Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2807139, 10 pages
https://doi.org/10.1155/2022/2807139

Research Article

FGSR: A Fine-Grained Ship Retrieval Dataset and Method in Smart Cities

Yunting Xian, Jin Xian, Lu Lu, and Ji Tang

School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

Correspondence should be addressed to Jin Xian; xi88@scut.edu.cn

Received 3 April 2022; Revised 24 April 2022; Accepted 26 April 2022; Published 30 May 2022

Academic Editor: Mu En Wu

Ship reidentification is an important part of water transportation systems in smart cities. Existing ship reidentification methods lack a large-scale fine-grained ship retrieval dataset in the wild and existing ship recognition solutions mainly focus on the ship target identification rather than the fine-grained ship reidentification. Furthermore, previous ship target identification systems are usually based on synthetic aperture radar (SAR) image, automatic identification system (AIS) data, or video streaming, which is confronted with expensive deployment costs, such as the installation cost of SAR and AIS, and the communication and storage overhead. Indeed, ship reidentification benefits for traffic monitoring, navigation safety, vessel tracking, etc. To address these problems, we propose a new large-scale fine-grained ship retrieval dataset (named FGSR) that consists of 30,000 images of 1000 ships captured in the wild. Besides, to tackle the difficulty of spatial-temporal inconsistency in ship identification in the wild, we design a multioriented ship reidentification network named FGSR-Net that consists of three modules to address different crucial problems. The pyramid fusion module was aimed at addressing the problem of variant size and shape of ship targets, the occlusion modules attempt to detect the unchangeable area of ship images, while the multibranch identity module generates discriminative feature representation for ship targets from different orientations. Experimental evaluations on FGSR dataset show the effectiveness and efficiency of our proposed FGSR-Net. The mean average precision of ship reidentification is around 92.4%, and our FGSR-Net proposed method only takes 3 seconds to give the retrieval results from 30,000 images.

1. Introduction

The intelligent water transportation system (IWTS) is to monitor and manage ships that sail on water [1]. The ship recognition that recognizes ships target sailing or parking on the water is the key component of IWTS in smart cities. In contrast to traditional intelligent road transportation system (IRTS) [2], IWTS is confronted with (1) various types of ships, (2) high deployment cost, and (3) complex settings. Thus, ship recognition is more challenging than vehicle recognition. To achieve ship recognition, a series of literature [3–6] have been investigated and explored.

In terms of the technology used in ship identification, it can be roughly divided into two categories: (1) ship recognition assisted by specialized equipment [3], such as Synthetic Aperture Radar (SAR), Automatic Identification System (AIS), and (2) ship recognition through common equipment [4], such as the camera. In general, the former can cover a wider scope and is suitable for marine or coastal settings, but is faced with a high deployment cost. Meanwhile, the latter does not require ships to install any extra equipment, which is expense-friendly for ships. Besides, some small vessels do not install AIS to save money or turn off AIS to avoid monitor. Furthermore, SAR is effective to capture images of large ships but is unable to capture images of small vessels. Thus, recent ship recognition systems [4, 5, 7, 8] employ cameras to capture images of ships and further adopts the computer vision technique [9, 10] to implement ship recognition.

Although ship recognition systems based on the computer vision technique have an advantage in deployment cost and small target recognition, they still provide coarse-grained services. In other words, prior ship recognition systems [4–8] only support ship target detection and classification, which fails to satisfy the increasing requirements of IWTS, such as tracking

a given ship. Ship tracking provides crucial on-site microscopic kinematic traffic information which benefits traffic flow analysis, ship safety enhancement, traffic control, etc. [11, 12].

Given a monitored person image, person reidentification (Re-ID) was aimed at retrieving the location of the specific person in images or videos given the query image. It can be treated as a subproblem of image retrieval [13] using computer vision technology to determine whether there is a specific person in an image or video sequence, which is regarded as a subproblem of image retrieval [13]. Inspired by person Re-ID [13] and vehicle Re-ID [14], we adopt the Re-ID technology to recognize ships and retrieve a given ship. Different from person reidentification or vehicle reidentification, which has been studied thoroughly before, ship reidentification (ship Re-ID) is still underexplored. The ship Re-ID is quite different from the person Re-ID and vehicle Re-ID. The differences can be concluded into three-folds: (1) The shape difference: different from the coherent shape of the human body and vehicle, the shape of different ships can be quite different, for example, the shape of different types of ships also varies greatly. (2) The size difference: in person Re-ID, human bodies are always about the same size, which can enable the input image to have same resolution. However, the sizes of the different ships can be largely different from each other. For example, the sizes of the canoe and steamship are quite different. This phenomenon makes the designed network learn difficultly with different resolution image inputs. (3) The appearance change: different from the human always on the same clothes captured by the cameras, the appearance of the ship can be changed in a short time. For instance, the facilities can be removed from their original position in a short space of time.

Considering the above-mentioned problem, we propose a Fine-Grained Ship Retrieval NETwork (FGSR-Net) to address the ship Re-ID problem. Specifically, to address the problem of difference in shape and size, the proposed FGSR-Net employs a pyramid structure to extract the feature of the input image, which can handle the different resolution inputs. With regard to the problem of appearance change, we designed an occlusion attention mechanism to produce an occlusion map to represent the region in which the two images remain consistent. With the help of the occlusion map, we can compute the similarity of the areas that have not changed, while we can compare the areas that have changed. Combining these two similarities, we can retrieve the corresponding ship was given a ship image.

To facilitate the research of the ship Re-ID, we also propose a new large-scale fine-grained ship retrieval dataset (named FGSR) that consists of 30,000 images of 1000 ships captured in the wild. We set up high-definition cameras on the banks of the river to capture images of ships passing by and sent out a communication request to the ships to ask if we can use their ship's images for research purposes only.

In a word, we have three contributions:

(i) We construct a new large-scale fine-grained ship retrieval dataset (named FGSR) that consists of 30,000 images of 1,000 ships captured in the wild for the intelligent water transportation system in smart cities

(ii) We propose a novel fine-grained ship reidentification network (FGSR-Net) to address the ship Re-ID problem. FGSR-Net adopts a pyramid structure to address the shape and size problem, while it also employs an occlusion attention module to tackle the problem of appearance change

(iii) Our extensive experiments show the outperformance compared with existing state-of-the-art methods from relevant fields

2. Related Work

The related work of this paper consists of ship target recognition and target reidentification. Thus, we introduce the existing schemes of ship target recognition and target reidentification.

2.1. Ship Target Recognition. Ship target recognition usually adopts SAR image, AIS data, or video streaming. Karabayır et al. [5] pointed out the importance of a training library and proposed a ship target recognition scheme based on k-nearest-neighbour (KNN) classifier. Chaturvedi et al. [3] integrated SAR image and AIS data to identify ships at an area, which required ships to install radar and AIS equipment. However, some illegal ships deliberately shut off the radar or AIS systems. Nowadays, researchers tried to utilize video streaming and deep learning to recognize ships. To overcome the influences of ship background, object occlusion, variations of weather, and light conditions for target ship recognition, Zhao et al. [4] proposed a two-stage neural network (DCNet) to detect and recognize ships from video streaming, where one neural network was used to detect ships and the other one was used to recognize ships. Cao et al. [6] also adopted two neural networks to recognize ships, where a convolutional neural network (CNN) was used to extract the ship image features and KNN-SVM was utilized to train to recognize ships. In contrast to DCNet, Fu et al. [7] realized the ship target recognition based on the single-stage neural network, faster regions with CNN (F-RCNN), which only required a single-stage to recognize a target. After that, Fu et al. [15] carefully improved the F-RCNN to improve the accuracy of detection. Specifically, in [15], they extract the feature of target using ResNet [16] and optimize the F-RCNN [17] using batch normalization layer. Considering a complex marine environment, Zou et al. [8] combined the hard example mining technology and F-RCNN and replaced VGG16 with the ResNet just like the work in [15]. Cao et al. [6] adopted the single-stage target recognition network, YOLO, to recognize ships and analyze the ship behaviours.

Different from these methods for ship target recognition, which only extracts the location information of ships from the input images, our method can differentiate different ships by reidentification.

2.2. Ship Reidentification. Ship reidentification is pretty similar to person reidentification (Re-ID). Geng et al. [13] considered the person Re-ID task a classification/identification task and a verification task and adopted classification/identification loss and verification loss to train a classification subnet and a verification subnet. Varior et al. [18] used the Siamese network to implement the person Re-ID by comparing two similarities of the two photos. Intuitively, it is easier to produce discriminative representations by combining the global feature and local feature of a person. Wang et al. [19] proposed, a multibranch deep network architecture, Multiple Granularity Network (MGN), which consists of a global branch and two local branches. MGN was regarded as the state-of-the- art method from Person Re-Id. Huang et al. [14] applied person Re-ID technology to vehicle reidentification and designed a deep feature fusion with multiple granularity (DFFMG) method. DFFMG consists of one branch for global feature representations, two for vertical local feature representations, and the other two horizontal ones.

In this paper, we extend the person Re-ID technology to ship reidentification and carefully design a novel multiple granularity networks. Different from existing Re-ID methods with prior knowledge on human, our method explicitly considers the characteristic (shape, size, and appearance) of ships in our model design.

3. Our Proposed Dataset and Method

3.1. FGSR: A New Large-Scale Fine-Grained Ship Retrieval Dataset

3.1.1. Camera Selection. In this work, to obtain a high-resolution image of ship targets, we select camera DS-2DY9240IX-A(T5) of Hikvision to capture the image. We report the key parameters of that camera in Table 1.

3.1.2. The Difficulties of Dataset Collection. In the ship Re-ID task, there is no available dataset to interstate this task. A large dataset is difficult to collect for three reasons. (1) The image of a specific ship is private, we can collect the image of ships only with their owner's permission. (2) It is a huge cost to collect ship images. Some of the ships are far from shore, so we often need some expensive high-definition cameras to capture them. (3) Time consuming, we need to set up the cameras in a specific position and take a snapshot of the ship on the water surface.

3.1.3. The Design Scheme of Dataset Collection. In the stage of dataset collection, to reduce labour costs, we design an automatic ship capture system (ASCS). We show the diagram of the ASCS in Figure 1. In ASCS, we deploy two cameras, i.e., global camera and local camera. Specifically, we first utilize the global camera to spy on the large area of the shipping lane. The local camera will focus on the detailed texture of the ship targets detected by the global camera. In this system, we employ the YOLO detector [20] as the ship detection engine in the global camera. In this system, the positions of two cameras (i.e., global camera and local camera) are carefully tuned to ensure the local camera can localize the ship position with the detected position of the ship provided by the global camera. In order to cover the entire waterway, we can set up more than one ASCS. The detailed ship images captured by local cameras are stored in our database for further processing.

3.1.4. The Detail of the Collected Dataset. Due to the lack of datasets for ship retrieval, we deploy 50 ASCS nearby different waterways and collect about one million images of ship targets for further usage. The captured images are in 1920×1080 resolution. We then manually crop the image regions of ships and collect them into our FGSR dataset. We obtain the identity information of each ship by correlating the AIS information with the ship image. After further processing of the collected image, our dataset contains 30,000 images of 1,000 ships in the wild. We show several samples in Figure 2. Our dataset contains various ship images across different light conditions and spatial locations. Besides, in our dataset, the visual appearance of the same ship will be different when it carries different cargoes. Also, different from existing datasets for vehicle retrieval [14] which contains cars with a unique licence plate in the back, our proposed FGSR has some very similar ships without any explicit prior knowledge such as text or licence plate. This makes our proposed dataset much more challenging and essential for real-world applications.

3.1.5. Experimental Setting. We extract the representations of these images using the proposed FSRN, after which we split these representations into two sets, i.e., query set and gallery set. In our experiment setting, the gallery set contains 23,156 ship targets, while there are 2748 feature vectors in the query target.

3.2. FGSR-Net: A Multioriented Ship Reidentification Network. With the challenges we mentioned above, we proposed a fine-grained ship reidentification network (FGSR-Net) that contains three modules to address these problems (i.e., the problem of variant size and shape, the area of ship image is changeable). And the last module, multibranch identity module, not only captures the global information of ship targets but also recognizes their details of them in both the horizontal and vertical directions.

3.3. Overview. In this work, we propose a novel fine-grained ship reidentity network (FGSR-Net) to address the problem of ship Re-ID. In our proposed FGSR-Net, the three-module consist of the main contribution of this paper. (1) In the pyramid fusion module, as the size and the shape of different ships are quite different, to enable the model handle the different resolution inputs, we utilize the feature extraction convolutional network to extract pyramid feature maps to represent the different levels of spatial information. After that, we employ a pyramid fusion module to aggregate these different size feature maps to obtain a semantically strong representation. (2) We also propose an occlusion module to predict an occlusion map to indicate the area that can be changed in different time slots. Then, we can mainly compare the unchangeable area of the ship. (3) Combine the physical

Table 1: The partial parameter of our adapted camera.

Camera	Image sensor	1/1.8 progressive scan CMOS
	Shutter speed	1/1 s ~1/30,000 s
	Focus	Auto; semiauto; manual
	WDR	140 dB WDR
	Digital zoom	16x
	Optical zoom	40x
Lens	Focal length	6.0 mm to 240 mm, 40 × optical
	Zoom speed	Approx.5.6 s
	Aperture	F1.2
Illuminator	IR distance	Up to 400 m IR distance
PTZ	Movement range (Pan)	360
	Movement range (Tilt)	+40 ~ -90
	Presets	300
	Patrol scan	8 patrols
	Pattern scan	Pattern scans

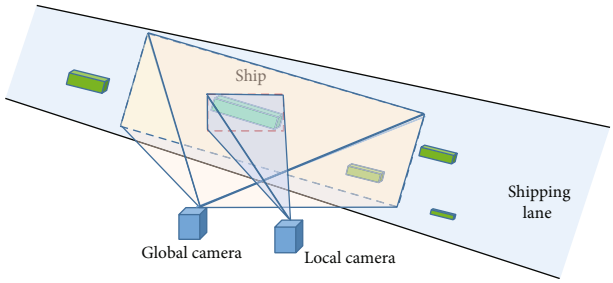

Figure 1: The illustration of our proposed automatic ship capture system.

Figure 2: The samples of our collected dataset.

shape of the ship, we employ a multibranch identity module to capture the texture information in different orientations, i.e., a global branch to capture the whole information of the ship target, a vertical branch, and a horizontal branch to recognize the details of ship targets from the horizontal and vertical perspectives, respectively.

3.4. Pyramid Fusion Module. As we mentioned above, the size and ship difference of ship targets are the main obstacles of ship Re-ID. Inspired by FPN [21], we can also employ a similar structure to address the different size and shape problems.

As shown in Figure 3, the feature extraction convolutional layers firstly extract a pyramid feature maps $\{P_i^0\}_{i=1}^N$, where N is the number of feature maps. To aggregate the different scare level features, we proposed a pyramid fusion module as shown in Figure 4. The pyramid fusion module consists of T stage, and each stage is derived.

From the former stage, while the first stage is the original feature maps $\{P_i^0\}_{i=1}^N$. We can formulate it as

$$P_i^t = \mathrm{UP}\left(P_{i+1}^t\right) + f_i^t\left(P_i^{t-1}\right), t = 1, \cdots, T\text{-}1, \quad (1)$$

$$F = \mathrm{AVG}\left(P_0^O, P_0^1, \cdots, P_0^{T-1}\right), \quad (2)$$

where the UP $(\cdot)$ means the upsample operator, while the f_i^t is a 1×1 convolutional layer to reduce channel dimensions.

With a coarser-resolution feature map P_{i+1}^t, we upsample the spatial resolution by a factor of 2 (using nearest neighbour upsampling for simplicity), which results in a feature map that has the same resolution of the P_i^t. The upsampled map is then merged with the corresponding bottom-up map (which undergoes a 1×1 convolutional layer to reduce channel dimensions) by element-wise addition. After that, we combine all last feature maps (i.e., $P_0^O, P_0^1, \cdots, P_0^{T-1}$ in four stages to generate the scale-invariance feature map F.

3.5. Occlusion Module. Different from the person Re-ID problem, the appearance of ship targets can be changed in different time slots. For example, the facilities can be moved in the ship and the cargo carried by ships may vary from time to time. This phenomenon makes ship retrieval become difficult. To tackle this problem, we propose an occlusion module to produce an occlusion map, which aims at identifying the areas of ship targets that are changeable.

Figure 5 shows the architecture of the proposed occlusion module. Our occlusion module consists of two major module: a transformer-based spatial feature extractor to extract long-range occlusion-aware features and a direction-aware attention module to model the spatial correlation of the input feature, from a direction-based perspective. The input feature of the occlusion module is first forwarded to a flatten layer to obtain N patch-wise embeddings, where N is equal to 1024 in our module. We then

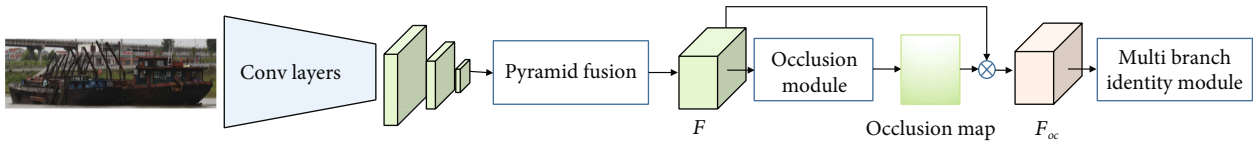

Figure 3: The framework of our fine-grained ship reidentification network. Firstly, we input the ship image into a convolutional network to obtain a pyramid feature map. In pyramid feature maps, all levels are semantically strong, including the high-resolution levels. Then, we apply a pyramid fusion module to aggregate all scare feature maps. To identify the moveable facilities on the ship, we design an occlusion module to estimate an occlusion map which identifies the areas that can be changed in other time slots. After that, we combine the occlusion map and the aggregated feature to produce an occlusive feature. Finally, the multibranch identity module takes the occlusive feature as input to produce the identity feature vector for the corresponding image.

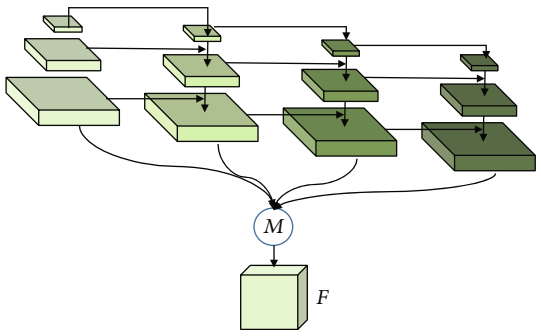

Figure 4: The illustration of proposed pyramid fusion module.

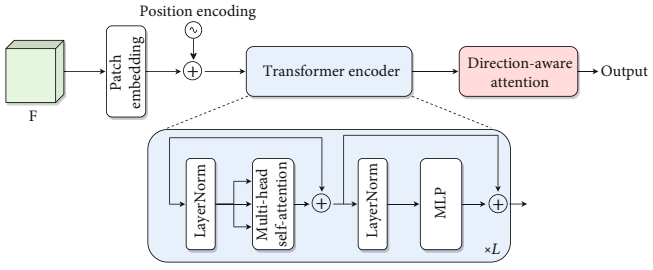

Figure 5: The illustration of the proposed occlusion module.

add a position embedding to each patch n. Different from classic transformer proposed in [22], which adopts a fixed positional embedding by two sine and cosine functions to model the relative positional information in frequency, our position embeddings are learnable parameters for a more flexible position encoding. The details of patch embeddings can be represented in

$$\mathrm{PatchEmbedding}(\mathrm{F}) = [\mathrm{F}_1, \mathrm{F}_2, \cdots, \mathrm{F}_\mathrm{N}] \in \mathrm{R}^{\mathrm{N}\times\mathrm{p}\times\mathrm{p}\times\mathrm{C}}, \tag{3}$$

where p and C are the size of the patch and the number of channels of the input feature F.

After obtaining the embedded patches, we feed all the patches to a transformer encoder, which consists of L transformer layers sequential with LayerNorm, multihead self-attention and multilayer perception (MLP). The details of the proposed transformer layer are shown as follows:

$$\mathrm{F}^\mathrm{m} = \mathrm{MSA}\left(\mathrm{LayerNorm}\left(\mathrm{F}^\mathrm{i}\right)\right) + \mathrm{F}^\mathrm{i}, \tag{4}$$

$$\mathrm{F}^\mathrm{o} = \mathrm{MLP}(\mathrm{LayerNorm}(\mathrm{F}^\mathrm{m})) + \mathrm{F}^\mathrm{m}, \tag{5}$$

where F^i, F^m, F^o, and MSA denote the input feature, the middle interim features, the output feature of the transformer encoder, and multihead self-attention in a single encoder layer of the transformer encoder, respectively. In our design, we sequentially stack L layers to form the entire transformer encoder, where L is set to 12 in our module.

After the long-range occlusion-aware features extracted, they are inputted into a direction-aware attention. As shown in Figure 6, we aim to produce a spatial occlusion map to indicate the changeable area of ship targets in the occlusion module. In the very beginning, we first utilize a 1×1 convolutional layer to reduce the dimensions of the input feature map:

$$\bar{\mathrm{F}} = \mathrm{f}_{\mathrm{smooth}}(\mathrm{F}), \tag{6}$$

where f_{smooth} is the 1×1 convolutional layer. In this way, we can obtain a feature map with dimensions as $1 \times H \times W$.

In our practice, the ship is photographed as a rectangular object. Thus, we can utilize the strip pooling technique [23] to average all the feature values in a row or a column as shown in Figure 6. Thus, the output of the horizontal strip pooling and vertical strip pooling can be written as

$$\begin{aligned} y_i^h &= \frac{1}{W} \sum_{0\le j<W} x_{i,j}, \\ y_i^v &= \frac{1}{H} \sum_{0\le i<H} x_{i,j}. \end{aligned} \tag{7}$$

With the obtained feature vectors, we apply a matrix multiplication to fuse them to produce the occlusion map that indicates the area that can be easy to change in other time slots:

$$\mathrm{M} = \sigma\left(\mathrm{y}^\mathrm{h} \times (\mathrm{y}^\mathrm{v})^\mathrm{T}\right), \tag{8}$$

where σ is a sigmoid function to ensure the result in the range [0, 1].

With the horizontal strip pooling and vertical strip pooling layers, the network can easily investigate the inherent knowledge of the ship target. Thanks to the long and narrow kernel shape, the produced occlusion map can focus on

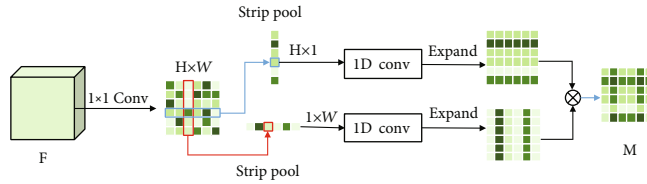

Figure 6: The illustration of the direction-aware attention in the proposed occlusion module.

capturing local details due to its narrow kernel shape along the other dimension.

3.6. Multibranch Identity Module. In the person Re-ID problem, the most important thing is to produce a discriminative representation for each instance. Inspired by MGN [19], we apply a multibranch to produce the representation with respect to different orientations, i.e., horizontal or vertical.

As shown in Figure 3, we mask the feature F with the occlusion map M to produce the occlusive feature map F_{oc}. We then feed the occlusive feature map F_{oc} into our multibranch identity module to obtain the discriminative feature representation.

We illustrate the structure of multibranch identity module in Figure 7. As shown in Figure 7, there are three branches in multibranch identity module, i.e., global branch, horizontal branch, and vertical branch. The global branch was aimed at capturing the global information of the ship targets, while the last two branches attempt to recognize the details of ship targets from the horizontal and vertical perspectives, respectively. Similar to the MGN [19], after obtaining the output of the former occlusion module, we feed it into three different branches, i.e., global branch, horizontal branch, and vertical branch.

Here, we report the settings of these three branches in Table 2.

In the global branch, we first utilize convolution layers to downsample the input feature; then, a global max-pooling layer is applied on the downsampled feature map, while a 1×1 convolution layer is used to reduce the dimension and smooth the output. Finally, a 256-dim global feature vector $\mathbf{f}_g^G$ is produced to represent whole information of the given image.

The second and last branches (horizontal branch and vertical branch) have a similar

structure to the global branch. Specifically, we do not downsample the input features but uniformly split them into several parts in horizontal/vertical orientation to maintain the proper areas of reception fields for local features. Then, we utilize the same following structure to learn the local feature representations as learning global features.

In addition to splitting the feature map in the beginning, we are still downsampling the feature map to obtain a global feature representation for the last two branches (i.e., $\mathbf{f}_h^G$ and $\mathbf{f}_v^G$). Therefore, we can enforce the consistency between the global representations in the three branches.

3.7. Objective Functions. Here, to boost the learning of discriminative feature representation, we mainly utilize the widely used loss functions in the reidentification task to act as our objective functions, i.e., the softmax classification loss L_{softmax} and triplet loss $L_{-\text{triplet}}$.

3.7.1. Softmax Classification Loss. At the very beginning, we classify the feature representations of i-th ship targets into i-th category. Here, we expect that each feature can be classified into the correct category. For i-th learned features $\boldsymbol{f}_i$, L_{softmax} is formulated as

$$L_{soft\ max} = -\sum_{i=1}^{B} \log \frac{e^{W_{y_i}^t f_i}}{\sum_{c=1}^{C} e^{W_c^t f_i}}, \tag{9}$$

where $\boldsymbol{W}_c$ is a weight vector for class c. B and C is the size of minibatch and the number of classes in training, respectively. We employ the softmax classification loss on both the global features (i.e., $\mathbf{f}_g^G$, $\mathbf{f}_h^G$, and $\mathbf{f}_h^G$) and local features (i.e., $\{\mathbf{f}_h^i|_{i=0}^2\}$ and $\{\mathbf{f}_v^i|_{i=0}^2\}$).

3.7.2. Triplet Loss. After obtaining the dimensional reduced global features in three branches, we apply a triplet loss on these three global features to learn a more diverse representation for each individual ship target. This loss function is formulated as follows:

$$L_{\text{triplet}} = -\sum_{i=1}^{P}\sum_{a=1}^{K}\left[\alpha + \max_{p=1\cdots K} \mathbf{f}_a^{(i)} - \mathbf{f}_p^{(i)}{}_2 - \min_{n=1\cdots K, j=1\cdots P, j\neq i} \mathbf{f}_a^{(i)} - \mathbf{f}_n^{(j)}{}_2\right], \tag{10}$$

where $\mathbf{f}_a^{(i)}$, $\mathbf{f}_p^{(i)}$, and $\mathbf{f}_n^{(j)}$ refer to the feature extracted from anchor, positive, and negative samples, respectively. The α is the margin hyperparameter to control the differences of intra and interdistances. Here, the positive sample is the ship targets with the same identify with the anchor, while the negative sample is for different identities. In each minibatch, there are P selected identities and we select K images from each identity to perform triplet loss.

3.7.3. Ship Retrieval. After obtaining the features of those three branches in FSRN, we concatenate them (i.e., $\mathbf{f}_g^G$, $\mathbf{f}_h^G\mathbf{f}_h^G$, $\{\mathbf{f}_h^i|_{i=0}^2\}$, and $\{\mathbf{f}_v^i|_{i=0}^2\}$) into a 2304-dim feature vector as the representation of specific ship target.

Let $G = \{\mathbf{f}_g^i\}_{i=0}^N$ denoted as the feature galley that contains the N features of ship targets in database. Given a ship target query $\boldsymbol{f}_q$, we aim to retrieve the most similar ship target in the galley:

$$\frac{\mathbf{f}_q \times \mathbf{f}_g^i}{\left\|\mathbf{f}_q\right\| \cdot \left\|\mathbf{f}_g^i\right\|}. \tag{11}$$

Here, we use cosine distance to measure the similarity of

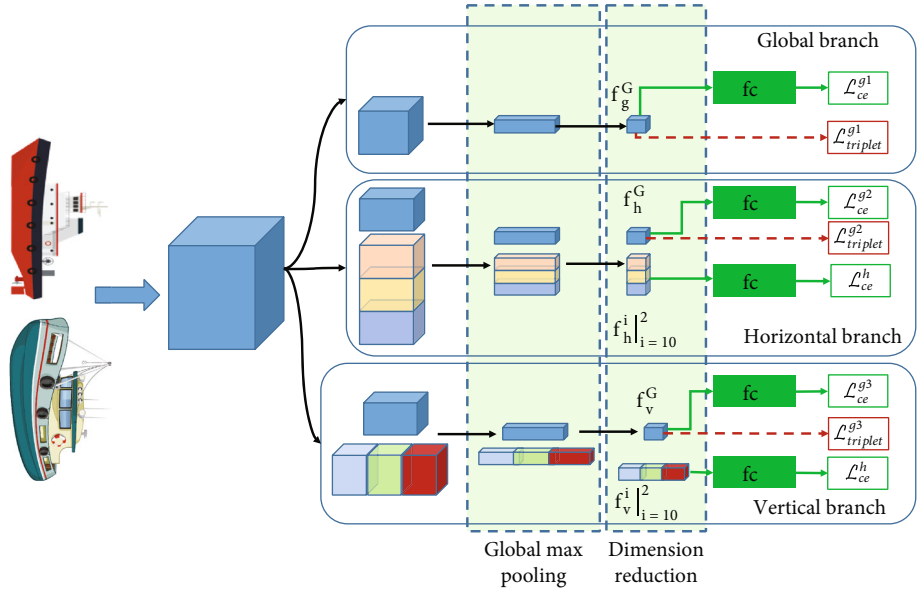

Figure 7: The illustration of proposed multibranch identity module.

two features. Thus, we can obtain the similarity between query and any feature in gallery $S = \{s_i\}_{i=0}^{N}$. Then, we can sort them to get the most similar ships.

3.8. Experimental Evaluations. In this work, we propose a large-scale fine-grained ship retrieval dataset (FGSR dataset) to evaluate our proposed method. In this section, we conduct massive experiments to verify the effectiveness of the proposed method.

3.8.1. Dataset and Metric. In our experiments, we follow previous Re-ID work [19] to report the cumulative matching characteristics (CMC) at rank-1, rank-5, rank-10, rank-20, and rank40 and mean average precision (mAP) on our proposed datasets.

3.8.2. Implementation Details. In this work, we implement our whole framework based on PyTorch. One GeForce RTX 3090 GPU is used to run all the experiments. In our proposed network, the convolutional layers before the pyramid fusion module are borrowed from the RestNet50 [16]. And we extract the output of each block of RestNet50 to form our pyramid feature maps. Their size are $128 \times 128 \times 128$, $256 \times 64 \times 64$, $512 \times 32 \times 32$, and $1024 \times 16 \times 16$. The kernel size of 1×1 convolutional layers in the pyramid fusion module and occlusion module is 3×3. In the training stage, we adopt the SGD optimizer with a learning rate of 0.0004 and a weight decay of 0.0005 and train 300 epochs.

3.9. Compare Methods. In this section, since there is no approach for ship Re-ID, we compare our method with several popular object Re-ID methods, i.e., MGN [19], OSNet [24], VANet [25], and VehicleNet [26].

(i) MGN. Multiple Granularity Network (MGN) contains three branches, a global branch to capture the global information of the human body, a horizontal branch, and a vertical branch to extract the local detail representations

(ii) OSNet. Omniscale network (OSNet) designs a residual block composed of multiple convolutional streams, each detecting features at a certain scale. Also, it introduces a unified aggregation gate mechanism to dynamically fuse multiscale features with input-dependent channel-wise weight

(iii) VANet. Viewpoint-aware network was aimed at learning two metrics for similar viewpoints and different viewpoints in two feature spaces, respectively. The former one (within-space constraint) forces the positive pairs closer than negative pairs in each feature space, while the latter one (cross-space constraint) does the same thing when pairs are in different feature spaces

(iv) VehicleNet. In VehicleNet, they design a simple yet effective two-stage progressive approach to learning more robust visual representation from their proposed dataset

Table 2: Structure of multibranch identity module. Here, our input image is 384×128. "Branch" refers to the name of branches. "Part No." refers to the number of partitions on feature maps. "Map Size" is the size of output feature maps. "Dims" indicates the number of the channel of output. "Feature" means the output of the corresponding branch.

Branch	Part no.	Map size	Dims	Feature
Global	1	12×4	256	$\mathbf{f}_g^G$
Horizontal	3	24×8	$256 * 3 + 256$	$\{\mathbf{f}_h^i \mid_{i=0}^{2}\}, \mathbf{f}_h^G$
Vertical	3	24×8	$256 * 3 + 256$	$\{\mathbf{f}_v^i \mid_{i=0}^{2}\}, \mathbf{f}_v^G$

Table 3: Comparison of our algorithm with other methods on the collected dataset.

Method	Rank-1	Rank-5	Rank-10	Rank-20	Rank-40	mAP
MGN [19]	83.6	85.4	87.5	93.6	96.9	83.5
OSNet [24]	86.3	88.4	93.2	96.7	98.3	85.4
VANet [25]	82.1	86.3	90.4	93.6	96.8	80.4
VehicleNet [26]	85.3	87.9	91.3	95.6	98.3	84.6
Ours	94.3	95.7	98.4	99.5	100	92.4

In this work, we train the above methods using our proposed dataset and report the results to compare them with our proposed methods.

3.10. Comparison to the State-of-the-Art Methods. In this section, we first compare our method with the current Re-ID methods (i.e., MGN [19], OSNet [24], VANet [25], and VehicleNet [26]) on our proposed dataset, and the results are reported in Table 3. We can see that our method obtains the best results on Rank-1 as well as mAP. It is notable that our method achieves 100% in Rank-40.

Compared with the original MGN, our method gains 8.9% in mAP, because we split the feature map into several stripes in two different orientations, while MGN only split feature maps in one orientation. Also, compared with all methods, our method can effectively explore the inherent knowledge of the ship target to obtain the best results. For the shape, these results verify the effectiveness of our proposed method.

Figure 8 shows the retrieval results from our proposed method from a query image. Our method can retrieve the same ship images from multiple views. This demonstrates the effectiveness of our proposed direction-aware modules.

3.11. Ablation Studies

3.11.1. The Effectiveness of Pyramid Fusion Module. To evaluate the effectiveness of the pyramid fusion module, we remove the pyramid fusion module from our whole framework, while the feature extraction convolutional layers only output one level feature map, which will be fed into an occlusion map. The results are reported in Table 4.

We can find that, the variance method without pyramid fusion module ("Ours w/o PFM") underperform our full method ("Ours") for all metrics, e.g., our full method achieves a significant improvement of 3.9% over the "Ours w/o PFM" on Rank-1. These results indicate that the pyramid fusion module plays an important role in our framework. With a pyramid fusion module, our approach can address the variant size and shape problem of ship Re-ID.

3.11.2. The Effectiveness of Occlusion Module. To better identify the identical ship targets, we proposed an occlusion module to estimate an occlusion map that represents the changeable areas of the ship image. By masking this area, the remaining areas are stable. Thus, we can rely on these areas to identify a ship image. We report the results of our method without the occlusion module. As shown in Table 5, the occlusion module can gain stable improvements for our method. These results indicate the effectiveness of our proposed occlusion module.

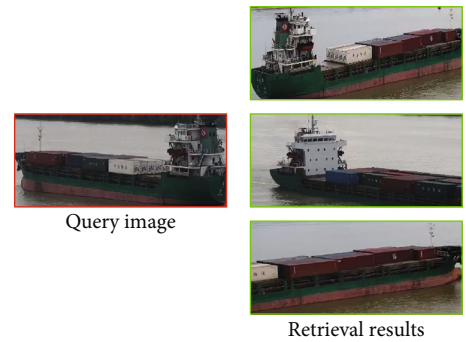

Figure 8: Retrieval results from our proposed method.

Table 4: The ablation study of our proposed method. We denote Trans. as the transformer module in our proposed MBIM.

Method	Rank-1	Rank-5	Rank-10	Rank-20	Rank-40	mAP
Ours w/o PFM	90.4	92.8	94.3	96.2	98.3	90.4
Ours w/o OM	91.8	92.4	95.6	97.2	99.3	91.3
Ours w/o MBIM	89.5	91.2	94.0	95.8	97.5	90.0
Ours w/o trans.	90.3	92.5	94.4	96.6	98.1	90.8
Ours	94.3	95.7	98.4	99.5	100	92.4

3.11.3. The Effectiveness of Multibranch Identity Module. To verify the multibranch identify module in our method, we replace it with 3 convolutional layers and finally output a 1024-dimension feature vector to represent the corresponding ship image. The results are shown in Table 4. We can find that the performance degrades if we remove the multibranch identify module from our full method. For example, the "mAP" becomes 90.0% from 92.4%.

These results show that extracting the representative feature vector from different orientations can better identify the different ships.

3.11.4. The Effectiveness of Each Subfeature. In this section, we report the results of variant combinations of extracted features, i.e., $\mathbf{f}_g^G$, $\mathbf{f}_h^G$, $\mathbf{f}_h^G$, $\{\mathbf{f}_h^i|_{i=0}^2\}$, and $\{\mathbf{f}_v^i|_{i=0}^2\}$. The results are shown in Table 5. We can find that the results using splitted features outperform the results using only global features. For example, the results of column "$\{\mathbf{f}_h^i|_{i=0}^2\}$" and "$\{\mathbf{f}_v^i|_{i=0}^2\}$" outperform the first three columns, i.e., columns "$\mathbf{f}_g^G$", "$\mathbf{f}_h^G$", and "$\mathbf{f}_v^G$". Moreover, we can obtain the best results when we concatenate all global features and all splitted features (as shown in the column "All" in Table 5). These results indicate that capturing the detailed information of ship targets can boost the performance of our ship retrieval system.

Table 5: Results with different settings on our proposed dataset. "+" means the concatenation operation, while "All" in the last column indicates that we concatenate all subfeatures to form the ship representation.

Representation	$\mathbf{f}_g^G$	$\mathbf{f}_h^G$	$\mathbf{f}_v^G$	$\{\mathbf{f}_h^i \mid_{i=0}^2\}$	$\{\mathbf{f}_v^i \mid_{i=0}^2\}$	$\mathbf{f}_g^G + \mathbf{f}_h^G + \mathbf{f}_v^G$	$\{\mathbf{f}_h^i \mid_{i=0}^2\} + \{\mathbf{f}_v^i \mid_{i=0}^2\}$	All
Rank-1	75.1	74.3	74.9	86.4	88.2	83.5	90.3	94.3
Rank-5	76.2	76.0	75.8	88.6	90.1	84.1	93.4	95.7
Rank-10	80.4	83.1	82.1	90.7	94.2	88.3	96.9	98.4
mAP	73.6	73.2	73.5	83.8	86.4	81.5	88.8	92.4

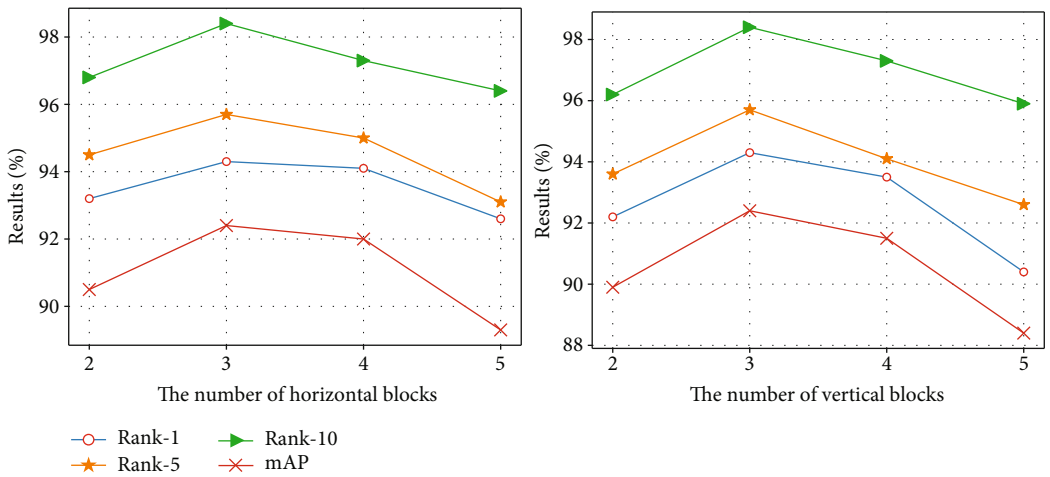

Figure 9: The evaluation of the number of blocks in two orientations.

Table 6: The speed of retrieving a ship target (Unit: second).

Scale	5 k	10 k	20 k	30 k
Rank-1	0.8	1.5	2.3	3.0
Rank-5	1.2	1.9	2.6	3.6
Rank-10	1.6	2.4	3.0	3.8

3.11.5. The Influence of Different Number Blocks. In this work, we mainly split the occlusive feature map into 3 blocks in two different orientations. To evaluate the effect of the number of blocks, we also conduct related experiments to investigate it. We show the results in Figure 9. The experiments show that we can obtain the best results when the number of blocks is 3 in two different orientations.

3.12. Speed Evaluation. In addition to the accuracy, we also evaluate the retrieval speed of our proposed method. We report the results on Table 6. From Table 6, we see that our retrieval method can search a similar target in the gallery with less time cost. For example, for Rank-1, our retrieval method takes around 3 s to give the retrieval result from 30,000 images. Thus, we argue that the proposed retrieval method is efficient.

4. Conclusion

In this paper, we aim to study the ship retrieval methods for intelligent water transportation system in smart cities. To achieve this goal, we construct a new large-scale fine-grained ship retrieval dataset (named FGSR) that consists of 30,000 images of 1,000 ships captured in the wild. Besides, we propose a novel fine-grained ship reidentification network (FGSR-Net) based on the MGN, which consists of three important modules: pyramid fusion module, occlusion module, and multibranch identify module. By applying our proposed method, we can address the variant size and shape problem and better produce discriminative feature representation. Our extensive experiments show the outperformance compared with existing state-of-the-art methods from relevant fields.

Data Availability

The data are available upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] M. Mohaimenuzzaman, S. M. Monzurur Rahman, M. Alhussein, G. Muhammad, and K. Abdullah al Mamun, "Enhancing safety in water transport system based on Internet of Things for developing countries," *International Journal of Distributed Sensor Networks*, vol. 12, no. 2, 2016.

[2] M. A. Sotelo, F. J. Rodriguez, and L. Magdalena, "Virtuous: vision-based road transportation for unmanned operation on urban-like scenarios," *In: IEEE Transactions on Intelligent Transporta- tion Systems*, vol. 5, no. 2, pp. 69–83, 2004.

[3] S. K. Chaturvedi, C. S. Yang, K. Ouchi et al., "Ship recognition by integration of SAR and AIS," *In: The Journal of Navigation*, vol. 65, no. 2, p. 323, 2012.

[4] H. Zhao, W. Zhang, H. Sun, and B. Xue, "Embedded deep learning for ship detection and recognition," *Future Internet*, vol. 11, no. 2, p. 53, 2019.

[5] O. Karabayır, U. Saynak, M. Z. Kartal et al., "Synthetic-Range-Profile-Based training library construction for ship target recognition purposes of scanning radar systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 4, pp. 3231–3245, 2020.

[6] X. Cao, S. Gao, L. Chen, and Y. Wang, "Ship recognition method combined with image segmentation and deep learning feature extraction in video surveillance," *Multimedia Tools and Applications*, vol. 79, no. 13-14, pp. 9177–9192, 2020.

[7] F. Huixuan, Y. Li, Y. Wang, and P. Li, "Maritime ship targets recognition with deep learning," in *In: 2018 37th Chinese control conference (CCC)*, pp. 9297–9302, Wuhan, China, 2018.

[8] J. Zou, W. Yuan, and Y. Menghong, "Maritime target detection of intelligent ship based on faster R-CNN," in *In: 2019 Chinese automation congress (CAC)*, pp. 4113–4117, Hangzhou, China, 2019.

[9] K.-K. Tseng, R. Zhang, C.-M. Chen, and M. M. Hassan, "DNetUnet: a semi-supervised CNN of medical image segmentation for super-computing AI service," *The Journal of Supercomputing*, vol. 77, no. 4, pp. 3594–3615, 2021.

[10] E. K. Wang, C.-M. Chen, M. M. Hassan, and A. Almogren, "A deep learning based medical image segmentation technique in Internet-of- Medical-Things domain," *In: Future Generation Computer Systems*, vol. 108, pp. 135–144, 2020.

[11] G. Vivone, P. Braca, and J. Horstmann, "Knowledge-based multitarget ship tracking for HF surface wave radar systems," *In: IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3931–3949, 2015.

[12] X. Chen, S. Wang, C. Shi, H. Wu, J. Zhao, and J. Fu, "Robust ship tracking via multi-view learning and sparse representation," *In: The Journal of Navigation*, vol. 72, no. 1, pp. 176–192, 2019.

[13] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," 2016, http://arxiv.org/abs/1611.05244.

[14] P. Huang, R. Huang, J. Huang et al., "Deep feature fusion with multiple granularity for vehicle re-identification," *In: CVPR Workshops*, pp. 80–88, 2019.

[15] F. Huixuan, Y. Li, Y. Wang, and L. Han, "Maritime target detection method based on deep learning," in *In: 2018 IEEE International Conference on Mechatronics and Automation (ICMA). IEEE*, pp. 878–883, Changchun, China, 2018.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *In: Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

[17] R. Girshick, "Fast R-CNN," in *In: Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, Santiago, Chile, 2015.

[18] R. R. Varior, M. Haloi, and G. Wang, "Gated Siamese convolutional neural network architecture for human re-identification," in *In: European Conference on Computer Vision*, pp. 791–808, Springer, Cham, 2016.

[19] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *In: Proceedings of the 26th ACM international conference on Multimedia*, pp. 274–282, New York, 2018.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *In: Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, Las Vegas, NV, USA, 2016.

[21] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *In: Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, Honolulu, HI, USA, 2017.

[22] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *In: Advances in neural information processing systems*, vol. 30, 2017.

[23] Q. Hou, L. Zhang, M. M. Cheng, and J. Feng, "Strip pooling: rethinking spatial pooling for scene parsing," in *In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4003–4012, Seattle, WA, USA, 2020.

[24] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *In: Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3702–3712, Seoul, Korea, 2019.

[25] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *In: Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8282–8291, Seoul, Korea, 2019.

[26] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, and T. Mei, "VehicleNet: learning robust visual representation for vehicle re-identification," *IEEE Transactions on Multimedia*, vol. 23, pp. 2683–2693, 2020.