Self-Learning Framework with Temporal Filtering for Robust Maritime Vessel Detection

Amir Ghahremani†, Egor Bondarev, and Peter H.N. de With

Eindhoven University of Technology, Eindhoven, the Netherlands
†A.Ghahremani@tue.nl

Abstract. With the recent development in ConvNet-based detectors, a successful solution for vessel detection becomes possible. However, it is essential to access a comprehensive annotated training set from different maritime environments. Creating such a dataset is expensive and time consuming. To automate this process, this paper proposes a novel self learning framework which automatically finetunes a generic pre-trained model to any new environment. With this, the framework enables automated labeling of new dataset types. The method first explores the video frames captured from a new target environment to generate the candidate vessel samples. Afterwards, it exploits a temporal filtering concept to verify the correctly generated candidates as new labels for learning, while removing the false positives. Finally, the system updates the vessel model using the provided self-learning dataset. Experimental results on our real-world evaluation dataset show that generalizing a fine-tuned Single Shot Detector to a new target domain using the proposed self-learning framework increases the average precision and the F1-score by 12% and 5%, respectively. Additionally, the proposed temporal filter reduced the noisy detections in a sensitive setting from 58% to only 5%.

Keywords: maritime surveillance, vessel detection, convolutional networks (CNN), ConvNet, self-learning, automated dataset creation

1 Introduction

With recent advances in automated surveillance systems, maritime and harbor authorities start actively exploiting machine vision techniques. In such an advanced systems, an important application is to monitor the maritime environment against contingent hazards jeopardized by unknown pathless watercrafts. In order to realize this, surveillance systems have to process and analyze the visual data collected by cameras deployed along the shorelines.

Surveying the conventional visual monitoring methodologies, object detection is routinely regarded as the first main task [1,2]. Common historic methods have achieved robust results by exploiting regional variations of pixels, as a distinctive indication of moving object presence (e.g. background subtraction approaches) [3,4]. However, the fluctuating nature of water as a dominant background for a typical maritime scene leads to failure when using such conventional

detection methods. Additionally, a maritime surveillance camera does not only capture water, but also land pieces and infrastructure. Consequently, irrelevant objects moving in non-water regions would initiate or cause false positives. As a strategy for handling this challenge, irrelevant objects could be neglected using methods proposed to extract clusters of water pixels as regions of interest [5]. However, falsely detected/missed regions triggered by complex scenes and scenarios still expose the system to detect unrelated objects. Moreover, maritime scenes often contain stationary vessels next to the shorelines, which are not detected either.

Contemporary development in convolutional neural networks (ConvNets) have substantially refashioned automated object detection procedures by deliberately seeking for the anticipated target patterns according to their inherent properties [6, 7]. ConvNet detectors principally pursue the following scheme: feature extraction, bounding box generation, and classification. Among the state-of-the-art detectors, Single Shot Detector (SSD) [8] exceedingly outperforms its competitors in terms of speed and achieves satisfactory detection accuracy. In [8], the network has been successfully evaluated on several classical benchmark image sets. Although the manifold categories of objects are covered in those datasets, after investigating the samples, one can notice that images have a low resolution and are predominantly encompassing the intended object. Moreover, the challenging outdoor surveillance cases like complex background, miscellaneous weather conditions, divergent occlusion scenarios, multiple various-sized objects, different object distances to cameras, are not represented in classical benchmark samples. However, our research is based on the European APPS research project, aiming at industry-oriented Advanced Plug & Play Smart surveillance systems, where we consider all previously mentioned complex maritime surveillance scenarios.

Within the discussed setting, our objective is to enhance object detection with improved analysis based on deep learning. However, in first experiments, testing a pre-trained ConvNet model on the scenes having different characteristics from the original training set often failed to provide acceptable results. Consequently, the development of scene-specific object detectors has recently emerged as an attractive research topic pursued in many state-of-the-art publications [9–12]. These specific methods commonly attempt to automatically assemble appropriate samples from a target domain and then re-train the available model. In accordance with this, our main contribution is to extend the SSD to the ship detection problem and design a new transfer learning framework to achieve high precision in detection at a low false negative rate.

In this paper, we aim at exploiting ConvNets to detect vessels on genuine maritime surveillance image sequences. Initially, we found out that specific datasets dedicated to one harbor often cover a few camera viewpoints only (sometimes even from the same location) and show vessel types that are partly restricted and dominant for the related specific industrial harbor area. As a consequence, the training with such datasets leads to a specific detector that is not suited for a broad set of harbor areas because of limitations in camera setup and intrinsic

parameters, which all leads to a lower performance for other environments.

In this paper, we therefore generalize the finetuned SSD on arbitrary scenes, scenarios and vessel types, and we propose a novel self-learning framework for maritime surveillance applications. The proposed system adopts the following blueprint. Firstly, it generates candidate samples from a new dataset using the finetuned vessel-oriented SSD architecture. These supplementary images are captured by a different camera in varying setups based on disparate locations and having various contexts. Secondly, the generated false positive candidates are discarded by endorsing the samples labeled as correct. To perform this, a dual-condition criterion is employed: a) evaluate the confidence score of detections, and b) apply a temporal filtering strategy to investigate the dynamics of the detected box over the sequence. Finally, the network enriches the verified sample set by adding images from successfully learned source data. This preserves the system from losing the already learned useful source information. Additionally, the model will be corrected on the source samples which unexpectedly prompt false detections. We evaluate the method on an annotated image set from various locations in several cities and suburbs in the Netherlands (including Amsterdam and Rotterdam) and Turkey (Istanbul). The images of this dataset are extracted from videos captured at different day/year-times. These images include objects with divergent size, captured from different camera positions and setups, appearing occlusions, object truncations, etc. Fig. 1 illustrates a few examples extracted from this dataset.

This paper is organized as follows. Section 2 provides an overview on related work. Section 3 explains the proposed method. Section 4 presents the experimental results and validation. Section 5 concludes the paper.

2 Related Work

This section provides a brief overview on the research work on ConvNet transfer learning that performs the finetuning to adapt the object detector to the specific scenes.

Deep-learning based classifiers are widely exploited in many practical applications [6] because of their advantageous. Furthermore, also for ConvNets, this development has encouraged industry-oriented researchers to deploy them even in products. As a basic requirement, ConvNets need to be trained by suitable training sets. However, challenging practical cases are often not represented in the data produced in a laboratory. If those datasets are limited, semi-supervised and weakly-supervised learning-based methods can be exploited.

Self-learning aims to automatically sample data from a target domain and finetune a detector for the specific visual patterns. The main assumption is that finetuning a generic pre-trained detector with automatically extracted labels from an arbitrary set of target domain images, would adapt the detector to the target environment conditions.In [9], the transfer learning methods tuning a detector to a specific target domain, are categorized into three main groups.

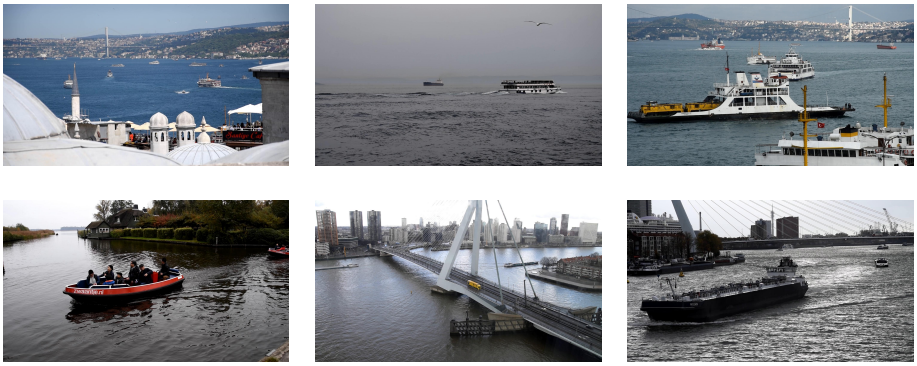

Fig. 1: Six example images from Evaluation set.

The methods falling into the first group [13], adjust the source learning parameters to enhance the model accuracy in an objective scope. These methods exploit prior knowledge about source data like visual information. The second group [14] aims at reducing the dissimilarities between the source and target domains by exploiting techniques for adapting distributions, i.e. to manipulate the marginal and conditional distributions to reduce the data dimension in both domains. In order to improve the ConvNet performance on a target domain, the third category of methods [9] enhances the training set by appending appropriate samples from the target domain. With this definition, our work belongs to the third group, since it automatically labels the data from the target domain.

Augmenting the complete source dataset with new samples extracted from a target environment increases the size of the training set and requires more iterations to convergence [13]. The work in [15] deploys a combination of the source samples together with new samples from a target scene. The method proposed in [11] collects only new samples from a target domain to produce the transfer learning dataset. Obviously, this method looses the advantageous information of the source data. In [16], the method gathers new samples from a target domain and combines those with the beneficial source samples only. Other methods exploit information like visual cues and contextual attributes, motion features, to enrich the training set by selecting useful samples from a target domain [17]. Additionally, the method in [10] deploys a sequential Monte Carlo filter to specialize a generic classifier to the specific scenes.

SSD (Single Shot Detector) is a feedforward ConvNet that evaluates the presence of an object instance in the pre-defined default bounding boxes, followed by a non-maximum suppression stage to produce the final detection. This detector allows to omit the region proposal generation stage, encapsulating all the computations in a single network. As stated in [8], SSD achieves 76.9% mAP on the PASCAL dataset. This ConvNet has also proved to be a fast general object detector, which is essential for real-time surveillance applications.

In this paper, we extend the SSD network to address multiple vessel detection

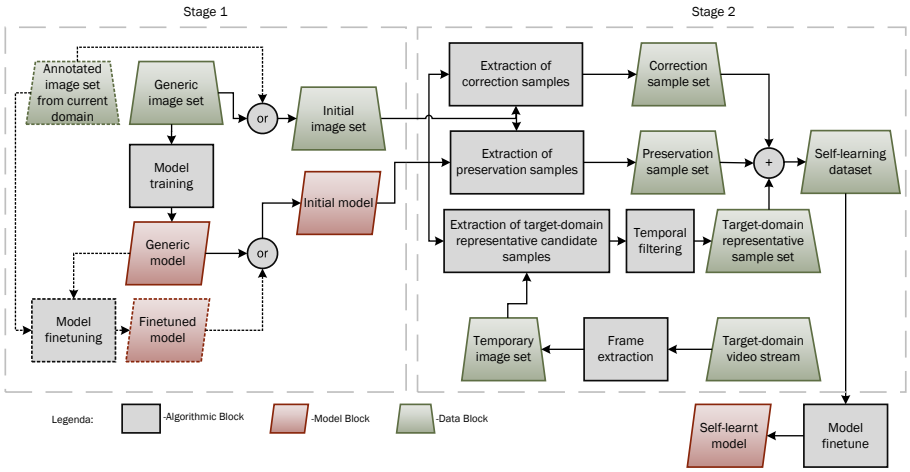

Fig. 2: Architecture of the proposed self-learning method.

problems. Firstly, the SSD network is pre-trained using the VGG-16 model [18] and will be finetuned on more than 48,000 maritime images. Then, we follow a bootstrapping procedure to improve the efficiency of the finetuned model on the training set. Additionally, we propose a novel target-domain specialization framework to automatically adapt the network to any dataset, captured by cameras with different intrinsic and setup. We also provide a challenging evaluation set, which consists of 1,041 annotated maritime surveillance images.

3 Architecture Pipeline

3.1 System Overview

Fig. 2 illustrates the architecture of the proposed system, containing two main stages. In the first stage, an initial deep model of vessels is trained on top of the SSD using an image set collected from the current domain. We refer to these images as the initial dataset in the remainder of this paper. Since the main idea of this work is to propose a robust self-learning framework to generalize an initial model (pre-trained on the current domain) to a new target domain, it is not important which initial dataset is used in the first stage. It can be a generic benchmark dataset (e.g. PASCAL), or it can be collected and labeled from the current domain by the user. As mentioned before, we specifically focus on the vessel detection problem as part of our research for maritime surveillance within the industry-oriented APPS project. Therefore, we generate our initial model by finetuning a pre-trained PASCAL-based VGG model, using a non-public vessel-oriented dataset.

The second stage provides a self-learning block, which automatically selects

the useful data from both the initial and the target-domain samples and finetunes the network. In the remainder of this paper, we will refer to these images as the "self-learning dataset". The second stage is divided into three sub-stages. The first and second sub-stage find those images from the initial dataset that result in false and true detections, respectively, to establish the self-learning set. The third sub-stage produces the candidate samples from the target domain and verifies them through a temporal filtering approach. Finally, the vessel model is updated with the self-learning set. The following subsections explain the dataflow and individual architectural modules in detail.

3.2 Deep Vessel Detection Model

A captured scene from a typical maritime environment contains various kinds of objects and structures (e.g. cars on the shorelines, bridges, buildings, etc.). In order to robustly detect all kinds of vessels independent from the surrounding environment, we employ ConvNet networks. For a surveillance application, the system should detect ships in real-time. Moreover, it should be robust against the real-world noisy data. Therefore, in this work, we have adopted the SSD detector. Here, the VGG model is used as the base model. We first finetune the network on a vessel-oriented image set. The system uses this as the initial model and supplies this to the second stage.

3.3 Self-Learning Process

The lack of labeled training data is a critical challenge for exploiting ConvNets in practical surveillance applications. Additionally, the performance of a trained deep detector often drops when testing for the new target domain, mainly when the scene structure or the camera characteristics change. To handle these challenges, self-learning has emerged as an interesting research topic among the state-of-the-art work.

In this second stage, we propose a novel self-learning data-augmentation framework for maritime surveillance applications. The input of this stage is the initial vessel-oriented model. Here, we aim at automated creation of the self-learning dataset to correct the initial vessel model and adapt the model to the new target domain. This second stage consists of three main parts, each appending new samples to the self-learning dataset.

Correction Samples The SSD is trained on the initial dataset during the first stage. However, when we apply the network to the initial vessel-oriented training images, we face high amounts of false positives/negatives. This is a common machine learning problem and often happens since both the labeled objects and the background produce similar features at the training phase. Nevertheless, after exploring the false detections, we have found several unexpected cases, where the vessel is clearly in front of the camera. Therefore, the system starts generating the self-learning dataset by adding randomly selected samples from the initial images causing false detections.

Useful Source Samples After applying the initial vessel model to the initial set, we observe plenty of images yielding just true positives. Obviously, the lack of those frames in the self-learning dataset deprives the system from useful information. Therefore, the system enriches the self-learning dataset using a random set of images including only those detected vessels. This ensures that the useful information from the source training set will be present among the self-training images.

Target-Domain Adaptation Samples Although the initial dataset includes more than 48,000 images covering several types of the weather conditions, all images share a similar background and are captured by the same camera. Consequently, both the precision and detection rate drop when the network is applied to a new target domain (i.e. images captured by a different camera or from a new environment containing new vessel types and/or background). In order to address this problem, we have employed a self-learning process.

First, we alternatively extract random frames from the new target domain to make another image set, which is referred to as the temporary set. The system automatically generates a large number of candidate labels from the temporary set for the target-domain adaptation purpose. Since the network is already fine-tuned for vessel detection, it detects most of the watercrafts located at a close distance to the camera. A small fraction of far vessels are also detected. Detection misses mostly occur on vessels positioning far away from the camera. However, we have noticed that the absence of missed detections in the self-learning dataset does not affect the finetuning performance, since those missed vessels are typically detected in the next frames.

An important aspect is temporal filtering of the candidate labels. Briefly, the proposed self-learning framework first applies the initial model to the new target domain to produce the annotation labels. However, since the initial model is trained on the initial dataset, it does not generate the label for most of the desired new target-domain objects. In order to increase the number of annotated objects, we perform the detection with a low confidence score. Nevertheless, such a score results in more noise regions. However, we have noticed that false positive detections rarely happen for water clusters. Many false positives arise on the shorelines, bridges, buildings, e.g. on the objects that have special vessel-like structures. In order to refute the false detections, we propose to extract and use information from the frame sequence. Since the near-shoreline regions remain mostly/partly stationary during short intervals (a few seconds), the proposed system removes those false detections by performing a temporal filtering technique to discard the detected boxes with non-dynamic content. The filter considers the average value of the subtracted pixels over the entire bounding box to investigate the variation in the region through the short-time intervals. If the calculated variation is higher than the noise threshold (T_N), the detected bounding box is maintained as a true label. Otherwise, the system neglects the box.

Although the proposed temporal filter removes most of the false detections,

still few labels of background regions remain among the correct labels with vessels. However, since the final signal-to-noise ratio is fairly high (see the Section 4), these defect labels do not considerably affect the framework performance. Additionally, the temporal filtering sometimes removes vessels moving at a far distance from the camera, since in short-time intervals those vessels' pixels do not change. This case can also happen for vessels standing next to the shorelines, especially when the water is not wavy. However, since we do not need to label all the visible vessels, this case is not critical. Concluding, according to the experimental results provided in the following section, the ratio between the correctly removed false detections and the incorrectly removed true positives, is rather high. Moreover, the ratio between removed background labels and the background labels remaining after the filtering, is also high.

In addition to the false detection boxes, the proposed system also ignores the objectless frames (i.e. the frames without detected vessels). At the end, the remaining frames will be added to the self-learning dataset. As the last step, the network will be finetuned with a low learning rate on the self-learning dataset.

4 Empirical Validation

This section begins with providing an overview on the experimental materials and process. Afterwards, we validate the proposed framework.

4.1 Experimental Process

A. Datasets

Botlek Dataset: In the first stage, the VGG-based SSD network is finetuned on our vessel-oriented initial dataset. Since this image set was captured from the Botlek region in the port of Rotterdam, we will refer to it as the Botlek dataset in the sequel of this paper. The Botlek dataset consists of 48,364 samples, which are extracted from videos shot in the Botlek region. The videos cover 6 different viewpoints on the region. Since the recordings were running for several months, a vast variety of weather conditions and daytimes are represented in this dataset. The camera model used in the recordings is Axis Q1604, which is a surveillance camera providing a resolution of $1,536 \times 2,048$ pixels, at 25 fps. Fig. 3 illustrates three Botlek example frames.

Evaluation Dataset: To improve robustness of the finetuned network, we have recorded several videos from various locations in many waterways (lakes, channels, rivers, sea sides) in the Netherlands (including Amsterdam and Rotterdam) and Turkey (Istanbul). These videos were recorded during different day/year-periods. The videos contain a vast variety of camera setups embracing different viewpoints and heights. Additionally, several vessel types and detection scenarios are represented, including multiple occluded vessels with divergent sizes and distances to the cameras. Furthermore, water region-types like rivers,

Fig. 3: Three example images from the Botlek dataset.

lakes, and under-bridges are covered. For the recordings, we have used the Canon D5500 camera with $1,080 \times 1,920$ pixel resolution. We have separated 50 videos for evaluation, randomly extracted 1,041 images and manually annotated those to make an evaluation dataset. It is important to mention that the videos used in the evaluation set are exclusively detached from the rest of the data.

Temporary Dataset: After separating the mentioned 50 videos for the evaluation set, we select another set of 20 videos from our new recordings to represent the new target domain. Then, we alternatively extract random frames with short-time differences from these videos to make the temporary set. Although a minority of these images contain similar background as the evaluation set samples, many different scenes are also included.

B. Architecture realization details

PASCAL-SSD: Our SSD network is configured based on an image resolution of 512×512 pixels. We compared the performance of the VGG-based SSDs pre-trained with the PASCAL VOC, COCO, and ILSVRC datasets on the evaluation set. Since the PASCAL-trained SSD produced the best detection results, we use this model as the basis for our work and will refer to this combination of network as PASCAL-SSD in the remainder of this paper. The model is pre-trained for 240,000 iterations.

Botlek-SSD: At the first stage, we finetune the PASCAL-SSD on the Botlek dataset for 196,855 iterations. A 25-% fraction of the images is used as test data and the rest of the images as training data. We start the finetuning at a learning rate of 0.001 and decrease the rate by a factor of 10 after 143,000 iterations. The finetuning is stopped when the final loss converges to 1.06. We call the resulting model Botlek-SSD, which will be supplied to the second stage.

Self-Learned SSD: As mentioned in the architecture section, generation of the triple self-learning training set is the main task of the second stage. To this end, first the Botlek-SSD performs the detection on the Botlek training set. Despite the 99-% precision rate with 59,759 correctly detected boxes, still 470 false positive and 10,438 false negative boxes appear among the results, spreading over 8,199 images. To represent the correction samples, the proposed system

randomly adds 911 images out of these frames to the self-learning dataset. Additionally, the system arbitrarily picks 3,089 samples out of the 40,165 images providing just true positives, to keep the useful information of the source data.

To complete the self-learning dataset, the Botlek-SSD is applied on the temporary set to automatically generate the target-domain adaptation candidate samples. In this step, the system detects the boxes that provide confidence scores higher than 0.1. Although this low threshold value seems to increase the risk of false detections, the subsequent temporal filtering approach automatically verifies the target-domain adaptation samples and removes the irrelevant boxes. However, these exceptional rare cases occur when the illumination of the scene suddenly changes, or the detector recognizes a water region as a vessel. Nevertheless, these cases occur at a negligible rate. Here, the system selected 2,205 samples out of the 5,480 temporary set images and added them to the self-learning dataset.

Finally, the target-domain adapted model is produced by finetuning the Botlek-SSD on the self-learning dataset for 2,000 iterations with the learning rate of 0.0001.

4.2 Validation Results

In this subsection, we compare the self-learned SSD with the PASCAL-SSD and the Botlek-SSD. We also investigate the temporal filtering performance.

Temporal Filtering Performance: The proposed self-learning framework applies a low confidence score to produce a high number of candidate samples. This low score often results in many false detections of irrelevant objects. Since the initial model is already trained on vessels, and vessels intrinsically expose a structure in pixels, these false detections rarely occur on structureless dynamic water pixels. According to our experiments, partly-stationary background areas are the most likely regions that cause this detection noise. Our detector often produces false detections on the bridges and vessel-like buildings. Consequently, the proposed framework uses the temporal information of the frame sequences to identify and ignore the falsely produced candidate labels through the temporal filtering approach. In this subsection, we provide the statistical analysis of the temporal filtering algorithm.

The statistical analysis is as follows. After applying the Botlek-SSD (initial model) to the 5,480 images of the temporary set, the images and their corresponding candidate labels are processed by the temporal filtering block. This filter removes 3,275 images, since each frame has neither a detection, nor one or more produced candidate labels surviving the temporal filtering. In order to provide a statistical analysis on the performance of the proposed temporal filtering method, we investigate the remaining images. Since manual validation of all the labels from 2,205 remaining filtered images is too labor intensive, we explore 250 randomly selected annotated images for an approximate analysis. Prior to the temporal filter, the selected images contain 1,276 candidate labels, including 743 noise labels and 533 vessel labels. The filter correctly removes 680 noise boxes, while falsely keeping 63 noise labels as a vessel, i.e. 91.52% of the noise

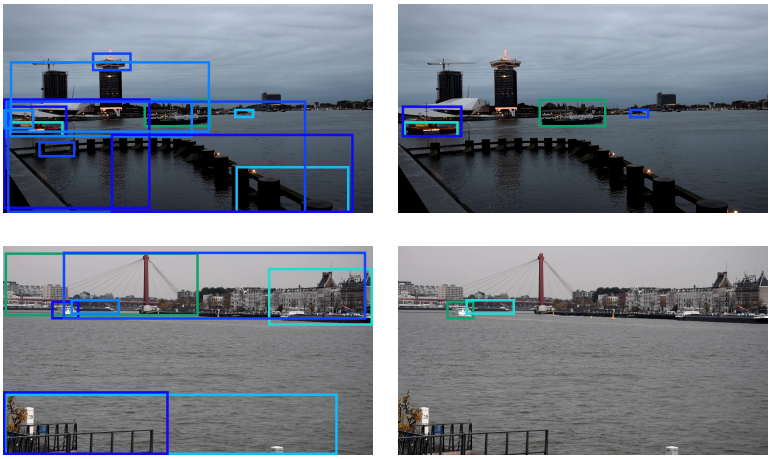

Fig. 4: Two temporal filtering output examples.

labels are correctly removed. Moreover, the filter correctly retains 454 vessel labels, which means only 14.82% of the vessels are removed by the filter. Overall, 58.23% of the provided candidates were noise labels prior to filtering and the temporal filter decreased this ratio to 4.94%. Fig. 4 illustrates two examples on how the temporal filtering removes the falsely detected candidate labels.

Self-Learning Framework Performance: Generally in a real-world outdoor monitoring application, items like the object size, distance to the camera, noise, occlusion, truncation, scene illumination and weather conditions, are considered when defining the performance expectations from the system. In order to accurately analyze the efficiency of the proposed target-domain adaptation approach, we select the vessel size, occlusion, and truncation as the criteria to derive the complete dataset into three versions of varying detection difficulty as follows:
-*Easy evaluation dataset*: the bounding-box size is more than 10,000 pixels, no occlusion, and no truncation;
-*Moderate evaluation dataset*: the bounding-box size is between 3,000 and 10,000 pixels, less than 30% of the vessel pixel area is occluded or truncated;
-*Hard evaluation dataset*: the bounding-box size is less than 3,000 pixels, more than 30% of the vessel pixel area is occluded or truncated.

We evaluate the previously introduced three methods on each level of difficulty. Tables 1, 2, and 3, illustrate the results. Each dataset class is tested three times with different Intersection-over-Union (IoU) thresholds. All the methods have performed the detection with the confidence score of 0.5. According to Table 1 (easy case), although the two vessel-adapted networks produce more true positives, the PASCAL-SSD surprisingly outperforms these methods in terms of the Average Precision (AP). However, by increasing the level of difficulty,

Table 1: Method comparison on Easy evaluation dataset.

	PASCAL-SSD			Botlek-SSD			Self-learned SSD		
IoU	0.3	0.4	0.5	0.3	0.4	0.5	0.3	0.4	0.5
TP	874	843	825	858	843	817	911	898	881
FP	621	625	640	874	889	915	760	773	790
FN	245	249	264	234	249	275	181	194	211
AP	0.58	0.57	0.56	0.50	0.49	0.47	0.55	0.54	0.53
F1	0.66	0.66	0.65	0.61	0.60	0.58	0.66	0.65	0.64

the vessel-adapted methods produce better results. On the Moderate evaluation dataset, the Botlek-SSD is still outperformed by the PASCAL-SSD in terms of AP by 10%. However, the self-learned SSD produced the same AP as the PASCAL-SSD while showing a 6% higher F1-score, since it provides 235 more correct detections and 88 less object misses.

We select the Hard evaluation dataset with IoU=0.5 as the criterion to compare the three methods in detail. For this dataset, the self-learned method outperforms the Botlek-SSD by 12% in terms of the AP. It also produces 279 less false detections, 118 less missed detections and 118 more true positives. Although the PASCAL-SSD is comparable with the proposed method in AP by producing a just 2% lower value, it provides 294 less correct detections, which means that network results in a 5.8% higher miss rate. Additionally, the proposed method outperforms both the PASCAL-SSD and the Botlek-SSD by 7% and 5%, respectively, for the F1-score. Fig. 5 provides a comparison of the outputs of the methods on four evaluation frames.

It is important to mention here that the initial model used in this paper is produced by finetuning the PASCAL-SSD on the non-public Botlek dataset. In case that an initial model is created only on the PASCAL dataset, e.g. without finetuning on a specific labeled maritime dataset, the performance results of the framework may become lower.

Table 2: Method comparison on Moderate evaluation dataset.

	PASCAL-SSD			Botlek-SSD			Self-learned SSD		
IoU	0.3	0.4	0.5	0.3	0.4	0.5	0.3	0.4	0.5
TP	1230	1214	1182	1361	1337	1286	1481	1459	1417
FP	495	511	543	830	854	905	567	589	631
FN	1196	1212	1244	1065	1089	1140	945	967	1009
AP	0.71	0.70	0.69	0.62	0.61	0.59	0.72	0.71	0.69
F1	0.59	0.58	0.57	0.59	0.58	0.56	0.66	0.65	0.63

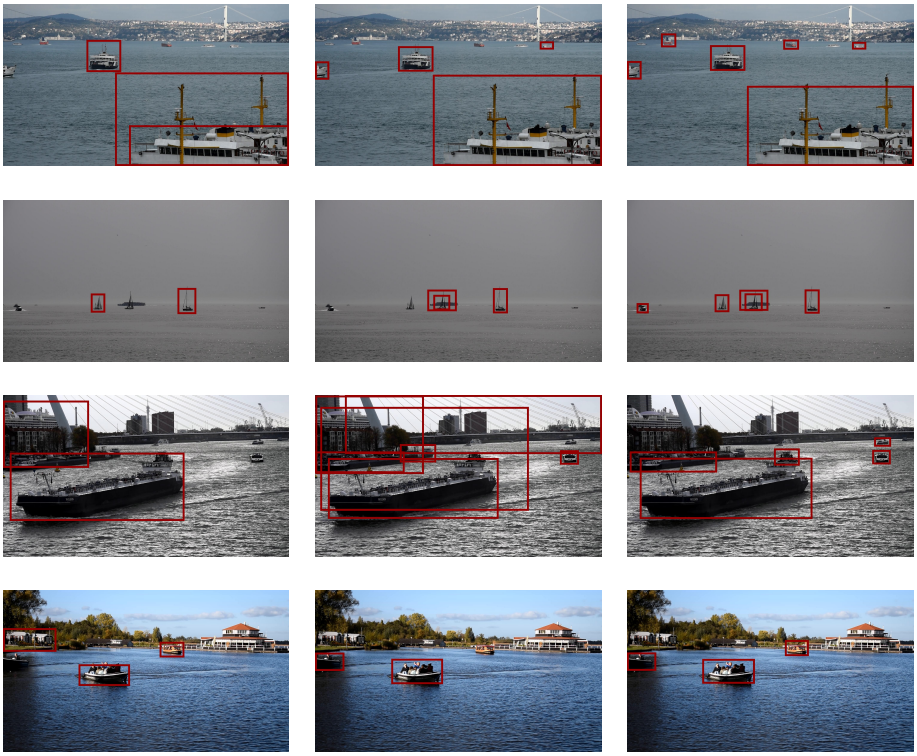

Fig. 5: Comparison of methods. From left to right, columns represent the outputs of the PASCAL-SSD, the Botlek-SSD, and the Self-learned SSD, respectively.

5 Discussions and Conclusions

The traditional object detection methods often fail to detect vessels under severe weather or raging water conditions. Additionally, the false negative probability of the detection of stationary vessels increases. However, a ConvNet-based system can enhance the possibilities of successfully addressing the vessel detection problem in the industry-oriented maritime surveillance applications, since ConvNets search for the desired objects independent of the surroundings.

In order to achieve robust results with ConvNets for a specific application, one needs to finetune a pre-trained model on a comprehensive annotated dataset collected from the desired target domain. However, by changing the location or capturing equipment, a system would need a new training dataset. Nevertheless, manual creation of a labeled training set is costly in terms of time. In order to solve this problem, the state-of-the-art methods are broadly exploiting semi-supervised techniques to design a framework that automatically finetunes a pre-trained ConvNet from the new raw data. Therefore, this paper has proposed a robust ConvNet self-learning framework for maritime vessel detection.

Table 3: Method comparison on Hard evaluation dataset.

	PASCAL-SSD			Botlek-SSD			Self-learned SSD		
IoU	0.3	0.4	0.5	0.3	0.4	0.5	0.3	0.4	0.5
TP	1479	1457	1417	1690	1660	1593	1784	1758	1711
FP	268	290	330	541	571	638	286	312	359
FN	3615	3637	3677	3404	3434	3501	3310	3336	3383
AP	0.85	0.83	0.81	0.76	0.74	0.71	0.86	0.85	0.83
F1	0.43	0.43	0.41	0.46	0.45	0.43	0.50	0.49	0.48

In this work, we first finetune a pre-trained single shot detector on an annotated maritime image set. This provides an initial vessel model, which is affected by the current domain characteristics. Second, we develop a self-learning framework which automatically generates the candidate labels from the target domain data, and performs a temporal filtering approach to verify the labeled samples. Finally, the system finetunes the model on the produced self-learning dataset. When applying this proposed framework to a SSD trained on a vessel-oriented dataset, the resulting network outperforms the initial model with a promising average precision of 83% at a high detection rate. This method also provides a 5% higher F-1 score. We have also presented an annotated evaluation dataset for the vessel detection problem, which contains challenging scenes and scenarios.

Future work will improve the proposed method in producing more verified labels from the target-domain data. Moreover, we plan to improve the detection efficiency on the vessel positioning at a far distance from the camera.

6 Acknowledgement

This work is supported by the European ITEA APPS project. We thank the company Vinotion for providing the Botlek dataset to us. We also show our gratitude to the company NVIDIA for granting us a “TITAN X PASCAL” GPU.

References

1. Khurana, P., Sharma, A., Narayan Singh, Ah., Kumar Singh, P.: A survey on object recognition and segmentation techniques. 3rd Int. IEEE Conf. on Computing for Sustainable Global Development (INDIACom), (2016).
2. Shantaiya, S., Verma, K., Mehta, K.: A Survey on Approaches of Object Detection. Int. Journal of Computer Applications (0975 – 8887) Volume 65– No.18, March (2013).
3. Bidyalakshmi Devi, R., Jina Chanu, Y., Manglem Singh, Kh.: A Survey on Different Background Subtraction Method for Moving Object Detection. Int. Journal for Research in Emerging Science and Technology, Vol. 3, Issue 10, Oct. (2016).
4. Abdul Malik, A., Khalil, A., Ullah Khan, H.: Object Detection and Tracking using Background Subtraction and Connected Component Labeling. Int. Journal of Computer Applications (0975 – 8887) Vol. 75, No. 13, August (2013).

5. Ghahremani, A., Bondarev, E., de With, P.H.N.: Water Region Extraction in Thermal and RGB Sequences Using Spatiotemporally-Oriented Energy Features. IS&T Electronic Imaging - Algorithms and Systems, USA, (2017).
6. Druzhkov, P.N., Kustikova, V.D.: A survey of deep learning methods and software tools for image classification and object detection. Pattern Recognition and Image Analysis, Vol. 26 No. 1 (2016).
7. Cabrera-Vives, G., Reyes, I., Forstert, F., Estevez, P.A., Maureira, J.C.: Supernovae detection by using convolutional neural networks. Int. Joint Conf. on Neural Networks, JCNN, (2016).
8. Liu, W., Anguelov, D., Erhan, D., Szegedy, Ch., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single Shot MultiBox Detector. European Conf. on Computer Vision - ECCV (2016).
9. Maâmatou, H., Chateau, T., Gazzah, S., Goyat, Y., & Amara, N. E. B.: Transductive Transfer Learning to Specialize a Generic Classifier Towards a Specific Scene. in VISIGRAPP (4: VISAPP). (2016).
10. Mhalla, A., Chateaub, T., Maâmatoua, H., Gazzaha, S., Amara, N.E.B.: SMC faster R-CNN: Toward a scene-specialized multi-object detector. Computer Vision and Image Understanding 000 1–13, (2017).
11. All, K., Hasler, D., Fleuret, F.: Flowboostappearance learning from sparsely annotated video. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1433–1440, (2011).
12. Wang, M., Li, W., Wang, X.: Transferring a generic pedestrian detector towards specific scenes. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 3274–3281, (2012).
13. Aytar, Y., Zisserman, A., Rasa, T.: Model transfer for object category detection. Int. IEEE Conference on Computer Vision, pp. 2252–2259, (2011).
14. Pan, S. J., Tsang, I. W., Kwok, J. T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE Trans. on Neural Networks, 199–210, (2011).
15. Li, X., Ye, M., Fu, M., Xu, P., Li, T.: Domain adaption of vehicle detector based on convolutional neural networks. Int. Journal of Control, Automation and Systems 13 (4) 1020–1031, (2015).
16. Wang, X., Hua, G., Han, T.X.: Detection by detections: Non-parametric detector adaptation for a video. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Pages 350-357, (2012).
17. Wang, X., Wang, M., Li, W.: Scene-Specific Pedestrian Detection for Static Video Surveillance. IEEE Trans. on Pattern Analysis and Machine Inteligence, Vol. 36, No. 2, February 2014.
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: NIPS. (2015).