Ship Landmark: An Informative Ship Image Annotation and Its Applications

Mingxin Zhang†, Qian Zhang†, Ran Song*, Paul L. Rosin, and Wei Zhang

***Abstract*—Visual perception of ships has been attracting increasing attention in the fields of computer vision and ocean engineering. Despite the extensive work related to landmark detection of common objects, the role of landmarks in ship perception has been overlooked. In this paper, we aim to fill this gap by focusing on ship landmarks. Specifically, we give a comprehensive analysis of both the physical structure and deep features of ships, which finds that highlighted areas in feature maps correspond with structurally significant parts of ships. By summarizing the locations of such areas in ships, we define 20 ship landmarks and build the Ship Landmark Dataset (SLAD), the first ship dataset with landmark annotations. We also provide a benchmark for ship landmark detection by evaluating state-of-the-art landmark detection methods on the newly built SLAD. Moreover, we showcased several applications of ship landmarks, including ship recognition, ship image generation, key area detection for ships, and ship detection. Project web page: https://vsislab.github.io/Ships_VSIS/.**

***Index Terms*—Ship Images, Landmark Detection, Maritime Transportation, Computer Vision**

I. Introduction

SHIPS play an irreplaceable role in transportation by carrying over 80% of the world's trade [1]. It is thus important to monitor ship positions and navigation status for safety and economic reasons. For a considerable period of time in the past, ship monitoring was mainly done by full-time maritime staff, who may be fatigued after working for a long time, resulting in the inability to achieve sustainable monitoring. Over the past decade, the burst of artificial intelligence techniques has led to the rapid development of unmanned means of ship monitoring, where ship perception usually takes an important role [2]–[5].

Nowadays, intelligent ship perception systems involve multiple sensors to capture as much information about ships as possible, including but not limited to the Automatic Identification System (AIS), synthetic aperture radar (SAR) systems, land-based radars, and cameras. Among these means of ship perception, researchers have given most attention to the vision-based techniques due mainly to the rapid development of computer vision and deep learning techniques, which have spawned a range of innovative algorithms, large-scale datasets, and influential academic competitions.

Researchers have invested much effort in brain-like neural networks to endow deep models with human-like visual perception capabilities [6]–[8], leading to great success in classifying and detecting visual objects primarily using texture information [9]–[11]. However, different from other common objects, visual textures of different ships share a rather high similarity due to the shared requirements imposed by the specific nature of marine navigation, while visual textures of the same ship vary significantly when the states of cargo, the viewpoints of observation, or the painting colours change. Therefore, texture-based methods [12], [13] cannot produce satisfactory results in some challenging ship perception tasks such as fine-grained ship recognition and ship key area detection, as they acquired the status of ships based mainly on instance-level or pixel-level texture information of ship images. In this case, exploring additional forms of information can provide a solution to achieve accurate ship perception.

As listed in Table I, there already exists a number of public ship datasets [14]–[39] that can be used for ship analysis. Most of them [14]–[31] were constructed for ship detection in the remote sensing scenario, where the bounding boxes of ships are annotated in optical or SAR satellite images. Although bounding box annotations in remote sensing datasets allow deep models to learn to locate ships from a bird's-eye view, they do not provide detailed information about ships. In contrast, ship images in RGB datasets [32]–[39] are usually captured from a horizontal perspective using cameras installed onshore or onboard, and thus they contain more visual clues about the ships.

In the field of visual perception, a landmark refers to a specific location or region in an image considered visually distinctive and semantically meaningful. In multiple computer vision tasks, localization and recognition of interest points have shown great potential in contributing to building a high-level semantic understanding of objects. For instance, facial interest points, also known as facial landmarks, play a crucial role in expression analysis [45] and computer-aided facial disease diagnosis [46]. With the skeleton defined by landmarks that correspond to the major joints of the human body, deep models perform well in human pose estimation, which has been successfully applied in action recognition [47], gesture analysis [48], [49], and human-computer interaction [50]. Moreover, some researchers utilized landmarks in the visual

This work was supported in part by the National Natural Science Foundation of China under Grants 61991411, U22A2057, and 62076148, in part by the National Science and Technology Major Project under Grant 2021ZD0112002, in part by the Shandong Excellent Young Scientists Fund Program (Overseas) under Grant 2022HWYQ-042, in part by the Young Taishan Scholars Program of Shandong Province No.tsqn201909029, and in part by Project for Self-Developed Innovation Team of Jinan City under Grant 2021GXRC038.

Mingxin Zhang, Qian Zhang, Ran Song, and Wei Zhang are with the School of Control Science and Engineering at Shandong University, China (Email: 95zhangmingxin@gmail.com; zhq9669@gmail.com; ransong@sdu.edu.cn; davidzhang@sdu.edu.cn).

Paul L. Rosin is with the School of Computer Science and Informatics at Cardiff University, UK (Email: RosinPL@cardiff.ac.uk).

†Mingxin Zhang and Qian Zhang contributed equally to this work.

*Corresponding author: Ran Song (Email: ransong@sdu.edu.cn).

TABLE I
COMPARISON OF PUBLICLY SHIP DATASETS.

Name	Acquisition Method	Images Source	Images	Instances	Categoreies	Years
WHU-RS19 [14]	Optical Radar	Open Source	1,005	50	1	2011
ImageNet 2012 [40]	Cameras	Open Source	16,114	—	12	2012
PASCAL VOC 2012 [41]	Cameras	Open Source	353	791	1	2012
NWPU VHR-10 [16]	Optical Radar	Open Source	800	302	1	2014
MARVEL [32]	Cameras	Open Source	400,000	—	26	2016
SMD [33]	Cameras	Shipboard cameras	157,547 (frame)	157,547	6	2016
HRSC2016 [17]	Optical Radar	Open Source	1,070	2,976	25	2016
RSC11 [18]	Optical Radar	Open Source	232	100	1	2016
MS COCO [42]	Cameras	Open Source	3,164	11,189	1	2017
SSDD [15], [43]	SAR	Open Source	1,160	2,456	7	2017
OpenSARShip [19]	SAR	Open Source	41	11,346	17	2017
NWPU-RESISC45 [20]	Optical Radar	Open Source	31,500	700	1	2017
DOTA [21]	Optical Radar	Open Source	2,806	2,702	1	2017
SeaShip [34]	Cameras	Shore-based cameras	31,455 (7,000)	40077 (9221)	6	2018
DIOR [16]	Optical Radar	Open Source	23,463	64,000	1	2018
HRRSD [22]	Optical Radar	Open Source	21,761	3,886	1	2018
Boat Re-ID [35]	Cameras	Shore-based cameras	5,523	5523	-	2019
AR-Ship-Dataset [23]	SAR	Open Source	210	43,819	-	2019
Airbus ship [26]	Optical Radar	Open Source	1,925,560	231,723	1	2019
MASATI [27]	Optical Radar	Open Source	6,212	3,313	1	2019
xView [28]	Optical Radar	Open Source	1,413	5,672	9	2019
McShips [36]	Cameras	Open Source	14,709	26,529	13	2020
HRSID [24]	SAR	Open Source	136	16,951	-	2020
LS-SSDD-v1.0 [25]	SAR	Open Source	15	9,000	-	2020
FGSD [29]	Optical Radar	Open Source	2,612	5,634	43	2020
ABOships [37]	Cameras	Shipboard cameras	9,880	41,967	9	2021
ShipRSImageNet [30]	Optical Radar	Open Source	3,435	17,573	50	2021
SeaSAw [44]	Cameras	Shipboard cameras	1.9M	14.6M	12	2022
LEVIR-Ship [31]	Optical Radar	Open Source	3,896	3,219	1	2022
VesselReID [38]	Cameras	Open Source	30,589	30,589	-	2023
SmartShip-HEU [39]	Cameras	Open Source	12,300	23,542	6	2023

perception of vehicles, where localizing such points helps in both pose estimation and 3D shape reconstruction of vehicles [51]. When humans perceive ships, their landmarks often exist near the corners or the intersections of specific areas including the bow, stern, and superstructure. These points help humans quickly grasp the structural characteristics of a ship and form the corresponding semantic concepts. Based on the successful application of landmarks in various scenarios and the human perception of ships, we argue that landmarks of ships are able to facilitate their visual perception. However, there is no work that applies landmarks for ship visual perception, due in part to the absence of ship datasets with landmark annotations.

This paper proposes the first ship dataset with landmark annotations by exploring the visual structure of ships. The initial step is to determine the positions and the number of ship landmarks. To accomplish this, we investigate the visual structure of ships and discover the presence of key areas on the ship's hull that contribute to ship perception. Specifically, we conduct an in-depth comparative analysis of deep feature maps for ship images with different views and categories, allowing us to figure out the locations of key areas. Then, taking into account both the experimental results and the expert knowledge of ships, we define a series of landmarks to annotate the ships for a new dataset named Ship Landmark Dataset (SLAD). Based on the newly built dataset, we present a benchmark with state-of-the-art algorithms for landmark detection to serve as baselines for further research on ship landmark detection. Furthermore, we discuss the potential applications of SLAD in the hope that it can provide valuable insights for researchers in the field of ship perception.

Overall, the contributions of this paper are threefold:

(1) We analyze the distribution of ship features generated by different recognition models to define 20 landmarks in a ship image, which provides structural information useful for various ship perception tasks.

(2) We create SLAD, the first public dataset of ship images with landmark annotations, and evaluate multiple landmark detection methods on SLAD to provide a benchmark for boosting further research in ship perception.

(3) We demonstrate the significance of SLAD by showcasing several applications of ship landmarks, including ship recognition, ship image generation, key area detection for ships, and ship detection.

II. RELATED WORK

This section reviews the related work, which can be grouped into three categories, including ship datasets, landmark datasets, and landmark detection methods.

A. Ship Datasets

In the past decade, the rapid development of computer vision techniques has spurred a range of research works in vision-based ship analysis, leading to a number of publicly available ship image datasets. According to the distance of image acquisition, these datasets can be divided into remote sensing datasets and ordinary datasets. The remote sensing datasets are usually composed of SAR images captured from SAR satellites, or optical images captured by optical radar. For

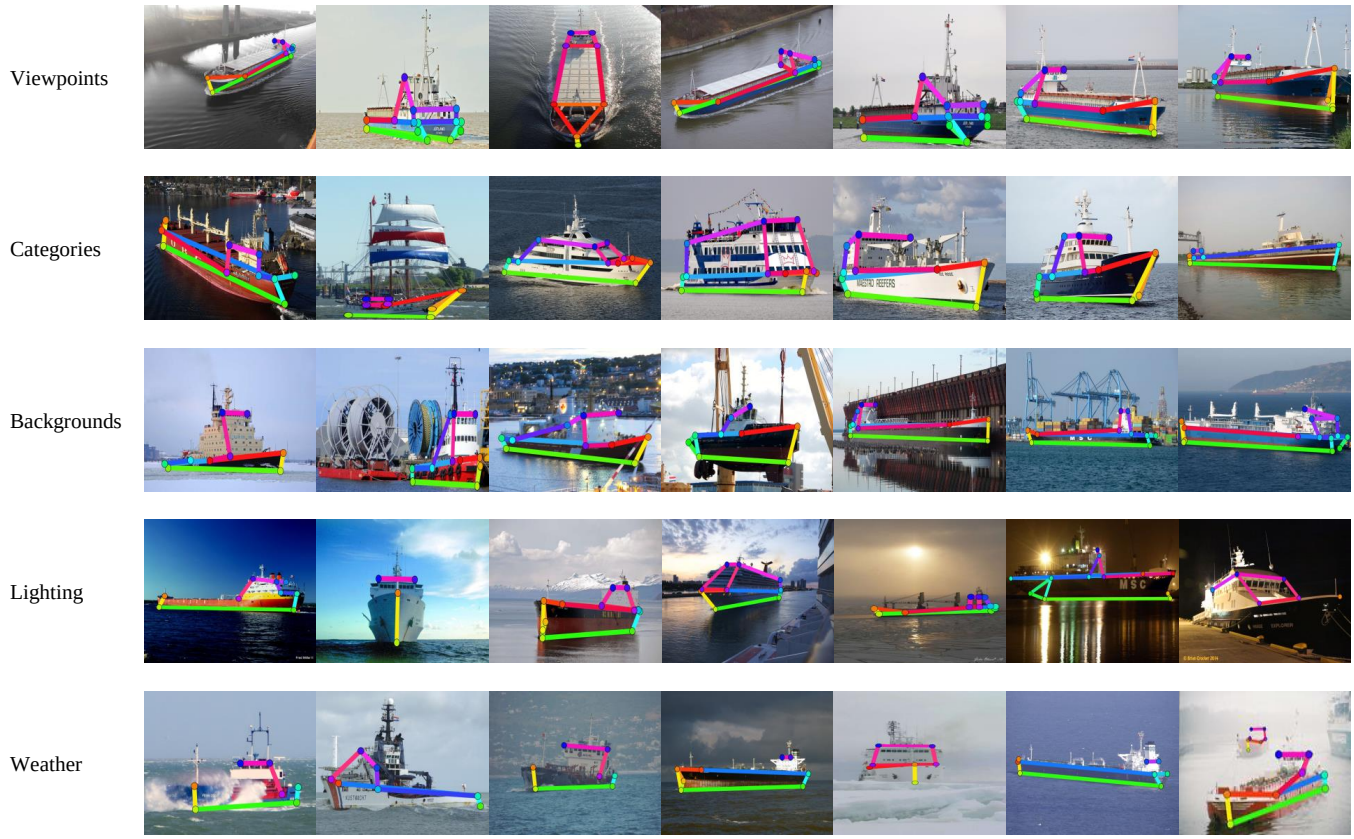


Fig. 1. Examplar ship images with landmark annotations in SLAD. The ship images differ in viewpoints, categories, backgrounds, lighting, and weather.

example, Gallego et al. [27] created the MASATI dataset with $6,212$ optical satellite images collected from Microsoft Bing maps. Each sample in this dataset was manually labelled with its category. Li et al. [15] publicly released the first ship SAR dataset, SSDD, which is further improved by Zhang et al. [43] for the official release. In order to facilitate deep learning-based SAR ship detection, Zhang et al. [25] introduced a large-scale ship SAR dataset named LS-SSDD-v1.0. Huang et al. [19] presented the OpenSARShip dataset consisting of $11,346$ SAR ship chips from 41 Sentinel-1 SAR images. Such kinds of images can provide a broad view and global coverage but lack detailed description of ships.

In contrast, the ordinary datasets include ship images captured by onshore or onboard cameras, which provide a close-up perspective of ships with more visual details. Gundogdu et al. [32] collected 2 million in ship images from a website where the users upload images with ship attributes. They categorized and annotated these images for building the large-scale MARVEL dataset. Iancu et al. [37] constructed the ABOships dataset with $9,880$ images and $41,967$ annotations to address ship detection in a number of operating scenarios. Although there have been several ordinary ship datasets that annotate the ships with their locations and categories, none of these datasets focus on specific parts to obtain the local visual details of ships. Therefore, we aim to address this limitation by annotating each ship with suitable landmarks to facilitate detailed perception of the ships.

B. Landmark Datasets

In the field of computer vision, landmarks have been attracting significant attention for a long time, resulting in numerous datasets for Landmark Detection [42], [52]–[55]. Most of these datasets [42], [52], [53] are designed for human pose estimation, which provide annotated images or videos specifically focused on landmarks of the human body. For example, in the MS COCO dataset [42], the persons in the images are annotated with 17 landmarks, covering the face and the major joints of the human body. PoseTrack [52] is a large-scale video-based dataset for human pose estimation, where 15 landmarks are defined to represent the pose of a person in video frames. There also exist some landmark-based datasets in other scenarios. Cao et al. [54] created the Animal-Pose dataset, in which not only the coordinates of landmarks were labelled, but also the "bones" were defined to build the skeletons of animals. Wang et al. [55] annotated the vehicles in VeRi-776 dataset [56] with 20 landmarks, from which orientation invariant features can be extracted for vehicle re-identification. However, there is no landmark dataset specifically created for ships, and the concept of ship landmarks has never been mentioned in previous works. Hence, we are committed to filling this gap by constructing a public ship dataset with landmark annotations.

C. Landmark Detection Methods

Landmark detection is a computer vision technique that aims to identify and locate distinct points of objects in

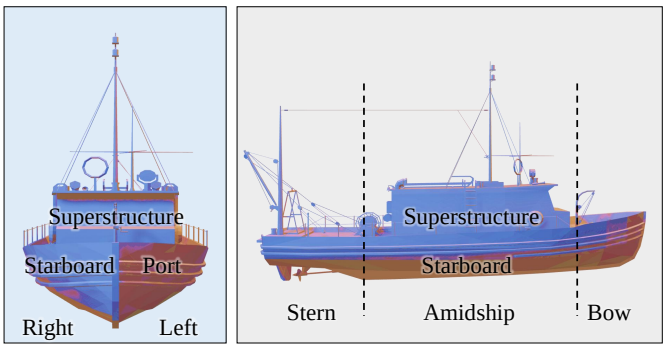


Fig. 2. Overview of ship structure. This figure illustrates the structural composition of the ship from two main perspectives.

images. Most existing methods for landmark detection focus on landmark-based human pose estimation due to its wide applications in action recognition and human computer interaction. For example, DeeperCut [57] was specially proposed for multi-person pose estimation, which applied image-conditioned pairwise terms between body parts to boost the performance of landmark prediction. Sun et al. [58] designed an effective backbone with multi-scale fusion modules to learn high-resolution representation for human pose estimation.

In general, existing landmark detection approaches can be roughly divided into two categories: top-down methods [58]–[67] and bottom-up methods [57], [68]–[76]. The top-down methods usually involve the bounding box of an object as prior knowledge, allowing the detector to place its focus on the object rather than the cluttered background. As such, the difficulty of landmark detection is significantly reduced. On the contrary, the bottom-up methods start with detecting all landmarks in the entire image, followed by a landmark clustering process to obtain the final detection results. Since dense landmark prediction is a necessary step in bottom-up landmark detection, such kind of methods require a large amount of annotated data to achieve satisfactory performance.

III. Ship Landmark Dataset

In this section, we elaborate the process of creating the Ship Landmark Dataset (SLAD), including data collection, ship landmark selection, and data annotation.

A. Data Collection

Considering the unique characteristics of maritime scenes, conducting on-ground photography to collect ship images is labour-intensive and potentially hazardous. Therefore, we download ship images from *Shipspotting*[1], where the basic information of ships, such as category, location, and International Maritime Organization (IMO) number, can be easily accessed. Then we filter out ships with less than 150 images, and select images of each ship with various external conditions while ensuring the class balance of ships. As a result, SLAD consists of $12,199$ images of 949 different ships. In Fig. 1, we show some representative ship images in SLAD, which vary considerably in terms of viewpoint, category, background, light, and weather.

To annotate the ship samples in SLAD with landmark information, we further build a dataset for identity-level ship recognition (referred to as Ship-ID) by filtering out the images with more than one ship from SLAD. This dataset facilitates the analysis of deep features of ships for landmark selection. Note that we only keep in the Ship-ID dataset the ships containing more than 8 images captured at different views to ensure 5 images per ship for training and 3 images per ship at least for testing. Moreover, the Ship-ID dataset consists of $11,079$ ship images, which are further split into training and testing subsets with $3,195$ and $7,884$ ship images, respectively.

B. Ship Landmark Selection

Similar to the landmarks of other objects, ship landmarks should refer to specific locations that hold special significance or importance. In this section, we aim to find out the significant parts of ships via conducting comparative analysis on the structural characteristics and the feature distributions of different ships.

1) Structural Analysis of Ships: Ship structure is an invariant property of ships that remains relatively stable in the face of changing factors such as viewpoint, illumination, and background. Since each type of ship has its specific structure [77], [78], humans can effectively distinguish between different types of ships relying on the invariant features of ship structures. Specifically, as shown in Fig. 2, a ship includes three main parts in the horizontal direction: the stern, which is the rear or aft section of the ship; the midship, which is the middle section; and the bow, which is the front or forward section. In the vertical direction, a ship is divided into two parts by the deck, namely the superstructure and the main hull. The superstructure refers to the upper part of the ship that includes various compartments like cabins, bridge, navigation areas, and other facilities. The main hull, on the other hand, is the lower part of the ship submerged in water and provides buoyancy and stability to the ship.

Different types of ships usually have different combinations of bow, stern and superstructure, which provide strong clues for visual perception of ships. For example, passenger ships often have a prominent superstructure located towards the midship or aft. They usually have multiple decks for accommodation, entertainment, and other facilities, and their bows are often sleek and rounded for better hydrodynamics. Cargo ships typically have a large, rectangular-shaped superstructure located towards the bow or midship. The stern of a cargo ship is usually flat and wide to facilitate loading and unloading of cargo. In Fig. 3, we present the structural sketches of several types of ships with different bow, stern, or superstructure. One can find that the relative positions of the vertices determine the structural property of ships and serve as distinctive features, which includes some key information that may be useful for ship perception. Therefore, it is feasible to define ship landmarks near these vertices.

[1]http://www.shipspotting.com

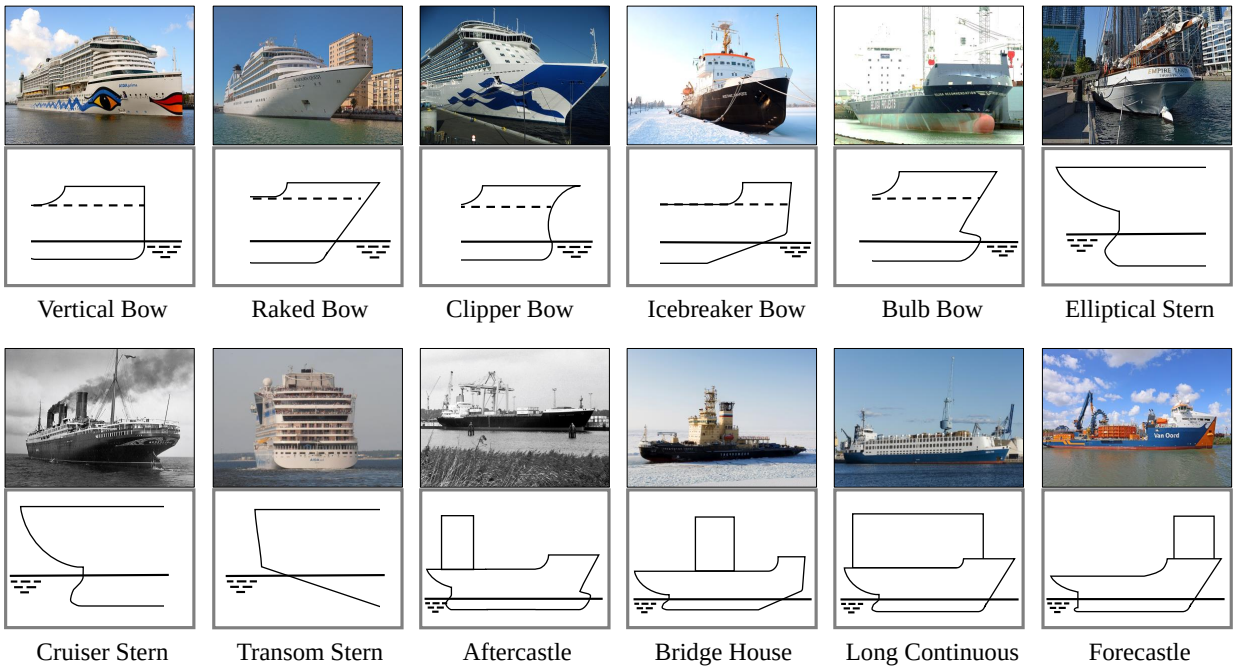


Fig. 3. Images and sketches of representative types of ship parts, including 5 types of bow, 3 types of stern, and 4 types of superstructure.

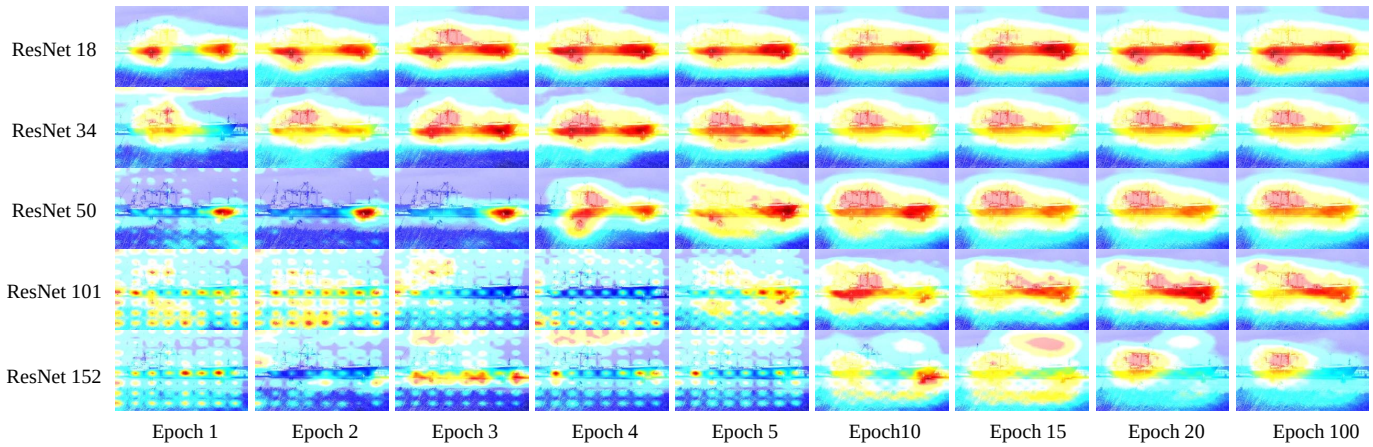


Fig. 4. CAMs for the same image (the one corresponding to 'Forecastle' in Fig. 3) generated by various models during the training process. Each row corresponds to a specific model.

2) Feature Analysis of Ships: Based on the Ship-ID dataset, we are able to train ship recognition models with the dataset and conduct an in-depth feature analysis. Specifically, we start with training the deep models with ResNet-based architectures but at different depths (i.e. numbers of layers). Compared to the original version of ResNet networks, we set the dimension of the last fully connected layer to be the same as the number of ship identities in the Ship-ID dataset, which means that we treat each ship as a unique class.

We use class activation map (CAM) to visualize ship features through the heatmaps of class-wise activation which have the same size as the input image. In a CAM, higher values indicate locations more important for the specific class. We employ Full-Grad CAM [80], which considers the gradient with respect to not only the input but also the bias term, to generate heatmaps with values ranging from 0 to 1. In Fig. 4, we show the CAMs of deep models with various depths during the training stage. It gives the evolution of the feature maps as the training progresses, from which we can obviously find that the recognition models gradually learn to focus their attention on key parts of ships. Such a finding is consistent with previous research [81], [82] that indicated the varying importance of different regions for object recognition. Fig. 5 shows that models with smaller sizes (18, 34, 50 layers) learn faster, so that they are capable of locating key parts of ships after only 5 epochs of training. It can be seen that the performance of the network varies from one model to another, but is similar overall in terms of attention to key regions.

Moreover, the CAMs in Fig. 4 have some local similarities, which suggests the existence of important areas in ship images critical for perception. Accordingly, we attempt to explore such areas commonly present in different ship images captured at

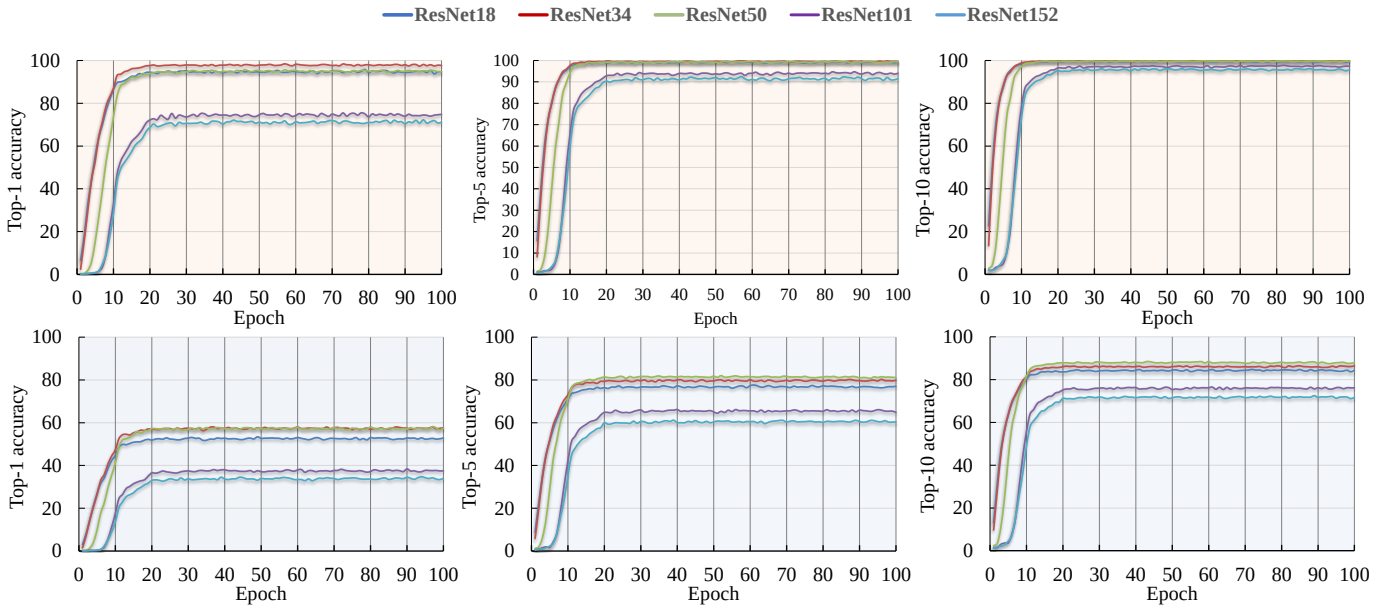


Fig. 5. Accuracy curves on the training (top row) and the testing (bottom row) sets during the training process. Each model is depicted with a distinct colour.

different viewpoints. Specifically, we defined the mean correct CAM for a ship image as an average of the CAMs generated with checkpoint models that predict the correct identity of the input ship image. Then, we extracted 3 sub-CAMs from each CAM according to the activation values. Fig. 6 shows the CAMs and sub-CAMs of 4 ships, each of which has 4 different views. In Fig. 6, the sub-CAMs gradually pay attention to the most important areas from left to right. The activation values for ship profiles are generally below 0.8, while the values higher than 0.8 mainly occur in some specific areas such as the bow, stern, and superstructure of the ships. The values higher than 0.95 tend to be located at the junctions of such areas. In addition, Fig. 7 shows the visualised results of the CAM images of ViT [79] which replaces the ResNet encoder in this test. We can observe that the attention of the Transformer-based ViT does not concentrate on a specific region but spreads over the image. In comparison, the CNN-based ResNet highlights locally the ship region.

By summarising the locations of the important areas highlighted in CAMs, we find that they show consistency with the physical structure of ships. In other words, the highlighted areas tend to appear in the parts that reflect the structural characteristics of the ships. Based on this finding, we define 17 landmarks for each ship based on the multi-view CAMs (as shown in Fig. 6) to represent its structure. Besides, considering that a ship with a particular structure produces a specific waterline (i.e. the line where the ship meets the water), we define 3 additional landmarks on the waterline, one at the bow and two at the stern, which fits the ship structure of sharp bow and flat stern. As a result, each ship in SLAD is annotated with 20 landmarks and Fig. 8 shows some examples. A detailed description of ship landmarks is shown in Table II. In addition to representing key characteristics of the ship, these landmarks can also be used for segmenting the functional areas of the ship, which is supposed to facilitate ship perception in many application scenarios.

C. Data Annotation and Statistics

1) Data Annotation: Once the ship landmarks are defined, we are able to annotate the collected ship images with the coordinates of landmarks. In total, there are three types of annotations in SLAD. As shown in Fig. 9, they include image-level categorical labels (only for images containing a single ship), instance-level bounding boxes, and coordinates of ship landmarks. As shown in Table III, when compared with the existing publicly available ship image datasets, SLAD stands out with its unique ship landmarks annotation.

The category annotations are derived from the category tags obtained when the images were downloaded from the web. For the bounding box and the landmark annotations, we use the format in the MS COCO dataset. Each bounding box is represented by a 4-dimensional vector [x, y, w, h], where the first two elements indicate the coordinates of the top-left corner of the bounding box, and the last two denote its width and height respectively. As for the landmark annotations, we utilize a 60-dimensional vector to record the statistics of 20 ship landmarks:

$$[x_1, y_1, \alpha_1, \cdots, x_{20}, y_{20}, \alpha_{20}], \tag{1}$$

where each x_i and y_i denote the coordinates of the i-th landmark. $\alpha_i \in \{0, 1, 2\}$ indicates the status of the i-th landmark. In specific, $\alpha = 0$ represents that the i-th landmark is not annotated and occluded, so that x_i and y_i are both set to 0 in that condition. $\alpha = 1$ indicates that the i-th landmark is annotated but occluded. $\alpha = 2$ represents that the i-th landmark is annotated and visible in the ship image.

2) Data Statistics: SLAD includes 12,199 images of 949 ships captured under different external conditions. Regarding the geographical attributes, the ships in the dataset belong to 88

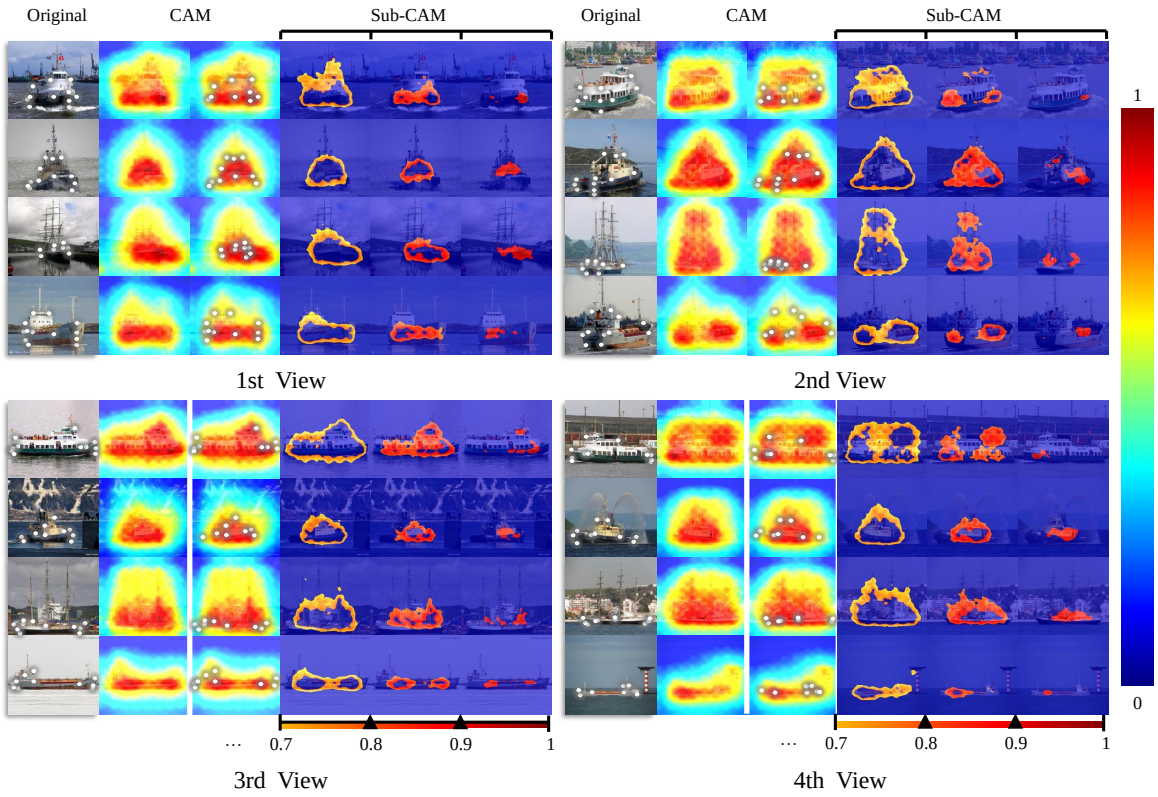


Fig. 6. Visualisation of the mean correct CAMs for 4 ships in 4 different views. Each quarter section shows the key areas of 4 different ships in the same view. To better visualize the distribution of the key areas of ships, 3 sub-CAMs are extracted from the mean correct CAM according to the activation values.

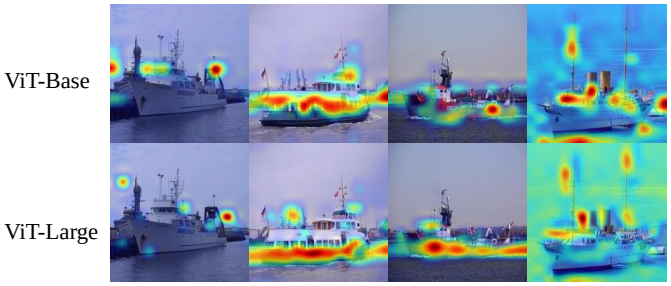


Fig. 7. Visualisation of the mean correct CAMs for 4 ships using ViT [79].

different countries or regions and have been captured in over 900 ports around the world. In terms of category, there are 121 different ship categories in our dataset, including cargo ship, passenger ship, barge ship, etc. This means that on average, there are only about 8 ships within each category, ensuring a diverse range of ship categories in the dataset.

Moreover, most of these images contain only one ship, while the remaining images contain multiple ships. For ease of use, we divide the dataset according to the ratio of $7/1/2$, resulting in $8,530$ images in the training set, $1,209$ images in the validation set, and $2,458$ images in the testing set.

IV. Ship Landmark Detection Benchmark

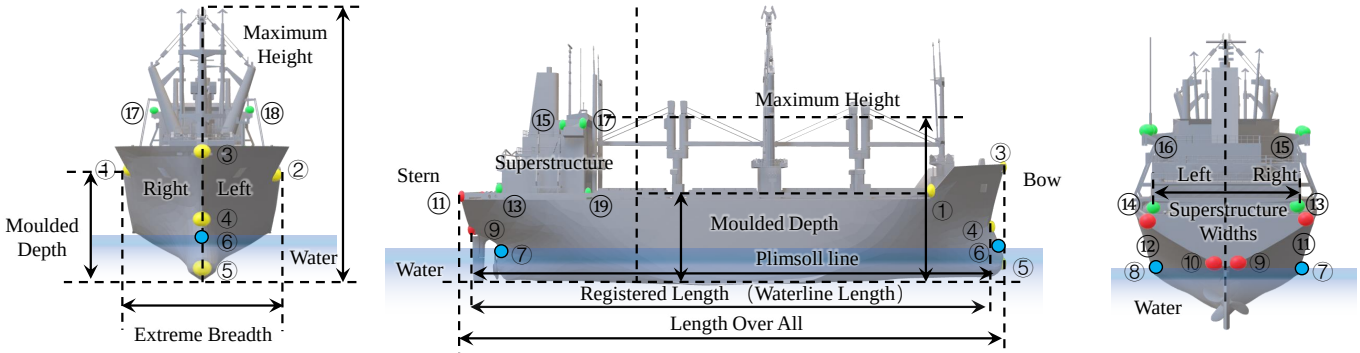
To boost the use of SLAD, we evaluate several state-of-the-art bottom-up landmark detectors on it.

A. Evaluation Metrics

Following the works in landmark-based human pose estimation, two widely used evaluation metrics, Average Precision (AP) and Average Recall (AR), are employed to estimate the performance of landmark detectors.

Precision measures how well the detected landmarks align with the ground truth or the expected locations of the landmarks. A high precision represents that the detected landmarks closely match the true landmarks, while a low precision indicates a larger deviation or error in the detection. Recall measures the completeness or the proportion of true landmarks successfully detected. A high recall score demonstrates that the model can accurately identify a large portion of the true landmarks, while a low recall means that some landmarks are missed or not detected.

In the case of landmark detection, the similarity between the detected and the ground-truth landmarks is usually estimated via Object Keypoint Similarity (OKS). If OKS is greater than a threshold, the sample is regarded as positive, and vice versa. As such, we can calculate AP and AR separately at various thresholds. Also, according to the size of the bounding box, we classify the ships in the SLAD into three types, small, medium, and large ships. We report AP and AR scores with different OKS thresholds and ships of different sizes. AP is the average of AP scores at 10 thresholds $(0.5, 0.55, \cdots, 0.9, 0.95)$ for all ships. AP50 and AP75 denote the AP scores with the thresholds of 0.5 and 0.75 respectively. AP_M and AP_L denote the AP scores for large and medium ships respectively. The



(a) Front view (b) Right View (c) Back View

Fig. 8. Overview of the ship landmarks defined in this paper. The subfigures illustrate the distribution of ship landmarks from three different perspectives. Yellow, red, green, and blue points denote landmarks in the bow, the stern, the superstructure respectively. The three blue points indicate the landmarks on the waterline. The numbered circles represent the point labels.

TABLE II
NAMES AND DESCRIPTIONS OF THE 20 SHIP LANDMARKS.

No.	Name	Description
1	Right Side Forecastle Deck	From stern to bow, start of right side foredeck
2	Left Side Forecastle Deck	From stern to bow, start of left side foredeck
3	Prow	The forward-most part of a ship's bow above the waterline.
4	Draft Mark Of Bow/Load Line Mark of Bow	Markings on ship's bow indicating draught, often with distinctive color.
5	Vice Bow Peak	For bows with two bulbs in part, the lower bump is the most forward.
6	Bow Waterline	The point at which the bow meets the sea surface closest to the ship's mid-profile.
7	Right Stern Waterline	The final intersection of the deck on the right side of the stern with the ocean water.
8	Left Stern Waterline	The final intersection of the deck on the left side of the stern with the ocean water.
9	Load Line Mark of Right Stern	Intersection of right stern load line with stern's end.
10	Load Line Mark of Left Stern	Intersection of left stern load line with stern's end.
11	Right Stern Peak	Uppermost vertex of the rearmost part of the stern with the right-hand freeboard deck.
12	Left Stern Peak	Uppermost vertex of the rearmost part of the stern with the left-hand freeboard deck.
13	Right Deck House Bottom Back	Lower right rear vertex of the largest continuous building in the superstructure.
14	Left Deck House Bottom Back	Lower left rear vertex of the largest continuous building in the superstructure
15	Right Deck House Top Back	Upper right rear vertex of the largest continuous building in the superstructure
16	Left Deck House Top Back	Upper left rear vertex of the largest continuous building in the superstructure.
17	Right Deck House Top Front	Upper right front vertex of the largest continuous building in the superstructure.
18	Left Deck House Top Front	Upper left front vertex of the largest continuous building in the superstructure.
19	Right Deck House Bottom Front	Lower right front vertex of the largest continuous building in the superstructure.
20	Left Deck House Bottom Front	Lower left front vertex of the largest continuous building in the superstructure.

(a). Image-level category (b). Instance-level bounding box (c). Landmarks annotations

Fig. 9. Examples of the three types of annotations in SLAD.

definitions of AR, AR50, AR75, AR_M, and AR_L scores are similar to those of the AP series.

B. Analysis of Results

In this paper, we build the benchmark by evaluating 6 landmark detection methods on SLAD, including CPM [59], SimpleBaseline [67], HRNet [58], HRNet+UDP [83], HRFormer [76], and SimCC [84].

- CPM [59], namely Convolutional Pose Machines, uses cascaded convolutional neural networks to progressively extract and refine landmark information of the human body.
- SimpleBaseline [67] provides a simple but effective baseline method for landmark detection. It employs a backbone network along with a few deconvolutional layers to predict heatmaps for the landmarks.
- HRNet [58] applies a parallel multi-scale fusion strategy to integrate multi-resolution representations for higher accuracy and robustness in landmark detection.
- HRNet+UDP [83] enhances HRNet with Unbiased Data Processing (UDP), and combined classification and regression encoding-decoding.
- HRFormer [76] is a high-resolution transformer architecture that utilizes the feature fusion strategy in HRNet to capture multi-scale information and employs self-attention modules to model long-range dependencies.
- SimCC [84] performs landmark detection via two coordinate classification branches, which predict the horizontal and vertical coordinates of the landmarks separately.

TABLE III
COMPARISON OF PUBLICLY AVAILABLE SHIP IMAGE DATASETS.

Name	Category	Bbox	Auxiliary	Images
ImageNet 2012 [40]	✓	✗	✗	16,114
PASCAL VOC 2012 [41]	✓	✓	✗	353
MARVEL [32]	✓	✗	✗	400,000
SeaShip [34]	✓	✓	✗	7000
Boat Re-ID [35]	✓	✓	✗	5,523
ABOships [37]	✓	✓	✗	9,880
VesselReID [38]	✓	✗	✗	30,589
SmartShip-HEU [39]	✓	✓	✗	12,300
MS COCO [42]	✓	✓	✓	3,164
SLAD	✓	✓	✓	12,197

In our experiments, we take ResNet-50 as the backbone network for both SimpleBaseline and SimCC unless otherwise specified. For HRNet, we report the performance of its two variants, i.e. HRNet-W32 and HRNet-W48, where the suffix indicates the width (or number of channels) of high-resolution sub-networks in the last three stages. Similarly, HRNet+UDP [83] also have two variants, which take HRNet-W32 and HRNet-W48 as backbone networks respectively. For HRFormer, here is the report on the base version of the implementation, the details of which can be found in [76]. Besides, it is worth noting that all the methods are conducted under two different settings of input size, i.e. 256×192 and 384. For training the above methods on SLAD, we apply Adam optimizer with a learning rate of 5e-4, betas (0.9, 0.999), and weight decay 0.01. The batch size is set to 32 for most of the experiments except for the HRFormer when the input size is 384. We reduce the batch size to 12 for that scenario due to higher computation complexity. All the experiments are performed on a Ubuntu18.04 system with 4 Nvidia Tesla V100 GPUs, the methods implementation above is based on MMPose [91].

Table IV lists the evaluation results, from which we obtain the following observations:

(1) CPM produces the worst performance in ship landmark detection, with the AP score being at least 19.14% lower than other methods. This is mainly because CPM is the only two-stage method among the referred methods in the table. This kind of method performs feature extraction and landmark estimation in two separate stages, leading to the extracted features from the backbone network potentially not being entirely suitable for the requirements of the landmark detection task.

(2) High-resolution input can improve the overall performance of ship landmark detection. For all the competing methods, increasing the input size from 256×192 to 384 leads to performance gains in terms of both precision and recall. This is not surprising as high-resolution input often provides more fine-grained details of ships, which contribute to extracting high-quality representations for landmark localization.

(3) SimpleBaseline and SimCC are not as effective as other end-to-end methods in ship landmark detection. Such results demonstrate that taking HRNet as a backbone is a better choice for ship landmark detection, due mainly to its ability in fusing multi-scale representations. Moreover, using UDP for data transformation can further boost the performance of ship landmark detection. For example, when the input size is 256×192, adding UDP improves the AP and the AR scores of HRNet-32 by 2.15% and 2.40% respectively. We also conducted comparative experiments of the baselines with transformers-based backbones (i.e. Swin-Tiny, Swin-Base and Swin-Large [90], HRFormer-Small and HRFormer-Large [76]) and listed the results in Table V. Since the task of landmark detection relies mainly on the localization capability of the network while Transformers have a strong ability of capturing global information, directly employing the normal transformer backbones does not yield promising results.

We also explore how the width and the depth of the backbone network affects ship landmark detection. By comparing the results of HRNet-W32 and HRNet-W48, we find that the wider the backbone is, the better the landmark detection performance. In Table V, we show additional experimental results of SimpleBaseline with various backbone networks. It is obvious that the deeper the backbone is, the higher the evaluation scores are. Besides, the ResNeXt series produce better results than the ResNet series. In conclusion, ship landmark detectors benefit a lot from a strong backbone network that can extract discriminative ship features.

To further estimate the generalization ability of the landmark detector trained on SLAD, we test the trained HRNet-W48 model on 4 public ship datasets. As the ship images in the datasets are not annotated with landmark information, we only present the qualitative results in Fig. 10. It can be seen that the HRNet-W48 model trained on SLAD can effectively detect the landmarks of unseen ships, which demonstrates strong generalization ability.

V. APPLICATIONS OF SHIP LANDMARKS

In this section, we showcase some applications of ship landmarks to several typical ship perception tasks, including ship recognition, ship image generation, key area detection for ships, and ship detection, where RGB images of horizontal views are taken as input. It is noteworthy that some methods using other types of data for ship perception have also shown promising results, such as SAR-based ship classification [92], [93], detection [94]–[96], and segmentation [97], [98].

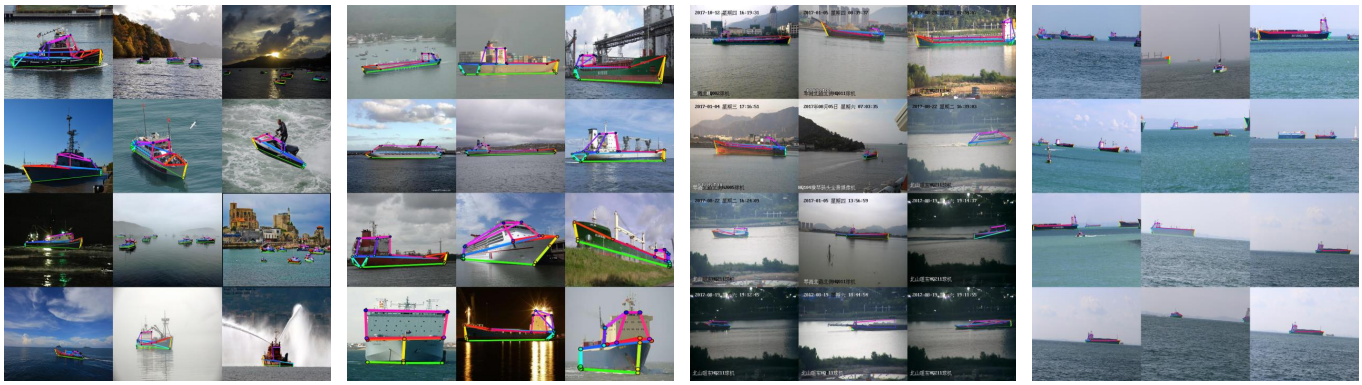
A. Ship Recognition

To explore how the landmarks of a ship affect its recognition, we propose to involve the locations of landmarks as extra information to guide the deep model to focus on informative areas. As shown in Fig. 11, we perform landmark-aware ship recognition by highlighting a small region around each landmark in the image. Specifically, we build a landmark image of the same size as the original ship image, where the RGB values are set to $[255, 255, 255]$ in a circular region with a radius of 20 pixels around each ship landmark, and the remaining pixels are set to $[0, 0, 0]$. Then, we generate the landmark-aware image as the weighted sum of the landmark image and the original image, expressed as

$$I_{ka} = I_{ori} + \beta I_{mark}, \tag{2}$$

TABLE IV
AVERAGE PRECISION (AP) AND AVERAGE RECALL (AR) OF VARIOUS LANDMARK DETECTION METHODS ON SLAD.

Methods	Input size	AP	AP50	AP75	AP_M	AP_L	AR	AR50	AR75	AR_M	AR_L
CPM [59]	256×192	0.2925	0.7738	0.1415	0.1033	0.3020	0.4051	0.8388	0.3396	0.1735	0.4162
	384×288	0.3807	0.8485	0.2825	0.1196	0.3923	0.4886	0.8824	0.4751	0.1735	0.5033
SimpleBaseline [67]	256×192	0.5033	0.9136	0.4896	0.1859	0.5166	0.5934	0.9221	0.6262	0.2653	0.6086
	384×288	0.5721	0.9145	0.5991	0.1879	0.5888	0.6558	0.9276	0.7033	0.2633	0.6743
HRNet-W32 [58]	256×192	0.5710	0.9340	0.6180	0.2258	0.5844	0.6565	0.9424	0.7204	0.3020	0.6732
	384×288	0.6218	0.9474	0.6842	0.2040	0.6417	0.7042	0.9579	0.7664	0.3102	0.7230
HRNet-W48 [58]	256×192	0.5982	0.9463	0.6594	0.2315	0.6125	0.6810	0.9533	0.7516	0.3163	0.6979
	384×288	0.6422	0.9458	0.7081	0.2277	0.6602	0.7217	0.9533	0.7897	0.3000	0.7416
HRNet-W32+UDP [58], [83]	256×192	0.5926	0.9324	0.6468	0.2410	0.6094	0.6805	0.9486	0.7438	0.3286	0.6976
	384×288	0.6406	0.9443	0.7253	0.2445	0.6572	0.7185	0.9517	0.7921	0.3224	0.7373
HRNet-W48+UDP [58], [83]	256×192	0.6139	0.9355	0.6757	0.2386	0.6299	0.6932	0.9463	0.7593	0.3245	0.7110
	384×288	0.6557	0.9467	0.7368	0.2540	0.6719	0.7304	0.9525	0.8022	0.3388	0.7491
HRFormer [76]	256×192	0.6070	0.9457	0.6779	0.2212	0.6216	0.6896	0.9541	0.7601	0.3020	0.7079
	384×288	0.6689	0.9579	0.7576	0.2620	0.6865	0.7422	0.9618	0.8162	0.3286	0.7616
SimCC [84]	256×192	0.5284	0.9131	0.5294	0.1952	0.5418	0.6164	0.9213	0.6589	0.2653	0.6331
	384×288	0.5818	0.9240	0.6193	0.2095	0.5988	0.6663	0.9377	0.7212	0.2918	0.6842



(a) MS COCO (b) Marvel 2016 (c) SesShips (d) SMD

Fig. 10. Generalization evaluation of HRNet-W48 trained on SLAD. This figure shows the qualitative results on 4 public datasets, including (a) MS COCO [42], (b) Marvel2016 [32], (c) Seaships [34], and (d) Singapore Maritime Dataset (SMD) [33].

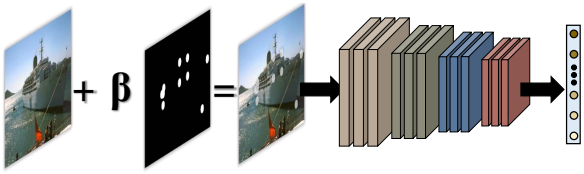


Fig. 11. Landmark-guided ship recognition. We replace the original input of the recognition model with a synthetic landmark-aware image where the regions around the landmarks are highlighted.

where I_{ori} and I_{mark} denote the original ship image and the landmark image respectively. To estimate the value of β, we start with training the network using the original images, and then test the trained network using the synthesized I_{ka} generated with different β. The experimental results indicate that a larger β value leads to greater differences in distribution between the new synthesized data and the source data, especially when $\beta > 0.2$. We additionally train the same network using the synthesized datasets generated with different β and then test the networks, and find that the trained network achieves the highest prediction accuracy when $\beta = 0.2$. Therefore, we set $\beta = 0.2$ in our experiments. Except for replacing the original input image with the landmark-aware one, the other settings for training the ship recognition model remain the same.

Fig. 12 shows ship recognition losses and accuracy with and without the landmark information. From the curves in Fig. 12, we find that it makes no difference when we use the landmark information to train a network with less than or equal to 50 layers. However, it significantly helps larger networks, such as ResNet101 and ResNet152, to reduce the recognition losses and improve the recognition accuracy. This is mainly because it is difficult to fully train large models with limited data, while the landmark information can provide additional constraints for training, thereby helping to improve the performance of ship recognition.

TABLE V
AVERAGE PRECISION (AP) AND AVERAGE RECALL (AR) OF SIMPLEBASELINE [67] WITH DIFFERENT BACKBONE NETWORKS ON SLAD.

Methods	Input size	AP	AP50	AP75	AP_M	AP_L	AR	AR50	AR75	AR_M	AR_L
SimpleBaseline	256×192	0.5033	0.9136	0.4896	0.1859	0.5166	0.5934	0.9221	0.6262	0.2653	0.6086
+ ResNet-50 [85]	384×288	0.5721	0.9145	0.5991	0.1879	0.5888	0.6558	0.9276	0.7033	0.2633	0.6743
SimpleBaseline	256×192	0.5088	0.9021	0.5032	0.1782	0.5222	0.5968	0.9151	0.6316	0.2612	0.6127
+ResNet-101 [85]	384×288	0.5883	0.9252	0.6319	0.2215	0.6044	0.6697	0.9369	0.7212	0.3041	0.6871
SimpleBaseline	256×192	0.5277	0.9240	0.5546	0.1747	0.5415	0.6160	0.9338	0.6706	0.2633	0.6326
+ResNet-152 [85]	384×288	0.6003	0.9370	0.6659	0.2088	0.6172	0.6802	0.9431	0.7477	0.2939	0.6985
SimpleBaseline	256×192	0.5130	0.9005	0.5188	0.1592	0.5280	0.6015	0.9151	0.6449	0.2306	0.6187
+ResNeXt-50 [86]	384×288	0.5800	0.9143	0.6278	0.2109	0.5975	0.6593	0.9245	0.7173	0.2612	0.6783
SimpleBaseline	256×192	0.5355	0.9122	0.5566	0.2102	0.5486	0.6180	0.9283	0.6690	0.2857	0.6336
+ResNeXt-101 [86]	384×288	0.6045	0.9260	0.6609	0.1943	0.6234	0.6833	0.9377	0.7438	0.2551	0.7033
SimpleBaseline	256×192	0.5477	0.9126	0.5690	0.1800	0.5633	0.6342	0.9283	0.6807	0.2571	0.6524
+ResNeXt-152 [86]	384×288	0.6083	0.9275	0.6699	0.2167	0.6270	0.6874	0.9354	0.7469	0.2694	0.7074
SimpleBaseline	256×192	0.4351	0.8721	0.3847	0.1467	0.4471	0.5255	0.8925	0.5327	0.2204	0.5397
+ MobileNet V2 [87]	384×288	0.5152	0.8993	0.5378	0.1724	0.5289	0.5988	0.9104	0.6511	0.2510	0.6150
SimpleBaseline	256×192	0.4104	0.8274	0.3518	0.1062	0.4247	0.4878	0.8528	0.4751	0.1571	0.5035
+ ShuffleNet V2 [88]	384×288	0.4657	0.8755	0.4479	0.1618	0.4785	0.5471	0.8933	0.5670	0.2327	0.5621
SimpleBaseline	256×192	0.5453	0.9116	0.5766	0.1797	0.5597	0.6268	0.9276	0.6838	0.2837	0.6437
+ VGG16 [89]	384×288	0.6140	0.9242	0.6762	0.1730	0.6333	0.6830	0.9330	0.7461	0.2327	0.7041
SimpleBaseline	256×192	0.4210	0.8576	0.3536	0.1364	0.4331	0.5153	0.8886	0.5164	0.2245	0.5287
+ Swin-Tiny [90]	384×288	0.5254	0.9114	0.5363	0.2310	0.5388	0.6097	0.9276	0.6526	0.3000	0.6243
SimpleBaseline	256×192	0.2528	0.6755	0.1315	0.0867	0.2601	0.3428	0.7492	0.2757	0.1224	0.3530
+ Swin-Base [90]	384×288	0.5261	0.9121	0.5401	0.1875	0.5413	0.6066	0.9213	0.6519	0.2551	0.6229
SimpleBaseline	256×192	0.4210	0.8576	0.3536	0.1364	0.4331	0.5153	0.8886	0.5164	0.2245	0.5287
+ Swin-Large [90]	384×288	0.3029	0.7248	0.2130	0.0775	0.3131	0.3776	0.7687	0.3364	0.1082	0.3899
SimpleBaseline	256×192	0.5705	0.9318	0.6187	0.2597	0.5850	0.6583	0.9494	0.7220	0.3612	0.6726
+HRFormer-Small [76]	384×288	0.6144	0.9464	0.6772	0.2391	0.6300	0.6978	0.9509	0.7609	0.3306	0.7154
SimpleBaseline	256×192	0.6070	0.9457	0.6779	0.2212	0.6216	0.6896	0.9541	0.7601	0.3020	0.7079
+HRFormer-Base [76]	384×288	0.6689	0.9579	0.7576	0.2620	0.6865	0.7422	0.9618	0.8162	0.3286	0.7616

B. Ship Image Generation

Ship landmarks can be seen as a simplified representation of ships, which is often sparse compared to visual textures. We believe that by learning the mapping between these two kinds of representations, it is possible to generate ship images using ship landmarks.

With the ship landmark annotations provided by SLAD, we can establish the structures of ships in the style of "ship skeleton". As shown in Fig. 13, the skeleton of a ship contains not only the coordinates of the ship landmarks but also the links among these landmarks. We apply the pix2pix [99] method based on Generative Adversarial Network (GAN) to achieve structure-to-texture generation. After the training with a small amount of data (less than $10,000$ structure-texture image pairs), the GAN-based model is able to map the structural input to its corresponding texture image. In Fig. 13(a), the top row shows the skeletons of different ships while their texture outputs are shown in the bottom row. It can be seen that the texture image of a ship can be well produced based on the ship skeleton. This indicates that the ship skeleton contains highly discriminative visual cues that can be used to convey important ship information.

Furthermore, we built the 3D skeleton of each ship in accordance with the physical symmetry prevalent in ships. Fig. 13(b) shows that by feeding the ship skeletons of different views into the trained structure-to-texture model, we can obtain the texture of images captured from different viewpoints of the same ship. Note that although such ship skeletons of different views are more likely to be out-of-distribution ones unseen during training, the GAN-based model still generates high-quality images of ships.

C. Key Area Detection for Ships

In terms of definition, a ship skeleton formed by landmarks is similar to the keypoint-based pose of the human body. While as a kind of the rigid object, ships exhibit relatively stable geometric structures compared to humans, providing a stronger sense of direction. With this knowledge, we are able to achieve key area detection on the basis of ship skeleton. Due to our careful consideration of the structural characteristics of ships when selecting ship landmarks, the skeleton of a ship naturally divides itself into several areas as illustrated in Fig. 14(a), including the deck area, bow area, stern area, and superstructure area. As a result, pixels on the

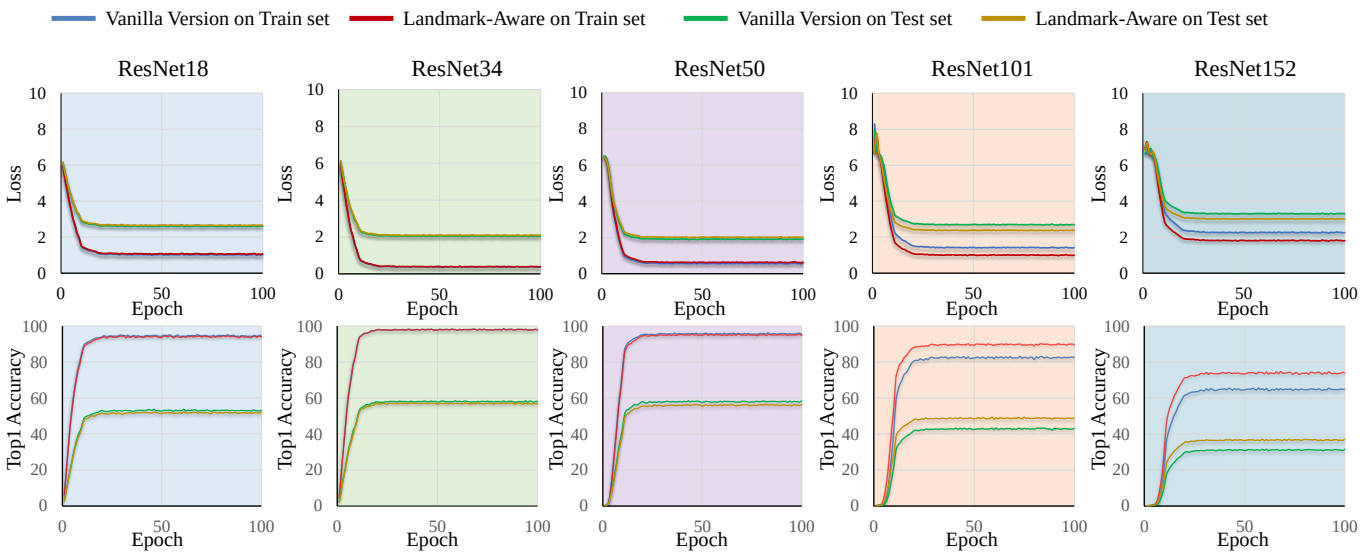
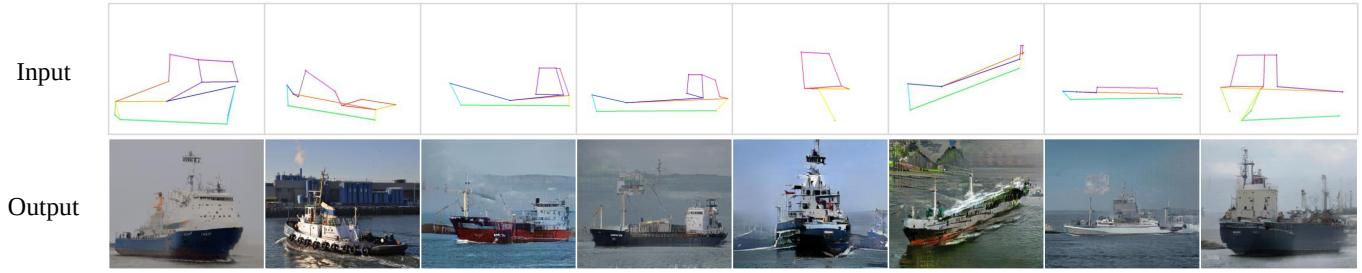


Fig. 12. Effect of landmark information on the performance of ship recognition. This figure shows the loss curves and average Top-1 performance of various methods with and without landmark information across 10 duplicate experiments.



(a) Ship Image Generation

(b) Multi-view Ship Image Generation

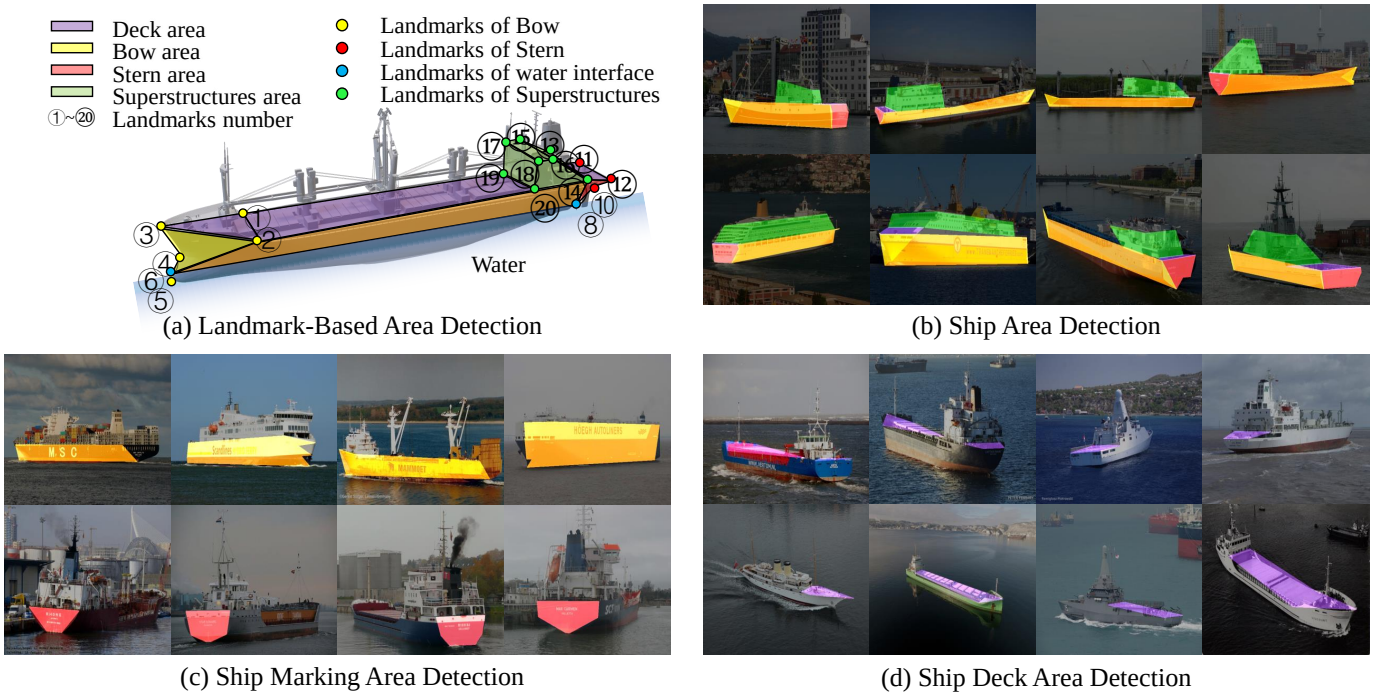
Fig. 13. Results for landmark-based ship image generation. (a) Ship image generation for different ships with corresponding ship skeletons. (b) Multi-view ship image generation for a single ship by rotating the 3D skeleton.

ship's surface are assigned with corresponding semantic labels, representing which specific region they belong to. We present some qualitative results of key area detection for ships in Fig. 14(b).

Key area detection provides insights into the detailed perception of specific areas of ships, which is of great significance in the scenario of ship management. For example, according to international maritime regulations, ships are required to display markings such as names, registries, and IMO numbers in visible areas on their hull. Thus a detailed perception of such areas aids in identifying different ships. As shown in Fig. 14(c), the parts containing the marking information can be easily segmented, which significantly reduces the difficulty of automatic ship identification. Simultaneously, this can also be used to determine whether a ship has displayed the required identification information in accordance with regulations. Furthermore, there are strict regulations regarding the cargo areas of a ship. As shown in Fig. 14(d), key area detection has the potential to enable intelligent ship monitoring systems to locate compliant cargo areas and make preliminary assessments of the cargo contents, which helps improve the safety of maritime transportation.

D. Ship Detection

Ship detection is one of the fundamental components of maritime surveillance, which is crucial for ensuring the safety



(a) Landmark-Based Area Detection

(b) Ship Area Detection

(c) Ship Marking Area Detection

(d) Ship Deck Area Detection

Fig. 14. Landmark-based key area detection for ships. (a) The schematic diagram of key area detection based on landmarks. (b) Examples of key area detection on real ship images. (c) detection of key areas with identification markings. (d) detection of key areas for cargo.

of navigation at sea. Since the ship landmarks defined in this paper are semantically meaningful points on the ship body, the spatial relationships among these landmarks remain fixed, allowing us to obtain the concepts related to ship orientation, such as stern or bow and left side or right side. Therefore, given the coordinates of ship landmarks and combing the characteristics of symmetry in the ship structure, it is possible to achieve 3D detection of ship objects using only a 2D ship image as input. Considering that most ships are symmetrically designed, most ship landmarks come in left-right pairs, allowing us to enclose the main hull with a hollow cuboid.

In Fig. 15, we present some examples of 3D object detection for 2D ship images. The 3D bounding boxes are drawn based on the extreme values of the landmarks' coordinates, combined with the symmetry of the ship structure. In particular, to showcase the spatial layout of the ships, we display the lines of the 3D bounding box in different colors, where green, red, and blue represent the right side, left side, and all the other lines, respectively. It can be observed that the 3D bounding boxes in the figure accurately locate the main body of the ships, indicating that the ship landmarks work well in facilitating monocular 3D ship detection.

VI. Conclusion

This paper introduces ship landmarks, which have not been explored in previous works in the fields of computer vision and marine engineering. The most distinct feature of this paper is that we present SLAD, the first ship dataset with landmark annotations. Specifically, we discover a strong correlation between high-value regions on the deep feature maps and the physical structure of ships, which enables us to define 20 landmarks to annotate the ships in maritime images for creating SLAD. It is composed of $12,199$ ship images captured under different external conditions. We evaluate several landmark detectors on SLAD to provide a benchmark for ship landmark detection in the hope that it can promote the development of this unexplored field. Moreover, we show the applications of SLAD in different fields of ship perception, including ship recognition, ship image generation, key area detection for ships, and ship detection.

However, the focus of our work is currently limited to the annotation of ship landmarks in RGB images that are captured in horizontal view. In the future, we shall extend it to other types of image data such as SAR images.

References

[1] L. M. Millefiori, P. Braca, D. Zissis, G. Spiliopoulos, S. Marano, P. K. Willett, and S. Carniel, "Covid-19 impact on global maritime mobility," *Scientific Reports*, vol. 11, no. 1, pp. 1–16, 2021.

[2] M. Lin, Z. Zhang, Y. Pang, H. Lin, and Q. Ji, "Underactuated USV path following mechanism based on the cascade method," *Scientific Reports*, vol. 12, no. 1, pp. 1–14, 2022.

[3] Y. Wu, X. Chu, L. Deng, J. Lei, W. He, G. Królczyk, and Z. Li, "A new multi-sensor fusion approach for integrated ship motion perception in inland waterways," *Measurement*, vol. 200, p. 111630, 2022.

[4] Z. Sui, Y. Wen, Y. Huang, C. Zhou, L. Du, and M. A. Piera, "Node importance evaluation in marine traffic situation complex network for intelligent maritime supervision," *Ocean Engineering*, vol. 247, p. 110742, 2022.

[5] J. Lin, P. Diekmann, C.-E. Framing, R. Zweigel, and D. Abel, "Maritime environment perception based on deep learning," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[6] M. Zhang, J. Feng, K. T. Ma, J. H. Lim, Q. Zhao, and G. Kreiman, "Finding any waldo with zero-shot invariant and efficient visual search," *Nature Communications*, vol. 9, no. 1, pp. 1–15, 2018.

[7] Y. Xu and M. Vaziri-Pashkam, "Limits to visual representational correspondence between convolutional neural networks and the human brain," *Nature Communications*, vol. 12, no. 1, pp. 1–16, 2021.

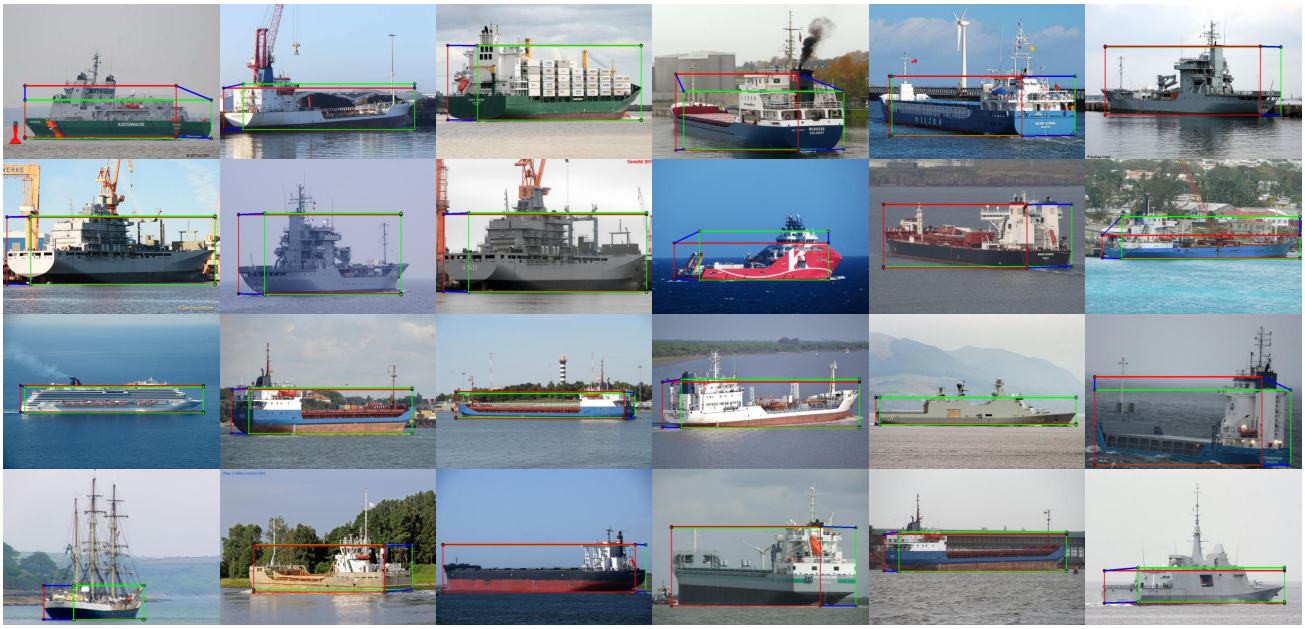


Fig. 15. Examples of landmark-based ship detection. Based on the landmarks, the bounding boxes provide information about the orientation of ships.

[8] Y. Wang, L. Zhang, R. Song, H. Li, P. L. Rosin, and W. Zhang, "Exploiting inter-sample affinity for knowability-aware universal domain adaptation," *International Journal of Computer Vision*, vol. 132, no. 5, pp. 1800–1816, 2024.

[9] N. Mei, R. Santana, and D. Soto, "Informative neural representations of unseen contents during higher-order processing in human brains and deep artificial networks," *Nature Human Behaviour*, vol. 6, no. 5, pp. 720–731, 2022.

[10] S. Madan, T. Henry, J. Dozier, H. Ho, N. Bhandari, T. Sasaki, F. Durand, H. Pfister, and X. Boix, "When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 146–153, 2022.

[11] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *CVPR*, 2019, pp. 8546–8556.

[12] R. Zhao, J. Wang, X. Zheng, J. Wen, L. Rao, and J. Zhao, "Maritime visible image classification based on double transfer method," *IEEE Access*, vol. 8, pp. 166 335–166 346, 2020.

[13] H. Xue, X. Chen, R. Zhang, P. Wu, X. Li, and Y. Liu, "Deep learning-based maritime environment segmentation for unmanned surface vehicles using superpixel algorithms," *Journal of Marine Science and Engineering*, vol. 9, no. 12, p. 1329, 2021.

[14] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 1, pp. 173–176, 2010.

[15] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *BIGSARDATA*, 2017, pp. 1–6.

[16] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.

[17] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *ICPRAM*, vol. 2, 2017, pp. 324–331.

[18] L. Zhao, P. Tang, and L. Huo, "Feature significance-based multibag-of-visual-words model for remote sensing image scene classification," *Journal of Applied Remote Sensing*, vol. 10, no. 3, p. 035004, 2016.

[19] L. Huang, B. Liu, B. Li, W. Guo, W. Yu, Z. Zhang, and W. Yu, "OpenSARShip: A dataset dedicated to sentinel-1 ship interpretation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 1, pp. 195–208, 2017.

[20] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[21] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *CVPR*, 2018, pp. 3974–3983.

[22] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

[23] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sensing*, vol. 11, no. 7, p. 765, 2019.

[24] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120 234–120 254, 2020.

[25] T. Zhang, X. Zhang, X. Ke, X. Zhan, J. Shi, S. Wei, D. Pan, J. Li, H. Su, Y. Zhou *et al.*, "LS-SSDD-v1.0: A deep learning dataset dedicated to small ship detection from large-scale sentinel-1 SAR images," *Remote Sensing*, vol. 12, no. 18, p. 2997, 2020.

[26] Airbus, "Airbus ship detection challenge," 2019. [Online]. Available: https://www.kaggle.com/c/airbus-ship-detection

[27] A.-J. Gallego, A. Pertusa, and P. Gil, "Automatic ship classification from optical aerial images with convolutional neural networks," *Remote Sensing*, vol. 10, no. 4, p. 511, 2018.

[28] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord, "xView: Objects in context in overhead imagery," *arXiv preprint arXiv:1802.07856*, 2018.

[29] K. Chen, M. Wu, J. Liu, and C. Zhang, "FGSD: A dataset for fine-grained ship detection in high resolution satellite images," *arXiv preprint arXiv:2003.06832*, 2020.

[30] Z. Zhang, L. Zhang, Y. Wang, P. Feng, and R. He, "ShipRSImageNet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8458–8472, 2021.

[31] J. Chen, K. Chen, H. Chen, Z. Zou, and Z. Shi, "A degraded reconstruction enhancement-based method for tiny ship detection in remote sensing images with a new large-scale dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[32] E. Gundogdu, B. Solmaz, V. Yücesoy, and A. Koc, "MARVEL: A large-scale image dataset for maritime vessels," in *ACCV*, 2017, pp. 165–180.

[33] 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 Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 1993–2016, 2017.

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

[35] P. Spagnolo, F. Filieri, C. Distante, P. L. Mazzeo, and P. D'Ambrosio,

"A new annotated dataset for boat detection and re-identification," in *AVSS*, 2019, pp. 1–7.
[36] Y. Zheng and S. Zhang, "McShips: A large-scale ship dataset for detection and fine-grained categorization in the wild," in *ICME*, 2020, pp. 1–6.
[37] B. Iancu, V. Soloviev, L. Zelioli, and J. Lilius, "ABOShips-an inshore and offshore maritime vessel detection dataset with precise annotations," *Remote Sensing*, vol. 13, no. 5, p. 988, 2021.
[38] Q. Zhang, M. Zhang, J. Liu, X. He, R. Song, and W. Zhang, "Unsupervised maritime vessel re-identification with multi-level contrastive learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 5, pp. 5406–5418, 2023.
[39] L. Su, Y. Chen, H. Song, and W. Li, "A survey of maritime vision datasets," *Multimedia Tools and Applications*, pp. 28 873–28 893, 2023.
[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[41] M. Everingham, S. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.
[42] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014, pp. 740–755.
[43] T. Zhang, X. Zhang, J. Li, X. Xu, B. Wang, X. Zhan, Y. Xu, X. Ke, T. Zeng, H. Su *et al.*, "Sar ship detection dataset (ssdd): Official release and comprehensive data analysis," *Remote Sensing*, vol. 13, no. 18, p. 3690, 2021.
[44] P. Kaur, A. Aziz, D. Jain, H. Patel, J. Hirokawa, L. Townsend, C. Reimers, and F. Hua, "Sea situational awareness (SeaSAw) dataset," in *CVPR*, 2022, pp. 2579–2587.
[45] C. Zheng, M. Mendieta, and C. Chen, "POSTER: A pyramid cross-fusion transformer network for facial expression recognition," in *ICCV*, 2023, pp. 3146–3155.
[46] Y. Xia, C. Nduka, R. Yap Kannan, E. Pescarini, J. Enrique Berner, and H. Yu, "AFLFP: A database with annotated facial landmarks for facial palsy," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 4, pp. 1975–1985, 2023.
[47] Z. Li, X. Gong, R. Song, P. Duan, J. Liu, and W. Zhang, "SMAM: Self and mutual adaptive matching for skeleton-based few-shot action recognition," *IEEE Transactions on Image Processing*, vol. 32, pp. 392–402, 2022.
[48] J. Liu, X. Wang, C. Wang, Y. Gao, and M. Liu, "Temporal decoupling graph convolutional network for skeleton-based gesture recognition," *IEEE Transactions on Multimedia*, 2023.
[49] Z. Li, Y. Zhong, R. Song, T. Li, L. Ma, and W. Zhang, "Detal: Open-vocabulary temporal action localization with decoupled networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
[50] T.-H. Tsai, C.-C. Huang, and K.-L. Zhang, "Design of hand gesture recognition system for human-computer interaction," *Multimedia Tools and Applications*, vol. 79, pp. 5989–6007, 2020.
[51] L. Ke, S. Li, Y. Sun, Y.-W. Tai, and C.-K. Tang, "GSNet: Joint vehicle pose and shape reconstruction with geometrical and scene-aware supervision," in *ECCV*, 2020, pp. 515–532.
[52] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, "PoseTrack: A benchmark for human pose estimation and tracking," in *CVPR*, 2018, pp. 5167–5176.
[53] E. Vendrow, D. T. Le, J. Cai, and H. Rezatofighi, "JRDB-Pose: A large-scale dataset for multi-person pose estimation and tracking," in *CVPR*, 2023, pp. 4811–4820.
[54] J. Cao, H. Tang, H.-S. Fang, X. Shen, C. Lu, and Y.-W. Tai, "Cross-domain adaptation for animal pose estimation," in *ICCV*, 2019, pp. 9498–9507.
[55] 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 *ICCV*, 2017, pp. 379–387.
[56] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *ECCV*, 2016, pp. 869–884.
[57] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *ECCV*, 2016, pp. 34–50.
[58] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019, pp. 5693–5703.
[59] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016, pp. 4724–4732.
[60] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *CVPR*, 2018, pp. 7103–7112.
[61] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017, pp. 2334–2343.
[62] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017, pp. 2961–2969.
[63] S. Huang, M. Gong, and D. Tao, "A coarse-fine network for keypoint localization," in *ICCV*, 2017, pp. 3028–3037.
[64] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "CrowdPose: Efficient crowded scenes pose estimation and a new benchmark," in *CVPR*, 2019, pp. 10 863–10 872.
[65] W. Liu, J. Chen, C. Li, C. Qian, X. Chu, and X. Hu, "A cascaded inception of inception network with attention modulated feature fusion for human pose estimation," in *AAAI*, 2018.
[66] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *CVPR*, 2017, pp. 4903–4911.
[67] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *ECCV*, 2018, pp. 466–481.
[68] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017, pp. 7291–7299.
[69] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, "ArtTrack: Articulated multi-person tracking in the wild," in *CVPR*, 2017, pp. 6457–6465.
[70] U. Iqbal, A. Milan, and J. Gall, "PoseTrack: Joint multi-person pose estimation and tracking," in *CVPR*, 2017, pp. 2011–2020.
[71] S. Jin, W. Liu, W. Ouyang, and C. Qian, "Multi-person articulated tracking with spatial and temporal embeddings," in *CVPR*, 2019, pp. 5664–5673.
[72] S. Jin, X. Ma, Z. Han, Y. Wu, W. Yang, W. Liu, C. Qian, and W. Ouyang, "Towards multi-person pose tracking: Bottom-up and top-down methods," in *ICCV PoseTrack Workshop*, vol. 2, no. 3, 2017, p. 7.
[73] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," *NeurIPS*, vol. 30, 2017.
[74] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *ECCV*, 2018, pp. 269–286.
[75] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
[76] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "HRFormer: High-resolution vision transformer for dense predict," *NeurIPS*, vol. 34, 2021.
[77] D. G. Watson, *Practical Ship Design*. Elsevier, 1998, vol. 1.
[78] A. F. Molland, *The Maritime Engineering Reference Book: A Guide to Ship Design, Construction and Operation*. Elsevier, 2011.
[79] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[80] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," *NeurIPS*, vol. 32, 2019.
[81] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
[82] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, neurIPS.
[83] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The devil is in the details: Delving into unbiased data processing for human pose estimation," in *CVPR*, 2020, pp. 5700–5709.
[84] Y. Li, S. Yang, P. Liu, S. Zhang, Y. Wang, Z. Wang, W. Yang, and S.-T. Xia, "SimCC: A simple coordinate classification perspective for human pose estimation," in *ECCV*, 2022, pp. 89–106.
[85] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[86] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[87] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
[88] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.
[89] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[90] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.
[91] M. Contributors, "OpenMMLab Pose Estimation Toolbox and Benchmark," https://github.com/open-mmlab/mmpose, 2020.
[92] T. Zhang, X. Zhang, X. Ke, C. Liu, X. Xu, X. Zhan, C. Wang, I. Ahmad, Y. Zhou, D. Pan *et al.*, "Hog-shipclsnet: A novel deep learning network with hog feature fusion for sar ship classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–22, 2021.
[93] T. Zhang and X. Zhang, "A polarization fusion network with geometric feature embedding for sar ship classification," *Pattern Recognition*, vol. 123, p. 108365, 2022.
[94] X. Xu, X. Zhang, and T. Zhang, "Lite-yolov5: A lightweight deep learning detector for on-board ship detection in large-scene sentinel-1 sar images," *Remote Sensing*, vol. 14, no. 4, p. 1018, 2022.
[95] X. Xu, X. Zhang, Z. Shao, J. Shi, S. Wei, T. Zhang, and T. Zeng, "A group-wise feature enhancement-and-fusion network with dual-polarization feature enrichment for sar ship detection," *Remote Sensing*, vol. 14, no. 20, p. 5276, 2022.
[96] T. Zhang, X. Zhang, C. Liu, J. Shi, S. Wei, I. Ahmad, X. Zhan, Y. Zhou, D. Pan, J. Li *et al.*, "Balance learning for ship detection from synthetic aperture radar remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 182, pp. 190–207, 2021.
[97] T. Zhang and X. Zhang, "A mask attention interaction and scale enhancement network for sar ship instance segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
[98] ——, "Htc+ for sar ship instance segmentation," *Remote Sensing*, vol. 14, no. 10, p. 2395, 2022.
[99] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

Mingxin Zhang is currently pursuing the Ph.D. degree in Pattern Recognition and Intelligent Systems from School of Control Science and Engineering, Shandong University, Jinan, China.

His research interests include computer vision, robot navigation and deep learning.

Qian Zhang received the B.Eng. degree from Northeastern University at Qinhuangdao, Qinhuangdao, China, in 2018, and the M.S. degree from Shandong University, Jinan, China, in 2021. She is currently pursuing the Ph.D. degree in Pattern Recognition and Intelligent Systems.

Her research interests include computer vision and machine learning.

Ran Song received the B.Eng. degree in telecommunications engineering from Shandong University, China, in 2005, and the Ph.D. degree in computer vision from the University of York, UK, in 2009.

He is currently a professor with the School of Control Science and Engineering, Shandong University. His research interests include 3D shape analysis and 3D visual perception. He has published more than 100 papers in peer-reviewed journals and international conferences.

Paul L. Rosin is a professor with the School of Computer Science and Informatics, Cardiff University, UK. Previously, he was a lecturer at Brunel University London, U.K., research scientist at the Institute for Remote Sensing Applications, Joint Research Centre, Ispra, Italy, and lecturer at Curtin University of Technology, Perth, Australia.

His research interests include image representation, semantic segmentation, low level image processing, machine vision approaches to remote sensing, methods for evaluation of approximation algorithms, medical and biological image analysis, mesh processing, non-photorealistic rendering, and analysis of shape in art and architecture.

Wei Zhang received the Ph.D. degree in electronic engineering from the Chinese University of Hong Kong in 2010.

He is currently a professor with the School of Control Science and Engineering, Shandong University, Jinan, China. His research interests include computer vision and robotics. Prof. Zhang has served as a program committee member and a reviewer for various international conferences and journals.