IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 19, 2021, accepted April 22, 2021, date of publication May 12, 2021, date of current version May 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3079132

LMD-TShip*: Vision Based Large-Scale Maritime Ship Tracking Benchmark for Autonomous Navigation Applications

YUNXIAO SHAN[1,4], SHANGHUA LIU[1], YUNFEI ZHANG[2,5], MIN JING[2], AND HUAWEI XU[3]

[1]Unmanned System Institute, Sun Yat-sen University, Guangzhou 510275, China
[2]Yunzhou Intelligent Technology Company Ltd., Zhuhai 519080, China
[3]The Fifth Research Institute of MIIT, Guangzhou 510610, China
[4]Shenzhen Research Institute, Sun Yat-sen University, Guangzhou 510275, China
[5]Southern Marine Science and Engineering Laboratory of Guangdong, Zhuhai 519080, China

Corresponding author: Huawei Xu (xuhuawei2018@163.com)

This work was supported in part by the Grant JCYJ20180508152434975, in part by the Key-Area Research and Development Program Grant of Guangdong Province under Grant 2018B010108004 and Grant 2019B090919003, and in part by the Fundamental Research Funds for the Central Universities Grant under Grant 19lgpy229.

ABSTRACT Accurate ship tracking is very important for the security of maritime activities, especially the raising requirements of autonomous navigation applications, e.g., autonomous surface vehicles (ASVs). Unlike deep-learning-based object-tracking methods are prevailing in autonomous driving because of good environmental robustness and high tracking accuracy, few deep-tracking models can be found for maritime ships. The main reason for that is the lack of qualified ship datasets, especially datasets with ship-based perspectives. Therefore, a large-scale, high-definition dataset for ship tracking, LMD-TShip (Large Maritime Dataset), is provided in this paper. In this dataset, five types of ships are included, from cargo ships, fishing ships, passenger ships, and speed boats to unmanned ships. Specifically, LMD-TShip consists of 40,240 frames in 191 videos, each of which is carefully and manually annotated with bounding boxes in YOLO format. Moreover, 13 attributes are used to label videos, e.g., scale variation (SV), occlusion (OCC), basically covering tracking challenges of maritime ship tracking. Next, a detailed analysis is carried out to demonstrate the characteristics of LMD-TShip. Finally, experiments with five baseline short-term tracking models on the dataset are performed, e.g., ECO, SiamRPN++, and the experimental results demonstrate its good evaluation ability, which will provide effective means for training and testing tracking models related to maritime ships.

INDEX TERMS Visual tracking, maritime environment, autonomous navigation, ship tracking dataset.

I. INTRODUCTION

As a part of autonomous driving, ASVs are arising as a new research field. Achieving autonomous navigation needs ASVs to accurately sense environment, and track surrounding objects, especially ships, to make sure sailing safety. In order to track objects, a variety of sensors are used to capture objects information. For example, traditional RADAR based ship trackers can sense targets even at long distance. However, the problem of blind spots is inevitable and the details of ships cannot be obtained, such as accurate shape, size, which are very important for the navigation of ASVs. To decrease blind area and provide more details, both thermal and visual cameras are widely used for this purpose. Due to the high energy consumption and long response time, thermal imaging cameras are more suitable for situations that do not require high real-time performance [1], [2], and visible light cameras with low energy consumption and fast imaging abilities are more preferred by ASVs [3], [4] [5]. However, with visible light cameras, unlike deep-learning-based trackers which are prevailing in autonomous driving, e.g., vehicle tracking [6] and pedestrian tracking [7], existing ship tracking methods for ASVs are still dominated by traditional methods [8], [9]. Developing deep tracking methods for AS-Vs in maritime environment faces many challenges, such as specifically-designed deep networks, possible network optimization according to characteristics of

The associate editor coordinating the review of this manuscript and approving it for publication was Huiping Li.

The dataset and benchmark can be found on https://yat-sen-robot.github.io/usilab-web/

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

ASVs, etc. Compared with these challenges, as data-driven methods, providing high-quality datasets should be a priority for developing deep trackers. However, as far as we know, open-source ship datasets with on-board perspectives which can be used to train and test deep trackers for ASVs are still unavailable by now.

There are several datasets for ship tracking. MarDCT [10] provided a video surveillance system and also constructed a dataset for building ship surveillance models. Nevertheless, only nine videos were included in the dataset, and the video resolution was relatively low. A larger dataset, the Seagull Dataset [11], was released in 2017, and included 150,000 images and was well labeled. However, most videos contain only one target with simple ocean backgrounds, and the ability of the tracking models to deal with complex tracking scenarios cannot be evaluated, such as cluttered backgrounds and occlusion. Moreover, the images of Seagull was captured in bird-eyes view, and it is difficult to train deep trackers used for applications with totally different perspectives, such as ASVs. To develop a dataset with ship-based perspectives, an abundant dataset, the Singapore Maritime Dataset [12] (SMD), which collected ship images on-shore and on-board, was considered. However, the available report was only made based on the evaluations of traditional ship-tracking methods, and all experimental results were performed with on-shore videos.

Therefore, a large-scale maritime ship dataset with ship-based perspectives, which can be used for training and testing deep trackers, is still urgently needed. However, building such a maritime ship dataset is a challenging task. First, the dataset should be large enough for training and testing deep trackers. In consideration of complex maritime environments, it is difficult to shoot a large number of videos, especially with on-board cameras. On-board videos require a data-collecting ship to carry equipment and an experienced captain to drive under appropriate weather. Nevertheless, a ship that is capable of withstanding various sea conditions is unaffordable for most research institutes, and maritime weather is always unpredictable. Second, the dataset should be challenging enough for evaluating SOTA tracking models, such as complex backgrounds and occlusion. Third, creating a qualified dataset will take both a significant number of workers and an amount of time to label and classify the videos. Moreover, a comprehensive analysis and test report regarding dataset performance are necessary for future users to train and test their models.

Aware of the aforementioned problems, the present authors have cooperated with a top R&D company of unmanned ships and took 1 year to build a short-term ship-tracking dataset, namely LMD-TShip. With the help of the company, both ships and captains became available for sailing with the proposed collecting equipment. In our efforts to collect images covering a wide range of maritime tracking scenarios, ship videos were continuously shot in multiple types of weather, time, backgrounds, and platforms. Afterward, targets were annotated with accurate bounding boxes manually and the challenging attributes of each video were labeled. Moreover, a comprehensive statistical analysis of dataset attributes was carried out to reflect the characteristics of the dataset. Concretely, 13 attributes are discussed, and main attributes are all included, such as scale variation, occlusion, multiple targets. Finally, five baseline tracking models are used to test the effectiveness of the dataset. Specifically, a maritime tracking dataset with the following contributions is provided.

1) The dataset contains 191 videos with an average frame count of 212 and basically covers the main characteristics of maritime ship tracking. In each video, every frame is labeled manually and carefully. To the best of our knowledge, the proposed dataset is one of the largest high-quality maritime tracking datasets with dense and accurate annotation, detailed attribute analysis (up to 13 attributes are discussed), and experimental tests.
2) Thirteen different challenging attributes of the dataset are analyzed in detail, which clearly shows the challenges of maritime ship-tracking tasks and provides possible heuristics for the follow-on development of SOTA maritime ship trackers.
3) Experiments were carried out with five kinds of baseline trackers to show the effects of the dataset and demonstrate their advantages and disadvantages in dealing with challenging attributes. Moreover, a training and testing protocol was established and validated so that future researchers can simply follow and then evaluate their specifically designed trackers.

The rest of this paper is organized as follows: In Section II, related literature is reviewed. In Section III, data acquisition is discussed. In Section IV, the details of the dataset are presented. The experimental results of five baseline tracking models on the dataset are provided in Section V. Finally, conclusions are summarized in Section VI.

A. OBJECT-TRACKING DATASETS

At present, the existing tracking datasets can be roughly divided into two categories, comprehensive datasets and ship datasets.

1) COMPREHENSIVE DATASETS

OTB100 [13] was proposed in 2015 with 100 videos in total, 25% of which are gray-scale sequences. Moreover, each video was labeled with attributes that represent the common difficulties in the field of target tracking. VOT [14] was released next and contained 60 videos with high-resolution color sequences. For OTB and VOT, they are not designed for specific objects, covering people, animals, vehicles, etc., and more emphasize the universality of trackers. Unlike OTB100 and VOT, which emphasize short-term challenging tracking tasks, LTTW [15] is specially designed for long-term tracking, covering human beings, vehicles, and animals in the wild, and comprises 366 sequences for a total of 4h of video. Recently, the largest densely annotated tracking dataset, LaSOT [16], was published, which consisted

TABLE 1. Comparison with other maritime tracking dataset.

Dataset	Videos	Frames	Annotation	Attributes	Types
MarDCT	9	23200	692	N/A	N/A
Seagull	19	150K	48408	N/A	6
SMD	36	17450	17450	N/A	7
LMD-TShip	191	200K	40240	13	5

of 1,400 sequences with more than 3.5M frames in total and contained most common tracking targets. In addition, LaSOT provided additional language specifications, considering the connections of visual appearance and natural language. Besides, recently, several new tracking datasets are released [17], [18]. Although these datasets nearly cover almost most of common tracking objects, the particularity of maritime ships cannot be reflected from these common scenarios, such as irregular violent shaking, remarkable scale variant, etc. Therefore, it is necessary to build a professional marine ship dataset for training and testing deep trackers for ship-based applications.

2) SHIP DATASETS

Most of the currently available maritime datasets are built for ship detection, such as Seaships [19], Marvel [20], MODD1 [21], and MODD2 [22]. Comparatively, there are few maritime datasets for tracking. MarDCT [10] was collected by a surveillance system and contained only nine videos of ships. The Seagull Dataset [11] provided a rich dataset for ship tracking. The videos in the seagull dataset were collected by an airplane, and although there were 19 videos covering six ship types, it had fewer objects in each category. Moreover, the viewpoints of the videos were static and were distributed between 150 and 300 m above the ocean surface. Recently, the Singapore Maritime Dataset [12] (SMD) was released, including 32 videos collected on-shore and four videos collected on-board. However, the lack of detailed analysis and performance tests on deep trackers have limited its wide application.

B. OBJECT-TRACKING METHODS

Compared with traditional generative models mentioned in [23], the discriminant tracking method transforms the object tracking problem into a binary classification problem which seeks the decision boundary between the tracking target and the background, and separates the foreground from backgrounds by maximizing the classification of targets. According to the different classification strategies proposed, discriminative tracking models can be divided into correlation-filter-related tracking models, and deep-learning-based tracking models [24].

The correlation-filter-based trackers compute confidence scores with cross-correlation between the template frame and current frame, and the highest score in confidence graphs corresponds to the predicted position of targets. MOSSE [25] was the first method that applied correlation filters for tracking. Based on MOSSE, CSK [26] introduced the kernel method and dense sampling single-channel gray features. KCF uses multi-channel HOG features. However, when the scale of a target becomes larger, KCF [27] can only detect parts of the target. For adaptation to different scales, DSST [28] was proposed to detect the targets with changing scales by adding scale filters. Although better performance has been validated, the lack of deep features limits its wide application. Recently, several correlation-filter-based models with deep features have been proposed, e.g., C-COT [29] and ECO [30].

With the development of deep-learning technology, deep-tracking models have gradually become the mainstream tracking methods, especially the siamese-network-based models. SiamFC [31] first introduced the siamese neural network into the field of target tracking, which inputs a template frame and current frame into a siamese network and then obtains two separate outputs. Nevertheless, both accuracy and efficiency are not satisfactory. To develop SOTA trackers with high speed, SiamRPN [32] was proposed with both siamese and region proposal subnetworks. In SiamRPN, the siamese subnetwork with shared parameters is used to extract the features of the template and current frame, and the region proposal subnetwork generates proposals by computing correlation both in the classification branch and regression branch. Then, an improved version of SiamRPN, SiamRPN++ [33], was advanced. To improve discriminative ability, SiamRPN++ used multilayer fusion to integrate shallow and deep features to obtain more details and semantic information. Aware of the advantages of siamese structures, many siamese-based models have emerged, such as SiamFC++ [34], SiamVGG [35], and DaSiamRPN [36]. Besides deep-learning based models, several correlation-filter-related tracking models combine learning network and correlation filters to complement each other and achieve SOTA tracking performance, e.g., ATOM (Accurate Tracking by Overlap Maximization) [37], DiMP (Learning Discriminative Model Prediction for Tracking) [38].

Compared with the rapid progress of deep-learning-based tracking methods in other fields, few works exist in the field of ship tracking. Although there were several traditional tracking methods, such as hidden Markov models [39], Kalman filters [10], [40], and optical flow [41], SOTA deep-learning-based trackers specially designed for ship tracking were rare due to the lack of an available dataset.

II. DATA ACQUISITION

In this section, the proposed video camera system and the strategy considered before collecting the dataset are introduced.

A. VIDEO CAMERA SYSTEM

Images are collected on-shore and on-board. With the assistance of data-collecting ships, two cameras are fixed on moving boats. Fig. 1 shows three types of ships used for collecting

FIGURE 1. Different ships equipped with sensors.

data. Moreover, on-shore data were collected by factory-built cameras and mobile phones (iPhone 5 and iPhone 5s) when weather conditions prohibited data collection. Therefore, due to the differences in acquisition equipment, the sizes of images in the dataset vary; Table 2 shows their distributions.

TABLE 2. The number and ratio of each image size.

Size	Number	Ratio
1920×1080 or 1080×1920	35161	0.874
640×480 or 480×640	2596	0.065
960×540	2308	0.057
1280×720	175	0.004

B. DATASET DIVERSITY

A good tracking dataset should be able to include various factors that can affect tracking performance. According to the characteristics of maritime ships, the following properties are included in the proposed tracking dataset:

1) SCALE

In the maritime environment, the appearance of different types of ships varies remarkably. Even the same ship type may be very different. Therefore, the dataset comprises different types of ships, as well as different appearances under the same type. To balance the dataset, the number of ships belonging to the same type of ship but with different appearances is basically the same.

2) WEATHER

Changing weather is one of the typical characteristics of the maritime environment, and different weather conditions can significantly affect the quality of images. Therefore, an attempt was made to collect images under different weather conditions.

3) SHAKING

Shaking is a normal state of targets on the sea surface, and the degree of shaking depends on different sea conditions. Moreover, the influence of shaking is not only manifested as the instability encountered in tracking targets, but also that of the acquisition equipment. Therefore, collecting sufficient and diversified shaking scenarios plays an important role in testing tracking models. In the proposed dataset, tracking videos with different degrees of shaking in shaking scenes were collected by driving the data-collecting ship with cameras at different speeds.

4) BACKGROUND

Although most of the tracking scenes on the sea are ocean and sky, they are not static but dynamic. In addition, coastlines are also an important background factor for maritime target tracking, such as buildings and other ships moored on the coast. Therefore, to test the ability of tracking models in complex backgrounds, an attempt was made to collect data from different backgrounds and reduce the proportion of simple backgrounds.

5) OCCLUSION

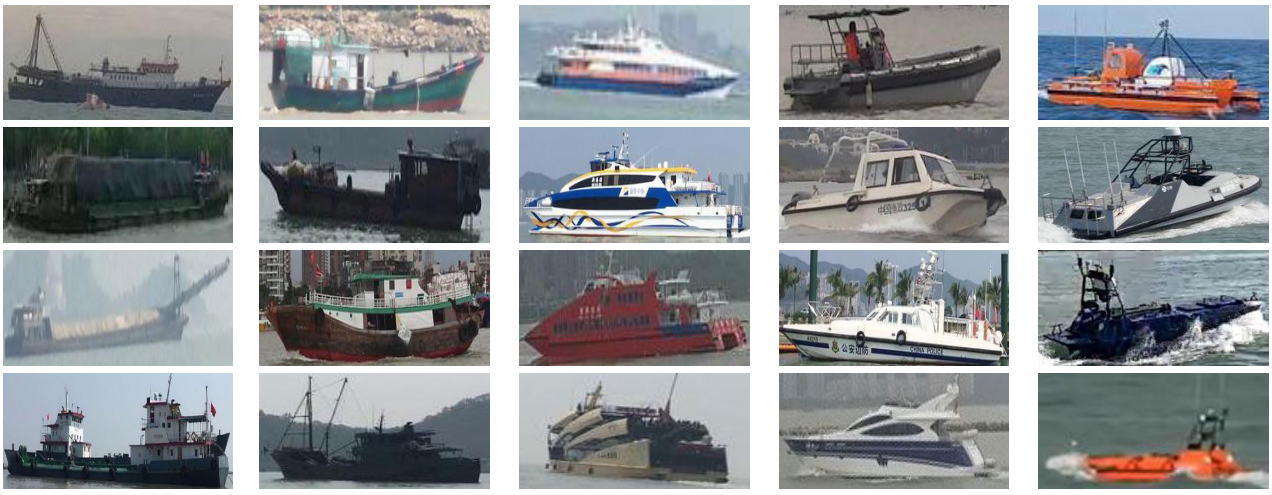
Occlusion is one of the most difficult problems in tracking tasks. For the case of maritime environments, the occlusion is most likely to occur in ports where there are many ships at anchor. To reflect actual occlusion situations, the proposed dataset includes different degrees of occlusion of different types of ships (from no occlusion to full occlusion).

6) SIMILARITY

The discrimination of similar targets is also an important indicator in evaluating the accuracy of trackers. In the maritime environment, the same types of ships may be very similar in size and color, and even human observers cannot identify them accurately. To make the dataset more challenging, a number of similar targets were collected, accounting for approximately 16% of the total number of videos.

C. ANNOTATION

To efficiently carry out annotation, cued by YT-BB [42], a labeling strategy to annotate every five frames of video sequences was devised. For the target-labeling process, a discriminative labeling method was developed by [16]. Specifically, if an object appears in the frame, a rectangular box is used to mark it; otherwise, if it is occluded, only the visible part is labeled. Although this method can be used directly, the annotation of ships will take extra effort. For example, the long and thin artificial constructed antennas and masts make the annotation of ships difficult. These parts can provide little information for localizing the ships and will introduce serious background noise. Therefore, they are carefully removed, as shown in Fig. (3a). Moreover, considerable time and manpower were directed to the process of double-checking and error-correcting. To achieve accurate annotation of the proposed dataset, the data-labeling team was divided into annotation and validation teams. The validation team reviewed all of the images annotated by the annotation team and verified the minimized area of the target box for every frame. Fig. (3b) shows a ship-overlap example that must be checked carefully. As a result of the aforementioned efforts, a well-labeled dataset was achieved.

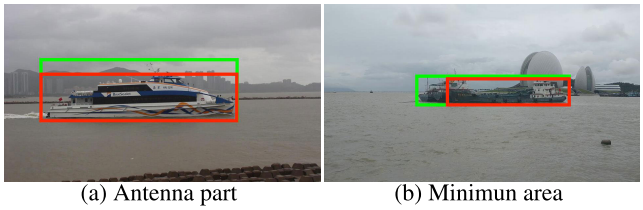


(a) cargo ship (b) fishing ship (c) passenger ship (d) speed boat (e) unmanned ship

FIGURE 2. Examples of five ship types.

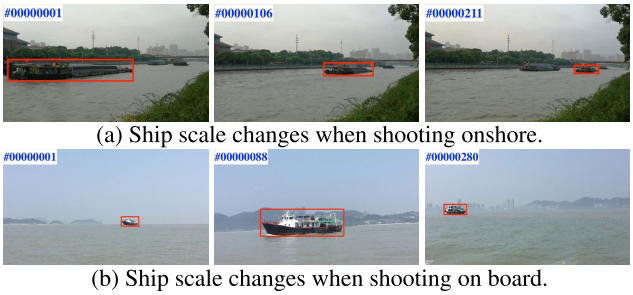
TABLE 3. The number and ratio of each ship category.

Ship Category	Frames	Videos	Mean Frames	Ratio
Cargo ship	7777	43	180	0.193
Fishing boat	11676	60	195	0.290
Passenger ship	7441	31	240	0.185
Speedboat	6284	18	349	0.156
Unmanned ship	7062	39	181	0.176
Total	40240	191	211	1.000



(a) Antenna part (b) Minimun area

FIGURE 3. Examples of double-check annotation. The initial annotation is in green and the fine-tuned annotation through double-check strategy is in red.



(a) Ship scale changes when shooting onshore.

(b) Ship scale changes when shooting on board.

FIGURE 4. Examples of scale change when shooting onshore and onboard.

III. DATASET DESIGN AND STATISTICS

A. DATASET DESIGN

To build the dataset, the collected data was categorized into five-ship categories: cargo ships, fishing boats, passenger ships, speedboats, and unmanned ships; these categories generally cover common ships. Fig. 2 shows examples of these ships and Table 3 the numbers of frames and videos in each ship category. Next, the attributes of every video were analyzed and appropriate attributes were then attached to the videos. To standardize the dataset, inspired by [16], [43], 13 attributes were deployed to describe the videos based on the characteristics of maritime ships, including *scale variation* (SV), *aspect-ratio change* (ARC), *fast and irregular motion* (FIM), *low resolution* (LR), *out-of-view* (OV), *illumination variation* (IV), *image quality* (IQ), *max scale variation* (SVM), *camera motion* (CM), *background clutter* (BC), *similar object* (SOB), *occlusion* (OCC), and *viewpoint change* (VC). Table 4 describes each attribute. However, in consideration of the particularity of the maritime environment, the representatives of several attributes are very different from other datasets. Specifically, the attributes are explained in detail.

1) SV AND SVM

The scale variation of the proposed dataset is mainly caused by two reasons: (1) the change of the distance between targets and cameras, and (2) the different sizes of ships. Fig. 4 shows the scale change of a ship in one video sequence caused by the change of distance when shooting on-shore and on-board separately. Additionally, the sizes of different ships vary remarkably. Fig. 5 shows the distribution of ship width and height. In general, the ratio (width/height) ranges from 0.75 to 9.21. The smallest ship occupies 13×2 pixels, but the largest ship occupies 1643×490 pixels. In particular, there are over

TABLE 4. Definition of 13 different attributes in LMD-TShip.

Attribute	Description	Attribute	Description
ARC	A certain ratio of ship aspect ratios is outside [0.5, 2] after 1s.	**OV**	A certain portion of the ship leaves the video frames.
IQ	A certain score of image is smaller than 10 with Laplacian sober [44].	**LR**	A certain area of ship is smaller than 2000 pixels.
FIM	A certain irregular motion of ship is larger than the size of ship.	**OCC**	At least some parts of ship are occluded.
SV	A certain ratio of bounding box area is outside [0.5, 2] after 1s.	**IV**	The illumination in the ship regions change greatly.
BC	The background has a similar appearance as the ship.	**SOB**	The target is less than 200 pixels away from similar ships.
VC	The appearance changes significantly due to viewpoint.	**CM**	Camera moves abruptly (decided by annotators).
SVM	The maximal ratio of bounding box area is outside [0.5, 2].		

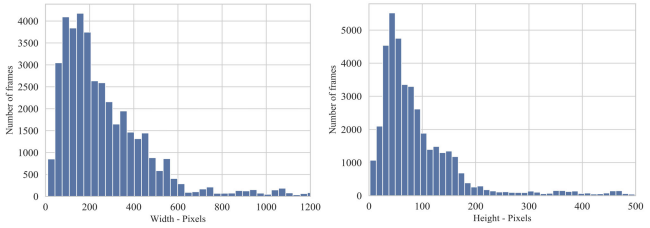


FIGURE 5. Distribution of width and height of bounding box.

(a) Foggy (b) Sunny (c) Cloudy (d) Rainy

FIGURE 6. Examples of different weathers and illuminations. Each column represents a weather with different illuminations (the top is brighter).

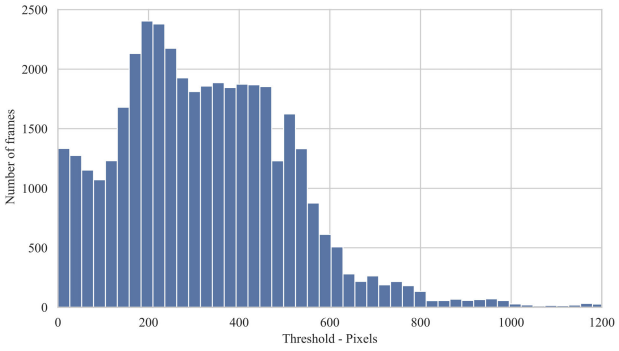


FIGURE 7. The distance of center points of ground truth boxes in two consecutive frames.

FIGURE 8. Examples of different background.

3,000 targets smaller than 2,000 pixels, and at least one frame measures less than 2,000 pixels in 32 videos, which makes the proposed dataset more challenging. Moreover, more than one-third of frames with a target smaller than 2,000 pixels belong to a speedboat.

2) IV AND IQ

Changing weather is a classical characteristic of the maritime environment, e.g., sunshine, cloudy, etc., which will degenerate the quality of images. Image quality is also affected by focus accuracy and exposure accuracy. Moreover, since different illuminations have different impacts on the quality of images, data were collected at noon with strong lighting conditions and at dusk with weak lighting conditions. Fig. 6 shows the different weather conditions and illumination conditions.

3) FIM, CM, AND ARC

The FIM, CM, and ARC attributes are related to motion. For the motion of ships, fast maneuvering and shaking are two main characteristics. Unlike land vehicles, which must travel on a structured road, ship maneuvering is more arbitrary, which may cause the change of ARC and become difficult to track. Moreover, the irregular motion caused by shaking must also be considered. To evaluate the motion attribute of the proposed dataset, the distance between the center of a target in consecutive frames was calculated. Fig. 7 shows different distance thresholds and the number of frames in given thresholds. A significant observation is that half of the frames (20,498 frames belonging to 148 videos) have a relatively large deviation compared to the previous frame.

4) BC

To obtain diversity of backgrounds, data were collected at different places on-shore and on-board. Fig. 8 shows several different backgrounds in the dataset.

5) OCC

In tracking scenes, occlusion has a significant impact on the recognition and tracking of objects. In the maritime


(a) Speedboat (Partially occluded)


(b) Cargo ship (Partially occluded)


(c) Passenger ship (Partially occluded)

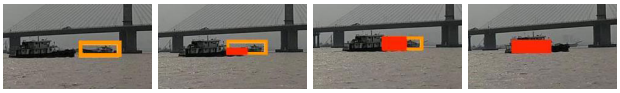

(d) Cargo ship (Fully occluded)

FIGURE 9. Examples of different degree of occlusion situation (color in red). Each row represents one video sequence. From left to right, each column shows that the target ship (color in orange) becomes more and more occluded over time.

FIGURE 10. Examples of vessels with same type and similar appearance. The target ship is in orange and other similar distractors are in blue and purple.

environment, especially at ports where ships are moored, the target ship is often occluded by other ships. More than 5% of the total videos present occlusion situations, and some videos even have been completely occluded. Fig. 9 shows different degrees of occlusion in different sequences, and particularly sequence *Cargo ship (Fully occluded)* in Fig. (9d) shows a complete occlusion situation. It is worth noting that we only annotate ships' visible areas.

6) SOB

Although there are great differences among different types of ships, the appearance of the same type of ship can be very similar, which brings great challenges to ship tracking. Therefore, the proposed dataset contains images of similar ships in unified scenes. Fig. 10 shows several scenes that may affect the performance of trackers.

7) LR, OV, AND VC

As described in Table 4, LR, OV, and VC are basically the same as in other datasets.

B. DATASET STATISTICS

To further demonstrate the characteristics of the proposed dataset, statistical analysis of the attributes of the proposed dataset was carried out and the weights of attributes are shown in Fig. 11. In Fig. 11, it can be observed that the tracking challenges of maritime object tracking are mainly the results of scale (SVM, SV, and ARC), motion (FIM and CM), resolution (LR and IQ), and background (BC and SOB), which means that a SOTA ship tracker should pay more attention to these attributes.

IV. EXPERIMENT

Five SOTA trackers were evaluated on the proposed dataset, including DSST [28], KCF [27], ECO [30], SiamRPN [32], and SiamRPN++ [33]. To evaluate LMD-TShip comprehensively, separate tests were carried out on the entire dataset (Protocol I) and testing dataset (Protocol II).

A. BASELINE TRACKERS

KCF is well-known for its tracking efficiency and ECO combines hand-crafted features with deep features and stands out in correlation filter trackers. DSST is an improved version of KCF and can adapt to the scale change of targets by combining a position filter and scale filter. Among the deep trackers used, SiamRPN was the winner of VOT [14] from 2015 to 2017, and SiamRPN++ improves the tracking accuracy of SiamRPN by introducing multilayer feature maps. To further investigate the impacts of LMD-TShip on deep-learning-based models, besides the original AlexNet [45] used in SiamRPN, two different backbones were deployed in this paper, including VGG16 [46] and ResNet50 [47], namely SiamRPN-VGG16 and SiamRPN-ResNet50. Moreover, in most tracking models a penalty module is widely used to punish dramatic displacement and scale change to improve the overall tracking accuracy. In the reference phase of SiamRPN, a penalty module consisting of a cosine window and scale-change penalty is used when selecting proposals. The cosine window is used to suppress the large displacement and the scale-change penalty is proposed to prohibit large changes in size and ratio. With this penalty module, SiamRPN re-ranks the scores of proposals and chooses the best one by NMS. Whether the penalty module is appropriate for maritime ship-tracking applications was a question to be answered in the experiments conducted in this work. For this purpose, "_No" was added to indicate the tracker without a penalty module.

The special appearance of a ship challenges the selection of anchor sizes. According to the method proposed in [23], eight appropriate aspect ratios listed in Section IV-B were obtained and applied to the anchor generation phase in all siamese trackers.

B. EXPERIMENTAL SETUP

CNN is initialized in siamese trackers without any pre-trained model. The learning rate was reduced by 10% for every 100 epochs. To initialize anchor parameters in generation phrase, eight anchors were set by the K-means method [23], namely {(470, 157), (83, 30), (40, 20), (119, 56), (177, 44), (894, 359), (337, 106), and (230, 75)}. Other parameters for

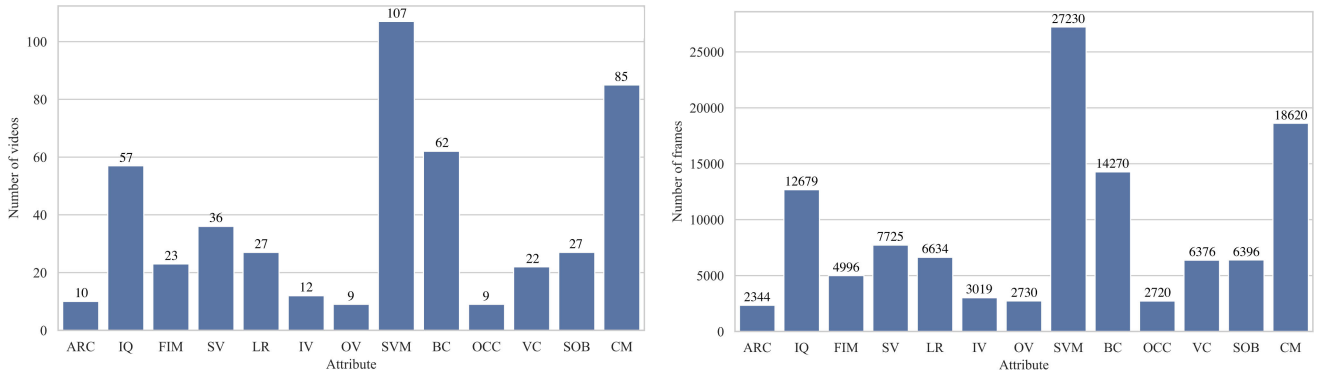


FIGURE 11. Distribution of videos and frames in each attribute.

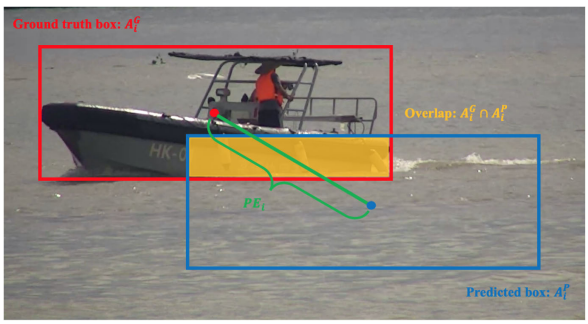


FIGURE 12. Pixel error and intersection over union.

training the network can be found in [32]. All experiments were conducted on Tesla V100s and an Intel(R) Xeon(R) Gold 6134 CPU@3.20 GHz.

C. EVALUATION METRIC

One-pass evaluation (OPE) - using groundtruth to initialize only the first frame of tracking sequences, and metrics were used to evaluate the tracking performance, which were proposed in OTB [13] and are briefly described below.

1) AVERAGE PIXEL ERROR

Average pixel error is denoted APE. This metric is used to measure the distance between the center point of the predicted bounding box and that of the annotated ground truth box, as shown in Fig. 12.

If (X_i^G, Y_i^G) is the center point of the ground truth box and (X_i^P, Y_i^P) that of the predicted box, then the calculation formula is expressed as follows:

$$PE_i = \sqrt{(X_i^G - X_i^P)^2 + (Y_i^G - Y_i^P)^2} \tag{1}$$

$$APE = \frac{1}{N}\sum_{i=1}^{N} PE_i \tag{2}$$

where PE_i is the Euclidean distance between the ground truth box and predicted bounding box at time step i. N is the number of frames.

2) AVERAGE OVERLAP

Average overlap is computed through the intersection over union (IOU). As shown in Fig. 12, IOU measures the overlap degree of the predicted pose A_i^P and ground truth pose A_i^G at time step i. It can be computed as follows:

$$IOU_i = \frac{A_i^G \bigcap A_i^P}{A_i^G \bigcup A_i^P} \tag{3}$$

$$AO = \frac{1}{N}\sum_{i=1}^{N} IOU_i \tag{4}$$

3) SUCCESS PLOT

When the IOU of a frame is higher than a given threshold (e.g., 0.7), this frame is regarded as successful. The success rate is the proportion of the number of successfully tracked frames to the number of all frames. The success plot can be drawn according to different thresholds.

4) FRAMES PER SECOND

Frames per second (FPS) denotes the number of images that can be processed by the tracking algorithm per second. This metric is a crucial protocol used to measure the real-time performance of trackers.

D. RESULTS AND ANALYSIS

The overall performance of all baselines in Section *Overall performance* was evaluated, and then the impacts of the 13 attributes on baselines in Section *Attribute performance* were investigated. The evaluation method in [16] was used for reference in the present work. Inspired by [16], two evaluation protocols established for the experiments.

Protocol I. In this protocol, 11 trackers on all 191 video sequences in the proposed dataset were evaluated. Protocol I aims at providing a large-scale evaluation.

Protocol II. Under this protocol, the proposed dataset was split into a training set and a testing set. The testing set selected randomly accounts for approximately 20% of total frames, and the rest comprises the training set. Protocol II aims at providing a set of sequences for training and testing the tracking models. By comparing the similar experimental

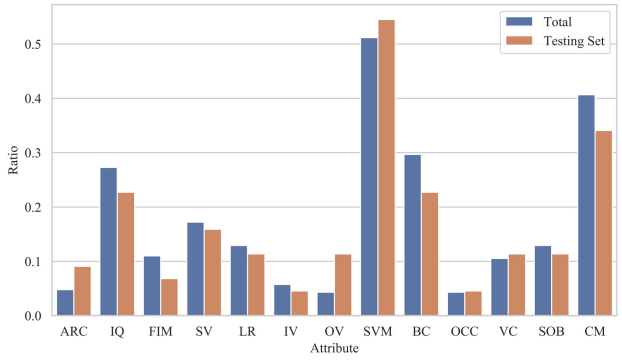


FIGURE 13. Distribution of attributes in total and testing set.

results in Protocols I and II, the rationality of dataset partitioning is verified and reference results are provided for future researchers. Table 5 shows the details of the training and testing sets and Fig. 13 demonstrates the distribution of the 13 attributes in the testing set and all sequences, which shows that the proportion of each attribute is approximately the same in the testing set and all sequences.

TABLE 5. Comparison between training and testing set.

	Videos	Frames	Max Frames	Mean Frames
Training Set	152	31527	1689	207
Testing Set	39	8713	984	223
Total	191	40240	1689	212

1) OVERALL PERFORMANCE

Under Protocol I and II, the AO and APE of each category and the average results were reported, as shown in Tables 6 and 7, respectively. Roughly speaking, the results of Protocol II are generally consistent with the results of Protocol I, which further validates the reasonableness of the dataset partitioning in Protocol II.

Upon careful inspection of Tables 6 and 7, first, a widely accepted result is observed, namely that the correlation-filter-based methods maintain high speed but poor accuracy, on average. Moreover, the AO value of ECO is higher than all original Siamese trackers(with Penalty Module), which shows that the combination of a filter and deep features has great advantages in the maritime environment, but the real-time performance is poor. The second observation is that the performance of Siamese trackers without a penalty module (SiamRPN_No, SiamRPN-VGG16_No, SiamRPN-ResNet50 _No, and SiamRPN++_No) outperform their respective original trackers, which demonstrates that the penalty module degrades the performance of siamese trackers in the maritime environment. The third observation is that, with the deepening layers of backbones, the AO score of the SiamRPN-based trackers (SiamRPN, SiamRPN-VGG16, and SiamRPN-ResNet50) increases, which proves that enhanced feature-extraction ability can improve tracking performance. In addition, SimRPN++_No achieves the highest AO scores of 0.755 and 0.733 in Protocols I and II, respectively. Without a penalty module, SimRPN++_No combines multi-layer feature maps to demonstrate SOTA ship tracking performance.

In addition to the AO value, the APE value also significantly reflects the aforementioned observations. Nevertheless, an interesting observation is that SiamRPN-ResNet50 trackers perform the worst on APE in both Protocols I and II, namely 255 for SiamRPN-ResNet50 and 383 for SiamRPN-ResNet50_No in Table 6 and 150 for SiamRPN-ResNet50 and 341 for SiamRPN-ResNet50_No in Table 7. The main reason for this is the poor feature-extraction ability of ResNet50 on small targets. SiamRPN-ResNet50 can extract the stronger feature by ResNet50 with a larger receptive field. Although a robust feature-extraction ability and large receptive field can improve accuracy for large-scale targets, the large receptive field may lead to a lower resolution and ignore the location of small targets when the target only occupies a few pixels. In LMD-TShip, most of the small targets are distributed over speedboats, as mentioned in *Section III-A1*. Therefore, both SiamRPN-ResNet50 and SiamRPN-ResNet50_No perform worst in terms of APE for speedboats, as shown in Tables 6 and 7.

2) ATTRIBUTE PERFORMANCE

Fig. 14 and 15 show the performance of 11 trackers on the attributes of *scale variation* (SV) and *max scale variation* (SVM) on all sequences and the testing set. In LMD-TShip, video sequences with the attributes of SV and SVM are mainly caused by the changing distance between cameras and targets, the target entering or leaving the images, and the change of viewpoint. In principle, DSST should perform the best since it is specially designed for adapting to changing scales. Nevertheless, the worst performance is observed for DSST, even worse than KCF and ECO. The main reason for this is that the huge scale gap between different ships and sudden location changes due to shaking make the scale-adaptation module in DSST unable to improve IOU with an equal scaling function, but increases the background in the corresponding bounding box, which leads to increased tracking-failure probability.

Fig. 16 and 17 show the performance of 11 trackers on the attributes of *illumination variation* (IV) and *image quality* (IQ) on all sequences and the testing set. The sequences with attributes IV and IQ are mostly caused by changing light conditions at sea and weather with low visibility, such as overcast and foggy days. One can observe the overriding advantage of deep features when dealing with complicated scenarios, which may degenerate the quality of images and make it difficult to complete HOG features. Poor HOG features will decrease the performance of correlation-filter-based models, e.g., KCF and DSST. However, deep trackers can still main-

TABLE 6. Quantitative evaluation of different tracking algorithms under protocol I. The best values for each metric and ship category (AO | APE) using bold font.

Model	Penalty module	AO ↑	APE ↓	Cargo ship	Fishing boat	Passenger boat	Speed boat	Unmanned ship	FPS ↑
DSST	-	0.329	111	0.322 | 199	0.264 | 126	0.424 | 52	0.298 | 165	0.372 | 90	84
KCF	-	0.573	84	0.618 | 70	0.552 | 110	0.625 | 46	0.476 | 139	0.588 | 48	**135**
ECO	-	0.649	61	**0.761** | **12**	0.714 | 46	0.674 | 27	0.414 | 190	0.599 | 59	9
SiamRPN	✓	0.496	92	0.454 | 74	0.562 | 67	0.565 | 115	0.460 | 170	0.398 | 62	74
	×	0.654	61	0.626 | 41	0.688 | 30	0.732 | 36	0.484 | 212	0.702 | 26	74
SiamRPN-VGG16	✓	0.570	82	0.540 | 54	0.558 | 123	0.590 | 89	0.554 | 85	0.626 | 35	57
	×	0.663	61	0.621 | 41	0.668 | 80	0.765 | **15**	0.571 | 121	0.679 | 49	57
SiamRPN-ResNet50	✓	0.587	255	0.592 | 58	0.596 | 663	0.571 | 118	0.508 | 166	0.654 | 25	33
	×	0.696	383	0.675 | 50	0.688 | 999	0.729 | 38	0.660 | 476	0.733 | **21**	33
SiamRPN++	✓	0.550	105	0.510 | 67	0.599 | 90	0.521 | 162	0.500 | 177	0.589 | 48	13
	×	**0.755**	**30**	0.708 | 31	**0.786** | **21**	**0.770** | 21	**0.716** | **62**	**0.775** | 25	13

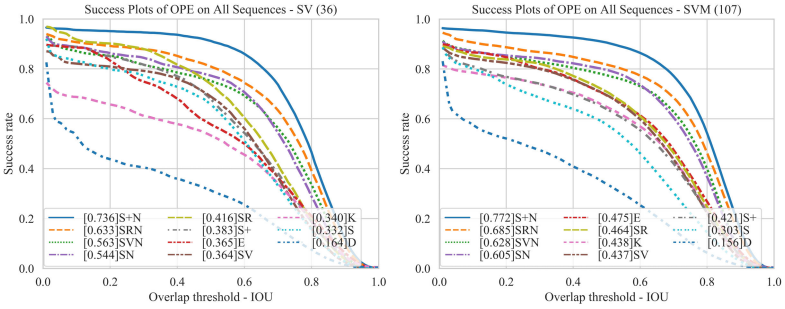
TABLE 7. Quantitative evaluation of different tracking algorithms under protocol II. The best values for each metric and ship category (AO | APE) using bold font.

Model	Penalty module	AO ↑	APE ↓	Cargo ship	Fishing boat	Passenger boat	Speed boat	Unmanned ship	FPS ↑
DSST	-	0.284	94	0.373 | 24	0.294 | 113	0.441 | 56	0.116 | 147	0.274 | 112	84
KCF	-	0.588	57	0.699 | 23	0.703 | 52	0.496 | 79	0.470 | 82	0.572 | 38	**135**
ECO	-	0.592	75	**0.783** | **6**	0.764 | 30	0.620 | 43	0.217 | 225	0.677 | 33	9
SiamRPN	✓	0.472	111	0.311 | 112	0.594 | 36	0.518 | 205	0.460 | 140	0.478 | 54	74
	×	0.592	78	0.605 | 25	0.673 | 32	0.728 | 12	0.410 | 217	0.636 | 35	74
SiamRPN-VGG16	✓	0.542	82	0.455 | 79	0.588 | 43	0.480 | 263	0.572 | **39**	0.607 | 41	57
	×	0.665	**48**	0.643 | 43	0.738 | 21	0.766 | 14	0.566 | 81	0.646 | 77	57
SiamRPN-ResNet50	✓	0.549	150	0.590 | 59	0.573 | 46	0.472 | 225	0.535 | 109	0.568 | 40	33
	×	0.689	341	0.782 | 13	0.708 | 31	0.762 | 20	0.55 | 1190	0.705 | **27**	33
SiamRPN++	✓	0.485	145	0.457 | 53	0.653 | 33	0.430 | 277	0.408 | 264	0.461 | 63	13
	×	**0.733**	51	0.737 | 28	**0.796** | **7**	**0.792** | **9**	**0.673** | 96	**0.724** | 87	13

tain good performance by extracting deep features when the features are weakened.

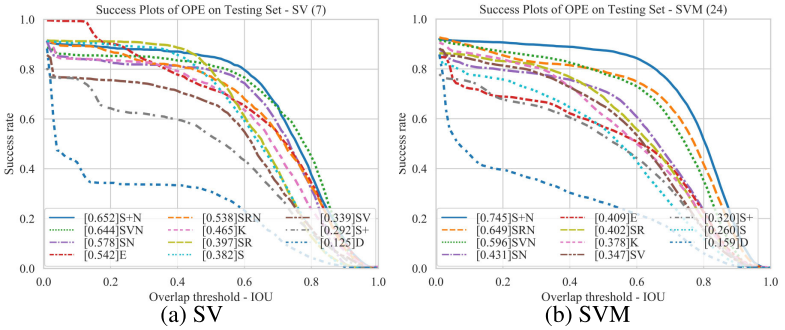
Fig. 18 and 19 show the performance of all trackers on the attributes of *fast irregular motion* (FIM), *camera motion* (CM), and *aspect ratio change* (ARC) on all sequences and the testing set. Most of the sequences with the attributes of FIM, CM, and ARC are caused by the fast movement and shaking of ships, which may result in boundary effects. For correlation-filter-based trackers, the boundary effects can be defined as follows: When the target moves to the boundary of the detection area, the cosine window will filter out the target pixels and fail to track. As shown in Fig. 18, the performance of both KCF and DSST is poor. For deep trackers, when the target leaves the search area, the trackers fail to locate the target. Compared with the success plots in Protocol I, the performance of baseline models on the attributes FIM, CM, and ARC decreases; the best success rate under Protocol II is 0.789, but on FIM, CM, and ARC it is 0.614, 0.732 and 0.581, respectively. The dramatic decrease indicates that FIM, CM, and ARC are the main challenges in maritime tracking.

Fig. 20 and 21 show the performance of all trackers on the attributes of *background clutter* (BC), *occlusion* (OCC),



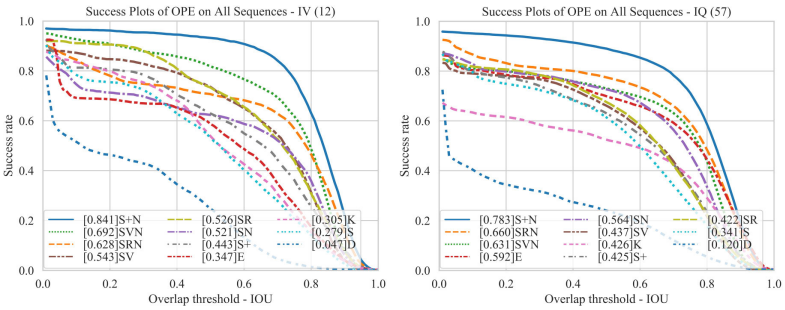
(a) SV (b) SVM

FIGURE 14. Results of scale variation under protocol I using success rate. For brevity, KCF, ECO, DSST, SiamRPN, SiamRPN-VGG16, SiamRPN-ResNet50, SiamRPN++, SiamRPN_No, SiamRPN-VGG16_No, SiamRPN-ResNet50_No and SiamRPN++_No are abbreviated as K, E, D, S, SV, SR, S+, SN, SVN, SRN and S+N respectively.



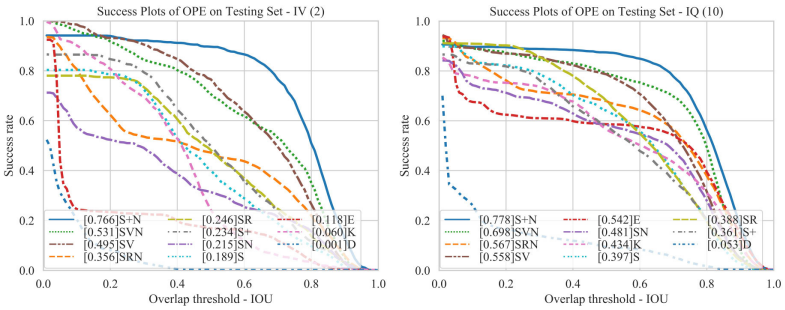
(a) SV (b) SVM

FIGURE 15. Results of scale variation under protocol II using success rate.



(a) IV (b) IQ

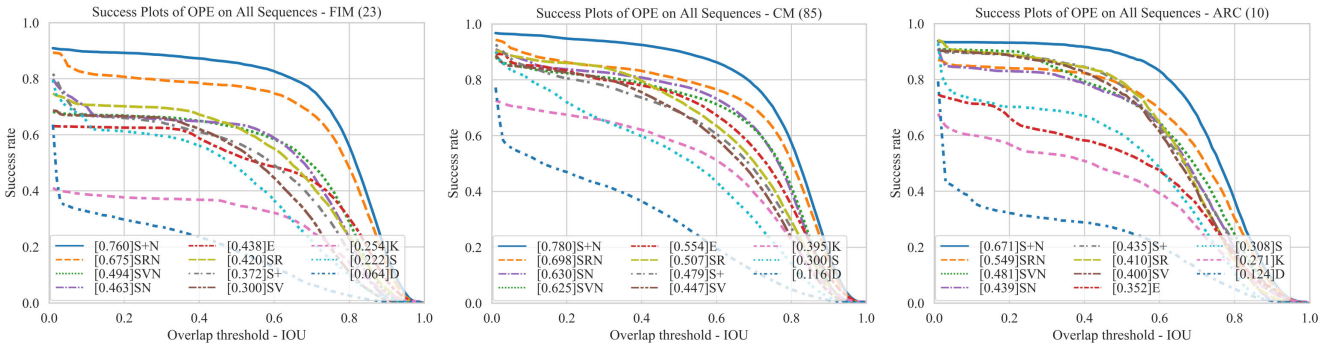
FIGURE 16. Results of weather variation under protocol I using success rate.



(a) IV (b) IQ

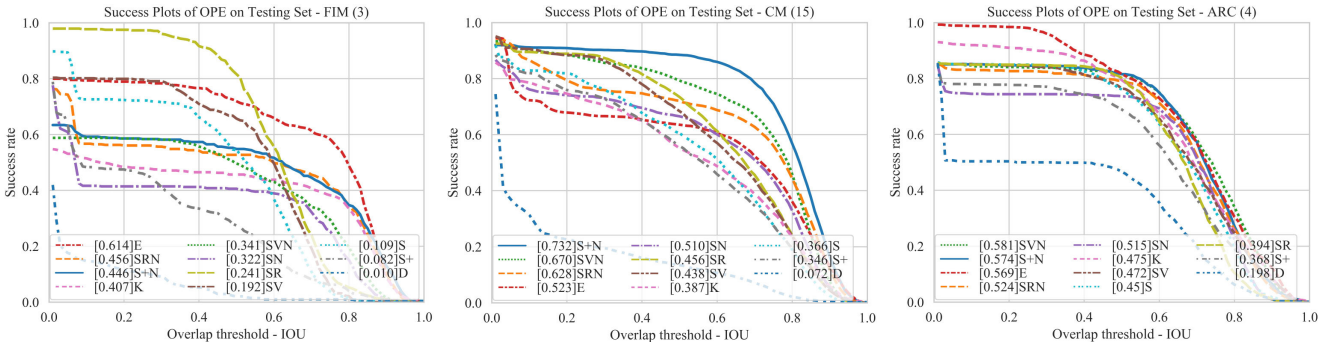
FIGURE 17. Results of weather variation under protocol II using success rate.

and *similar object* (SOB) on all sequences and the testing set. When the target stops at a port or shore, a large number of ships and buildings form a similar background. At the same time, due to the movement of other ships, occlusion will be caused and nearby similar objects will present. With the extraction of deep features, deep trackers perform well



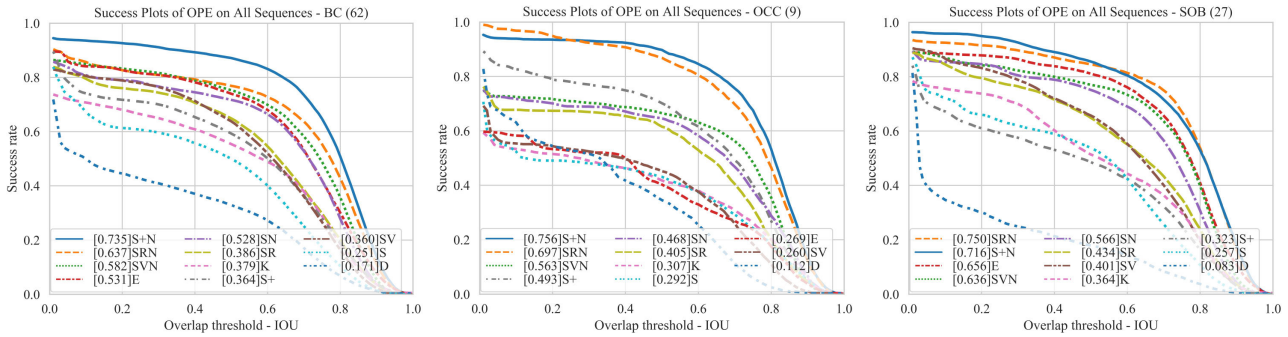
(a) FIM (b) CM (c) ARC

FIGURE 18. Results of shaking variation under protocol I using success rate.



(a) FIM (b) CM (c) ARC

FIGURE 19. Results of shaking variation under protocol II using success rate.

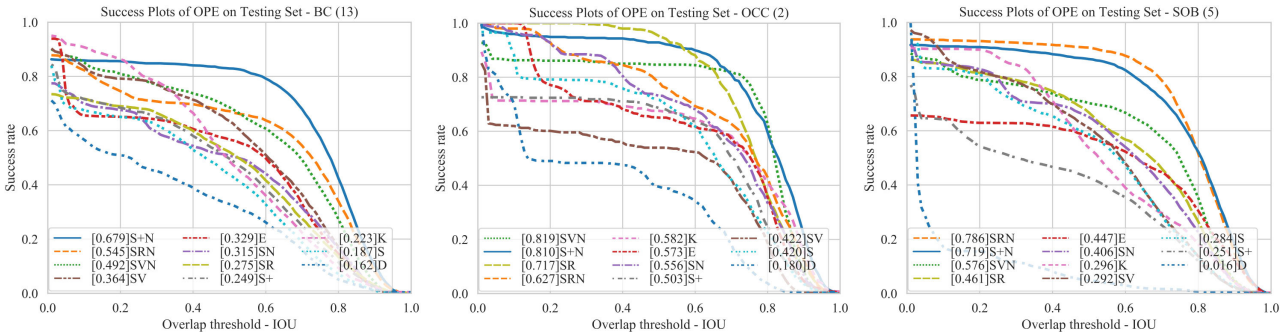


(a) BC (b) OCC (c) SOB

FIGURE 20. Results of occlusion and similarity under protocol I using success rate.

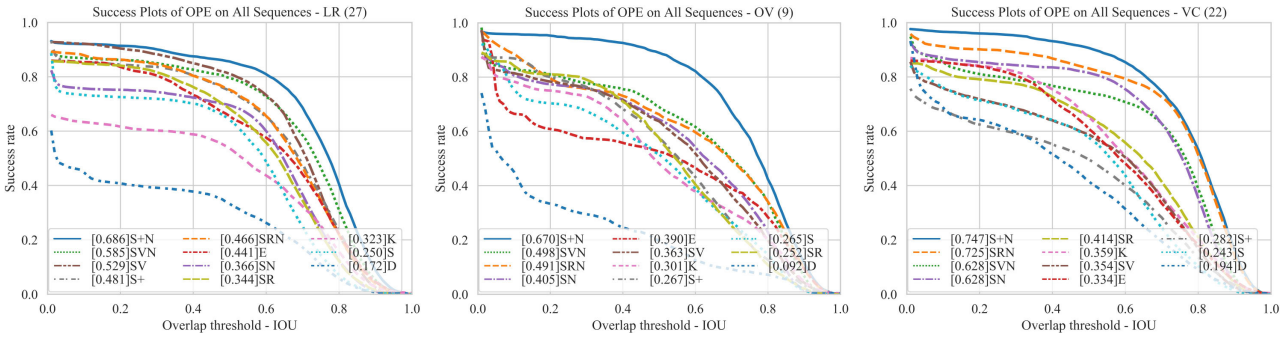
in the above challenges and show basically consistent results between the entire sequence and testing set. For correlation filters, KCF outperforms ECO and DSST. A reasonable explanation for this is that KCF can accurately predict occluded targets by linear interpolation to update model parameters. Nevertheless, DSST may change the bounding box to adapt to the occlusion or background, which may result in tracking failure.

Fig. 22 and 23 show the performance of all trackers on attributes of *small bounding box* (LR), *out-of-view* (OV), and *viewpoint change* (VC) on all sequences and the testing set. In Fig. (22), small targets are susceptible to environmental influences, such as overlapping of ship wakes, which will result in contaminated sample boxes for small targets and tracking failures. The results in Fig. (22) are consistent with the aforementioned discussions. Regarding the attribute OV, it is found that all correlation-filter-based and siamese trackers with penalty modules are ranked in the bottom six. When some portion of the ship leaves the frame, the cosine window in correlation filters will filter most pixels and the location



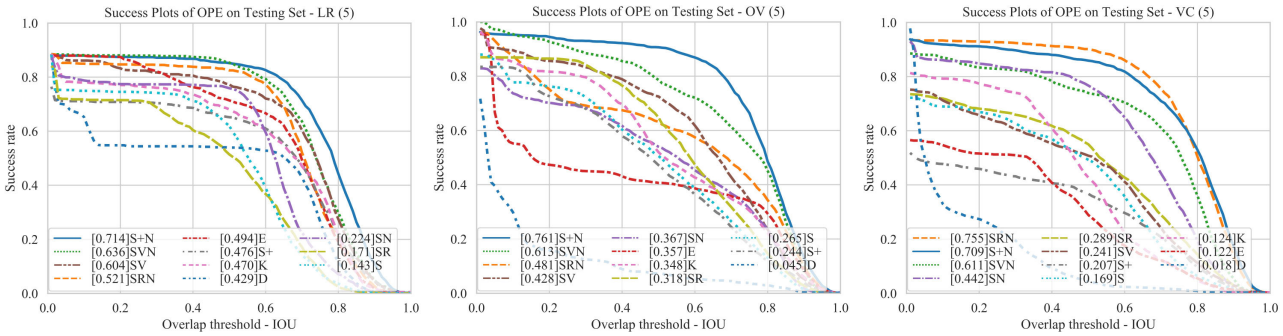
(a) BC (b) OCC (c) SOB

FIGURE 21. Results of occlusion and similarity under protocol II using success rate.



(a) LR (b) OV (c) VC

FIGURE 22. Results of LR, OV and VC under protocol I using success rate.



(a) LR (b) OV (c) VC

FIGURE 23. Results of LR, OV and VC under protocol II using success rate.

may not be the global maximum response value. Meanwhile, the cosine window in the penalty module suppresses the changes in the scale of the ship, which vary remarkably when leaving the frames. On the attribute VC, the appearance of the target may change remarkably when the viewpoint changes; for example, capturing ships from bow to hull. From the success plots on the attribute VC, the advantages of deep-learning trackers in extracting deep features are re-verified, and the worst performance of all correlation filters in the testing set is a strong argument for the statement.

In the overall performance experiment, SiamRPN++ with a multi-layer structure achieves the best result, which means that the strong discriminative ability is also the core of developing a ship tracker. Moreover, among all 13 attributes, FIM, CM, and ARC are the most important challenges for maritime ship tracking, and future trackers should pay more attention to these attributes.

V. CONCLUSION

A dataset for maritime ship tracking is presented in this paper. Compared with other existing datasets for maritime ship trackers, the proposed dataset has larger scale images, more challenging scenarios, and more comprehensive data analysis and test results reporting. Based on the detailed attribute analysis and experiments presented herein, researchers can deeply learn the characteristics of maritime ship tracking

and design deep tracking models accordingly. Moreover, for ease of use, a training and testing protocol is proposed and validated by a number of experiments to simplify the later implantation process. By providing this dataset, our hope is to diversify the development of deep maritime ship-tracking methods and promote scientific research progress in the maritime ship-tracking field.

REFERENCES

[1] X. Kong, L. Liu, Y. Qian, and M. Cui, "Automatic detection of sea-sky horizon line and small targets in maritime infrared imagery," *Infr. Phys. Technol.*, vol. 76, pp. 185–199, May 2016.

[2] S. Thome, N. Scherer-Negenborn, and M. Arens, "Visual tracker fusion and outlier detection on thermal image sequences," *Proc. SPIE*, vol. 10796, Oct. 2018, Art. no. 107960Q.

[3] B.-S. Shin, X. Mou, W. Mou, and H. Wang, "Vision-based navigation of an unmanned surface vehicle with object detection and tracking abilities," *Mach. Vis. Appl.*, vol. 29, no. 1, pp. 95–112, Jan. 2018.

[4] A. J. Sinisterra, M. R. Dhanak, and K. Von Ellenrieder, "Stereovision-based target tracking system for USV operations," *Ocean Eng.*, vol. 133, pp. 197–214, Mar. 2017.

[5] M. T. Wolf, C. Assad, Y. Kuwata, A. Howard, H. Aghazarian, D. Zhu, T. Lu, A. Trebi-Ollennu, and T. Huntsberger, "360-degree visual detection and target tracking on an autonomous surface vehicle," *J. Field Robot.*, vol. 27, no. 6, pp. 819–833, Nov. 2010.

[6] J. Lou, T. Tan, W. Hu, H. Yang, and S. J. Maybank, "3-D model-based vehicle tracking," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1561–1569, Oct. 2005.

[7] J. Liu, K. Chen, and X. Song, "Pedestrian tracking framework based on deep convolution network and SIFT," *Scientia Sinica Inf.*, vol. 48, no. 7, pp. 841–855, Jul. 2018.

[8] S. Fefilatyev, D. Goldgof, M. Shreve, and C. Lembke, "Detection and tracking of ships in open sea with rapidly moving buoy-mounted camera system," *Ocean Eng.*, vol. 54, pp. 1–12, Nov. 2012.

[9] Z. L. Szpak and J. R. Tapamo, "Maritime surveillance: Tracking ships inside a dynamic background using a fast level-set," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 6669–6680, Jun. 2011.

[10] D. Bloisi and L. Iocchi, "Argos—A video surveillance system for boat traffic monitoring in venice," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 7, pp. 1477–1502, 2009.

[11] R. Ribeiro, G. Cruz, J. Matos, and A. Bernardino, "A dataset for airborne maritime surveillance environments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2720–2732, Sep. 2017.

[12] D. K. Prasad, D. Rajan, L. Rachmawati, E. Rajabally, and C. Quek, "Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 8, pp. 1993–2016, Aug. 2017.

[13] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[14] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2137–2155, Nov. 2016.

[15] J. Valmadre, L. Bertinetto, F. J. Henriques, R. Tao, A. Vedaldi, W. M. A. Smeulders, H. S. P. Torr, and E. Gavves, "Long-term tracking in the wild: A benchmark," in *Computer Vision—ECCV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 692–707.

[16] H. Fan, H. Ling, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, and C. Liao, "LaSOT: A high-quality benchmark for large-scale single object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5374–5383.

[17] L. Huang, X. Zhao, and K. Huang, "GOT-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1562–1577, May 2021.

[18] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "TrackingNet: A large-scale dataset and benchmark for object tracking in the wild," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 300–317.

[19] Z. Shao, W. Wu, Z. Wang, W. Du, and C. Li, "SeaShips: A large-scale precisely annotated dataset for ship detection," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2593–2604, Oct. 2018.

[20] E. Gundogdu, B. Solmaz, V. Yücesoy, and A. Koç, "Marvel: A large-scale image dataset for maritime vessels," in *Computer Vision–ACCV*, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham, Switzerland: Springer, 2017, pp. 165–180.

[21] M. Kristan, V. S. Kenk, S. Kovačič, and J. Perš, "Fast image-based obstacle detection from unmanned surface vehicles," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 641–654, Mar. 2016.

[22] B. Bovcon, R. Mandeljc, J. Perš, and M. Kristan, "Stereo obstacle detection for unmanned surface vehicles by IMU-assisted semantic segmentation," *Robot. Auton. Syst.*, vol. 104, pp. 1–13, Jun. 2018.

[23] Y. Shan, X. Zhou, S. Liu, Y. Zhang, and K. Huang, "SiamFPN: A deep learning method for accurate and real-time maritime ship tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 315–325, Jan. 2021.

[24] L. Meng and C. Li, "Brief review of object tracking algorithms in recent years: Correlated filtering and deep learning," *J. Image Graph.*, vol. 24, no. 7, pp. 1011–1016, 2019.

[25] S. D. Bolme, J. Ross, B. Bruce, A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550

[26] F. J. Henriques, C. Rui, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 702–715.

[27] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[28] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–11.

[29] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Computer Vision—ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 472–488.

[30] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6638–6646.

[31] L. Bertinetto, J. Valmadre, F. J. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Computer Vision—ECCV*, G. Hua and H. Jégou, Eds. Cham, Switzerland: Springer, 2016, pp. 850–865.

[32] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.

[33] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4282–4291.

[34] Y. Xu, Z. Wang, Z. Li, Y. Ye, and G. Yu, "SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 12549–12556.

[35] Z. Zhang, H. Peng, and Q. Wang, "Deeper and wider siamese networks for real-time visual tracking," in *Proc. CVPR*, vol. abs/1901.01660, Mar. 2019. [Online]. Available: https://dblp.org/rec/conf/cvpr/ZhangP19.html?view=bibtex

[36] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Computer Vision—ECCV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 103–119.

[37] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4660–4669.

[38] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6182–6191.

[39] P. Westall, J. J. Ford, P. O'Shea, and S. Hrabar, "Evaluation of maritime vision techniques for aerial search of humans in maritime environments," in *Proc. Digit. Image Comput., Techn. Appl.*, 2008, pp. 176–183.

[40] S. Fefilatyev, D. Goldgof, and C. Lembke, "Tracking ships from fast moving camera through image registration," in *Proc. 20th Int. Conf. Pattern Recognit.*, 2010, pp. 3500–3503.

[41] W.-C. Hu, C.-Y. Yang, and D.-Y. Huang, "Robust real-time ship detection and tracking for visual surveillance of cage aquaculture," *J. Vis. Commun. Image Represent.*, vol. 22, no. 6, pp. 543–556, Aug. 2011.

[42] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "YouTube-BoundingBoxes: A large high-precision human-annotated data set for object detection in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5296–5305.

[43] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "TrackingNet: A large-scale dataset and benchmark for object tracking in the wild," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 300–317.
[44] J. L. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia, "Diatom autofocusing in brightfield microscopy: A comparative study," in *Proc. 15th Int. Conf. Pattern Recognit.*, 2000, pp. 314–317.
[45] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25, no. 2, pp. 1097–1105.
[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

YUNXIAO SHAN received the M.Sc. degree in machinery manufacturing and automation from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2018. From 2015 to 2016, he trained at Rutgers University, New Brunswick, NJ, USA. He currently works as an Associate Researcher with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He has contributed more than 20 publications on the research of autonomous systems. His research interests include the planning and control of autonomous vehicles, water surface intelligence, and research on autonomous surface vehicles.

SHANGHUA LIU received the B.S. degree in computer science and technology from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, in 2020. He is currently pursuing the M.Sc. degree with Columbia University. His research interests include image processing and object detection and tracking.

YUNFEI ZHANG received the Ph.D. degree from The Hong Kong University of Science and Technology. He is currently the Founder and the Chairman of Zhuhai Yunzhou Intelligence Technology Company Ltd. He was awarded the Professor-Level Senior Engineer. In 2014, he was rated as "the most beautiful young scientific and technological worker" by the Communist Youth League Central Committee and selected as the "World Youth Innovation 100 People".

MIN JING received the B.S. and M.S. degrees from the School of Computer Science, NUDT, Changsha, China, in 2001 and 2005, respectively, and the Ph.D. degree in military operation research from NDU, Beijing, China, in 2011. He is currently working as the Chief Project Engineer at Yunzhou-Tech Company Ltd., Zhuhai, China. He has hosted and participated in more than ten key projects and published more than 20 academic articles. His research interests include unmanned system swarm, simulation and emulation, and wireless ad-hoc networks.

HUAWEI XU received the Ph.D. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. Since 2012, he has been working with the Fifth Research Institute of MIIT, Guangzhou, China. He has contributed more than 20 publications on the automatic driving test technology.