilize the SMD for deep learning, a training, validation, and test set is needed. As there is no literature available providing such a split, we propose it in this section. Table 2 shows the training/validation/test split for all videos captured in the VIS spectrum. As the number of videos is rather small, the videos captured on-shore and on-board are combined. As there are several videos containing the same object, those videos should be in the same subset, i.e. either training or test. The grouped videos highlighted in blue color in Table 2 contain identical objects. Although there is a large number of videos with identical objects, it is possible to create a test set that contains no videos with identical objects and covers the available illumination conditions: hazy, daylight, and dark/twilight. Furthermore, it is possible to add videos with many (ten or more) and few (less than ten) objects per frame to each subset.

Table 3 shows the proposed data split for the NIR videos. There are videos containing identical objects, too. Those are again grouped and highlighted in blue color. Weather conditions are not taken into account here as it is difficult to determine them visually from the monochromatic images. As the number of NIR videos is even less compared to the VIS videos, no validation set is provided.

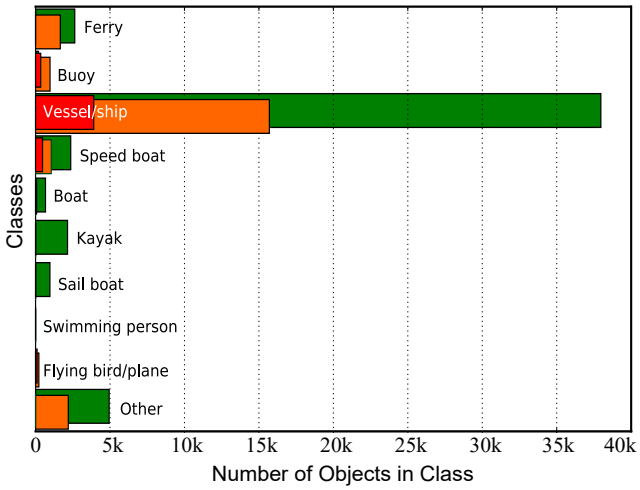

Figure 1. Class distribution for training (green), test (orange), and validation (red) set.

To analyze the quality of training and test set it is necessary to check if the classes are equally distributed. Figure 1 shows the training, validation, and test set's distribution of classes for the VIS data. Actually, there is a strong class imbalance. Even worse, there are classes that occur in only one of the two sets. Due to this strong class imbalance it is doubtful that a DCNN to detect multiple classes can be trained successfully. Another important property for object detection is object size. Figure 2 shows the bounding box areas present in the dataset. The majority of objects has a small size of 4,000 pixels or less. Hence, the dataset is more challenging as small objects are more difficult to detect [36]. The test set has a similar size distribution compared to the training set. Finally, the width-to-height aspect ratio distribution is shown in Fig 3. There is a similar distribution of aspect ratios for training and test set.

Table 2. Proposed split into training, validation, and test data for the VIS videos. Grouped videos in blue color contain identical objects.

Set	Subset	Video Name	Condition	Frames	Labels	Objects
Training	**On-Shore**	MVI_1451	hazy	439	3,270	8
		MVI_1609	dark/twilight	505	9,072	20
		MVI_1452	hazy	340	1,700	5
		MVI_1610	daylight	543	3,166	6
		MVI_1478	daylight	477	2,901	7
		MVI_1479	daylight	206	1,271	7
		MVI_1481	daylight	409	3,095	9
		MVI_1482	daylight	454	2,460	6
		MVI_1584	dark/twilight	550	7,320	14
		MVI_1613	daylight	626	6,574	12
		MVI_1614	daylight	582	6,957	13
		MVI_1615	daylight	566	3,843	8
		MVI_1617	daylight	600	5,940	14
		MVI_1619	daylight	473	2,838	6
		MVI_1620	daylight	502	3,012	6
		MVI_1583	dark/twilight	251	3,186	13
		MVI_1622	daylight	309	1,103	4
		MVI_1623	daylight	522	3,094	6
		MVI_1587	dark/twilight	600	8,858	15
		MVI_1624	daylight	494	1,976	4
		MVI_1625	daylight	995	8,111	11
		MVI_1592	dark/twilight	491	3,629	8
		MVI_1644	daylight	252	1,764	7
		MVI_1645	daylight	535	3,210	6
		MVI_1646	daylight	520	4,533	9
	On-Board	MVI_0801	daylight	600	919	2
Validation	**On-Shore**	MVI_1469	daylight	600	5,947	11
		MVI_1578	dark/twilight	505	3,535	7
	On-Board	MVI_0790	daylight	1,010	597	1
Test	**On-Shore**	MVI_1448	hazy	604	5,443	10
		MVI_1474	daylight	445	6,674	15
		MVI_1484	daylight	687	2,748	4
		MVI_1486	daylight	629	6,713	11
		MVI_1582	dark/twilight	540	6,480	12
		MVI_1612	daylight	261	2,514	10
		MVI_1626	daylight	556	5,329	12
		MVI_1627	daylight	600	4,200	7
		MVI_1640	daylight	310	2,183	9
	On-Board	MVI_0797	daylight	600	1258	3

3.2. Instance Segmentation Labels

Multitask learning considering object detection and instance segmentation is beneficial for both tasks' performance [22]. To utilize this promising approach, an instance segmentation GT is needed, which is not included in the SMD's annotations. In recent literature [41, 28, 34], it was proposed to generate this instance segmentation GT semi-automatically: GrabCut and Multiscale Combinatorial Grouping (MCG) are used to create pixelwise instance segmentations from bounding box annotations. However, our experiments showed that GrabCut does not work well for images captured in maritime environments since a large

Table 3. Proposed split into training and test data for the NIR videos. Grouped videos in blue color contain identical objects.

Set	Video Name	Frames	Labels	Objects
Training	MVI_1523	600	5,960	11
	MVI_1524	579	6,028	28
	MVI_1525	566	3,562	7
	MVI_1526	600	2,154	4
	MVI_1463	317	6,324	20
	MVI_1527	602	5,864	14
	MVI_1528	600	2,207	7
	MVI_1532	295	852	3
	MVI_1529	478	2,868	6
	MVI_1530	497	2,485	5
	MVI_0895	440	3,201	9
	MVI_1538	417	1,599	4
	MVI_1539	601	3,606	6
	MVI_1541	508	7,789	16
	MVI_1552	799	2,584	4
	MVI_1550	534	3,738	7
	MVI_1551	520	4,680	9
Test	MVI_1468	349	3,032	9
	MVI_1520	541	2,573	5
	MVI_1521	600	3,600	6
	MVI_1522	262	2,519	10
	MVI_1545	307	3,483	14
	MVI_1548	274	2,466	9

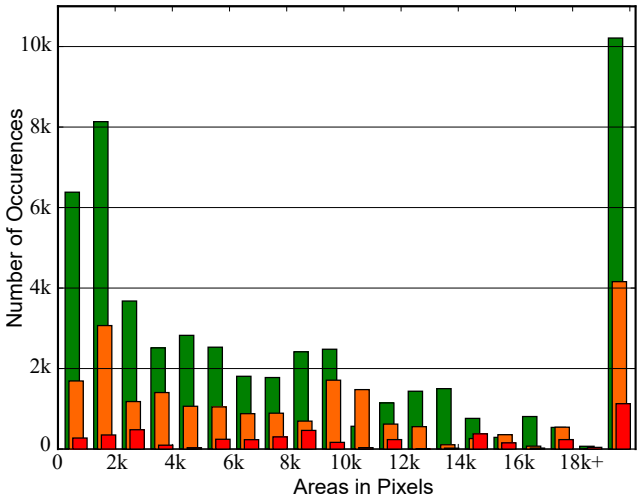

Figure 2. Distribution of bounding box areas for training (green color), test (orange color), and validation (red color) set.

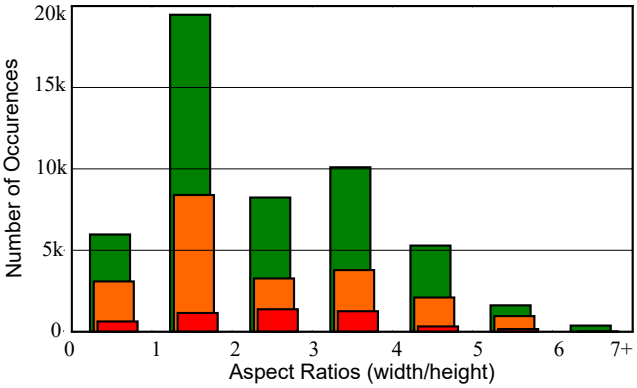

Figure 3. Distribution of bounding box aspect ratios for training (green), test (orange), and validation (red) set.

Algorithm 1: Local background subtraction to create instance segmentation labels from GT boxes.

Input: image, ground truth boxes
Output: pixelwise instance segmentation

annotations $\leftarrow \varnothing$;
foreach box_{gt} **do**
  $x_{min},\ y_{min},\ y_{max} \leftarrow$ `extrema` (box$_{gt}$);
  box$_{low} \leftarrow$ `box` (x_{min}, y_{max}, w, h_c);
  box$_{up} \leftarrow$ `box` ($x_{min}, y_{min} - h_c, w, h_c$);
  $\mu_{low},\ \sigma^2_{low} \leftarrow$ `gauss` (box$_{low}$, image);
  $\mu_{up},\ \sigma^2_{up} \leftarrow$ `gauss` (box$_{up}$, image);
  $d_{min} \leftarrow$ `dist_thresh` ();
  for $i \leftarrow 0$ **to** *max_iter* **do**
    $d_{min} \leftarrow d_{min} - i \cdot x$;
    for $p \leftarrow 0$ **to** `area` *(box$_{gt}$)* **do**
      mask[p] $\leftarrow 0$;
      $d_{low} \leftarrow$ `dist` (image[p], $\mu_{low}, \sigma^2_{low}$);
      $d_{up} \leftarrow$ `dist` (image[p], μ_{up}, σ^2_{up});
      if $d_{low} > d_{min}$ **and** $d_{up} > d_{min}$ **then**
        mask[p] $\leftarrow 1$;
      end
    end
    mask $\leftarrow$ `morph_opening` (mask, kernel);
    mask $\leftarrow$ `morph_closing` (mask, kernel);
    contour $\leftarrow$ `contour` (mask);
    if `area` *(contour)* $> 0.1 \cdot$ `area` *(box$_{gt}$)*
      then break;
  end
  annotations $\leftarrow$ `add_contour` (contour);
end
return annotations

number of pixelwise False Positive (FP) annotations occurs as seen in Fig. 4.

Instead, we propose Algorithm 1 to create instance segmentation labels for the SMD. For each GT box, we create an adjacent upper and lower box of constant height h_c. Within each box, we calculate the pixel value mean and variance (three dimensional in RGB for VIS). This multivariate Gaussian normal distribution for each box represents a simple background model for sky and sea. Each pixel in-

side the GT box is compared to each of the two background models using the Mahalanobis distance. We mark the pixel as foreground if both distances are sufficiently large. Clustering foreground pixels gives us the object's instance segmentation. GrabCut can be initialized with a set of pixels that definitely contain foreground [46]. Although Algorithm 1 tends to produce rather FN pixels than FP pixels, it turned out that it is still impossible to use the GrabCut for object segmentation even with an initialization using our instance segmentation. Hence, we directly use the clusters as instance segmentation labels.

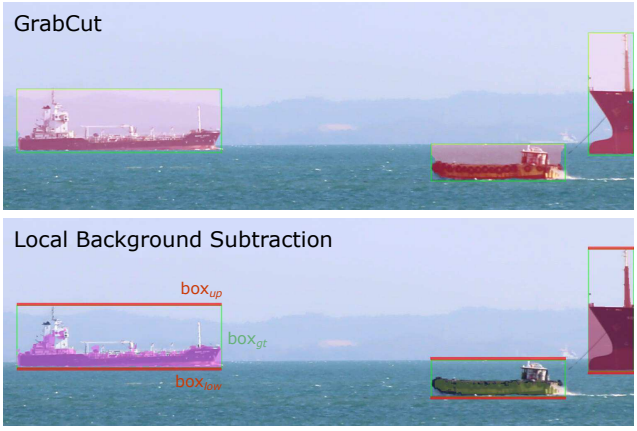

Figure 4. Instance segmentation labels created by GrabCut and Algorithm 1. While GrabCut is prone to produce FP segmentations right above the vessels, Algorithm 1 tends to produce FNs. The depicted boxes correspond to the ones mentioned in Algorithm 1.

Figure 4 shows the improvements by using the introduced approach to create instance segmentation labels. The red boxes above and underneath each green GT bounding box show the areas used for modeling the background. The new approach significantly reduces the number of FP pixels compared to GrabCut.

4. Experiments and Results

In this paper, Faster R-CNN and Mask R-CNN are evaluated for object detection in maritime environments. Both DCNNs use ResNet-101 [23] as backbone and are pre-trained using ImageNet and COCO that both contain many maritime objects. The models are taken from and trained using the Detectron [17] framework. We replace the fully-connected output layer of each DCNN with two fully-connected output neurons for the two classes *object* and *background*. We also tested using eleven neurons with the ten classes provided by SMD and background. However, since the f-score was about 0.03 lower on average compared to the two-class problem, we discarded this approach.

Three different width/height aspect ratios are available for the anchors provided by the pre-trained object detection DCNNs: 0.5, 1, and 2. In order to analyze if these ratios fit

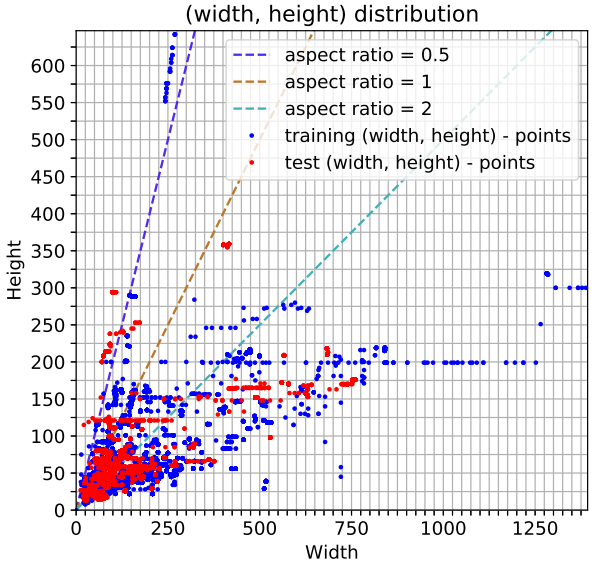

Figure 5. Training and test samples plotted as width/height points and the three RPN anchor aspect ratios depicted as straight lines.

well to the dominant aspect ratios of the objects within the dataset, we plot the training and test samples given by the width and height of their related bounding boxes in Fig. 5. The three anchor ratios are visualized as three straight lines since we apply them scale invariant using a Feature Pyramid Network (FPN) [35]. The object aspect ratios within the dataset are covered well by the provided anchors. However, it could be beneficial to use a fourth anchor aspect ratio of 3 to improve the detection of objects with a much larger width than height. We tried to introduce such an anchor but discovered that the dataset is definitely not sufficient to train the large number of new parameters that appear due to the introdution of the fourth anchor. As a result, the detection rate dramatically dropped and thus we discarded this approach for manipulating the anchors.

We evaluate using precision, recall, and f-score [27]. The Faster R-CNN is then evaluated for **(1)** fine-tuning only (FRCNN). As Lin *et al*. [35] demonstrated that FPNs improve object detection performance especially for small objects in the image, we not only apply an FPN within the Mask R-CNN but also within the Faster R-CNN. To evaluate the influence of model weights pre-trained for object detection and instance segmentation, **(2)** the Mask R-CNN is used (MRCNN w/o segm.). As the SMD does not provide instance segmentation labels for fine-tuning, we disable the segmentation branch for this training. To further improve the results, the semi-automatically determined instance segmentation labels are used to train the Mask R-CNN including its segmentation branch. This DCNN is evaluated for **(3)** fine-tuning with re-initializing only class dependent lay-

ers (MRCNN finetuned) and for **(4)** fine-tuning with re-initialization of the segmentation branch's deconvolution layer together with the fully connected layers (MRCNN re-init). Other combinations of re-initialization were tested but appeared not to be promising. For the last experiment, **(5)** the weakly supervised recursive training approach of Khoreva *et al*. [28] is adapted (MRCNN recursive). This approach is related to the Expectation-Maximization (EM) algorithm and iteratively improves the DCNN model and the instance segmentation labels simultaneously. As this approach is applied to object detection in the maritime domain for the first time, we perform an ablation study for weighting the detection and the segmentation loss within the loss function during training. As we cannot be sure about the quality of the semi-automatically determined instance segmentation labels, we need to assure that the segmentation loss is not dominating and thus biasing the training process. Figure 6 shows the resulting precision-recall curves. We evaluate three different weightings: Loss Ratio (LR) 1/2 means that the detection loss is weighted half compared to the segmentation loss. This is recommended for faster training [28]. For LR 2/1, the segmentation loss is weighted half compared to the detection loss. The f-scores show no significant performance difference. For comparison, we also evaluate a naïve initialization for instance segmentation using the entire GT rectangle (rect. init.). We did not evaluate GrabCut as we discovered that regularly either all pixels or no pixel inside a GT box was segmented as foreground. There is no significant difference between the approaches regarding the maximum f-score. As a consequence, weakly supervised recursive training seems to be equally harmed by either FPs or FNs. Since seg. init. LR 1/2 consistently performed best, we choose this approach and this ratio for the next experiments.

All experiments are performed using the same hyper-parameters as introduced by Goyal *et al*. [18]. Epoch dependent hyper-parameters are linearly scaled. The best result of each experiment is shown in Figure 7. Mask R-CNN with its segmentation branch disabled (colored blue) performs best overall with a rather large margin for one epoch of training and no additional re-initialization. This is remarkable since it shows that the Mask R-CNN is well-generalizing across datasets without the necessity of extensive fine-tuning. One reason could be that the maritime objects contained in ImageNet and COCO provide a good basis. Training the Mask R-CNN for five epochs with re-initialized fully connected and deconvolution layers (colored purple) performed second best overall. Interestingly, the Faster R-CNN trained for ten epochs and without re-initialization of the fully connected layers (colored red) is able to outperform the Mask R-CNN without re-initialization trained for five epochs (colored orange). Training the Mask R-CNN using the weakly supervised recursive training strategy [28] with ten iterations is not able to outperform any of the other results using this approach.

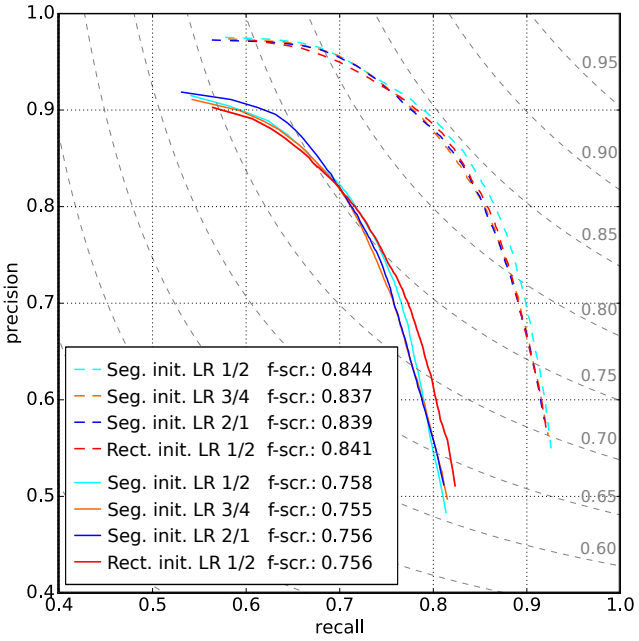

Figure 6. Evaluation of three different detection and segmentation loss ratios (LR) for weakly supervised recursive training [28] (MRCNN recursive) initialized with the instance segmentation labels (seg. init.) created by Algorithm 1. Rect. init. represents the naïve initialization using the entire GT box. Solid curves represent results for an IoU threshold of 0.5 and dashed curves for 0.3.

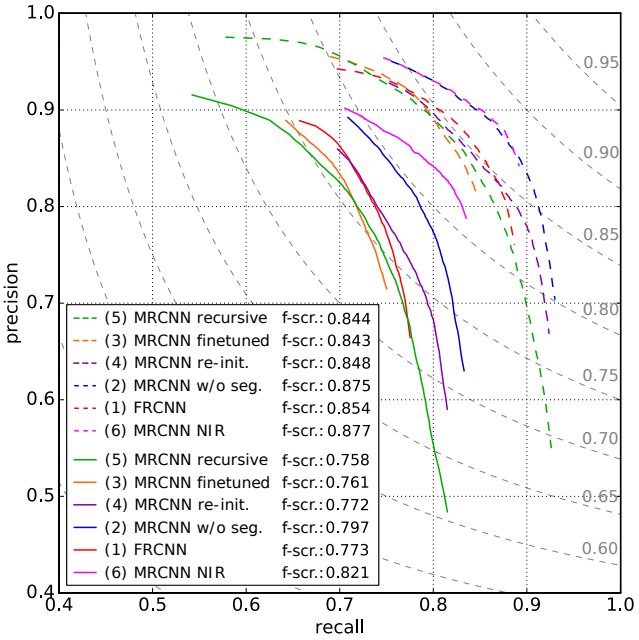

Figure 7. Best results for each experiment represented by precision-recall curves and maximum f-score. Solid curves represent results for an IoU threshold of 0.5 and dashed curves for 0.3. The dashed grey curves show the f-score.

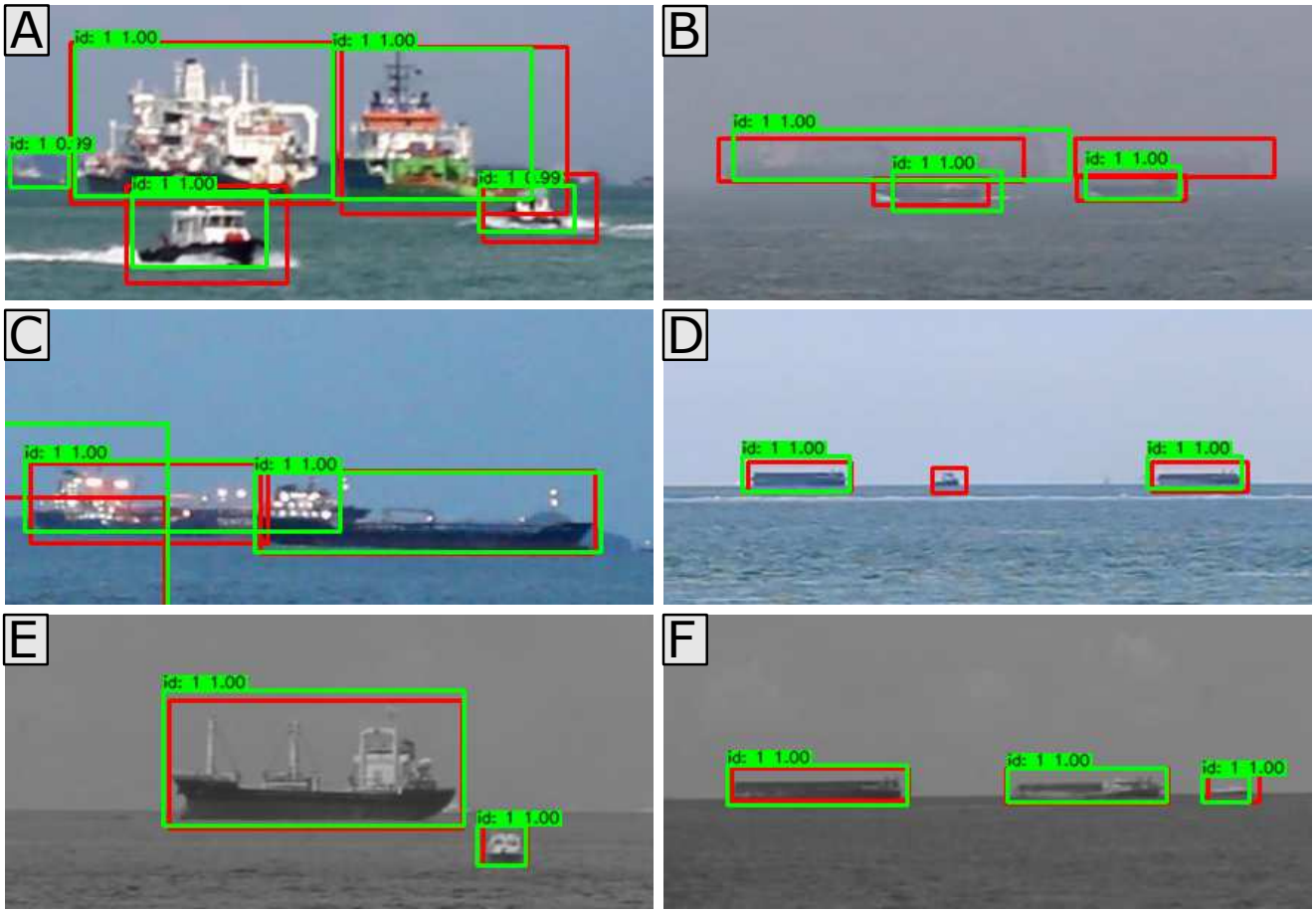

Figure 8. Qualitative evaluation using the Mask R-CNN with disabled segmentation branch for VIS and NIR images.

To evaluate **(6)** the object detection performance for the NIR spectrum, the best performing parameterization is adopted from training in the VIS spectrum (MRCNN NIR). The result is shown in Fig. 7 (colored pink). As the images from both spectra are not aligned the results are not comparable and multispectral detection is not possible. Nevertheless, the performance on NIR images is surprisingly good without any hyper-parameter tuning.

Figure 8 shows crops of four VIS and two NIR images for the qualiative evaluation. Detection bounding boxes are colored green and Ground Truth (GT) bounding boxes red. All detection bounding boxes are taken from inferencing the Mask R-CNN. As seen in Fig. 8 (B), hazy images are very challenging and one of the main reason for the occurrence of FNs. Tiny objects are another issue and produce FNs as seen in Fig. 8 (D). Figure 8 (E) shows a detection result for the NIR spectrum. The contrast between objects and background is rather low. Nevertheless, the DCNN is able to detect the objects correctly. In general, tiny and occluded objects are most challenging especially under hazy conditions and produce most of the FN detections in both the VIS and the NIR spectrum.

5. Conlusions

In this paper, we presented a novel benchmark for deep learning based object detection in the maritime domain utilizing the SMD. Drawbacks of the current SMD annotations were discussed and extensions suggested and published. Furthermore, a data split into training, validation, and test data was proposed for the VIS and the NIR spectrum of the SMD. Using a novel object segmentation algorithm tuned specifically for maritime scenes, instance segmentation labels were generated semi-automatically for weakly supervised recursive Mask R-CNN training. Best performance with a maximum f-score of 0.875 and 0.877 in the VIS and the NIR spectrum, respectively, was achieved by a fine-tuned Mask R-CNN with disabled instance segmentation branch. This actually is nothing else than a Faster R-CNN initialized with Mask R-CNN weights and subsequently fine-tuned. The result indicates that the Mask R-CNN is not only a powerful but also a well-generalizing DCNN for object detection. Our new benchmark [1] aims at improving reproducibility and comparability of research for object detection in maritime environments.

References

[1] Maria Andersson, Ronnie Johansson, Karl-Göran Stenborg, Robert Forsgren, Thomas Cane, Grzegorz Taberski, Luis Patino, and James Ferryman. The IPATCH System for Maritime Surveillance and Piracy Threat Classification. In *European Intelligence and Security Informatics Conference (EISIC)*, 2016. 1, 2

[2] Anurag Arnab and Philip H. S. Torr. Pixelwise Instance Segmentation with a Dynamically Instantiated Network. In *CVPR*, 2017. 1

[3] Domenico D. Bloisi, Luca Iocchi, Michele Fiorini, and Giovanni Graziano. Automatic maritime surveillance with visual target detection. In *Proc. of the International Defense and Homeland Security Simulation Workshop (DHSS)*, pages 141–145, 2011. 2

[4] Domenico D. Bloisi, Luca Iocchi, Andrea Pennisi, and Luigi Tombolini. ARGOS-Venice Boat Classification. In *IEEE AVSS*, 2015. 1, 2

[5] Henri Bouma, Dirk-Jan de Lange, Sebastiaan van den Broek, Rob Kemp, and Piet Schwering. Automatic detection of small surface targets with electro-optical sensors in a harbor environment. In *Proc. of SPIE Vol. 7114*, 2008. 2

[6] Borja Bovcon, Rok Mandeljc, Janez Perš, and Matej Kristan. Stereo obstacle detection for unmanned surface vehicles by imu-assisted semantic segmentation. *Robotics and Autonomous Systems*, 104:1–13, 2018. 2

[7] Jörg Brauchle, Steven Bayer, and Ralf Berger. Automatic ship detection on multispectral and thermal infrared aerial images using macs-mar remote sensing platform. In *Pacific-Rim Symposium on Image and Video Technology*, 2017. 1

[8] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating pedestrians via simultaneous detection & segmentation. In *ICCV*, 2017. 1

[9] Massimo Caccia. Autonomous surface craft: prototypes and basic research issues. In *Mediterranean Conference on Control and Automation*, 2006. 1

[10] Sable Campbell de Oliveira, Wasif Naeem, and George W. Irwin. A review on improving the autonomy of unmanned surface vehicles through intelligent collision avoidance manoeuvres. *Annual Reviews in Control*, 36(2):267–283, 2012. 1

[11] Giuseppe Casalino, Alessio Turetta, and Enrico Simetti. A three-layered architecture for real time path planning and obstacle avoidance for surveillance usvs operating in harbour fields. In *IEEE Oceans*, 2009. 1

[12] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE TPAMI*, 34(4):743–761, 2012. 1

[13] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *IJCV*, 111(1):98–136, 2015. 1

[14] Sergiy Fefilatyev, Volha Smarodzinava, Lawrence O. Hall, and Dmitry Goldgof. Horizon detection using machine learning techniques. In *International Conference on Machine Learning and Applications (ICMLA)*, 2006. 2

[15] Duncan Frost and Jules-Raymond Tapamo. Detection and tracking of moving objects in a maritime environment using level set with shape priors. *EURASIP Journal on Image and Video Processing*, 2013(1), 2013. 2

[16] Ross B. Girshick. Fast R-CNN. In *ICCV*, 2015. 2

[17] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 2, 6

[18] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017. 7

[19] Erhan Gundogdu, Berkan Solmaz, Veysel Yücesoy, and Aykut Koc. Marvel: A large-scale image dataset for maritime vessels. In *ACCV*, 2016. 1, 2

[20] Jarmo Hakala. Object Recognition for Maritime Application using Deep Neural Networks. Master's thesis, Tampere University of Technology, 2018. 2

[21] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 1

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, 2017. 2, 4

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[24] Jan Hendrik Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. Taking a deeper look at pedestrians. In *CVPR*, 2015. 3

[25] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In-So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, 2015. 1

[26] Chiyoon Jeong, Hyun S. Yang, and KyeongDeok Moon. A novel approach for detecting the horizon using a convolutional neural network and multi-scale edge detection. *Multidimensional Systems and Signal Processing*, 24(4):181, 2018. 1, 2

[27] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina N. Korzhova, and Jing Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: data, metrics, and protocol. *IEEE TPAMI*, 31(2):319–336, 2009. 6

[28] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017. 1, 2, 4, 7

[29] Kwanghyun Kim, Sungjun Hong, Baehoon Choi, and Euntai Kim. Probabilistic ship detection and classification using deep learning. *Applied Sciences*, 8(6), 2018. 2

[30] Daniel König, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *CVPR Workshops*, 2017. 1

[31] Matej Kristan, Vildana Sulic Kenk, Stanislav Kovacic, and Janez Pers. Fast image-based obstacle detection from unmanned surface vehicles. *IEEE Transactions on Cybernetics*, 46(3):641–654, 2016. 1, 2

[32] Kristoffer Kleven Krossholm. Unsupervised Object Detection in Images from Maritime Environments. Master's thesis, Norges teknisk-naturvitenskapelige universitet, 2017. 1, 2

[33] Maxime Leclerc, Ratnasingham Tharmarasa, Mihai Florea, Anne-Claire Boury-Brisset, Thia Kirubarajan, and Nicolas Duclos-Hindie. Ship classification using deep learning techniques for maritime target tracking. In *International Conference on Information Fusion (FUSION)*, 2018. 2

[34] Qizhu Li, Anurag Arnab, and Philip H.S. Torr. Weakly- and Semi-Supervised Panoptic Segmentation. In *ECCV*, 2018. 4

[35] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 6

[36] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312, 2014. 1, 3

[37] Vincent Marie, Ikhlef Bechar, and Frdric Bouchara. Real-time maritime situation awareness based on deep learning with dynamic anchors. In *IEEE AVSS*, 2018. 1, 2

[38] Anurag Mittal and Nikos Paragios. Motion-based background subtraction using adaptive kernel density estimation. In *CVPR*, 2004. 2

[39] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 2

[40] Francesco Palmieri, Francesco Castaldo, and Guglielmo Marino. Harbour surveillance with cameras calibrated with AIS data. In *IEEE Aerospace Conference*, 2013. 1

[41] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015. 4

[42] Dilip K. Prasad, C. Krishna Prasath, Deepu Rajan, Lily Rachmawati, Eshan Rajabaly, and Chen Quek. Challenges in video based object detection in maritime scenario using computer vision. In *arXiv:1608.01079*, 2016. 2

[43] Dilip K. Prasad, Deepu Rajan, Lily Rachmawati, Eshan Rajabally, and Chai Quek. Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 18(8):1993–2016, Aug 2017. 1, 2

[44] Lei Ren, Chaojian Shi, and Xin Ran. Target detection of maritime search and rescue: saliency accumulation method. In *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2012. 2

[45] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 2015. 2

[46] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut": interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH*, 2004. 6

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 2

[48] Thomas Statheros, Gareth Howells, and Klaus McDonald Maier. Autonomous ship collision avoidance navigation concepts, technologies and techniques. *The Journal of Navigation*, 61(1):129–142, 2008. 1

[49] Da Tang, Gang Sun, Ding-He Wang, Zhao-Dong Niu, and Zeng-Ping Chen. Research on infrared ship detection method in sea-sky background. In *Proc. of SPIE Vol. 8907*, 2013. 2

[50] Espen Johansen Tangstad. Visual Detection of Maritime Vessels. Master's thesis, Norges teknisk-naturvitenskapelige universitet, 2017. 1, 2

[51] Michael Teutsch and Wolfgang Krüger. Classification of small boats in infrared images for maritime surveillance. In *International Conference on WaterSide Security (WSS)*, 2010. 2

[52] Tanja van Valkenburg-van Haarst and Krispijn Scholte. Polynomial background estimation using visible light video streams for robust automatic detection in a maritime environment. In *Proc. of SPIE Vol. 7482*, 2009. 2

[53] Xiaoping Wang and Tianxu Zhang. Clutter-adaptive infrared small target detection in infrared maritime scenarios. *SPIE Optical Engineering*, 50(6), 2011. 2

[54] Hai Wei, Hieu Nguyen, Prakash Ramu, Chaitanya Raju, Xiaoqing Liu, and Jacob Yadegar. Automated intelligent video surveillance system for ships. In *Proc. of SPIE Vol. 7306*, 2009. 2

[55] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, 2016. 1

[56] Dian Zhang, Edel O'Connor, Kevin McGuinness, Noel Edward O'Connor, Fiona Regan, and Alan Smeaton. A visual sensing platform for creating a smarter multi-modal marine monitoring network. In *ACM International Workshop on Multimedia Analysis for Ecological Data*, 2012. 2

[57] Mabel M. Zhang, Jean Choi, Kostas Daniilidis, Michael T. Wolf, and Christopher Kanan. VAIS: A Dataset for Recognizing Maritime Imagery in the Visible and Infrared Spectrums. In *CVPR Workshops*, 2015. 2

[58] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. CityPersons: A Diverse Dataset for Pedestrian Detection. In *CVPR*, 2017. 1, 3

[59] 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 Transactions on Geoscience and Remote Sensing*, 48(9):3446–3456, 2010. 2