IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 19, 2019, accepted January 20, 2020, date of publication January 24, 2020, date of current version February 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969231

Marine Vessel Re-Identification: A Large-Scale Dataset and Global-and-Local Fusion-Based Discriminative Feature Learning

DALEI QIAO[1,2], GUANGZHONG LIU[1], FENG DONG[1,3], SHE-XIANG JIANG[1], AND LIKUN DAI[2]

[1]College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
[2]College of Information Engineering, Jiangsu Maritime Institute, Nanjing 211100, China
[3]College of Information Engineering, Shaoyang University, Shaoyang 422000, China

Corresponding author: Guangzhong Liu (gzhliu@shmtu.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 61202370, in part by the Postdoctoral Science Foundation of China under Grant 2019M651844, in part by the Qinglan Project and Advanced Study and Research of Jiangsu Province under Grant 2018GRF016, in part by the Scientific Research Fund of the Hunan Provincial Education Department under Grant 15C1241, and in part by the Project of the Qianfan Team, Innovation Fund, and the Collaborative Innovation Center of Shipping Big Data Application of Jiangsu Maritime Institute.

ABSTRACT A marine vessel re-identification system has to determine whether or not different images represent the same vessel. Accurate vessel re-identification improves onshore closed-circuit television monitoring in a vessel traffic services system as well as onboard surveillance of surrounding vessels. However, because ships are rigid bodies and the marine environment is harsh, the accurate re-identification of vessels at sea can be very difficult. We describe a marine vessel-re-identification framework, Global-and-Local Fusion-based Multi-view Feature Learning (GLF-MVFL), which is based on a combination of global and fine-grained local features. GLF-MVFL combines cross-entropy loss with our newly-developed orientation-guided quintuplet loss. We exploit intrinsic features of marine vessels to optimize multi-view representation learning for re-identification. GLF-MVFL uses ResNet-50 as the backbone network to extract features for simultaneous quintuple input. It detects and discriminates between features and estimates viewpoints to form a comprehensive re-identification framework. We created an annotated large-scale vessel retrieval dataset, VesselID-539, which contains images from viewpoints similar to those of an autonomous surface vessel, to use in evaluating the performance of the model. Extensive experiments and analysis of the results obtained from using VesselID-539 demonstrate that our approach significantly increases the accuracy of vessel re-identification and is more effective and robust for images from different viewpoints than other approaches.

INDEX TERMS Autonomous surface vessel (ASV), maritime surveillance, VesselID-539 dataset, multi views, vessel re-identification (V-ReID).

I. INTRODUCTION

An autonomous surface vessel (ASV) is a robotic agent that must sense its surroundings in real-time and identify where shores, islands, or other vessels around it are located. In practice, the obstacles most likely to be encountered by an ASV at sea are nearby vessels, making it important that the ASV is able to detect and track them in real-time. The ASV must recognize essential cues to support the advanced driver assistance system (ADAS) in avoiding collisions and making decisions concerning compliance with the *International Regulations for Preventing Collisions at Sea* (COLREGs) [1]. The ASV must be aware of the movement of surrounding vessels when making decisions for collision avoidance, possibly in case of emergency. In the past, detection and tracking of targets at sea has depended on radar and automatic identification systems (AIS), which treat a vessel as a point. However, this approach led to uncertainty because it ignores the size of a vessel. To determine the size of a vessel, it is necessary for the identification system to detect and re-recognize the vessel

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

from multiple camera frames that may not be consecutive. An ASV must determine whether or not frames identify the same vessel and then decide whether to associate them with existing tracklets or initiate a new track. By doing this, the problem of repetitive tracklet initiation is overcome.

Re-identification of vessels is similar to the re-identification of pedestrians [2]–[6] or vehicles [12], [14]–[17]. It is important for security surveillance, and in the creation of intelligent transportation systems (ITS) at sea [10]. Fig. 1 shows the four principal stages of vessel re-identification: vessel detection, feature extraction, feature transformation, and creation of a similarity metric.

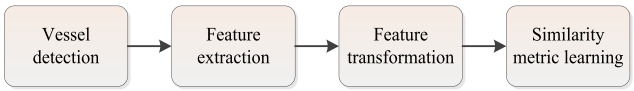


FIGURE 1. Stages in the vessel re-identification system.

Re-identification of people is significantly different from re-identification of marine vessels. There are three categories of problems that vessel re-identification must address. Vessels in the same class are rigid and homogeneous, and therefore highly similar, thus making it difficult to detect the subtle differences between them. Large ships vary significantly in appearance when the viewpoint changes. And, as far as we know, there is no available large-scale vessel re-identification dataset, whereas datasets for person re-identification are readily available. The extensive research on person and vehicle recognition can be used as a guide for vessel re-identification, although the conditions under which an onboard camera must operate are much harsher than those of the other two applications. A generalized model based on the visual appearance of vessels is urgently needed for vessel re-identification. To address these challenges and facilitate future research, we built a large-scale image dataset of marine vessels. The images have a variety of angles of view and are mainly from ship-borne cameras.

A. CONTRIBUTIONS

This study was developed in three stages. We first created a large-scale well-annotated dataset both for this study and future research needs. We then developed a deep learning framework for vessel re-identification that included a feature detection and discrimination module. The module extracts fine-grained local features and weights them according to global and local conditions, thereby ranking their importance. Finally, we derived a novel loss function and an orientation-guided quintuple loss function, based on viewpoint estimation and sample data mining.

In the remainder of this paper, Section II reviews the related literature. Section III gives a detailed description of the VesselID-539 dataset, including the methods we used to collect and annotate images. Section IV describes our methodology, and Section V presents our experimental designs and our analysis of the results.

II. RELATED WORK

In this section, we briefly review related work in generic object re-identification (emphasizing person and vehicle re-identification) and marine vessel detection and identification. We then outline developments in feature identification and discrimination and viewpoint estimation.

A. GENERIC OBJECT RE-IDENTIFICATION

1) PERSON RE-IDENTIFICATION

Person re-identification is intended to identify the same pedestrian from different camera views. Zajdel *et al*. were the first to address pedestrian re-identification in solving the cross-camera data association problem using multi-target multi-camera (MTMC) tracking [2]. Zheng *et al*. divided person re-identification into two stages: person detection and person re-identification, which latter is in turn divided into two steps: feature extraction and estimating feature similarity [3]. Hermans *et al*. developed TriHard loss to enable training networks to learn from hard samples [4]. Cai *et al*. created an attention network built on multi-scale and multi-part masks [5]. Heo *et al*. developed a teacher–student-based semi-supervised framework to estimate a person's attitude and orientation [6]. However, these state-of-the-art part-based and fine-grained methods for person re-identification do not work well on marine vessels well because the attitude of a ship can change greatly from different viewpoints. There are many widely-used person re-identification datasets, including Market1501 [7], MARS [8], DukeMTMC-reID [9] and CUHK-SYSU [11].

2) VEHICLE RE-IDENTIFICATION

Vehicle re-identification is fundamental to automatic vehicle control and has been extensively researched. Liu *et al*. created a large-scale dataset for vehicle re-identification and introduced a mixed difference network (MDNet) for vehicle recognition and re-identification [12]. Xiang *et al*. developed a global topological constraints network for fine-grained vehicle recognition [13]. They modeled the interactions between components using global constraints and incorporated them into a unified CNN network. Bai *et al*. developed a group-sensitive triplet embedding process (GS-TRE) to extract fine-grained vehicle features to resolve the problem of variation in features between different classes of vehicle [14]. Guindel *et al*. estimated vehicle orientation when locating the vehicle in an image using Faster R-CNN [15]. Zhang *et al*. developed a partially guided attention mechanism to locate regions for discrimination and combined it with global features [16]. Vehicle re-identification datasets include VehicleID [12], VeRi-776 [17], VERI-Wild [18], and CityFlow [19].

B. MARINE VESSEL DETECTION AND IDENTIFICATION

There have been many recent developments in ship detection and recognition based on remote sensing images [20]. The major drawbacks of using remote sensing images are that they

are taken from overhead and that they cannot be processed in real-time due to the long period between sensing and re-use. The development of edge discrimination enables us to create improved re-identification techniques for ASVs. Ward *et al*. synthesized a NAVHAZ dataset to avoid collisions at sea [21]. They generated 20 equally-spaced angles of the headway of each ship under different weather conditions and different sea states. Heyse *et al*. developed a method of identifying marine vessels using multi-level descriptions and created a refined multi-level classifier based on deep features [22]. Qiao *et al*. addressed ship re-identification for the long-term tracking of vessels at sea by considering each image as a set of visual cues [23]. Tian *et al*. used a word bag model to recognize depth features [24] and demonstrated its use on a large image gallery. Hilton *et al*. investigated the use of a capsule network to address viewpoint invariance in the classification of marine vessels [25]. Jinwen *et al*. introduced a graphic model based on energy loss in the metric-learning phase of a ship recognition algorithm [10]. To our knowledge, there is only one small dataset that has been created for ship re-identification [26].

C. VIEWPOINT ESTIMATION AND DISCRIMINATIVE FEATURE-BASED MODELS

1) VIEWPOINT ESTIMATION

Variation in the viewpoint of a vessel makes vessel re-identification a difficult task, while viewpoint can be estimated accurately in person and vehicle re-identification. Estimation of viewpoint is critical in both predicting vessel trajectory and vessel re-identification. Current efforts in this field concentrate on the use of convolutional features for estimating the viewpoint of the object. These attempts can be categorized into two types: direct estimation from features derived from the appearance of the vessel, and calculation from predicted key points. Saquib *et al*. developed an attitude-sensitive embedding network model which took account of a person's attitude and orientation; they incorporated the individual's features by weighting three-way views [27]. Ghahremani *et al*. utilized a single-shot detector to combine classification of the vessel type with an estimation of the viewing angle [28]. Li *et al*. introduced a viewpoint discernibility matrix to resolve viewpoint ambiguity caused by poor light or adverse weather conditions [29]. Wang *et al*. proposed an orientation-invariant feature-embedding framework for vehicle re-identification, which aggregated viewpoint-based features extracted from 20 predefined points in four orientations [30].

2) DISCRIMINATIVE FEATURE-BASED MODELS

Recent studies in person re-identification and vehicle re-identification have shown that global features alone are insufficient to differentiate near-identical objects because they lack fine-grained features necessary for individual discrimination [31]. Extracting partial features from multiple images has been shown to be effective; it significantly improves recognition and is increasingly used in fine-grained object recognition. Sun *et al*. developed a feature-based convolutional baseline approach which divides the entire image of a pedestrian into fixed equal parts in the horizontal direction, assigns a soft weight to the spatial distribution of each part, and eventually aligns them [32]. Tan *et al*. built on previous work to develop the multiple granularities network for person re-identification [33] by using two horizontal stripes and two vertical stripes (i.e., 2×2 grids) to semantically characterize vehicle features [34]. He *et al*. improved the detection of local features by combining partial and global features during the training phase [31]. Tan *et al*. developed the EfficientNet scaling method by uniformly scaling up CNN width, depth and resolution using a compound scaling method [35]. We used EfficientNet as the feature extractor to locate discriminative features in the images.

III. THE MARINE VESSEL RE-IDENTIFICATION DATASET

We built and annotated a large-scale image database, VesselID-539, to use in evaluating the model we propose and to support future research into vessel re-identification or fine-grained feature recognition. To the best of our knowledge, VesselID-539 is the largest corpus to date for marine vessel re-identification. In this section, we describe the collection and annotation of the dataset.

A. DATA COLLECTION

The marine vessel images dataset VesselID-539 was created using images from the website Marine Traffic (www.marinetraffic.com) for the period 2019-03-13–2019-03-16. (The download links for raw images and the processing and annotation script will be made public after this paper is accepted.) The raw vessel image dataset contains over 149 465 images of 511 vessels. These images were captured mainly by professional photographers around the world, from onboard or onshore cameras, at different times and locations. Each ship in the VesselID-539 dataset is represented by numerous images from different viewing angles showing different aspects. Fig. 2 shows some sample images for four different challenging scenarios.

B. ANNOTATION DESCRIPTION

We used YOLOv3 [36] to automatically locate the bounding box (BBox) of a vessel by using the pretrained weight from ImageNet and then fine-tuning it on the MARVEL dataset [37]. Table 1 displays the feature map and anchor box of the VesselID-539 dataset.

We manually corrected a number of annotations mislabeled by YOLOv3, relabeled some missing BBoxes, and identified the vessel of interest if there were two or more vessels in the image. The images of some vessels cover periods of up to several years, so we consider the same vessel with different colors or different loading conditions (e.g., container ships under full load or ballast conditions) as different vessels. Fig. 3 shows the statistics for VesselID-539, which indicates the number of vessels falling into a certain interval,

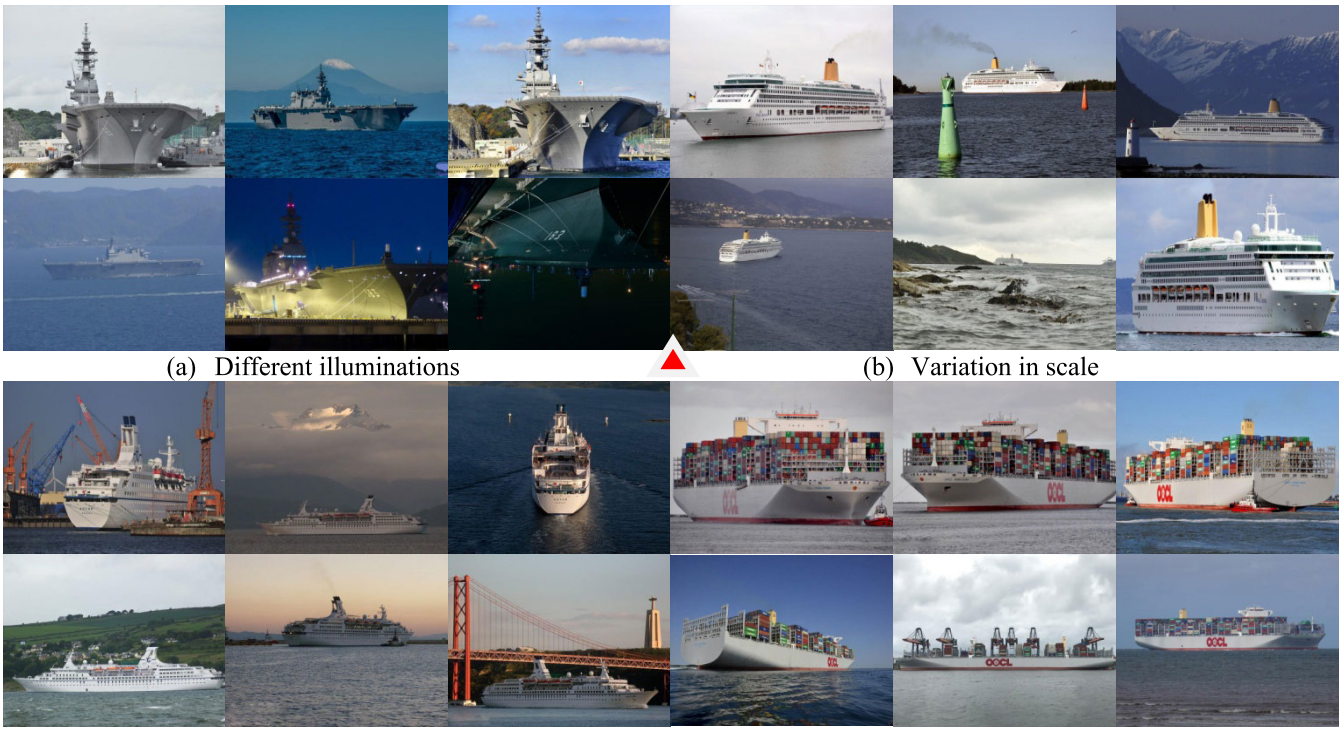


FIGURE 2. Examples of challenging images in the VesselID-539 dataset. Each quadrant (of six images) shows the same vessel in a different challenging scenario (clockwise, from top left: different Illuminations, variation in scale, change in background, and different viewpoints).

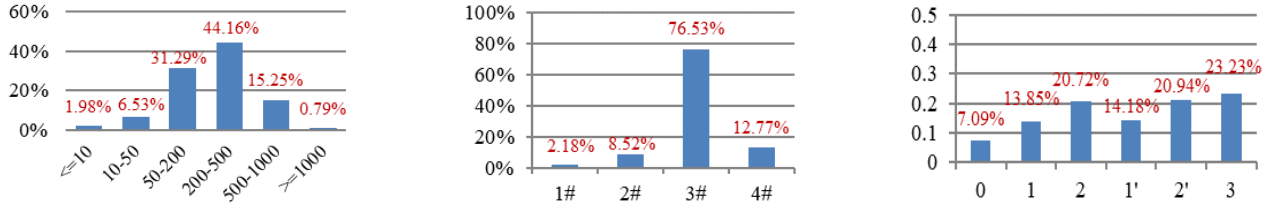


FIGURE 3. Statistical distributions of the VesselID-539 dataset.

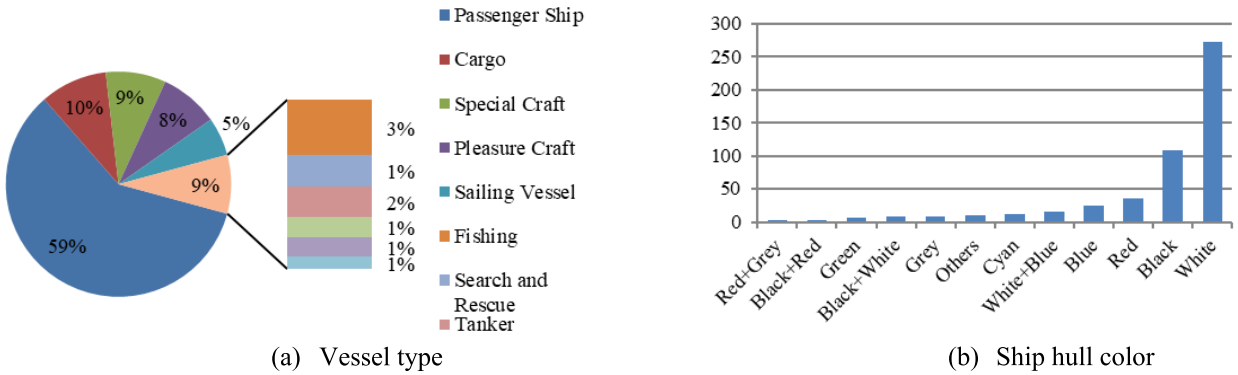
the number of discriminative features for each vessel, and the distribution of the viewpoints for each vessel. The predefined orientation bins are discussed in Section IV.

After data cleaning and relabeling, we had acquired 149 363 vessel images belonging to 539 vessels. There are on average 277 images per vessel in VesselID-539; maximum images per vessel are 1414 and minimum images per vessel are 3, as shown in Fig. 3(a). The work was done by six volunteers over two weeks. The statistical distributions of vessel type and hull color are shown in Fig. 4.

We cropped each image to the size of the ship's BBox. To exploit more information for vessel re-identification, we annotated each image with rich attribute labels such as ship name, color of the hull (e.g., white, grey, black, and red) and vessel type (e.g., passenger ship, tug, cargo carrier and special craft) as well as the angle of view of the vessel (orientation); we refer to these as multi-features. Although additional annotation increases the complexity of re-identification, it increases the flexibility necessary for real-time marine surveillance. For each cropped image of the vessel, we further manually annotated corresponding sub-BBoxes as multiple discriminative features, varying from one to four. A total of 447 926 sub-BBoxes were labeled in our VesselID-539 dataset. This work was done by ten volunteers over six weeks.

C. DATASET PARTITIONING

We divided VesselID-539 into training and test datasets using an 80/20 ratio, according to the empirical rule. The training set contained 104 554 images with 377 IDs, and the test set included 44 809 images with 162 IDs. The test set was further split into a probe set (20% of the IDs) and a gallery set (80% of the IDs). The partitioning of the dataset is shown in Fig. 5.



(a) Vessel type (b) Ship hull color

FIGURE 4. The statistical distribution of the VesselID-539 dataset: (a) the distribution of vessel types; (b) the distribution of vessel colors (specifically, the color of the ship hull above the waterline and below the main deck).

TABLE 1. The feature map and prior anchors of the VesselID-539 dataset. Using the K-means algorithm, nine prior anchors are clustered from the ground truth BBox of the training set. The aim is to constrain the range of the predicted vessels and achieve multi-scale learning.

Feature map	13×13			26×26			52×52		
Receptive field	Big			Medium			Small		
Prior anchors	367×175	355×248	349×117	346×337	216×311	153×172	234×86	265×180	96×47

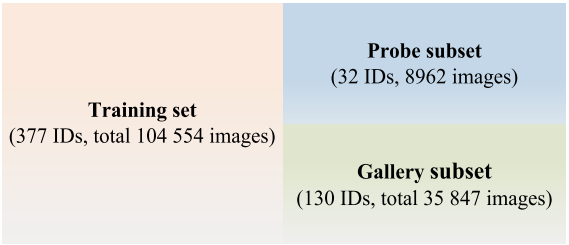


FIGURE 5. The partition diagram of the VesselID-539 dataset.

TABLE 2. Comparison results with other vessel datasets.

Dataset	IDs	Total Images	Average Images	Attr.	Parts	Orient.
VesselReID	733	4616	6	√	×	√
VesselID-539	539	**149 363**	**277**	√	√	√

Table 2 shows a comparison of our VesselID-539 dataset with the only publicly available dataset [26]. The VesselID-539 dataset contains more images in total and on average for each vessel, and thus more samples from diverse viewpoints; it also provides semantic annotations of each vessel's characteristics.

IV. APPROACH

In this section, we describe the GLF-MVFL model in detail. The architecture of the model is shown in Fig. 6. In the GLF-MVFL pipeline, we identify the discriminative features and then extract local features from the fine-grained images and aggregate them. Global features and local features are concatenated for discriminative feature learning.

When determining the similarities in images of a ship captured from different viewpoints, or when comparing an image of part of a ship with an image of the entire ship, the non-intersecting regions will be a distraction for the GLF-MVFL model. We developed a partition-and-aggregate strategy in response to this challenge, which is magnified by intra-class variation and inter-class similarity. Note that we train the detection model and the re-identification model separately; this paper focuses on viewpoint estimation and the vessel re-identification model.

A. PRELIMINARIES

1) PROBLEM STATEMENT

Vessel re-identification has similar goals to person and vehicle re-identification. For a given pair of input images taken from different viewpoints, the output is a similarity score indicating whether the two input images represent the same object. More formally, given a training set $\mathcal{T} = \{(\mathcal{I}_t, \mathcal{Y}_t)\}_{t=1}^{N}$, where I_i represents the input vessel image and Y_i is its identity label, then for any two vessel image pairs, a distance metric function is defined by $\mathcal{D}(\mathcal{I}_a, \mathcal{I}_b) : \mathcal{R}^{\mathcal{D}} \times \mathcal{R}^{\mathcal{D}} \rightarrow \mathcal{R}$ [38]. The key task of the V-ReID model is to cluster vessel images in the feature space F and to find an optimal mapping function $f(\mathcal{I}_t, \Theta)$ through training by minimizing a predefined loss function, where Θ indicates the parameters of $f(\cdot)$ to be learned. If two images I_a and I_b are of the same vessel from different viewpoints, they are clustered together and the similarity score $\mathcal{S}(\mathcal{I}_a, \mathcal{I}_b) \rightarrow 1$; if they are different vessels, separate them and $\mathcal{S}(\mathcal{I}_a, \mathcal{I}_b) \rightarrow 0$. In summary, $\mathcal{D}(\mathcal{I}_a, \mathcal{I}_b)$ satisfies the following conditions:

$$\underset{\substack{\mathcal{Y}_a==\mathcal{Y}_b \\ \mathcal{V}_a \approx \mathcal{V}_b}}{\mathcal{D}}(\mathcal{I}_a^{\mathcal{V}_a}, \mathcal{I}_b^{\mathcal{V}_b}) \leq \underset{\substack{\mathcal{Y}_a==\mathcal{Y}_b \\ \mathcal{V}_a <> \mathcal{V}_b}}{\mathcal{D}}(\mathcal{I}_a^{\mathcal{V}_a}, \mathcal{I}_b^{\mathcal{V}_b}) << \underset{\mathcal{Y}_a \neq \mathcal{Y}_b}{\mathcal{D}}(\mathcal{I}_a^{\mathcal{V}_a}, \mathcal{I}_b^{\mathcal{V}_b}) \tag{1}$$

where the superscripts V_a, V_b denote the viewpoints of the input vessel(s).

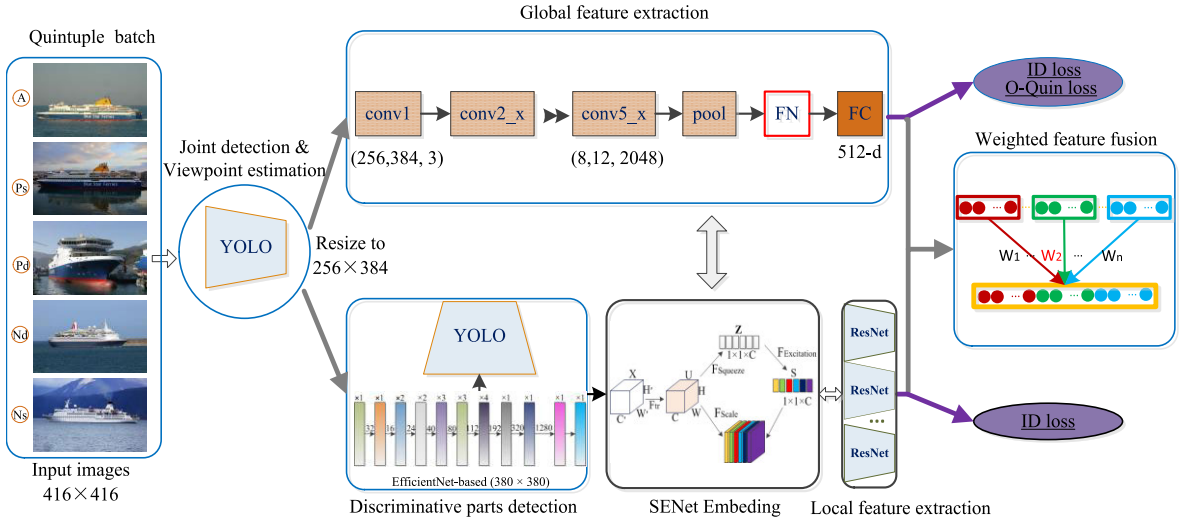


FIGURE 6. Illustratio. of the overall framework of the GLF-MVFL model, which consists of five major components: global feature learning module, discriminative feature detection module, local feature extraction module, viewpoint estimation module, and weighted feature fusion module. Items contained in the blue boxes are the focus of this paper.

2) DEFINITION OF MULTIPLE VIEWS AND SEMANTIC FEATURES

We divide the ship hull into a few semantic features, using a global-and-local fusion method: $I_i = \{P_i^k\}, k = 1, 2, \ldots, \mathcal{M}$, where P_i^k devotes the k th part of vessel image I_i with M parts in total. For simplicity, we treat all enclosing structures above the upper deck as a single superstructure. We defined four vessel parts for our model: stem, stern, side freeboard, and superstructure; in Fig. 7, they are marked with yellow lines. We can dynamically adjust the number of multiple views and defined vessel parts. In this study, we set them to a typical value of 4 because our goal is to demonstrate the effectiveness of GLF-MVFL.

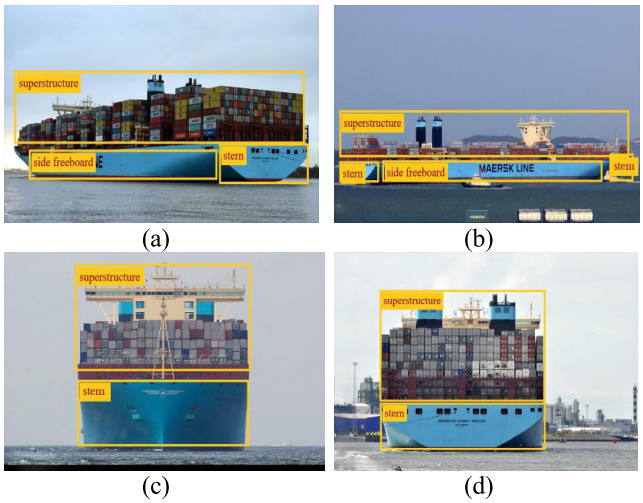


FIGURE 7. Th multiple views and multiple parts of our model (the images are of a container ship drawn randomly from VesselID-539): (a) port side viewpoint; (b) starboard side viewpoint; (c) stem viewpoint; (d) stern viewpoint. The yellow boxes indicate the detected discriminative features. We observe that a vessel's name is usually marked on each of its sides, and both name and port of registry are marked on the stern.

B. GLOBAL FEATURE EXTRACTION

In the re-identification stage, we first resize all the input images to the size of 256 × 384 and use an output feature map size of 2048 × 8 × 12. The input images are fed into the residual neural network (ResNet-50) for training and further feature extraction. In the global feature extraction stage, we introduce feature normalization (FN) to impose influence on the final feature representation layer of our vessel re-identification model. We do this to increase the discrimination capability of learned features and to mitigate the effects of unnormalized features during the loss calculation phase. FN, a technique suggested by Hasnat *et al.* [39], is a special form of BN and compatible with a normal distribution to ensure each feature contributes equally to the cost function. We set the scaling and translation parameters $\beta = 0$ and $\gamma = 1$ and substitute them into BN, allowing us to write the forward propagation formula of the FN network layer as:

$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{(\sigma_{\mathcal{B}})^2 + \varepsilon}}, \quad y_i \leftarrow \hat{x}_i \equiv \underset{\beta=0,\gamma=1}{\mathcal{BN}}(x_i) \tag{2}$$

where $\mu_{\mathcal{B}} \leftarrow \frac{1}{n}\sum_{i=1}^{n} x_i$, $\mathcal{B} = \{x_1, x_2, \ldots, x_n\}$ is the mean of the mini-batch, and $(\sigma_{\mathcal{B}})^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - u_{\mathcal{B}})^2$ is the variance of the mini-batch.

C. DETECTION OF DISCRIMINATIVE FEATURES AND FEATURE EXTRACTION

We devised the discriminative feature detector using the state-of-the-art framework YOLO (You Only Look Once) [36], which incorporates the multi-scale concept; its accuracy is on a par with ResNet-101 but it has an obvious advantage in speed. We modified YOLO by replacing the original

Darknet53 with EfficientNet [35] as the backbone network to improve the feature extraction capability of the detector. We merged the parameters of the batch normalization (BN) layer [40] into the convolution layer to improve the forward inference speed of the YOLO detector. The merging process of the convolution layer is:

$$\begin{aligned}\hat{x}_i &= \frac{\gamma\left(\sum_{i=0}^{n}(x_i^* w_i)-\mu\right)}{\sqrt{\sigma^2+\varepsilon}}+\beta \\ &= \sum_{i=0}^{n} x_i^* \gamma \frac{w_i}{\sqrt{\sigma^2+\varepsilon}} - \gamma\frac{\mu}{\sqrt{\sigma^2+\varepsilon}}+\beta \end{aligned} \tag{3}$$

In Eq. (3), we put $\eta = \gamma(\sigma^2+\varepsilon)^{-1/2}$ and eventually obtain the merged weight parameter $w_{merged} = \eta \times w_i$ and the bias parameter $\beta_{merged} = \beta - \eta \times \mu$; for the YOLO-based detector, the constant ε is set to 0.00001 to ensure numerical stability.

We use ResNet-50 as the local feature extractor, as we did to extract global features. Feature learning for each detected feature is facilitated by using the vessel ID as a label and the softmax function with cross-entropy loss as the classification supervisor. The total local branch loss is expressed as:

$$\begin{aligned}\mathcal{L}_{\text{parts}} &= \sum_{m=1}^{\mathcal{M}} \lambda_m \mathcal{L}_S^{(m)} \\ &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{\mathcal{M}}\left(\lambda_i g_n^{(m)} \log\left(\hat{g}_n^{(m)}\right)\right)\end{aligned} \tag{4}$$

where $\mathcal{L}_S^{(m)} = -\frac{1}{N}\sum_{n=1}^{N} g_n^{(m)} \log\left(\hat{g}_n^{(m)}\right)$ is the cross-entropy loss of the m th feature [41], and M and N are the total numbers of vessel identities and features. As will be detailed in Section IV, λ_i is the weight the cross-entropy loss of each feature.

D. VIEWPOINT ESTIMATION APPROACH

Zhang *et al*. suggested a method of resolving the problem of estimating viewpoint by discretizing the viewpoint space and predicting the probability for each orientation bin [16]. Accordingly, viewpoint estimation is performed as a classification task. The procedure for precisely aligning the matched parts of two vessel images is as follows. We quantize the viewpoint space into bins; the space is divided into four types, based on the symmetry of the hull (Fig. 8). Following [22] and [28], we divide the full range of viewpoints (360°) into $N_v = 16$ bins (itemized, starting at 1), with each bin representing a 22.5° angle. For each bin $\mathcal{O}_n (n = 1, 2, \ldots, N_v)$, the following condition is satisfied:

$$\mathcal{O}_n = \left\{\theta_n \in [0, 360) \middle| \frac{360}{N_v}\times(n-1) \le \theta_n < \frac{360}{N_v}\times n\right\} \tag{5}$$

where θ_n falls into a half-open interval.

Each vessel image with orientation bin O_b is assigned to its viewpoint $V(O_b)$. As can be seen from Fig. 8, we take the symmetry of the hull into consideration, and regard port and starboard as belonging to the same bin. Details are:

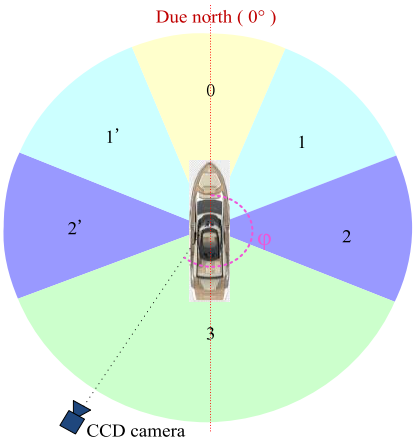


FIGURE 8. Illustration of the four predefined viewpoint bins. The sectors represented in the figure are: 0 stem viewpoint, 1 stem-starboard side viewpoint, 2 starboard-side viewpoint, and 3 stern viewpoint. Sectors 1 and 2 include symmetrical starboard and port sides. It should be note that φ gives an indication of viewpoint angle, which is calculated from the north as the 0° starting point.

$$\mathcal{V}(\mathcal{O}_b) = \begin{cases} \text{stem}, & \text{where } \mathcal{O}_b \in \{1\}\cup\{16\} \\ \text{stem -starboard}, & \text{where } \mathcal{O}_b \in \{2,3\}\cup\{14,15\} \\ \text{starboard}-\text{side}, & \text{where } \mathcal{O}_b \in \{4,5\}\cup\{12,13\} \\ \text{stern}, & \text{where } \mathcal{O}_b \in \{6,7,8,9,10,11\} \end{cases} \tag{6}$$

We classify the bin to which the vessel's viewpoint belongs indirectly, similar to [42]. By maximizing the classification interval in cosine space [43], this problem can be addressed for each orientation class by cosine cross-entropy loss:

$$\mathcal{L}_{\text{ve}} = -\frac{1}{N_s}\sum_{i=1}^{N_s}\log \times \left(\frac{\exp\left(s\cdot\cos\left(\theta_{y_i},i\right)-m\right)}{\exp\left(s\cdot\cos\left(\theta_{y_i},i\right)-m\right)+\sum_{j=1,j\neq j_i}^{N_v}\exp\left(s\cdot\cos\left(\theta_j,i\right)\right)}\right) \tag{7}$$

$$\text{s.t. } \cos\left(\theta_j,i\right) = W_j^T x_i \tag{8}$$

where θ represents the angle between weight vector W_j and input vector x_i, s is a scale factor, and m is a margin parameter which controls the distance (cosine margin term) and satisfies the decision boundary condition $s\cdot(\cos\theta_i - m - \cos\theta_j) = 0$. Note that both W and x are normalized in Eq. (8) to encourage the CNN network to focus on the task of optimizing the estimated viewpoint.

E. ORIENTATION-GUIDED QUINTUPLET LOSS

Current efforts in the re-identification of persons, vehicles, and even wild animals (e.g., Amur tiger re-identification [47])

focus on the exploitation of contrastive loss [44], triplet loss [45], and some improved variants of triplet loss, such as batch-hard triplet loss [4] or quadruplet loss [46].

Large vessels are more sensitive to a change in viewpoint. Therefore, for similarity metric learning, we add two extra image samples (one positive and one negative) to the triplet to constitute the quintuplet, as shown in Fig. 9. The orientation-guided quintuplet loss (O-Quin) is given by:

$$\begin{aligned}\mathcal{L}_{\text{Quin}} &= \left[\mathcal{D}\left(\mathcal{I}_a, \mathcal{I}_{ps}\right) - \mathcal{D}\left(\mathcal{I}_a, \mathcal{I}_{ns}\right) + \alpha\right]_+ \\ &+ \left[\mathcal{D}\left(\mathcal{I}_a, \mathcal{I}_{pd}\right) - \mathcal{D}\left(\mathcal{I}_a, \mathcal{I}_{nd}\right) + \beta\right]_+ \\ &\text{s.t. } \alpha >> \beta \end{aligned} \tag{9}$$

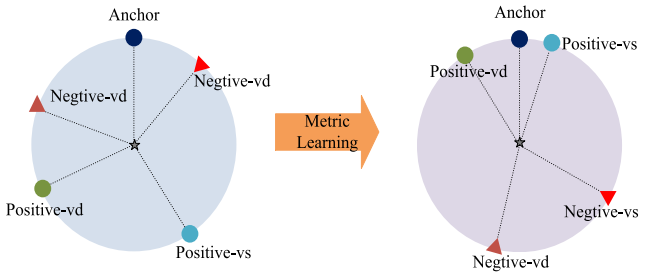


FIGURE 9. Metric learning of orientation-guided quintuplet loss. Given the input quintuples $\{< \mathcal{A}, \mathcal{P}_{\mathcal{VS}}, \mathcal{P}_{\mathcal{VD}}, \mathcal{N}_{\mathcal{VS}}, \mathcal{N}_{\mathcal{VD}} >\}$, which are anchor image $\mathcal{A}$, positive image from the same viewpoint $\mathcal{P}_{\mathcal{VS}}$(Positive-vs), positive image from a different viewpoint $\mathcal{P}_{\mathcal{VD}}$ (Positive-vd), negative image from the same viewpoint $\mathcal{N}_{\mathcal{VS}}$ (Negative-vs), and negative image from a different viewpoint $\mathcal{N}_{\mathcal{VD}}$ (Negative-vd).

where $[\bullet]_+ = Max[0, \bullet]$. In Eq. (9), α and β are margin values and satisfy the constraint that $\alpha \gg \beta$, meaning that α has a strong pull and β a weak pull. Here, $\mathcal{D}(\mathcal{I}_i, \mathcal{I}_j)$ is the cosine distance function, which is:

$$\mathcal{D}\left(\mathcal{I}_i, \mathcal{I}_j\right) = 1 - \cos\theta = 1 - \frac{f\left(\mathcal{I}_i\right) \cdot f\left(\mathcal{I}_j\right)}{\|f\left(\mathcal{I}_i\right)\|_2 \left\|f\left(\mathcal{I}_j\right)\right\|_2} \tag{10}$$

where θ is the angle between feature vectors $f(\mathcal{I}_i)$ and $f(\mathcal{I}_j)$, and $\|\bullet\|_2$ is an operator that denotes the L2 norm; $d(\bullet)$ is in the range [0, 2], with 0 being the most similar.

According to [4], learning only from simple samples limits the capacity of a trained network to generalize. We classify samples as hard according to the estimated viewpoint to incorporate hard sample mining in learning. As shown in Fig. 9, $\mathcal{P}_{\mathcal{VD}}$ represents a selected image with an almost completely different viewpoint from the anchor image, such as the stem rather than the stern (i.e., a very different positive sample). $\mathcal{N}_{\mathcal{VS}}$ indicates the most difficult negative samples to pick; they have a relatively small distance for feature vectors compared with the anchors. For simplicity, we ignore the effects of $\mathcal{D}(\mathcal{I}_a, \mathcal{I}_{ps})$, $\mathcal{D}(\mathcal{I}_a, \mathcal{I}_{nd})$ and β, and rewrite Eq. (9) as:

$$\begin{aligned}\mathcal{L}_{\text{Quin}} = \frac{1}{\mathcal{V} \times \mathcal{K}} \sum_{\mathcal{I}_a \in batch} &\left[\max_{\mathcal{T}_{pd} \in A} \mathcal{D}\left(\mathcal{I}_a, \mathcal{I}_{pd}\right)\right. \\ &\left. - \min_{\mathcal{I}_{ns} \in B} \mathcal{D}\left(\mathcal{I}_a, \mathcal{I}_{ns}\right) + \alpha\right]_+ \end{aligned} \tag{11}$$

where the batch consists of V vessels (each with a unique ID) and K different images of each vessel, for a total of V $\times$K images.

F. FUSION OF GLOBAL AND LOCAL FEATURES

The aspect of a vessel varies greatly from different viewpoints, and each aspect may contain different salient features. Over 90% of the vessel images in the dataset have more than three discriminative features. Thus we need to consider the weights of differentiated features in the process of feature fusion. Given one global feature f_g and a number of partial features $\{f_{p_i}\}_{i=1}^{n}$, we first aggregate these partial features into a local feature by:

$$f_{\text{loal}} = \alpha_1 f_1 \oplus \alpha_2 f_2 \oplus \cdots \oplus \alpha_P f_P \tag{12}$$

where the $\alpha_i, i = 1, 2, \ldots, P$, are the soft attention weights calculated by ResNet combined with a Squeeze-and-Excitation (SE) module [48], referred to as ResNet-SE.

Since partial features can be a perfect complement to global features, we combined the concentrated f_{local} with the global feature f_{global} as follows:

$$f_{fitsion} = \left[\lambda f_{global}, (1-\lambda)\tanh\left(f_{local} \odot W + B\right)\right] \tag{13}$$

where weight vectors W and bias terms B are learnable parameters; λ is a hyperparameter to weight partial features and global features that has an optimized value $\lambda = 0.4$.

A goal of training is to minimize the total loss function, which is the sum of Eq. (4), Eq. (7) and Eq. (11), by varying the weights of each component. We formulated the objective function as:

$$\mathcal{L}_{roid} = \underbrace{\beta_1 \mathcal{L}_{\text{Quin}} + \beta_2 \mathcal{L}_{\text{ve}}}_{\mathcal{L}_{glotal}} + \underbrace{\beta_3 \mathcal{L}_{\text{ve}} + \beta_4 \mathcal{L}_{\text{parts}}}_{\mathcal{L}_{Local}} \tag{14}$$

where β_1, β_2, β_3 and β_4 are hyperparameters used to balance the four different loss functions.

V. EXPERIMENT AND RESULT ANALYSIS

This section describes how we conducted a series of comparative and reductive experiments to evaluate the effectiveness of the GLF-MVFL model.

A. IMPLEMENTATION DETAILS

1) NETWORK ARCHITECTURE AND TRAINING

We used a residual convolutional network, ResNet-50 [49], as the backbone network for partial feature extraction, with a mini-batch size of 32. We ran 100 epochs with an initial learning rate of 0.001 divided by a factor of 5 every 20 epochs after the 40th epoch. Because of the constraints of graphics memory, we set $\mathcal{V} = 32$ and $\mathcal{K} = 8$ (i.e., Batch = 32×8).

2) BASELINE METHOD

The ResNet-50 model with cross-entropy combined with hard triplet loss was our baseline for comparison with GLF-MVFL [50]. Two other methods were selected to conduct the comparison experiments: the embedding network

(Cross-Entropy (Xent) + triplet loss, sometimes called multi-loss, which also uses ResNet-50 as the backbone network) was used from Bai's repository [14]; MDNet was also used, using the VGG_CNN_M_1024 backbone network, with coupled cluster triplet loss [12].

All algorithms in our experiments were processed on a 2.1 GHz Intel Xeon silver 4116 CPU with 64 GB of memory and two NVIDIA RTX2080Ti GPUs (11GB frame buffer) using the Pytorch deep learning framework (version 1.2).

B. EVALUATION PROTOCAL

We assume that each vessel in the query set will lead to the retrieval of a similar vessel from the gallery set. Therefore we used cumulative match characteristic (CMC), mean average precision (mAP) and mean average orientation similarity (mAOS) as performance metrics to evaluate our approach and to enable us to compare our results with other state-of-the-art methods.

1) CUMULATIVE MATCH CHARACTERISTIC

Rank-k indicates the probability that the top-k images in the search results (with the highest confidence) contain the correct result. In this study we evaluated several typical rank -k rankings (i.e., $k = 1, 5, 10,$ or 20). Cumulative match characteristic curves are drawn from rank-k values and are commonly used as evaluative indicators for closed-set testing. Assuming that querying and sorting operations are performed on the probe set consisting of Q vessels, the sorting results of each query are expressed in $k = (k_1, k_2, \ldots, k_Q)$ and CMC can be defined as:

$$\mathrm{CMC}(\mathcal{K}) = \frac{1}{\mathrm{Q}} \sum_{i=1}^{\mathrm{Q}} \begin{cases} 0, & k_i > \mathcal{K} \\ 1, & \text{others} \end{cases} \tag{15}$$

2) MEAN AVERAGE PRECISION

The average precision (AP) is given by the area enclosed by a precision-recall (P–R) curve and the coordinate axis. It indicates the performance of a Re-ID model. The AP of each query q can be calculated by:

$$AP(q) = \sum_{k=1}^{N} P(k) \times \Delta r(k) \tag{16}$$

where N is the total number of images in the test set, $P(k)$ is the precision when the k th image can be identified, and $\Delta r(k)$ indicates the recall value change in the number of images identified from $k-1$ to k. For a total of Q queries, mAP is given by:

$$mAP = \frac{1}{Q} \sum_{q=1}^{\mathrm{Q}} AP(q). \tag{17}$$

3) MEAN AVERAGE ORIENTATION SIMILARITY

We use mAOS to evaluate the performance of viewpoint estimation [16]:

$$mAOS = \frac{1}{Q} \sum_{q=1}^{\mathrm{Q}} AP(q) \times \frac{1 + cos(\Delta\varphi_q)}{2} \tag{18}$$

where $\Delta\varphi_q$ is the angular difference between estimated and ground truth orientation of detection.

C. COMPARISON TO STATE-OF-THE-ART RESULTS

Using the VesselID-539 dataset, we compared the performance of our approach with several state-of-the-art methods: baseline, EmbeddingNet, MDNet [12], and IORNet [26]. The results are summarized in Table 3.

From Table 3, it is evident that our GLF-MVFL approach provided the best results. When compared to the baseline pipeline results, GLF-MVFL showed 9.2%, 5.5% and 6.6% increases in mAP, Rank-1 and Rank-5. Note that after discarding the last down-sampling operation of ResNet-50 by modifying the last stride from 2 to 1, both mAP and Rank-1 improvements are even greater, increasing 3.2% and 2.8% of the original GLF-MVFL results. We customized a lightweight model, GLF-MVFL (Lite), in which the backbone network uses the same RestNet-50 as the baseline network to better verify the effect of O-Quin. Table 3 shows that GLF-MVFL (Lite) exceeded the baseline by 7% on mAP; it also showed a significant improvement in Rank-1/5/10/20.

D. VISUALIZATION OF RESULTS

In this section we provide some typical visualization results to show intuitively the accuracy of our vessel re-identification model as following.

Fig. 10 shows the top-10 retrieval results showing diverse viewpoints for the same vessel in our GLF-MVFL model. In contrast, the results retrieved by the baseline method are monotonous. We infer that features of the same vessel can be

TABLE 3. Results (%) compared with other state-of-the-art methods.

Method	Loss type	Backbone	mAP	Rank-1	Rank-5	Rank-10	Rank-20
Baseline	Xent+TriHard	ResNet-50	65.7	55.9	83.0	88.0	90.7
EmbeddingNet	Xent+Triplet	ResNet-50	66.4	56.8	82.4	86.8	91.3
MDNet	ArcFace	VGGM_1024	52.3	42.8	66.8	—	—
IORNet	Xent+Triplet	ResNet-50	71.5	61.2	—	95.5	—
GLF-MVFL (Lite)	Xent+O-Quin	ResNet-50	72.7	58.6	87.0	92.3	95.1
GLF-MVFL	Xent+O-Quin	ResNet50-SE	74.9	61.4	89.6	95.0	98.0
GLF-MVFL (last stride=1)	Xent+O-Quin	ResNet50-SE	**78.1**	**64.2**	**91.0**	**96.4**	**98.9**

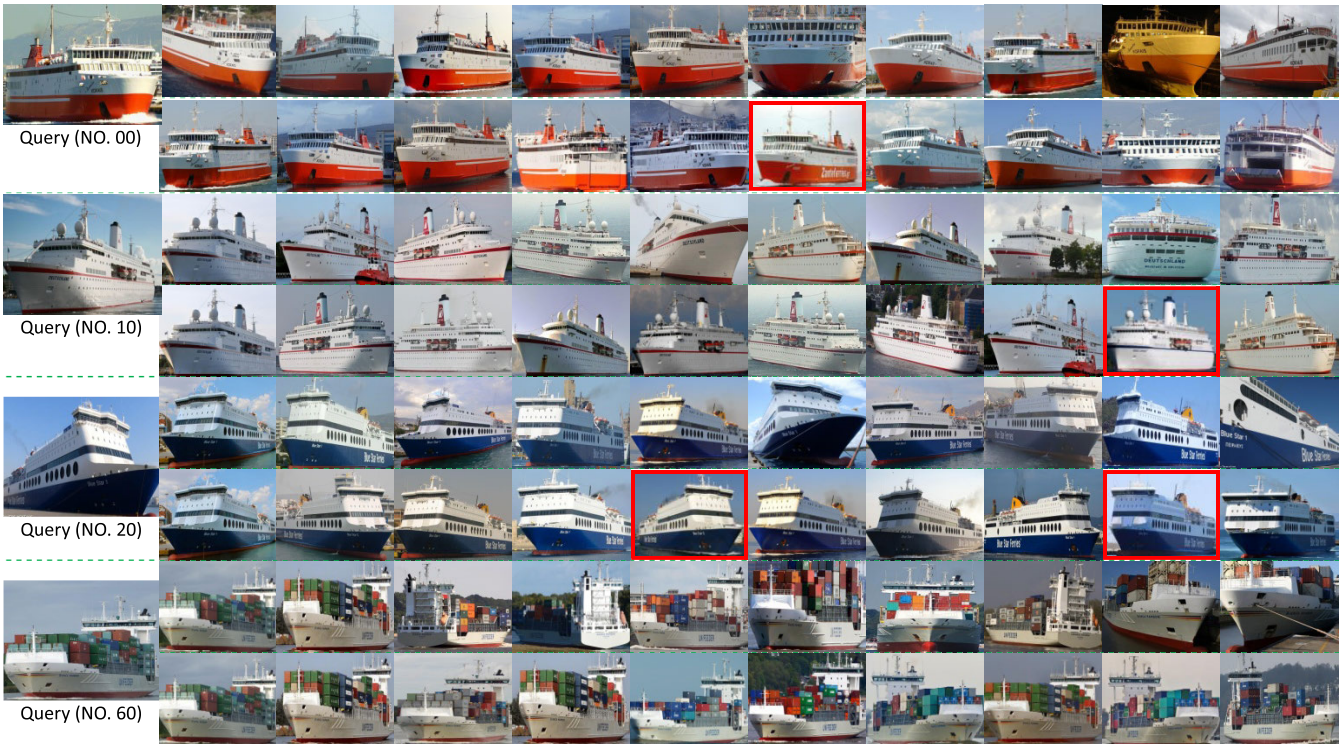


FIGURE 10. Illustration of the top-10 ranking list for retrieval results (four query images in total) from our VesselID-539 datasets. For each query image, the first row shows the top 10 retrieval results of our GLF-MVFL model; and the second row for the baseline model. The red bounding boxes show false positives.

TABLE 4. Performance (%) of different detectors and the effects of different vessel detectors. Note that we also use floating point operations per second (FLOPs) and frames per second (FPS) to indicate the complexity and inference speed of the model.

Vessel Detector							Vessel ReID			
Detector	Backbone	mAP	mAOS	GFLOPs	Params (M)	FPS	mAP	Rank-1	Rank-5	Rank-10
Faster RCNN	ResNet-101	75.2	77.6	98	64.3	5	73.8	59.5	88.3	93.6
SSD	MobileNetv2	41.2	57.5	**1.3**	22.1	59	55.5	44.7	66.4	70.4
YOLOv3	MobileNetv2	61.9	70.5	19.5	**9.3**	68	68.1	54.9	81.4	86.4
YOLOv3	Darknet53	55.3	61.6	65.7	59.2	46	59.4	47.9	71.1	75.4
YOLOv3	**EfficientNet**	**75.7**	**78.2**	18.6	19.3	56	**74.9**	**61.4**	**89.6**	**95.0**

clustered together using our GLF-MVFL model, no matter what the orientation of the vessel.

E. ABLATION STUDY AND DISCUSSION

1) EFFECT OF VESSEL DETECTOR

We tested the GLF-MVFL model to see if it could be developed further, using the viewpoint estimation approach described in Section IV. We replaced the detectors with Faster R-CNN [51], SSD [28], YOLOv3 (with MobileNetv2 as the backbone) and native YOLOv3 with Darknet53 [36] and conducted comparative experiments. The results are shown in Table 4.

Table 4 shows that our method gives the best tradeoff between mAP and FPS using the VesselID-539 dataset, and reduces both the time complexity (GFLOPs) and space complexity (Params). The key mAP and mAOS indexes increased by 20.4% and 16.6% over native YOLOv3. The best result is then given in the subsequent V-ReID, as shown on the right side of Table 4. The results also show that our viewpoint estimation approach is robust and compatible with many other state-of-the-art detectors and has better scalability. Table 4 also shows that there is a positive correlation between mAOS and mAP for V-ReID.

2) EFFECT OF GLOBAL-AND-LOCAL FUSION

To evaluate the effect of our global-and-local fusion mechanism, we conducted an ablation study by removing the global-and-local fusion and attention mechanisms. Table 5 shows the results.

A comparison of the results in Table 5 shows that when we combined partial features with global features, we improved the mAP and Rank-1 to 74.9% and 61.4%, which are the highest values of all methods. Global+ Partial shows results that show large increases over Global Only, particularly the 9.5% increase for Rank-20. For our overall framework

TABLE 5. Results of our method and three other variants (%).

Method	mAP	Rank-1	Rank-5	Rank-10	Rank-20
Partial Fusion	61.9	49.9	74.0	78.5	81.0
Global Only	66.9	53.9	80.0	84.9	87.5
Global+Partial	74.2	59.8	88.7	94.1	97.0
Global+Partial+Attention	**74.9**	**61.4**	**89.6**	**95.0**	**98.0**

TABLE 6. Effects of different resolutions (%).

Input size	mAP	Rank-1	Rank-5	Rank-10	Rank-20
64×196	66.7	53.9	83	90.6	95
128×256	71	58.5	87.2	94	96.9
128×384	72.1	59.9	87.3	91.9	94.8
256×384	74.9	61.4	**89.6**	**95.0**	**98.0**
256×768	**75.9**	**65.7**	88.6	93.7	96.7

(Global + Partial + Attention), the gain comes mainly from the extraction of fine-grained partial feature and the embedding of the channel attention module. These results also suggest that using our global-and-local fusion mechanism provides a greater capacity to discriminate than the use of global features only.

3) EFFECT OF DIFFERENT RESOLUTIONS

We conducted experiments using different resolutions to evaluate the effect of input size on mAP and CMC. Table 6 shows the results of using typical input size ranges from 64×196 to 256×768 pixels.

Table 6 shows that the best results are obtained when the input image size is 256×384, especially for mAP and Rank-1, which increased by 8.2% and 7.5% over the 64×196 input size. These data lead us to conclude that higher resolution improves performance. However, there are some bottlenecks: when further increasing resolution to 256×768, Rank-5, Rank-10 and Rank-20 show no improvement. Taking these results into account, we resized the input to 256×384 (with a height to width ratio of 2:3) for the rest of this study.

4) EFFECT OF BACKBONE NETWORK

To ensure a fair comparison between different backbone networks, we resized all images to 256×384 to compare our model with other state-of-the-art backbone models, ResNet-34/50/101/152, DenseNet, and SE_DenseNet, using the VesselID-539 dataset.

Table 7 shows that the combination of SE and ResNet-50 gave significantly better results and increased mAP by 2.3%. In contrast, the combination of SE and DenseNet gave only a fairly minor improvement. The reason for this result is that SE reduces the redundancy of ResNet, giving a more diverse internal structure, whereas DenseNet cannot be further optimized.

In summary, our GLF-MVFL model has the advantage that, without using any temporal information (e.g., RNN) and without the need for extra re-ranking, it achieved state-of-the-art results in marine vessel re-identification. The results demonstrate that the use of different detectors has significant effects on vessel re-identification. Thus, using an ingeniously-designed detector that can estimate the viewpoint is preferable to the use of off-the-shelf detectors. Previous experiments also demonstrate the universality of the vessel re-identification framework that we built.

VI. CONCLUSION

We have presented a global-and-local feature fusion vessel re-identification model which combines metric learning and representation learning for network training and incorporates global and local features. Experimental results demonstrate that our model accurately extracts discriminative features from ship images. This capability gives the model state-of-the-art performance in vessel re-identification. We also conducted ablation studies to identify the contribution of each component of our model to the overall performance. The VesselID-539 dataset that we created uniquely provides a large-scale dataset for marine vessel re-identification. VesselID-539 is well annotated and rich in attributes. In addition to its use for vessel re-identification, VesselID-539 can be generalized for other vision tasks at sea, such as fine-grained ship classification. Our future work has two directions. We will exploit this novel method to locate unannotated discriminative features and combine vessel re-identification with MTMC tracking for marine surveillance; and we will combine RGB images with infrared images, AIS, and video echoes from navigational radar. Our ultimate goal is to realize multi-modal re-identification of marine vessels. Due to the lack of other available large-scale datasets, we have to await the emergence of more vessel re-identification datasets to further evaluate our proposed model.

TABLE 7. Performance evaluation results with different backbone networks (%).

Backbone	mAP	Rank-1	Rank-5	Rank-10	Rank-20	Param. (Millions)	GFLOPs
ResNet-34	67.7	55.6	81.6	89.7	94.8	21.8	7.2
ResNet-50	72.6	60.5	88.7	94.1	97.0	25.6	8.1
SE_ResNet-50	**74.9**	61.4	**89.6**	**95.0**	**98.0**	25.6	8.8
ResNet-101	71.1	59.4	85.7	92.0	95.9	44.5	15.4
ResNet-152	72.4	61.0	85.9	91.4	94.5	60.2	22.7
DenseNet-121	73.7	62.1	89.3	94.9	96.8	**8.0**	**5.7**
SE_DenseNet	73.7	**62.3**	89.4	94.8	96.9	**8.0**	6.2

REFERENCES

[1] W. Naeem, G. W. Irwin, and A. Yang, "COLREGs-based collision avoidance strategies for unmanned surface vehicles," *Mechatronics*, vol. 22, no. 6, pp. 669–678, Sep. 2012.

[2] W. Zajdel, Z. Zivkovic, and B. Krose, "Keeping track of humans: Have I seen this person before?" in *Proc. IEEE Int. Conf. Robot. Autom.*, Jan. 2006, pp. 2081–2086.

[3] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*. [Online]. Available: https://arxiv.org/abs/1610.02984

[4] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: https://arxiv.xilesou.top/abs/1703.07737

[5] H. Cai, "Multi-scale body-part mask guided attention for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019. [Online]. Available: http://openaccess.thecvf.com/content_CVPRW_2019/html/TRMTMCT/Cai_Multi-Scale_Body-Part_Mask_Guided_Attention_for_Person_Re-Identification_CVPRW_2019_paper.html

[6] D. Heo, J. Nam, and B. Ko, "Estimation of pedestrian pose orientation using soft target training based on teacher-Student Framework," *Sensors*, vol. 19, no. 5, p. 1147, Mar. 2019.

[7] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[8] L. Zheng, "Mars: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Compu. Vis. (ECCV)*, Oct. 2016, pp. 868–884.

[9] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Compu. Vis. (ECCV)*, Oct. 2016, pp. 17–35.

[10] J. Lv, X. Chen, and M. Salah, "Intelligent re-recognition algorithm for specific ship target in busy waters under the actual scene," *J. Intell. Fuzzy Syst.*, vol. 35, no. 4, pp. 4433–4443, Oct. 2018.

[11] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. *End-to-End Deep Learning for Person Search*. Accessed: 2016. [Online]. Available: http://www.ee.cuhk.edu.hk/~xgwang/PS/paper.pdf

[12] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2167–2175.

[13] Y. Xiang, Y. Fu, and H. Huang, "Global topology constraint network for fine-grained vehicle recognition," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[14] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L.-Y. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2385–2399, Sep. 2018.

[15] C. Guindel, D. Martin, and J. M. Armingol, "Fast joint object detection and viewpoint estimation for traffic scene understanding," *IEEE Intell. Transp. Syst. Mag.*, vol. 10, no. 4, pp. 74–86, Sep. 2018.

[16] X. Zhang, R. Zhang, J. Cao, D. Gong, M. You, and C. Shen, "Part-guided attention learning for vehicle re-identification," 2019, *arXiv:1909.06023*. [Online]. Available: https://arxiv.xilesou.top/abs/1909.06023

[17] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[18] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "VERI-Wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3235–3243.

[19] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J.-N. Hwang, "CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8797–8806.

[20] U. Kanjir, H. Greidanus, and K. Oštir, "Vessel detection and classification from spaceborne optical images: A literature survey," *Remote Sens. Environ.*, vol. 207, pp. 1–26, Mar. 2018.

[21] C. M. Ward, A. G. Corelli, and J. D. Harguess, "Leveraging synthetic imagery for collision-at-sea avoidance," in *Proc. 8th Geospatial Informat., Motion Imag., Netw. Anal.*, May 2018, Art. no. 1064507.

[22] D. Heyse, N. Warren, and J. Tesic, "Identifying maritime vessels at multiple levels of descriptions using deep features," in *Proc. Artif. Intell. Mach. Learn. Multi-Domain Oper. Appl.*, May 2019, Art. no. 1100616.

[23] D. Qiao, G. Liu, J. Zhang, Q. Zhang, G. Wu, and F. Dong, "M^3C: Multimodel-and-multicue-based tracking by detection of surrounding vessels in maritime environment for USV," *Electronics*, vol. 8, no. 7, p. 723, Jun. 2019.

[24] C. Tian, J. Xia, J. Tang, and H. Yin, "Deep image retrieval of large-scale vessels images based on BoW model," *Multimed. Tools Appl.*, vol. 17, pp. 1–15, May 2019.

[25] C. Hilton, S. Parameswaran, M. Dotter, C. M. Ward, and J. Harguess, "Classification of maritime vessels using capsule networks," in *Proc. 15th Geospatial Informat.*, Jun. 2019, Art. no. 109920E.

[26] A. Ghahremani, Y. Kong, E. Bondarev, and P. H. N. de With, "Towards parameter-optimized vessel re-identification based on IORnet," in *Proc. Int. Conf. Comput. Sci.*, Jun. 2019, pp. 125–136.

[27] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 420–429.

[28] A. Ghahremani, Y. Kong, E. Bondarev, and P. H. D. With, "Multi-class detection and orientation recognition of vessels in maritime surveillance," *Electron. Imag.*, vol. 2019, no. 11, pp. 266-1–266-5, Jan. 2019.

[29] Z. Li, Y. Wang, and X. Ji, "Monocular viewpoints estimation for generic objects in the wild," *IEEE Access*, vol. 7, pp. 94321–94331, 2019.

[30] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 379–387.

[31] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3997–4005.

[32] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2017, pp. 480–496.

[33] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. ACM Multimedia Conf. Multimedia Conf. (MM)*, 2018, pp. 274–282.

[34] X. Tan, "Multi-camera vehicle tracking and re-identification based on visual and spatial-temporal features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 275–284.

[35] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114.

[36] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.xilesou.top/abs/1804.02767

[37] E. Gundogdu, B. Solmaz, V. Yücesoy, and A. Koç, "MARVEL: A large-scale image dataset for maritime vessels," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2016, pp. 165–180.

[38] Q. Xiao, H. Luo, and C. Zhang, "Margin sample mining loss: A deep learning based method for person re-identification," 2017, *arXiv:1710.00478*. [Online]. Available: https://arxiv.xilesou.top/abs/1710.00478

[39] A. Hasnat, J. Bohne, J. Milgram, S. Gentric, and L. Chen, "DeepVisage: Making face recognition simple yet with powerful generalization skills," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1682–1691.

[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2015, pp. 448–456.

[41] Z. Zhang and M. Huang, "Person re-identification based on heterogeneous part-based deep network in camera networks," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 4, no. 1, pp. 51–60, Feb. 2020.

[42] S. Tulsiani and J. Malik, "Viewpoints and keypoints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1510–1519.

[43] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

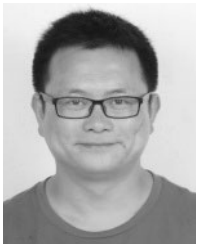
[44] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process Syst. (NIPS)*, 2014, pp. 1988–1996.

[45] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[46] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 403–412.
[47] S. Li, J. Li, W. Lin, and H. Tang, "Amur tiger re-identification in the wild," 2019, *arXiv:1906.05586*. [Online]. Available: https://arxiv.org/abs/1906.05586
[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.
[49] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[50] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," May 2019, *arXiv:1905.00953*. [Online]. Available: https://arxiv.org/abs/1905.00953
[51] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

DALEI QIAO received the B.S. and M.S. degrees from the China University of Mining and Technology, in 2004 and 2008, respectively. He is currently pursuing the Ph.D. degree with the College of Information Engineering, Shanghai Maritime University, China. His research interests include unmanned surface vessel, computer vision, and embedded systems.

GUANGZHONG LIU received the B.S. degree from Southwest Jiaotong University and the Ph.D. degree from the China University of Mining and Technology. He is currently a Professor of computer science and engineering with Shanghai Maritime University. His specific research interests include underwater acoustic communication technology, mobile networking, wireless communication, and intelligence information systems.



FENG DONG received the B.S. degree from the Hunan University of Science and Technology, in 2002, and the M.S. degree from the Huazhong University of Science and Technology, in 2008. He is currently pursuing the Ph.D. degree with the College of Information Engineering, Shanghai Maritime University, China. His research interests include computer vision and edge artificial intelligence.

SHE-XIANG JIANG was born in 1981. He is currently pursuing the D.Sc. degree with Shanghai Maritime University. His research interests include quantum image processing, quantum communication, quantum teleportation, and remote state preparation.

LIKUN DAI received the B.S. degree from Dalian Maritime University, in 1999, and the M.S. degree from Nanjing University, in 2005. He is currently an Associate Professor of network engineering with Jiangsu Maritime Institute. His research interests include unmanned surface vessel, deep learning, and embedded systems.