Size-invariant Detection of Marine Vessels From Visual Time Series

Tunai Porto Marques[1] Alexandra Branzan Albu[1] Patrick O'Hara[2] Norma Serra[1] Ben Morrow[1]
Lauren McWhinnie[3] Rosaline Canessa[1]

[1] University of Victoria, BC, Canada
[2] Canadian Wildlife Service, Environment and Climate Change Canada
[3] Heriot-Watt University, Edinburgh, Scotland

Abstract

Marine vessel traffic is one of the main sources of negative anthropogenic impact upon marine environments. The automatic identification of boats in monitoring images facilitates conservation, research and patrolling efforts. However, the diverse sizes of vessels, the highly dynamic water surface and weather-related visibility issues significantly hinder this task. While recent deep learning (DL)-based object detectors identify well medium- and large-sized boats, smaller vessels, often responsible for substantial disturbance to sensitive marine life, are typically not detected. We propose a detection approach that combines state-of-the-art object detectors and a novel Detector of Small Marine Vessels (DSMV) to identify boats of any size. The DSMV uses a short time series of images and a novel bi-directional Gaussian Mixture technique to determine motion in combination with context-based filtering and a DL-based image classifier. Experimental results obtained on our novel datasets of images containing boats of various sizes show that the proposed approach comfortably outperforms five popular state-of-the-art object detectors. Code and datasets available at `https://github.com/tunai/hybrid-boat-detection`.

1. Introduction

Anthropogenic activities in coastal areas (*e.g.* vessel traffic, fishing, recreation) can inflict long-lasting harmful effects on oceans. Given that sound travels five times faster in water than in air [1], noise pollution is correlated to behavioural disturbance in marine species [41, 40] and interferes with animal vocalization [7, 23]. The ecological footprint from increased marine vessel traffic observed over the last few decades [17] is clearly demonstrated [59].

Automatic Identification System (AIS) data from marine vessels is used for estimating vessel densities as spatially explicit proxies for stressors such as noise, disturbance, vessel-strike, and discharge of harmful substances [4, 39]. However, AIS was not designed as a tool for research and conservation [49], thus impact estimates based solely on it often ignore contributions from smaller vessels (which typically do not broadcast AIS data [21]). This poses a challenge for understanding the contribution of smaller vessels to the marine soundscape, particularly in densely populated coastal regions like the Salish Sea.

Among the several cetacean species found in the inshore waters of the Salish Sea are the endangered [42, 14] Southern Resident Killer Whales (SRKW); the majority of the Salish Sea is currently designated as a SRKW critical habitat. This area is highly trafficked by small whale-watching, fishing, research and recreational boats. Research shows that these vessels emit high acoustic energy in the mid- and high-frequency ranges, and are more likely to negatively interact with sensitive marine life [44]. Therefore small boat traffic has wide-ranging ecological impacts [44, 25, 60].

Optical systems [52, 18, 30] complement well AIS-based monitoring because of their ability to detect both AIS and non-AIS boats, their non-invasive nature and low-cost [27]. Visual sightings might also provide additional information such as the type of interaction a vessel engages in with the environment. However, the interpretation of these visual data is time-consuming [45, 48]. Large amounts of monitoring data [27] require manual detection in tedious and often error-prone routines, the reason why numerous recent works offered automatic vessel detection frameworks [10, 29, 56, 2, 55, 35, 65].

While recent DL-based object detectors [32, 6, 47] are efficient in identifying large-sized or near-shore marine vessels, they often miss small boats (either farther away from the camera or because of their actual size) [21]. For this reason, our study aims for the automatic detection of marine vessels of *any* size in visual data obtained at two sites inside the SRKW's critical habitat (see 4.1). The remainder of this article is structured as follows. Section 2 discusses works related to the proposed system. Section 3 details our approach for the detection of boats. In Section 4 we introduce two annotated datasets of images from monitoring sites in the Salish Sea and use them to evaluate the proposed system

with respect to state-of-the-art object detectors [32, 6, 47].

2. Related Works

Relevant works to our approach include custom detectors of marine vessels and generic DL-based object detectors.

Marine Vessel Detection. Methods that perform the visual detection of boats have been proposed for a number of monitoring configurations, such as satellites, unmanned aerial vehicles, boat-attached cameras and fixed-position cameras. Elvidge *et al*. [10] used infrared satellite images to detect boats in the day/night band based on the assumption that the lighting sources generated by fishing boats can be identified as intensity spikes. Using a different optical system layout, Kruger and Orlov [29] mounted a thermal imaging system on autonomous platforms and used its data to detect small vessels. Their method first estimates horizon lines, then performs detection in the vicinity of these lines, followed by tracking the identified objects.

Tran and Le [56] performed boat detection in sequences of images from a fixed location. Their first step, *temporal attention*, executes a background subtraction that isolates only moving elements of the sequence. A parallel step, *spatial attention*, looks for the salient regions of the image sequence. The final output is a weighted linear combination of both steps. The authors note that the segmentation of aquatic background (so that the foreground highlights only marine vessels) on *in-situ* surveillance images or videos is a challenging task, given the waters dynamic nature. Bao *et al*. [2] propose to first use a graph-based segmentation to detect water, followed by a saliency-based vessel detection. Bloisi *et al*. [5] offered a method that discretizes an unknown distribution, ultimately aiming to describe highly dynamic backgrounds (such as the surface of the water).

So far DL has been only sparsely used for marine vessels detection. Tang *et al*. [55] used a custom neural network to extract and classify candidate ship features. Liu *et al*. [35] proposed rotated region convolutional neural networks (RR-CNN) to identify marine vessels in satellite images. RR-CNN are able to efficiently encompass rotated targets under regions of interest (RoI), but none of the vessels in the dataset used (HRSC2016[1]) are small (see dataset **D2** in 4.1). Zhang *et al*. [65] extracted handcrafted vessel features from satellite images using line segments and saliency maps, and employed CNN to classify them.

Generic Object Detection. Object detectors perform both localization and classification tasks. Until 2012, most object detection methods ("traditional detection methods" [68]) extracted and used handcrafted features such as the Histogram of Oriented Gradients [8], multiple feature extractors [36, 3, 50], or class- and application-specific features [57, 58]. The Deformable Part-based Model (DPM) developed by Felzenszwalb *et al*. [11] and its further developments [12, 13] are considered the best-performing among the traditional detection methods.

AlexNet, a CNN-based image classifier proposed by Krizhevsky *et al*. [28], demonstrated the potential of using CNNs to extract generic and highly discriminative features from large sets of data [9]. Subsequently, a number of works proposed CNN-based object detectors capable of delimiting *where* target objects are, and exactly *what* class they belong to: R-CNN [16], Fast R-CNN [15], Faster R-CNN [47], Cascade R-CNN [6] and RetinaNet [32], among others [46, 34, 19, 31]. This process typically involves the training of CNN-based modules that perform localization and classification individually ("two-stage detectors") or under the same trainable network ("one-stage detectors"). Lin *et al*. proposed Feature Pyramid Networks (FPN) [31], capable of integrating the representation under multiple scales of objects during the CNN training process. When used with end-to-end object detectors, FPN significantly increases the final detection performance.

To the best of our knowledge, no work has used state-of-the-art object detectors pre-trained on large datasets [33] as part of a marine vessel detection framework. Moreover, we propose the first system that also focuses on the identification of small vessels observed in land-based visual data.

3. Proposed Approach

We propose a hybrid marine vessel detection system that combines state-of-the-art object detectors and a novel Detector of Small Marine Vessels (DSMV). The DSMV uses short time series of images (*i.e*. three images at a time) to detect small vessels. We use the term *small* to refer to vessels that have approximately 80 pixels of area, while *medium* and *large* vessels are those that occupy approximately 800 and $1,500$ pixels of area, respectively (see 4.1). The proposed system does not require a (often most error-prone) sea-land segmentation step as so other works [55, 2, 5]. It thus contributes not only to the environmental monitoring area, but also addresses the more general challenge of *small object detection*. Figure 1 visually summarizes the proposed hybrid detection approach.

Figure 2 highlights a fundamental challenge for the identification of small boats: given their size and changing appearance, a regular CNN-based feature extractor and classifier might not be able to distinguish them from a highly dynamic background that includes water surface perturbations, sunlight reflection or weather elements, floating driftwood and kelp. Indeed, the visual structure of these background elements is nearly identical to that of boats in many instances. We thus rely upon temporal information (*i.e*. movement conveyed by multiple images/frames of the same scene) as a cue for the presence of boats. We assume that a boat is going to move in a roughly horizontal manner (con-

[1] www.kaggle.com/guofeng/hrsc2016

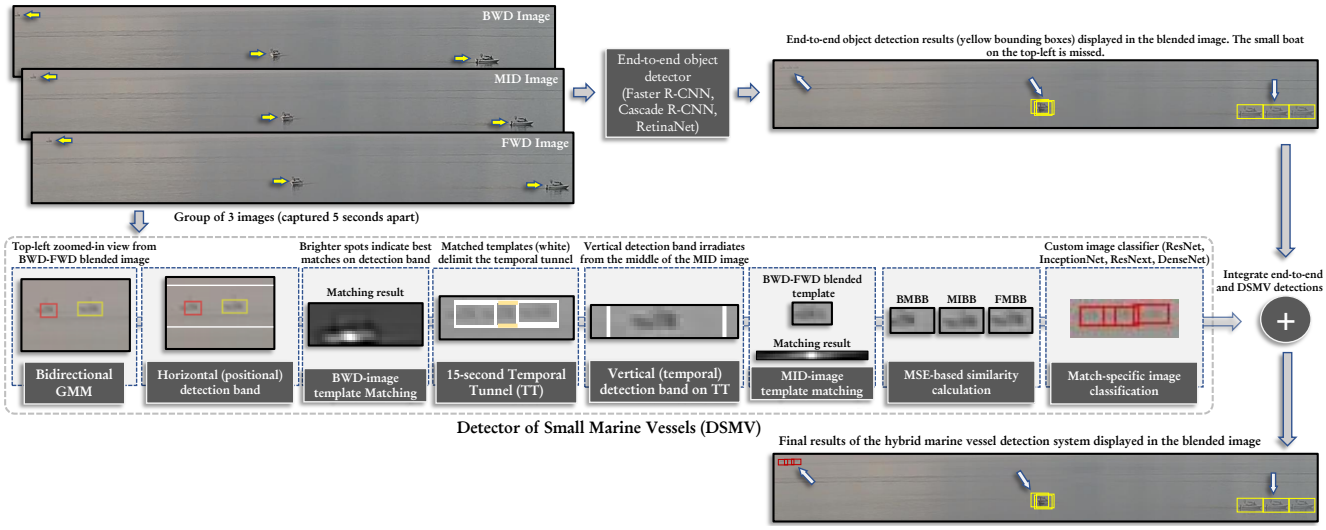


Figure 1: Computational pipeline of the hybrid marine vessel detector. The detection results of the end-to-end object detector and the DSMV are combined for enhanced detection capabilities. We invite the reader to zoom-in on this and the other images in the manuscript.

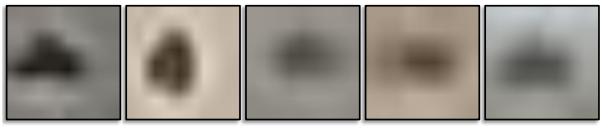


Figure 2: Examples of small marine vessels (mean area of 79 pixels) resized to 224×224 pixels. See 4.1 for details.

sidering fixed-position cameras) and that its visual features are not going to change during a small time window.

DSMV considers only three 5-second-apart images of the same location (henceforth referred to as "BWD", "MID", and "FWD" images) obtained from a land-mounted camera. This temporal window is chosen so that the DSMV can be deployed in remote sites where only limited data storage and transmission capabilities are available. Based on a thorough analysis of environmental monitoring data, we define four assumptions that help distinguish small marine vessels from false positives: **A1**) vessels that appear small on monitoring images are those farther away from the camera, and thus they should only move horizontally in a 15-second time window; **A2**) a boat identified in the MID image will remain inside a sub-region of this image bounded by the same boat identified in the BWD and FWD images; **A3**) the boat in the MID image is going to be positioned roughly in the middle of the aforementioned sub-region; **A4**) the boats identified in the three images present similar visual appearance.

These four assumptions encode contextual information acquired from the study of monitoring data. We explicitly incorporate them in the DSMV with the use of traditional computer vision methods, as detailed in the remainder of this Section. In order to further refine the detection results, we include a custom DL-based image classifier at the end of the DSMV. As a result, DSMV performs robust detection combining specific, contextual data, and generic visual features learned via the training of CNN-based frameworks. Notes about the implementation of the proposed hybrid detection system are concentrated in 4.3.

Bidirectional GMM. Traditional GMM-based systems [53] typically create background models based on inputs I ranging from $I_1, ..., I_{t-1}$, and perform foreground detection on the current input, I_t, in what we will henceforth refer to as *forward motion*. In the proposed system we create an exclusive GMM for each group of three images (*i.e*. BWD, MID and FWD), allowing for systems with low monitoring frequency to still use the proposed DSMV, as each group of inputs is processed independently. During the *forward motion*, the BWD and MID images are used to model the background, thus detecting the foreground on the FWD. Similarly, in our novel *backward motion* we use the FWD and MID images to detect motion on the BWD.

A similar approach was proposed by Shimada *et al*. [51], where two models are derived from distinct groups of "past" and "future" frames that do not overlap. Also, Minematsu *et al*. [38] uses two models derived from the same group of past frames which are analyzed in regular and backward chronological order. Our approach is different because we consider all temporal information by modeling two GMM out of an overlapping frame (*i.e*. MID image), and use both forward and backward motions.

Figure 3 illustrates our bidirectional GMM approach. In the forward motion, a set of motion-triggered connected components indicate the pixels that deviated from the back-

ground models created with the BWD and MID images (Figure 3b). Since the vessels are expected to create larger groups of quasi-connected components, we filter the results using morphology: an opening with a 3×3 ellipse followed by a dilation with a 5×5 ellipse. As shown in Figure 3c, this filtering eliminates small motion-triggered outputs (mostly noise) and combine the remaining pixels into compact groups. Figure 3d illustrates the result from the same process for the backward motion. The sets of connected components from each motion are used to delimit bounding boxes (BB) in the FWD and BWD images (Figure 3e), named forward-motion bounding boxes (FMBB) and backward-motion bounding boxes (BMBB). Following assumption **A1**, we set a horizontal detection band (Figure 3f) inside which the marine vessel is expected to travel through the group of three images. The height of this band is a product of a configurable parameter ψ_{hdb} by the height from each FMBB (see Figure 4). Each FMBB determines a horizontal detection band where template matching is performed.

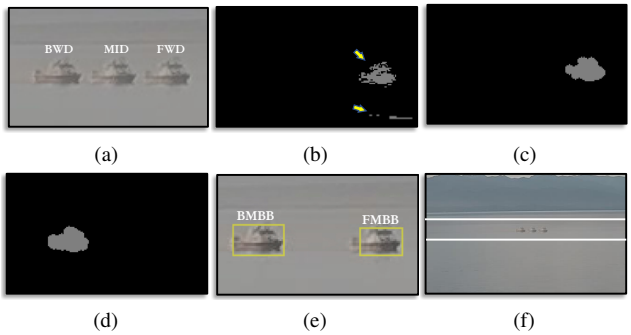


Figure 3: Bidirectional GMM strategy proposed. **(a)** Blended section from group of three images. **(b)** Forward motion raw output. **(c)** Forward motion filtered output. **(d)** Backward motion filtered output (BWD image). **(e)** Bounding boxes encompassing motion-triggered results. **(f)** Horizontal (positional) detection band.

Template matching. A FMBB verifies **A1** if a BMBB exists on the BWD image inside the horizontal (positional) detection band set by its position, height and a given ψ_{hdb} (see Figure 4). Template matching operations are only performed on pairs of valid FMBB/BMBB positioned inside the horizontal detection band, as illustrated by BMBB match candidates 1 and 2 of Figure 4. There are a number of approaches to follow when a query image has to be found/matched in another image. Most commonly, one would start by determining features using a feature extractor (*e.g.* SIFT [36], SURF [3], ORB [50]) and then match features between queries and candidates. However, since 1) we only compare each FMBB with a few potential BMBBs placed inside a reduced detection band, and 2) small regions representing boats often do not generate any output from regular visual feature extractors; we use template matching as in Kaehler and Bradski [26], which is simple and fast. The dimensions of a matched BMBB are adjusted to be equal to its template FMBB.

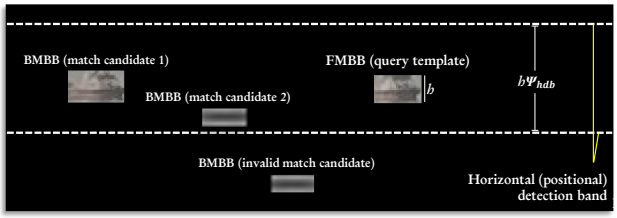


Figure 4: FWD-BWD images template matching. Each FMBB determines a single horizontal detection band based on their position, height h and parameter ψ_{hdb}. If one or more BMBB are positioned inside this band, the one that better matches the current FMBB is considered as its match.

Temporal Tunnel (TT). Once a FMBB is matched with a BMBB, assumption **A2** states that another match of that same marine vessel exists in the MID image, placed inside a TT delimited by the FMBB/BMBB pair (see Figure 5a). More specifically (**A3**), the vessel in the MID image should sit roughly in the middle of the TT. The width of this valid matching area in the middle of the TT is a product of width w from the FMBB by a configurable parameter ρ_{vdb} (see Figure 5c). A blended template (Figure 5b) is created by combining the BMBB and FMBB to prevent eventual occlusions in either reference BB from interfering with the MID-image matching. The blended template is used in a matching process that covers the entire TT (not only the valid matching area), resulting in a best match for MID-image bounding box (MIBB). If the MIBB falls inside the valid match region (green portion of Figure 5c), it is considered to be a valid candidate, as illustrated in Figure 5d.

Similarity Criteria. Assumption **A4** is based upon the empirical observation that within a 15-second time window all sightings of the same vessel should present similar appearance. This helps to further distinguish valid detection from false positives triggered by weather or incorrect matching results (see Figure 6 right). Each group of BMBB-MIBB-FMBB is used as the input of a Mean Squared Error (MSE)-based similarity analysis. The MSE between the blended FMBB/BMBB template and the MIBB is measured, and if it does not exceed a threshold, MSE_{th}, this group of three bounding boxes is further analyzed by an image classifier (see Figure 6).

Image classification. The last step of the DSMV uses a custom-trained DL-based classifier (we evaluate six state-of-the-art options, see 4.2) to classify each individual bounding box in a group of BMBB-MIBB-FMBB. A group of BMBB-MIBB-FMBB is only deemed as valid if the content of all three BBs is classified as an object of the *vessel* class. The custom-trained DL-based system performs

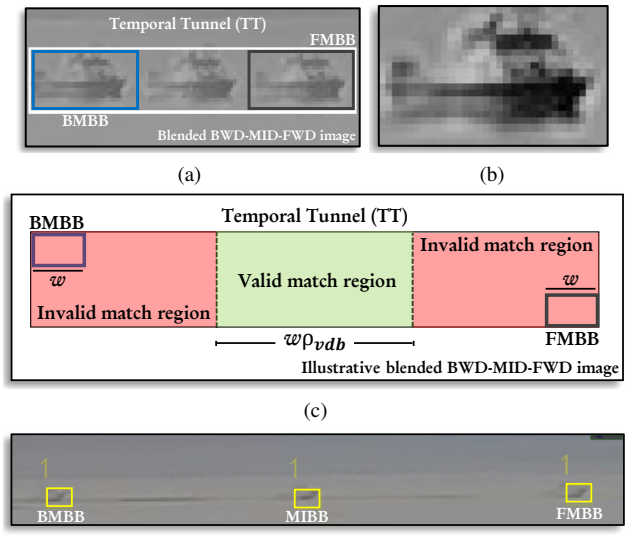


Figure 5: Temporal Tunnel (TT). **(a)** A 15-second TT bounded by a FMBB/BMBB pair. **(b)** Blended template composed by the FMBB and BMBB contents. **(c)** Only the template matching results inside a sub-region of the TT delimited by parameter ρ_{vdb} and width w are valid (**A3**). **(d)** BMBB-MIBB-FMBB matching output.

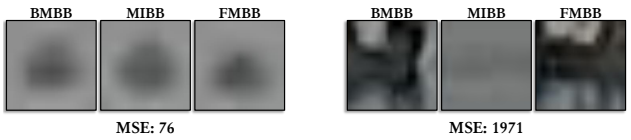


Figure 6: MSE-based similarity calculation. **Left:** Valid group of detection resulting in a low MSE. **Right:** An invalid group of detection candidates identified by a higher MSE. The similarity is measured between the contents of the blended FMBB/BMBB template and the MIBB.

a binary classification where each image patch is classified as either *background* or *vessel*. We train the DL classifiers by running the DSMV (without this last step) on $1,644$ monitoring images (Figure 7a), and manually distinguish between vessels and background in the resulting BMBB-MIBB-FMBB groups. These manually-curated image patches are resized to comply with each classifier's CNN layout (*i.e.* either 224×224 pixels or 299×299 pixels) and used in the training routine, as shown in Figure 7b. This training process uses images obtained off the coast of Vancouver Island, Canada, during the years of 2019 and 2020 (see Table 1). In total we used $1,879$ vessel image patches ($1,544$ train/335 validation) and $2,264$ background image patches ($1,633$ train/629 validation). Figure 7c illustrates a scenario where false positives are accurately classified as background (yellow bounding boxes).

End-to-end object detection. State-of-the-art object detectors are commonly trained on large datasets (*e.g.* COCO

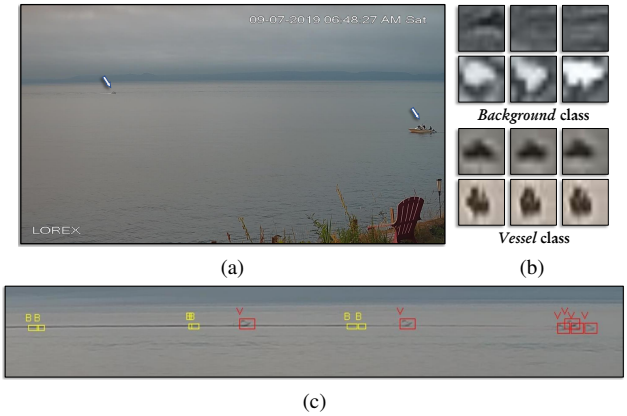


Figure 7: Image classification. **(a)** Image from the training set with vessels highlighted. **(b)** Resized patches used in the training of image classifiers. **(c)** DSMV results with a custom-trained ResNet-50 [20] distinguishing between groups of vessels (red BBs) and background (yellow BBs).

[33], ImageNet [9]) that include one or more classes related to marine vessels. Thus we use pre-trained end-to-end object detectors (see Section 4) to find easier-to-identify boats, represented by medium- and large-sized vessels. Given the efficacy of such systems, we combine the output of object detectors (medium- and large-sized boats) and the DSMV (small-sized boats) into a robust hybrid detector capable of identifying boats of any size (see Figure 1).

Smaller vessels represent a challenging task to end-to-end object detectors because their feature extractors are based on multiple layers of sequential convolutions that generate feature maps of progressively smaller dimensions. Small visual targets in the original image disappear during the feature extraction process (*i.e.* they are represented by less than a pixel in the feature map at a certain depth in the CNN), preventing their localization and classification. Figure 9b shows that end-to-end object detectors can efficiently identify bigger vessels, but often miss smaller ones.

4. Results and Discussion

This section details the experimental settings and results from a performance evaluation where two novel datasets are introduced and used to compare our proposed method with five state-of-the-art end-to-end object detectors.

4.1. Datasets

The images used in this project were obtained by University of Victoria's Coastal and Ocean Resource Analysis Laboratory (CORAL)[2] using optical cameras focused offshore to the south and west of southern Vancouver Island, BC, Canada, during the years of 2019 and 2020. These

[2]www.coral.geog.uvic.ca

monitored regions of the Salish Sea are classified as SRKW critical habitats. A LOREX® pan-tilt-zoom (PTZ) camera was installed at two fixed positions on a headland overlooking a major vessel traffic thoroughfare and configured to continuously capture three 1920×1080 pixels photos in the first 15 seconds of each minute. This time-lapsed configuration allows for the inference of vessel movement, directionality and behaviour.

We manually annotate (*i.e*. BBs are drawn around marine vessels) and make publicly available[3] two datasets used to evaluate the proposed hybrid detector under two conditions: **D1**) 633 images containing boats of various sizes (mean vessel area of 953 pixels); **D2**) 138 images presenting only small boats with a mean area of 79 pixels. **D2** highlights the capabilities of the DSMV, as most of its marine vessels are missed by the state-of-the-art object detectors. While creating both **D1** and **D2** we selected images under different weather conditions (see Figure 8a) and vessel layouts, so that all monitoring scenarios are well represented. Note that a dataset of vessel patches, **D3**, is also created exclusively for training the DSMV (see Figure 7b). Samples from the training and testing datasets are never obtained from the same day (see Table 1), avoiding any data contamination during evaluation.

(a) Hazy-day image from **D1** showing two medium-sized vessels.

(b) Clear-day image from **D2** containing only a small vessel.

Figure 8: Datasets used for testing.

4.2. Experimental results

The performance evaluation uses **D1** and **D2** and starts by employing five state-of-the-art end-to-end object detectors: Cascade R-CNN [6], Faster R-CNN [47] (with three different feature extraction networks), and RetinaNet [32]. All detectors used employ FPN [31] in their feature extraction routines. These detectors are pre-trained on the COCO [33] dataset, and since one of its 80 classes represents marine vessels ("boats" class), the initial set of experiments takes advantage of the pre-trained weights of these object detectors (see Table 2, configurations 1-5), looking only at detection of this class. Transfer learning experiments where we re-trained only part of these detectors using our custom detasets could not surpass the performance of the pre-trained weights. Thus the following results for end-to-end object detectors reflect the use of such weights.

The second part of the experiments ("*hybrid*" layout) evaluates the performance of the hybrid detector proposed. Given that numerous smaller vessels are missed by the object detectors, combining their output with those from DSMV greatly enhances the detection performance, especially for **D2** (where the vessels are particularly small).

We report the average precision (AP) in range $[0, 1]$ for three different intersection-over-union (IoU) thresholds $\in [0.2 : 0.1 : 0.4]$ (see Table 2). The decision of using lower-than-usual thresholds is based on the fact that the monitoring systems expected to use the proposed hybrid detector do not prioritize a precise fit around the visual targets, but rather a robust identification of their presence. Each hybrid layout explores the performance of the DSMV using one of six custom-trained image classifiers: ResNet-50 [20], Inception-V3 [54], DenseNet-201 [24], ResNext-50 and ResNext-101 [62], and Wide ResNet 50-2 [64]. The detection time per image using a PC equipped with an Intel® Core i7-9700 CPU, 32 GB of RAM memory and a GeForce® GTX 1660 Ti GPU is approximately 0.2 seconds when using only end-to-end object detectors, and roughly 0.4 seconds for the entire hybrid approach.

Table 2 presents the detection results for **D1** and **D2** for both end-to-end object detectors and the proposed hybrid approach. Due to space constraints we present only the best-performing results out of the 35 configurations tested. The first five configurations use only object detectors, and among these RetinaNet performed significantly better for vessels of various sizes (**D1**). The best-performing object detector for IoU thresholds 0.3 and 0.4 using **D2** was Faster R-CNN, showing that the dataset composition and IoU threshold must be considered when choosing to use a pre-trained object detector. Since **D2** presents boats on average 12 times smaller than those in **D1**, the detection task becomes much more challenging, as reflected by the lower performance of the pre-trained object detectors in **D2**.

The proposed hybrid approach (*i.e*. configurations 6-26 on Table 2) improved the performance from all state-of-the-art object detectors when corresponding stand-alone and hybrid layouts are compared. For example, the performance on dataset **D1** of Cascade R-CNN using ResNet-50 and IoU threshold 0.2 (configuration 4) was boosted in 16.45% by the addition of the DSMV employing ResNext-101 as backbone (configuration 14). Although a better performance is always provided by the proposed hybrid approach in **D1**, the detection gains are smaller than those observed in **D2** be-

[3] https://github.com/tunai/hybrid-boat-detection

Dataset	Purpose	Description	Images	Vessels count	Dimensions	Dates
D1	Testing	BBs around vessels of various sizes	633	1056	1920×1080	2019: Sep 1,2,4-5,20 Aug 18
D2	Testing	BBs around small vessels	138	165	1920×1080	2019: Jul 30-31
D3	Training/Validation	Square vessel patches for training	$3,177/964$	$3,177$	CNN-dependent†	2019: Jul 27-29 Sep 6-11,16-19 2020: Jan 1

†: Patches of 299×299 pixels for Inception [54], 224×224 pixels for DenseNet [24], ResNet [20], ResNext [62] and Wide ResNet [64] (see Figure 7b and Figure 2).

Table 1: Datasets for training the DSMV image classifier and evaluating the overall hybrid detection system proposed.

cause, as mentioned, the end-to-end object detectors work well for detecting medium- and large-sized vessels.

The potential of the DSMV is more explicit when using dataset **D2**, where the improvement in AP of the best-performing configurations for IoU threshold 0.3 (configurations 2 and 14) is 112.1%. On average, the boost in performance when considering best-performing configurations for all three IoU thresholds and **D2** is 89.28%. The AP drops significantly on all configurations when the IoU threshold is increased because the small vessels of **D2** are hard to precisely encompass in either manual annotation or autonomous detection, as their visual limits are often blurry (see Figure 2).

Figure 9 illustrates detection results under different layouts. On the second row of Figure 9a, the output of Faster R-CNN using ResNext-101 show that the object detector missed all the small boats contained in these six excerpts from images of **D2**. The output from the proposed hybrid approach (first row of Figure 9a) highlight the ability of DSMV to identify extremely small vessels that were initially missed. Results in **D2** also show that the proposed detector is robust to non-horizontal movements (*i.e.* assumption **A1**), given that some boats in it display a mostly concave trajectory. The results of the proposed system shown on Figure 9b distinguish between detection made by the DSMV (red BBs) and otherwise (yellow BBs). Note that the object detector (Faster R-CNN with ResNext-101) correctly identified medium- and large-sized boats, while most of the small-sized boats are identified only by the DSMV.

4.3. Implementation Details

We implemented the proposed system on Python and PyTorch [4] using pre-trained weights and object detector implementations from Detectron2 [61]. When using our system, the user can set an $x-$ and $y-$axis range of valid detection. This one-time setting allows regions of the image with static content (*e.g.* manufacturer logo and date on Figure 9b) to be ignored during the detection process. The $y-$axis range set for **D1** and **D2** tests were, respectively, $[281, 850]$ and $[650, 896]$. Additionally, an $x-$axis range of $[132, 1920]$ is set on **D2** tests to ignore a ladder (see Figure 8b).

The vertical (temporal) band is delimited by a ρ_{vdb} of 2, while the horizontal (positional) band is set by a ψ_{hdb} of 1.4. Once a BMBB candidate is associated with a FMBB, its content is removed from further template matching tasks, thus the content of a BMBB can only be matched with a single FMBB. The MSE threshold (MSE_{th}) used in the experiments is 600. In order to avoid a large group of invalid candidates in the initial phases of the DSMV (sometimes triggered by sunlight reflections), we limit the maximum number of FMBB considered. We start with a GMM threshold (squared Mahalanobis distance [37] between a pixel and the Gaussian distributions) of 120, and if the initial number of FMBB obtained is higher than 18, we increase this threshold by 100 and calculate the group of FMBB again. Given that vessels typically create more pronounced deviations from background distributions, the progressively higher thresholds of the proposed algorithm eventually filters the invalid FMBB. We use Zivkovic's [66, 67] background subtraction method as implemented in OpenCV[5] as a basis for our bidirectional GMM.

We prioritize the detection output (*i.e.* BB dimensions, position and score) of the object detectors when there exists an overlap with the output of the DSMV. Moreover, we set a standard detection score (for AP calculation purposes) for the DSMV of 0.91. Changing this value modifies the relevance assigned to DSMV detection, and marginally changes the overall AP of each configuration.

5. Conclusion

Our hybrid marine vessel detector uses short time series to identify boats of any size, shape, and under different viewing conditions. The proposed DSMV uses a combination of a novel bidirectional GMM strategy, classical computer vision methods and custom-trained DL-based classifiers for identifying challenging small vessels. Extensive experiments show that our hybrid approach outperforms five state-of-the-art object detectors on two datasets we make publicly available.

The proposed detector fulfills real-world automated processing needs of data managers and governance [43], in particular in critical habitats such as the Salish Sea. Its fast (approximately 0.4 seconds per image) and efficient detection enables the timely interpretation of monitoring data to support conservation and research efforts. We also provide novel visual datasets of AIS and non-AIS vessel fleets in important ecological areas to promote further research.

Our approach is based on a set of four assumptions that might not be representative of all monitoring layouts, thus one must be aware of them before employing our proposed system. Small adaptations (*e.g.* camera tilting, image pre-

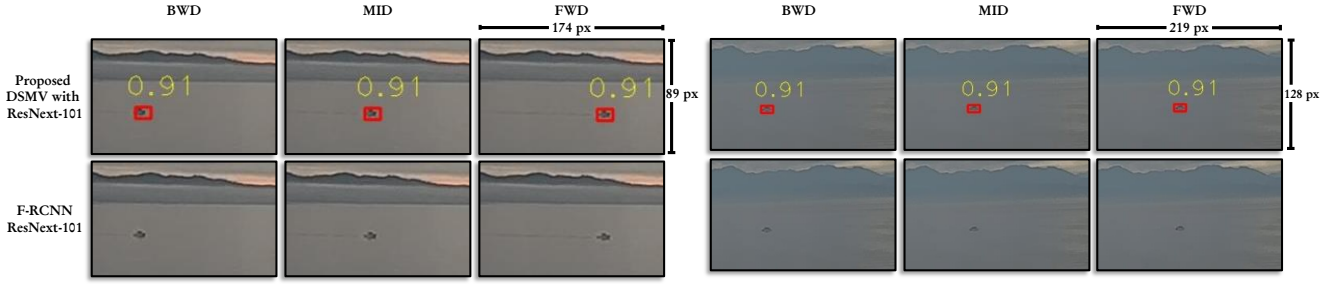
[4] www.pytorch.org

[5] www.opencv.org

			Dataset **D1** (various vessel sizes)			Dataset **D2** (small vessels)		
#	Configuration	DSMV Backbone	AP @ 0.2	AP @ 0.3	AP @ 0.4	AP @ 0.2	AP @ 0.3	AP @ 0.4
1	End-to-end: F-RCNN R-101 [1]	N/A	0.680	0.675	0.653	0.297	0.248	0.179
2	End-to-end: F-RCNN R-50 [2]	N/A	0.654	0.634	0.621	0.283	**0.256**	0.164
3	End-to-end: F-RCNN X-101 [3]	N/A	0.703	0.689	0.659	0.337	0.232	**0.186**
4	End-to-end: Cascade R-CNN R-50 [4]	N/A	0.699	0.680	0.662	0.271	0.229	0.151
5	End-to-end: RetinaNet R-101 [5]	N/A	**0.787**	**0.761**	**0.704**	**0.359**	0.240	0.134
6	Hybrid: F-RCNN R-101	ResNet-50	0.787	0.772	0.738	0.541	0.462	0.217
7	Hybrid: F-RCNN X-101	ResNet-50	0.774	0.756	0.720	0.570	0.457	0.295
8	Hybrid: Cascade R-CNN R-50	ResNet-50	0.798	0.771	0.736	0.553	0.487	0.242
9	Hybrid: RetinaNet R-101	ResNet-50	0.809	0.780	0.714	0.557	0.438	0.210
10	Hybrid: F-RCNN R-50	Inception-V3	0.750	0.730	0.700	0.445	0.408	0.241
11	Hybrid: F-RCNN X-101	Inception-V3	0.765	0.752	0.716	0.499	0.398	0.293
12	Hybrid: Cascade R-CNN R-50	Inception-V3	0.791	0.774	0.735	0.462	0.403	0.234
13	Hybrid: F-RCNN X-101	ResNext-101	0.785	0.767	0.729	**0.619**	0.506	**0.341**
14	Hybrid: Cascade R-CNN R-50	ResNext-101	0.814	0.787	**0.749**	0.608	**0.543**	0.282
15	Hybrid: RetinaNet R-101	ResNext-101	**0.833**	0.804	0.736	0.608	0.489	0.261
16	Hybrid: F-RCNN R-50	ResNext-50	0.747	0.726	0.701	0.464	0.420	0.224
17	Hybrid: F-RCNN X-101	ResNext-50	0.765	0.751	0.718	0.525	0.409	0.269
18	Hybrid: Cascade R-CNN R-50	ResNext-50	0.791	0.773	0.739	0.480	0.409	0.210
19	Hybrid: RetinaNet R-101	ResNext-50	0.826	0.800	0.736	0.514	0.389	0.201
20	Hybrid: F-RCNN X-101	Wide ResNet 50-2	0.779	0.761	0.723	**0.619**	0.506	**0.341**
21	Hybrid: Cascade R-CNN R-50	Wide ResNet 50-2	0.808	0.781	0.743	0.602	0.536	0.290
22	Hybrid: RetinaNet R-101	Wide ResNet 50-2	0.822	0.793	0.725	0.608	0.489	0.261
23	Hybrid: F-RCNN R-50	DenseNet-201	0.766	0.745	0.713	0.508	0.465	0.261
24	Hybrid: F-RCNN X-101	DenseNet-201	0.777	0.763	0.725	0.557	0.442	0.302
25	Hybrid: Cascade R-CNN R-50	DenseNet-201	0.804	0.786	0.747	0.526	0.457	0.251
26	Hybrid: RetinaNet R-101	DenseNet-201	0.832	**0.807**	0.738	0.533	0.408	0.224

[1,2,3]: Pre-trained Faster R-CNN [47] using FPN [31] with Resnet-101 [20], Resnet-50 [20] and ResNext-101 [62] as feature extractors, respectively.
[4,5]: Pre-trained Cascade R-CNN [6] and RetinaNet [32] using FPN [31] with Resnet-50 [20] and Resnet-101 [20] as feature extractors, respectively.

Table 2: Average Precision results for configurations combining pre-trained end-to-end object detectors, DSMV with custom-trained image classifiers, and different intersection-over-union thresholds (see 4.2 for details). Best results for each layout and dataset are highlighted in bold. Our hybrid approach outperforms all corresponding stand-alone pre-trained object detectors.



(a) Detection results on dataset **D2** for end-to-end object detector (second row) and proposed hybrid layouts (first row). Red BBs highlight DMSV detection.

(b) Detection results on dataset **D1** for the proposed hybrid approach (*i.e.* pre-trained object detector output combined with the DSMV output) using Faster R-CNN and ResNext-101. Yellow BBs indicate object detector-only results while red BBs show the DSMV-generated output.

Figure 9: Detection results of our hybrid detection system and stand-alone object detectors (second row of (**a**)).

processing) can assist in ensuring that these assumptions are valid for the visual data being processed. Different image frame rates can be employed by enlarging the time window considered by the bidirectional GMM. Boats that move towards the camera or are partially occluded (*e.g.* by other vessels) might result in false negatives from the DSMV.

Future work will involve ablation studies and the use of different background modelling strategies (*e.g.* Bloisi *et al.* discrete distribution [5]) in the first stages of the DSMV. Other methods to encode temporal information (*e.g.* Long Short-Term Memory networks [22]) and object tracking strategies [63] are also going to be considered.

References

[1] Ameer Abdulla. *Maritime traffic effects on biodiversity in the Mediterranean Sea. Volume 1: review of impacts, priority areas and mitigation measures*, volume 1. IUCN, 2008.

[2] Xinfeng Bao, Svitlana Zinger, Rob Wijnhoven, et al. Ship detection in port surveillance based on context and motion saliency analysis. In *Video Surveillance and Transportation Imaging Applications*, volume 8663, page 86630D. International Society for Optics and Photonics, 2013.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[4] Stefania Bertazzon, Patrick D O'Hara, Olesya Barrett, and Norma Serra-Sogas. Geospatial analysis of oil discharges observed by the national aerial surveillance program in the canadian pacific ocean. *Applied Geography*, 52:78–89, 2014.

[5] Domenico D Bloisi, Andrea Pennisi, and Luca Iocchi. Background modeling in the maritime domain. *Machine vision and applications*, 25(5):1257–1269, 2014.

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[7] Christopher W Clark, William T Ellison, Brandon L Southall, Leila Hatch, Sofie M Van Parijs, Adam Frankel, and Dimitri Ponirakis. Acoustic masking in marine ecosystems: intuitions, analysis, and implication. *Marine Ecology Progress Series*, 395:201–222, 2009.

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Christopher Elvidge, Mikhail Zhizhin, Kimberly Baugh, and Feng-Chi Hsu. Automatic boat identification system for viirs low light imaging data. *Remote Sensing*, 7(3):3020–3036, 2015.

[11] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[12] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. Cascade object detection with deformable part models. In *2010 IEEE Computer society conference on computer vision and pattern recognition*, pages 2241–2248. IEEE, 2010.

[13] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

[14] United States Fish and Wildlife Service. Endangered Species Act of 1973, 1973. SRKW listed as endangered in a 2005 amendment. Available at https://laws.justice.gc.ca/eng/acts/S-15.3/; accessed 25 September 2020.

[15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[17] Benjamin S Halpern, Melanie Frazier, Jamie Afflerbach, Julia S Lowndes, Fiorenza Micheli, Casey O'Hara, Courtney Scarborough, and Kimberly A Selkoe. Recent pace of change in human impact on the world's ocean. *Scientific reports*, 9(1):1–8, 2019.

[18] Bruce W Hartill, Stephen M Taylor, Krystle Keller, and Marc Simon Weltersbach. Digital camera monitoring of recreational fishing effort: Applications and challenges. *Fish and Fisheries*, 21(1):204–215, 2020.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Line Hermannsen, Lonnie Mikkelsen, Jakob Tougaard, Kristian Beedholm, Mark Johnson, and Peter T Madsen. Recreational vessels without automatic identification system (ais) dominate anthropogenic noise contributions to a shallow water soundscape. *Scientific reports*, 9(1):1–10, 2019.

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[23] Marla M Holt, Dawn P Noren, Val Veirs, Candice K Emmons, and Scott Veirs. Speaking up: Killer whales (orcinus orca) increase their call amplitude in response to vessel noise. *The Journal of the Acoustical Society of America*, 125(1):EL27–EL32, 2009.

[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[25] Frants Havmand Jensen, Lars Bejder, Magnus Wahlberg, N Aguilar Soto, M Johnson, and Peter Teglberg Madsen. Vessel noise effects on delphinid communication. *Marine Ecology Progress Series*, 395:161–175, 2009.

[26] Adrian Kaehler and Gary Bradski. *Learning OpenCV 3: computer vision in C++ with the OpenCV library*, pages 397–404. " O'Reilly Media, Inc.", 2016.

[27] 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.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[29] Wolfgang Krüger and Zigmund Orlov. Robust layer-based boat detection and multi-target-tracking in maritime environments. In *2010 International WaterSide Security Conference*, pages 1–7. IEEE, 2010.

[30] Darienne Lancaster, Philip Dearden, Dana R Haggarty, John P Volpe, and Natalie C Ban. Effectiveness of shore-based remote camera monitoring for quantifying recreational fisher compliance in marine conservation areas. *Aquatic Conservation: Marine and freshwater ecosystems*, 27(4):804–813, 2017.

[31] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[34] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[35] Zikun Liu, Jingao Hu, Lubin Weng, and Yiping Yang. Rotated region based cnn for ship detection. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 900–904. IEEE, 2017.

[36] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[37] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

[38] Tsubasa Minematsu, Atsushi Shimada, and Rin-ichiro Taniguchi. Background initialization based on bidirectional analysis and consensus voting. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 126–131. IEEE, 2016.

[39] Linda M Nichol, Brianna M Wright, Patrick O'Hara, and John KB Ford. Risk of lethal vessel strikes to humpback and fin whales off the west coast of vancouver island, canada. *Endangered Species Research*, 32:373–390, 2017.

[40] Douglas P Nowacek, Lesley H Thorne, David W Johnston, and Peter L Tyack. Responses of cetaceans to anthropogenic noise. *Mammal Review*, 37(2):81–115, 2007.

[41] Stephanie M Nowacek, Randall S Wells, and Andrew R Solow. Short-term effects of boat traffic on bottlenose dolphins, tursiops truncatus, in sarasota bay, florida. *Marine Mammal Science*, 17(4):673–688, 2001.

[42] Government of Canada. SARA (Species at Risk Act): An act respecting the protection of wildlife species at risk in Canada, 2002. S.C. 2002, c. 29. SRKW listed as endangered in a 2010 amendment. Available at https://laws.justice.gc.ca/eng/acts/S-15.3/; accessed 25 September 2020.

[43] Government of Canada Department of Fisheries and Oceans. Review of the effectiveness of recovery measures for southern resident killer whales, November 2016. Online; accessed 25 September 2020.

[44] Matthew K Pine, Andrew G Jeffs, Ding Wang, and Craig A Radford. The potential for vessel noise to mask biologically important sounds within ecologically significant embayments. *Ocean & Coastal Management*, 127:63–73, 2016.

[45] Tunai Porto Marques and Alexandra Branzan Albu. L2uwe: A framework for the efficient enhancement of low-light underwater images using local contrast and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 538–539, 2020.

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[47] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[48] Alireza Rezvanifar, Tunai Porto Marques, Melissa Cote, Alexandra Branzan Albu, Alex Slonimer, Thomas Tolhurst, Kaan Ersahin, Todd Mudge, and Stéphane Gauthier. A deep learning-based framework for the detection of schools of herring in echograms. *arXiv preprint arXiv:1910.08215*, 2019.

[49] MD Robards, GK Silber, JD Adams, J Arroyo, D Lorenzini, K Schwehr, and J Amos. Conservation science and policy applications of the marine vessel automatic identification system (ais)—a review. *Bulletin of Marine Science*, 92(1):75–103, 2016.

[50] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[51] Atsushi Shimada, Hajime Nagahara, and Rin-ichiro Taniguchi. Background modeling based on bidirectional analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1979–1986, 2013.

[52] CB Smallwood, KH Pollock, BS Wise, NG Hall, and DJ Gaughan. Expanding roving-aerial surveys to include counts of recreational shore fishers from remotely-operated cameras: benefits, limitations and cost-effectiveness. *North American Journal of Fisheries Management*, 32:1265–1276, 2012.

[53] Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 2, pages 246–252. IEEE, 1999.

[54] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE con-*

ference on computer vision and pattern recognition, pages 2818–2826, 2016.

[55] Jiexiong Tang, Chenwei Deng, Guang-Bin Huang, and Baojun Zhao. Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine. *IEEE Transactions on Geoscience and Remote Sensing*, 53(3):1174–1185, 2014.

[56] Thanh-Hai Tran and Thi-Lan Le. Vision based boat detection for maritime surveillance. In *2016 International Conference on Electronics, Information, and Communications (ICEIC)*, pages 1–4. IEEE, 2016.

[57] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[58] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[59] Robert Williams, Andrew J Wright, Erin Ashe, LK Blight, R Bruintjes, R Canessa, CW Clark, S Cullis-Suzuki, DT Dakin, Christine Erbe, et al. Impacts of anthropogenic noise on marine life: Publication patterns, new discoveries, and future directions in research and management. *Ocean & Coastal Management*, 115:17–24, 2015.

[60] Michael J Williamson, Ailbhe S Kavanagh, Michael J Noad, Eric Kniest, and Rebecca A Dunlop. The effect of close approaches for tagging activities by small research vessels on the behavior of humpback whales (megaptera novaeangliae). *Marine Mammal Science*, 32(4):1234–1253, 2016.

[61] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[62] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[63] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4):13–es, 2006.

[64] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[65] Ruiqian Zhang, Jian Yao, Kao Zhang, Chen Feng, and Jiadong Zhang. S-cnn-based ship detection from high-resolution remote sensing images. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 41, 2016.

[66] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 28–31. IEEE, 2004.

[67] Zoran Zivkovic and Ferdinand Van Der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, 27(7):773–780, 2006.

[68] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.