

Article

Research on Key Technology of Ship Re-Identification Based on the USV-UAV Collaboration

Wenhao Dou ¹, Leiming Zhu ¹, Yang Wang ^{1,*} and Shubo Wang ^{2,*}

¹ School of Electronic and Information Engineering, Harbin Institute of Technology (Shenzhen), Shenzhen 518071, China; 20b952021@stu.hit.edu.cn (W.D.); 19s152103@stu.hit.edu.cn (L.Z.)

² School of Automation, Institute of Intelligent Unmanned System, Qingdao University, Qingdao 266071, China

* Correspondence: yangw@hit.edu.cn (Y.W.); shubowang@qdu.edu.cn (S.W.)

Abstract: Distinguishing ship identities is critical in ensuring the safety and supervision of the marine agriculture and transportation industry. In this paper, we present a comprehensive investigation and validation of the progression of ship re-identification technology within a cooperative framework predominantly governed by UAVs. Our research revolves around the creation of a ship ReID dataset, the development of a feature extraction network, ranking optimization, and the establishment of a ship identity re-identification system built upon the collaboration of unmanned surface vehicles (USVs) and unmanned aerial vehicles (UAVs). We introduce a ship ReID dataset named VesselID-700, comprising 56,069 images covering seven classes of typical ships. We also simulated the multi-angle acquisition state of UAVs to categorize the ship orientations within this dataset. To address the challenge of distinguishing between ships with small inter-class differences and large intra-class variations, we propose a fine-grained feature extraction network called FGFN. FGFN enhances the ResNet architecture with a self-attentive mechanism and generalized mean pooling. We also introduce a multi-task loss function that combines classification and triplet loss, incorporating hard sample mining. Ablation experiments on the VesselID-700 dataset demonstrate that the FGFN network achieves outstanding performance, with a Rank-1 accuracy of 89.78% and mAP of 65.72% at a state-of-the-art level. Generalization experiments on pedestrian and vehicle ReID datasets reveal that FGFN excels in recognizing other rigid body targets and diverse viewpoints. Furthermore, to further enhance the advantages of UAV-USV synergy in ship ReID performance, we propose a ranking optimization method based on the homologous fusion of multi-angle UAVs and heterologous fusion of USV-UAV collaborative architecture. This optimization leads to a significant 3% improvement in Rank-1 performance, accompanied by a 73% reduction in retrieval time cost.

Keywords: ship re-identification; ranking optimization; USV-UAV collaboration



Citation: Dou, W.; Zhu, L.; Wang, Y.; Wang, S. Research on Key Technology of Ship Re-Identification Based on the USV-UAV Collaboration. *Drones* **2023**, *7*, 590. <https://doi.org/10.3390/drones7090590>

Academic Editor: Sanjay Sharma

Received: 1 August 2023

Revised: 14 September 2023

Accepted: 18 September 2023

Published: 20 September 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The increasing importance of the ocean economy in global business activities and the subsequent rise in shipping intensity have presented substantial challenges to navigating safety and ship supervision, both inland and offshore. Traditionally, collision avoidance has relied on Automatic Identification Systems (AIS) and shipboard radar. These systems use GPS to transmit dynamic information, such as the ship's position, speed, and heading, and static information, such as the ship's name and call sign, to the surrounding area. Despite being widely used, AIS has limitations regarding communication quality, update frequency, and ship density, which can result in system instabilities [1,2]. In addition to these technical constraints, improper crew operation and untimely avoidance can also contribute to severe collisions on the water. Therefore, strengthening the active identification of ships is significant in regulating the order and safety of maritime economic activities such as transportation and fisheries.

In contrast to AIS and radar-based safety warning systems, optical detection technology provides more visual and richer visual information for identifying targets on the

sea, including the type of vessel, vessel number, motion information, etc. The visual ReID ranks the gallery set images by similarity with the query image to match the target's ID information from the query image. While current researchers focus more on ReID for pedestrians [3,4] and vehicles [5,6], they mainly combine the target's visual features around the three aspects of feature representation, metric loss, and ranking optimization.

In ReID, feature representation could be global, local, or auxiliary. Global feature representation extracts global features from each target image to create a feature vector without additional annotation information. The ID-discriminative Embedding (IDE) model [7] treats each ID sample as an independent class, constructing the training process as a multi-class classification problem. Local features are generally fused with global features to obtain the final feature vector. Ref. [8] first detects different body parts of the pedestrian and combines their corresponding local features with global features. Ref. [9] divides the query image into several horizontal blocks to extract the related local features. Ref. [10] proposes a local feature matching strategy to improve the robustness of ReID by opportunistically weighting three types of complementary feature information; namely, the overall chromatic content, the spatial arrangement of colors into stable regions, and the presence of recurrent local motifs with high entropy under the symmetric and asymmetric perceptual principles. Auxiliary features such as gender, hair, and clothing optimize feature representation learning [11,12]. In vehicle ReID tasks, auxiliary features such as car model and color are generally used. Those auxiliary features help to achieve a finer-grained classification and construct a more reasonable metric space for the model.

The ReID employs metric loss to guide feature learning. The commonly used loss functions include identity loss [13,14], verification loss [3,15], and triplet loss [16–18]. Identity loss treats the training process of the ReID model as an image classification problem, where each identity ID is a different class [7]. Verification loss evaluates the consistency of two input samples under contrastive [19] or binary verification loss [3]. On the other hand, triplet loss considers the training process of the ReID model as a retrieval ranking problem, where the distance between positive sample pairs should be smaller than between negative sample pairs [18].

Ranking optimization is an important method to improve the retrieval performance of re-identification. The basic idea of re-ranking is to utilize the gallery-to-gallery similarity to optimize the initial ranking list. For instance, ref. [20] proposed the top-ranked similarity-pulling and bottom-ranked dissimilarity-pushing methods. Another widely used method, k-reciprocal re-ranking [21], mines the contextual information to improve the ranking list.

However, ship ReID still faces numerous problems due to the feature differences between ships and pedestrians or vehicles [22–24]. Complex observation conditions and significant feature differences of the different sizes of ships in multiple viewpoints reduce the re-recognition performance in traditional methods. Specifically, Figure 1a shows the intra-class differences caused by viewpoint changes for the same ship. Figure 1b shows the inter-class similarity for different ships of the same type. Meanwhile, there is a lack of extensive research and specific public datasets for ship ReID, and the data sources in the current research are generally web images [23], shore-based camera shots [25], and UAV shots [26].

Despite the recent growth in this field, studies on ship ReID remain relatively scarce. IORnet [27] proposes a TriNet loss function based on the CNN method to enhance vessel identification. This approach focuses on improving the similarities between vessel images belonging to the same vessel identity within the feature space. Additionally, it provides an annotated harbor vessel re-identification dataset. MVR-net [24] proposed and tested a multi-branch feature extraction backbone on the self-built public dataset VR-VCA. The framework employs two separate height-wise and width-wise branches to extract a more representative vessel representation in spatial dimensions. Moreover, ref. [22] proposes a dynamic alignment warships re-identification method that incorporates transfer learning with statistical and geometric feature transformations. A special dataset was constructed and tested to account for the unique sea sway characteristics of vessels. In another vein,

ref. [27] introduced an identity-oriented re-identification model that combines triplet loss and cross-entropy loss, using ResNet50 as the essential feature extraction network. Additionally, ref. [25] extended triplet loss by employing multiple query strategies.

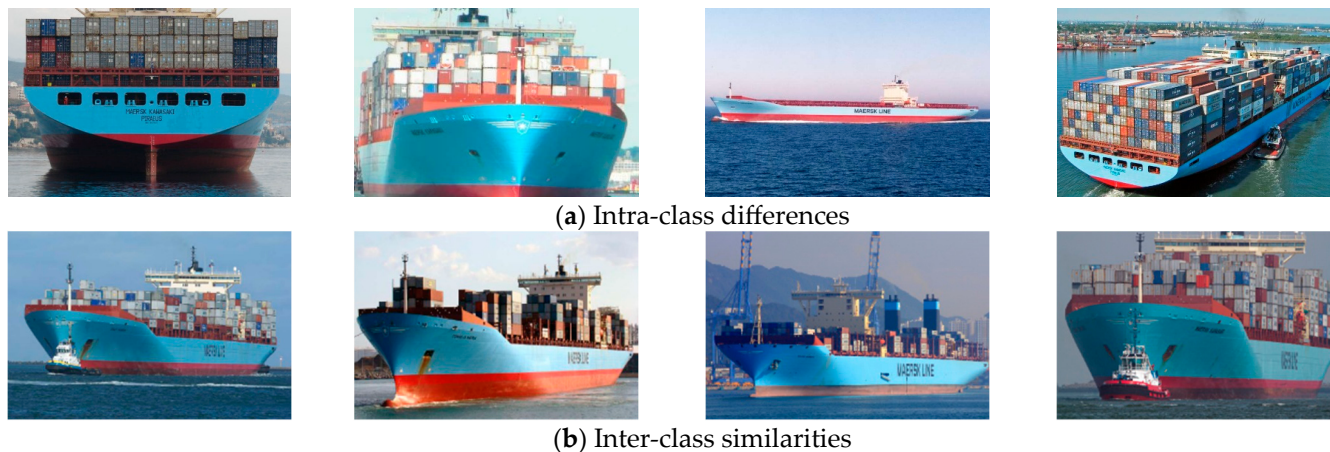


Figure 1. Intra-class differences (a) and inter-class similarities (b) for ships.

With the development of group intelligence technology, the cost and efficiency advantages of both homogeneous and heterogeneous multi-intelligence collaborative systems have been initially verified in various fields, including surveying and mapping and the military industry. Research on multi-intelligence cooperative systems has encompassed system integration design [28,29], cooperative control [30–33], and diverse cooperative task realization [34]. For instance, ref. [35] provides a water pollution monitoring method based on a USV-UAV system. Ref. [36] investigates a cooperative USV-UAV system for marine search and rescue with visual navigation and control. They designed an RL-based USV controller embedded with the UAV-based visual navigation under twin critic networks and actor networks. Ref. [37] presents a coastal management plan that divides tasks between UAVs and USVs, with UAVs responsible for numerous coastal target detection tasks and providing mission instructions to the USVs.

However, a substantial portion of existing research [30,38,39] remains centered on labor division and collaborative control within multi-intelligence systems. This often pertains to UAV formation strategies, water landing procedures for UAVs, joint trajectory control for UAVs and USVs, and similar aspects. Unfortunately, there is a noticeable scarcity of studies addressing the beneficial impact of UAVs' maneuverability advantages and collaborative intelligence on maritime tasks.

In this paper, we address the prevalent data problem in the current ship ReID task by constructing a ship ReID dataset, VesselID-700. To tackle the issue of the similar appearance of ship targets of the same model, we propose a fine-grained feature network called FGFN, which utilizes ResNet50 as the feature backbone network. We conduct ablation and generalization experiments on the relevant dataset to evaluate the effectiveness of our proposed approach. We propose a practical USV-UAV collaborative re-identification processing framework that can be applied in practical scenarios. We integrate this framework with a multi-view fusion ranking optimization method, resulting in a UAV-driven ReID process. Finally, we validate the ranking performance on an additional small 3D model dataset.

2. Materials and Methods

2.1. Datasets

The SeaShips Dataset [40] and Singapore Maritime Dataset (SMD) [41] are publicly available datasets commonly used for ship target detection training. The SMD dataset comprises 4085 images captured from shipboard video data in Singapore waters, containing a total of 31,614 targets classified into nine categories, including ferries, buoys, vessels,

speed boats, boats, kayaks, sailboats, persons, aircraft, and others. To effectively assess the accuracy of ship re-identification, the dataset must adhere to the following requirements, like those of pedestrian and vehicle ReID datasets:

1. Contain multiple image samples under the same ID label.
2. Include the same ID label images captured from multiple views.
3. Each image sample should feature a complete ship target.
4. Image samples of the ship target should maintain similar main features.
5. Query images should involve as many angles of the ship target as possible.

However, the SeaShips Dataset and SMD Dataset are unsuitable for the ReID task because they lack ship ID annotations, cannot classify the query and selected images, and fail to meet the aforementioned evaluation requirements. As of now, there is no publicly available dataset that supports ship ReID. Therefore, this paper constructs VesselID-700, a dataset for the ship ReID task. The dataset is created by cropping, grouping, and labeling ship images provided by ship photographers.

2.1.1. Dataset Collection

This paper collected raw data for the dataset from ShipSpotting [42], an international website for shipping photography. The website allows photographers to tag and upload images, which volunteers approve. The images on this website are labeled with the ship's ID, making them suitable for use as raw data for the ship ReID dataset. The VesselID-700 dataset in this paper comprises 56,069 images of seven common types of ships: Container-ships, General Cargo ship, Passenger vessels, Tankers, Tugs, Bulkers, and Fishing vessels.

2.1.2. Dataset Processing

The test set in VesselID-700 needs to divide the images into two categories: query and gallery images. In the pedestrian/vehicle ReID dataset, each image has a camera ID (CID). When testing the model performance, it is necessary to avoid query images matching the same object under the same CID to demonstrate the ability of the ReID model to recognize across cameras. Furthermore, for the raw data of VesselID-700, each image of the same ship is taken randomly from different angles. In order to constrain the dataset, all images of the same object in the VesselID-700 dataset need to be classified by angle and thus assigned the angle ID (AID).

We propose using the Histogram of Oriented Gradient (HOG) descriptor [43] as the basis for angle classification features. The HOG features describe the gradient direction of the pixel values. The global gradient features of the entire image are constituted by first calculating the local gradient direction features, followed by global statistics. The effectiveness of the HOG feature extraction method for ship image samples is demonstrated in Figure 2, where the orientation of each pixel in the HOG feature corresponds to the orientation of the texture in the original image.

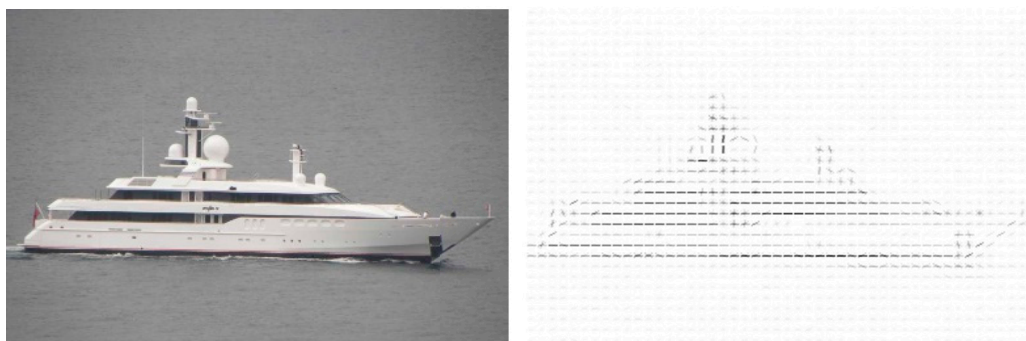
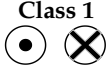



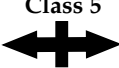






















Figure 2. HOG feature.

With the basic orientation of the ships in the picture obtained using the HOG feature descriptor, we deployed the K-means clustering algorithm [44] to label the AIDs for our dataset. Considering the relative shooting angles and heights of the original images in the VesselID-700 dataset, we categorized the basic orientation of the ships in the images into five categories: flat view, overhead view, left tilt, right tilt, and side view, which implies a hyperparameter $k = 5$ for the K-means. The schematic effect of the ship orientation clustering classification is presented in Table 1.

Table 1. Ship orientation clustering classification.

Class 1	Class 2	Class 3	Class 4	Class 5
				
				
				
				
				

To minimize the impact of the environmental background on ship features, we trained a vanilla YOLOv3 [45] model using the SeaShips and SMD datasets to crop the ship images in the VesselID-700 dataset. A newer and lighter version of YOLO allows for a faster and better cropping process.

We mixed the two datasets to obtain 12,075 images and randomly divided them into training and validation sets in a ratio of 8:2 to ensure representative training and testing results. After completing the training, the pre-trained YOLOv3 was used to crop the ship-bounding boxes, resulting in uniform ship images with a pixel size of 384×256 , as illustrated in Figure 3.



Figure 3. Example of the VesselID-700 dataset.

We compared the dataset VesselID-700 with other target types, and the results are shown in Table 2. The ship ReID video dataset VesselReID [25] was collected using two

cameras set up on the same riverbank, with periodic shooting angle adjustments, resulting in a small number of each ID and close angles. However, the VesselReID dataset is not available for download. Surveillance cameras captured the VeRI-776 vehicle ReID dataset, while the vehicle ReID dataset UAV-VeID was captured using drones and included parked vehicles in parking lots and vehicles moving on the road.

Table 2. Comparison of properties of ReID datasets.

Dataset	Target	ID Volume	Dataset Scale	Angle of View
VesselID-700	Vessel	700	56,069	Five angle types with random multi-angle
VesselReID	Vessel	733	4616	Random multi-angle
Market-1501	Person	1501	32,643	Six fixed angles
VeRI-776	Vehicle	776	51,035	Sixteen fixed angles
UAV-VeID	Vehicle	4601	58,767	Random multi-angle

2.2. Fine-Grained Feature Network Design

Compared with pedestrian and car ReID tasks, different ships tend to have a similar shape and paint, making it difficult to distinguish between classes, especially since the same ship may present different appearance features due to differences in camera angles, resulting in more significant intra-class differences. In contrast, people are easier to differentiate as they have more distinct features, including their faces, clothing, and accessories. While pedestrians may vary in their pose, their overall longitudinal features remain consistent under viewpoint changes. Ship ReID is more akin to the vehicle in terms of target features. However, boats have more categories than vehicles, and there is a more significant difference in volume between different categories of boats. Conversely, vehicles possess more detailed features useful for recognition, such as the labels affixed to windshields and the shapes of their wheels or lights.

Therefore, to improve ship ReID, the feature extraction network must pay greater attention to detailed features while ensuring the effective extraction of global features. The loss function should also consider the inter-class and intra-class differences. We propose the fine-grained feature network (FGFN) and a multi-task loss function designed to extract detailed image features without compromising global information and to cooperate with the metric loss function to distinguish high-similarity negative samples. The overall structure of FGFN is presented in Figure 4. ResNet50 [46] serves as the feature backbone network, supplemented by a Non-local module [47] to facilitate self-attentive and non-local feature connections. GeM Pooling [48] with learnable hyperparameters balances local and global relationships. Finally, Cross-entropy Loss is responsible for classification, and TriHard Loss [49] is responsible for retrieval ranking, combined into a multi-task loss function.

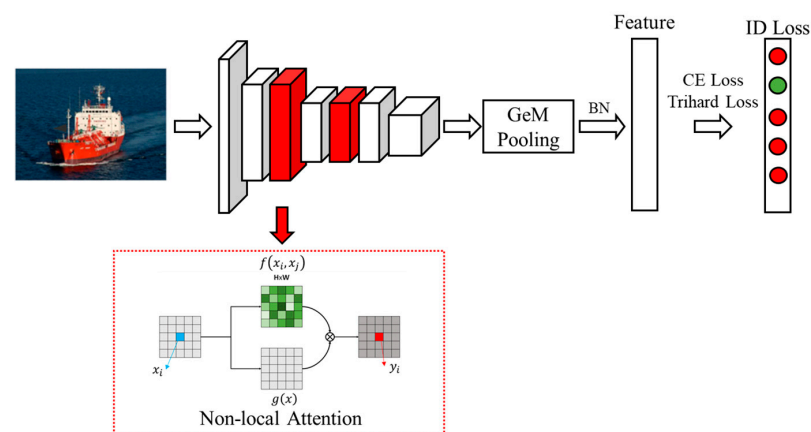


Figure 4. Fine-grained feature network and a multi-task loss function.

2.2.1. Non-Local Module

Based on the previous analysis, it is evident that due to the nature of ship images, ensuring that the feature extraction network can effectively differentiate positive samples from highly similar negative samples using a priori information is challenging. A spatial attention mechanism [50] that directs the feature extraction network to focus on local features is a suitable solution for this problem. Another way to guide the model's attention towards local features is the self-attention mechanism [51,52]. Moreover, typical convolutional perceptual fields in feature extraction networks are 3×3 or 5×5 in size. Increasing the perceptual field necessitates a bigger kernel size or the incorporation of deeper convolutional layers, making it difficult for the model to converge. To overcome this challenge, we utilize the Non-local module self-attention mechanism to provide cross-space information association and enhance the model's focus on relevant detailed feature locations. The non-local module computation process is shown in the red box in Figure 4. Non-local operations capture long-range dependencies directly by computing the interactions between any two positions without restrictions to adjacent points, which is equivalent to constructing a convolution kernel as large as the size of the feature map, thus allowing more information to be captured.

The general formula for the non-local module is expressed in Equation (1):

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j), \quad (1)$$

where x denotes the input feature map, y denotes the output feature map, $f(x_i, x_j)$ denotes the similarity between the feature x_i at the position i of x and the feature x_j at position j , while we use the embedded Gaussian as the pairwise function f . Additionally, $g(x_j)$ denotes the output of the feature map at position j in the form of a linear embedding, and $C(x)$ denotes the normalization factor.

2.2.2. GeM Pooling

The purpose of the pooling layer is to combine the multi-channel 2D feature maps produced by the backbone network into global features. The pooling helps to achieve invariance to local variations. In ship ReID, global features distinguish global differences such as ship structure and body color. In contrast, local features are used to distinguish the similar details of highly similar negative samples. Therefore, a pooling method for ship ReID is required to capture local and global features in the feature extraction network.

The pooling layer takes $X \in \mathbb{R}^{W \times H \times C}$ as the input and produces the output vector $f \in \mathbb{R}^{1 \times 1 \times C}$, where W, H, C denote the width, height, and number of channels of the feature map, respectively. Two types of pooling are commonly used: Max Pooling and Average Pooling. Generalized-mean pooling (GeM Pooling) [48,53] is a method that combines both Max and Average Pooling. It is a unified form of the two methods mentioned above. When the hyperparameter $p = 1$, GeM Pooling degenerates to mean pooling. On the other hand, when p tends to infinity, it represents maximum pooling. Adjusting the parameter p allows a balance between localization and the feature map response's globalization. As p increases, the response of the feature map becomes more localized.

$$f_{MaxPooling} = \max_{x \in X}(x), \quad (2)$$

$$f_{AvgPooling} = \frac{1}{|X|} \sum_{x \in X} x, \quad (3)$$

$$f_{GeMPooling} = \left(\frac{1}{|X|} \sum_{x \in X} x^p \right)^{\frac{1}{p}}, \quad (4)$$

2.2.3. Multi-Task Loss Function

The loss function, which serves as the goal function during convolutional neural network training, is one of the key aspects for the overall model to be effective. In the ship ReID task, we also face complicated ship class discriminations and the need to distinguish ship IDs. More critically, there are many similar appearance features between the same model of ships. To optimize those problems, multiple task loss functions are required to simultaneously constrain the model. For example, classifying ships into different classes as a classification task while discriminating each ID requires measuring the distance between each sample. Therefore, we must design multi-task loss functions to handle fine-grained classification and high-similarity negative sample problems.

The classification loss and metric loss are shown in Figure 5. Figure 5a shows the schematic diagram of classification loss, and the yellow dashed line is the hyperplane through which the classification loss tries to partition different classes into different subspaces. Figure 5b shows that the metric loss reduces the intra-class distance and increases the inter-class distance. Figure 5c shows the joint action of classification and metric loss.

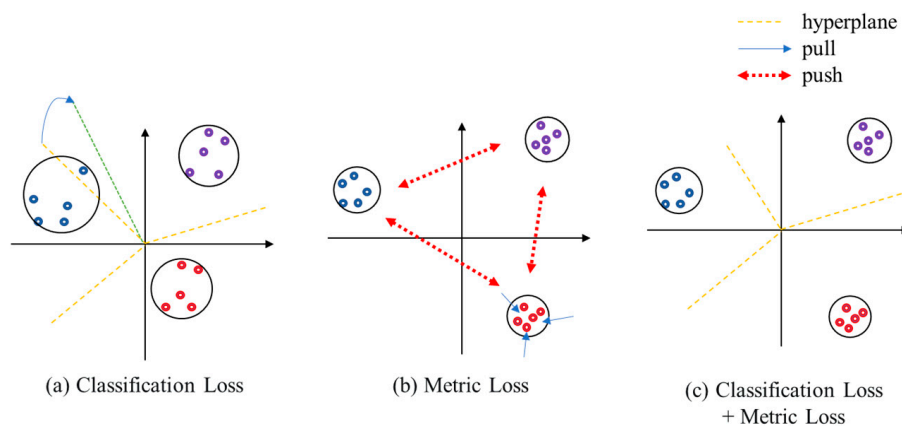


Figure 5. Schematic diagram of the multi-task loss.

1. Classification loss

Classification loss treats the training process of ship re-identification as an image classification problem, i.e., each ID is a different category; hence, it is also known as ID loss. Assuming that the label of the input image x_i is y_i and $p(y_i | x_i)$ denotes the probability that the softmax function will identify x_i as the category y_i , the cross-entropy calculates the classification loss in Equation (5) as:

$$L_{id} = -\frac{1}{n} \sum_{i=1}^n \log(p(y_i | x_i)). \tag{5}$$

2. Metric Loss

The metric loss plays a crucial role in the training process of ship ReID as it treats it as a retrieval ranking problem. One of the most used loss functions in retrieval tasks is the triplet loss, which consists of an anchored sample x_a , a positive sample x_p with the same identity, a negative sample x_n from a different identity, and a marginal distance parameter m . The central concept is that the distance between x_a and x_p plus m should be smaller than the distance between x_a and x_n , as shown in Equation (6):

$$d(x_a, x_p) + m < d(x_a, x_n), \tag{6}$$

then, triplet loss can be expressed as follows:

$$L_{tri} = \max(d(x_a, x_p) + m - d(x_a, x_n), 0). \tag{7}$$

During the actual training process, triplet loss is commonly implemented by randomly sampling a triplet of samples. However, random sampling cannot significantly constrain the model when the distances between samples are different, especially when there are many simple samples. Batch hard triplet (TriHard) loss [18] with batch hard sample mining has been proposed. In TriHard loss, P samples of IDs are selected for each training batch, and each sample randomly selects K different images, resulting in a training batch of $P \times K$ images. For each image x_a in the training batch, a positive sample with the farthest from x_a and a negative sample with the closest distance and x_a could form a triplet, the TriHard loss is denoted as:

$$L_{TH} = \frac{1}{P \times K} \sum_{a \in batch} (d_{max}(x_a, x_p) - d_{min}(x_a, x_n) + m). \quad (8)$$

2.2.4. Evaluation Metric for ReID

The most common metrics used to evaluate a ReID system are Cumulative Matching Characteristics (CMC) and average accuracy mAP. Rank- k (also known as CMC- k matching accuracy) indicates the probability of a correct match occurring in the top k retrieval results. For any query image q , Rank- k is shown in Equation (9) below, where $gt(q, k) = 1$ represents that the correct matching target of query image q appears before the k -th position of the retrieval sequence; otherwise, it is taken as 0.

$$\text{Rank}(k) = \frac{\sum_{q=1}^Q gt(q, k)}{Q}. \quad (9)$$

In general, Rank-1 can effectively reflect the retrieval performance of the ReID system when the query image has only one truth label in the candidate image library. Nevertheless, the actual situation is that the candidate library is likely to contain multiple truth labels for the query ID. The combined assessment of the model's retrieval ability for simple and complicated samples requires the aid of the mAP evaluation model. The mAP evaluates the average retrieval performance in the presence of multiple truth-valued labels. The mAP can be a guideline when two ReID systems do equally well on Rank-1. For calculating mAP, we first calculate the retrieval accuracy AP for each query image q , n denotes the total number of all images in the candidate library, and N denotes the total number of images in the candidate library that can be correctly matched. $P(k)$ denotes the accuracy of the first k results in the retrieval results, and $gt(k)$ denotes whether the k -th retrieval result is correctly matched. After obtaining AP, the retrieval accuracy of all query sets Q is averaged to obtain mAP.

$$\text{AP} = \frac{\sum_{k=1}^n P(k) \times gt(k)}{N}, \quad (10)$$

$$\text{mAP} = \frac{\sum_{q=1}^Q \text{AP}(q)}{Q}. \quad (11)$$

2.3. Multi-View Ranking Optimization Based on the USV-UAV Collaboration

In offshore regulatory scenarios, identifying non-cooperative targets using only AIS is impossible, and shore-based surveillance is ineffective in identifying over-the-horizon targets. Therefore, we propose a unified ReID processing system using the USV and the UAV, as illustrated in Figure 6. This system relies on the USV-UAV collaborative supervision platform to identify the target ships in the over-the-horizon range and decompose the ReID process. The UAV receives control instructions from the USV, leverages its mobility and multi-view angle capabilities, and is responsible for image acquisition and target ship detection. In contrast, the USV can be equipped with stronger computing power to take advantage of the high load and strong arithmetic power. It is responsible for feature extraction, retrieval, and communication with shore-based stations. For the ship ReID task based on the USV-UAV collaboration, the focus should be on reducing information

redundancy in the detection process and improving the efficiency of image acquisition by the UAVs. The images collected by the UAV are multi-angle data around a specified target, which may generate data redundancy and arithmetic power waste under a continuous angle view, causing a more significant impact on the unmanned platform with limited energy consumption.

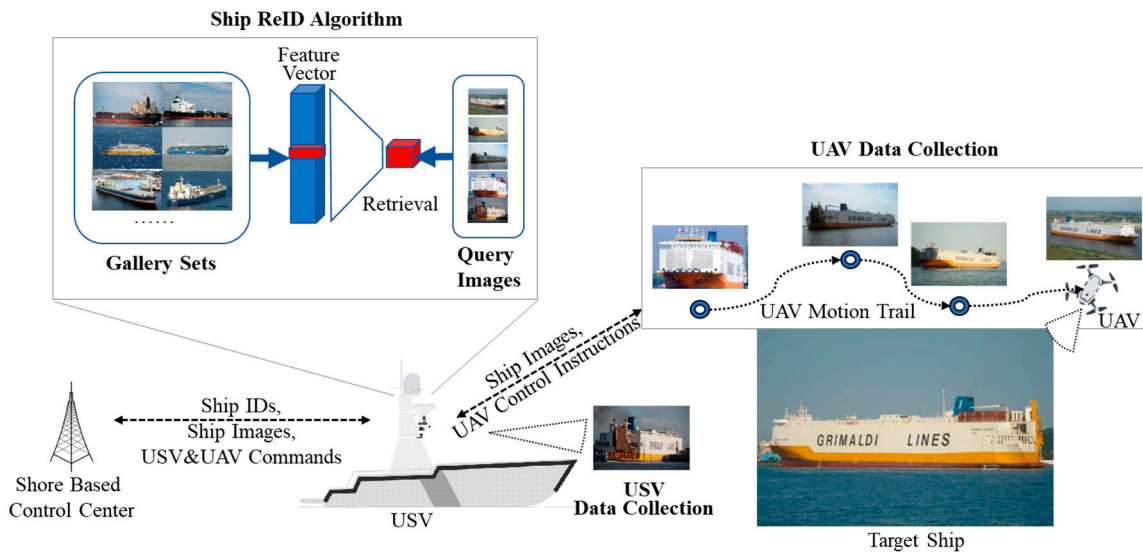


Figure 6. The USV-UAV collaborative ReID System.

We optimize the maritime ship re-identification retrieval ranking based on the USV-UAV collaboration, giving full play to the collaboration between the UAV and the USV. Aiming at the advantage of UAV maneuverability with multi-frame continuous angle image sequences as the query input, we propose FGFN-based ranking optimization to improve recognition accuracy by using multi-view information and, simultaneously, optimize the query process to reduce computational effort and time delay.

Considering the multi-frame ReID problem, we conduct a performance analysis on a small ship ReID test set from 3D ship CAD models with the pre-trained FGFN model on the VesselID-700 dataset. The test set consists of 10 3D Bulker models, with 360° image acquisition at 20° intervals for each model from a top-down view of 45°, as illustrated in Figure 7. The candidate image gallery set comprises single-shot images, meaning that each Bulker ID has only one corresponding image in the gallery set. Additionally, the gallery set is augmented with interference data, including more 3D model ship images at random angles and some authentic ship images from the VesselID-700 dataset. Ultimately, we obtain a test set containing 170 images in the query set and 3010 in the gallery set.

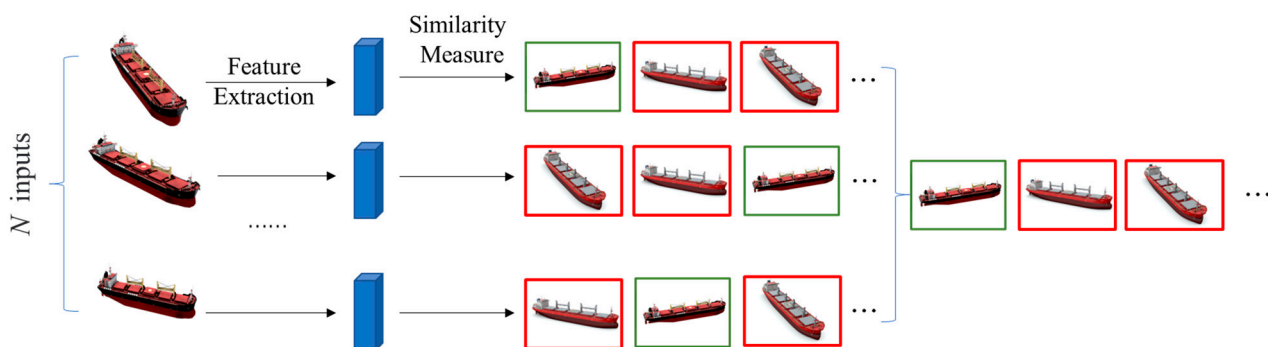


Figure 7. Continuous angle query images independent retrieval ranking schematic.

The multi-frame ReID process for a UAV flying around the target ship is shown in Figure 7. N homologous continuous angle query images are subjected to feature extraction by the FGFN model separately. The ranking sequence of each input is obtained after a similarity measurement, and the ranking result with the relatively highest accuracy is taken as the final output result (as shown with the green bounding). The independent retrieval process does not take advantage of the additional effective information from the continuous angle query images. Nevertheless, it pays for the computational effort and computational time consumption for the redundant information in the image sequence.

The optimized recognition process based on feature fusion is shown in Figure 8. After the same N times of feature extraction, the N feature vectors are fused into one feature vector, and then the candidate images are sorted. To demonstrate that the proposed method can reduce the computation, let the feature extraction computation be a , the feature similarity measure and ranking computation be b , and the feature fusion computation be c . Then, the total computation of the original method (Figure 7) is $N \times (a + b)$, and the total computation of the proposed feature fusion-based optimal retrieval method is $N \times (a + c) + b$. Then, when the fusion of N features computation $N \times c$ is less than the computation of $N - 1$ times similarity measure and ranking $(N - 1) \times b$, the proposed method has an advantage in computation. It is evident that the computational amount of feature fusion is related to the size of the feature, which is generally 2048 dimensions. Moreover, the computation amount of similarity measure depends on the size of the gallery set, while it should be in the order of tens of thousands of images.

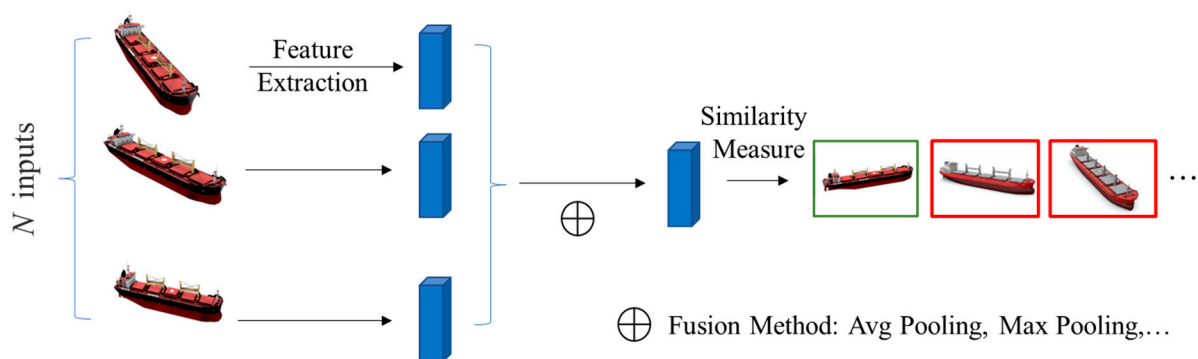


Figure 8. Continuous angle query images fusion retrieval ranking schematic.

In any case, in the collaborative platform, the USV can collect images of the target ship from a horizontal or upward view, which can also serve as supplementary data for the ship ReID task. Therefore, we propose a heterogeneous multi-view feature fusion retrieval ranking method, which incorporates the detection images from the USV as additional query images to the above homogeneous multi-frame fusion ranking. We validate the effectiveness of this method in subsequent experiments.

3. Results

In this section, we provide some typical visualization results, ablation experiments, and generalization experiments to show intuitively the accuracy and effectiveness of the proposed method. We also discuss the positive effect of collaborative ranking optimization on homologous multi-frame fusion and heterogeneous multi-view fusion on ReID accuracy and timeliness.

3.1. Implementation Details

The proposed model was implemented in Pytorch and all experiments were run on Linux with Intel(R) Xeon(R) Platinum 8153 CPU and an NVIDIA Tesla P100 with 12 GB GRAM. The stochastic gradient descent strategy is used as the optimizer with a momentum

of 0.9 and a weight decay rate of 0.001. The learning rate strategy is applied with a base learning rate of 0.01, minimum attenuation of 0.0001, and power of 0.9.

3.2. Comparison with the State-of-the-Art and Ablation Experiment

We conducted ablation experiments for the FGFN model on VesselID-700 ship ReID to demonstrate the model's effectiveness, and the test results are shown in Table 3. Compared with IORNet [27] and GLF-MVFL [23] with the same backbone network of ResNet50, the FGFN based on Non-local and GeM Pooling achieves a significant Rank-1 performance improvement. Under the ablation experiment with the feature extraction network structure fixed to ResNet50, adding the Triplet loss function as the multitask loss function can obtain large performance gains, including a 3.5% Rank-1 improvement and a 16% mAP improvement. Replacing the Triplet loss function with the TriHard Loss for batch-hard sample mining resulted in a 0.5% Rank-1 and 1.4% mAP performance gain. The combined effect of the Non-local and GeM Pooling modules results in about 2% Rank-1 gain and about 4% mAP gain, respectively.

Table 3. Comparison and ablation experiment results.

Method	Loss Type	Rank-1 (%)	mAP (%)
Baseline: ResNet50	CE	83.10	42.33
IORNet [27]	CE + Triplet	85.76	56.63
Base-GLF-MVFL [23]	CE + TriHard	84.14	48.78
GLF-MVFL [23]	CE + O-Quin	88.72	62.19
ResNet50	CE + Triplet	86.57	58.60
ResNet50	CE + TriHard	87.09	60.35
ResNet50 + Non-local	CE + TriHard	88.99	64.36
ResNet50 + GeM Pooling	CE + TriHard	89.05	64.09
FGFN (ResNet50 + Non-local + GeM Pooling)	CE + TriHard	89.78	65.72

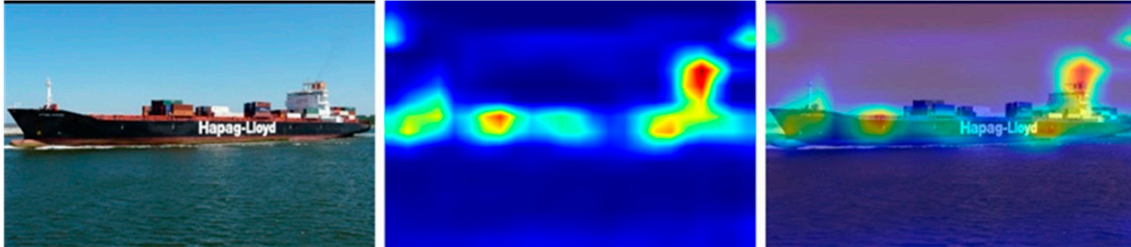
The feature response heat map in Figure 9 visually demonstrates which areas of the image are more useful for detection when features are aggregated. Compared to vanilla ResNet50, adding the Non-local attention module allows the fixed-depth backbone network to capture more local features under the short- and long-term relations [4]. The GeM Pooling module, on the other hand, effectively expands the perceptual field. When comparing Baseline and FGFN, the combined effect of attention and pooling extends the local features significantly to the global and obtains a broader and stronger feature activation.

The ReID results are visualized in Figure 10, where the left panel shows the query input, and the right panel displays the top ten query results sorted by similarity. Green boxes represent the correct IDs, while red boxes indicate the inconsistent ReID results. It is evident from the visualization that the FGFN model outperforms the Baseline in terms of ReID performance. In the case of the easy sample of the general cargo ship, the FGFN not only retrieves the correct ID from the gallery set but also ensures a high similarity ranking. FGFN achieves the unique correct ID with the first hit for the hard sample of the containership.

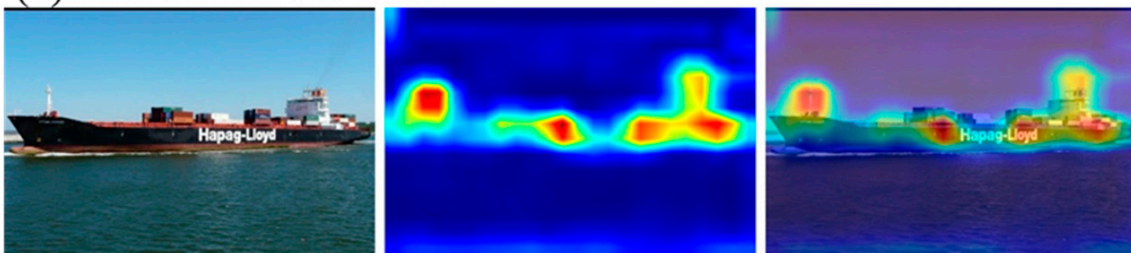
Figure 11 displays the distribution of cosine distances between positive and negative sample pairs, which were obtained by inference from different models. The horizontal axis represents the cosine distance of the sample pairs, and the vertical axis denotes the number of normalized distributions of the sample pairs. The intersection area between the positive sample pair (purple) and the negative sample pair (blue) represents the easily confused samples. In other words, the intersection area contains samples that could be either positive or negative sample pairs. A smaller area of this confusion region indicates a better metric space constructed by the model. Figure 11a shows the distance distribution obtained from the baseline model. Compared with the results obtained from the FGFN model in Figure 11b, the FGFN model exhibits a larger interval between the distance distributions of

positive and negative sample pairs. This larger interval is attributed to the balance of the TriHard metric for inter-class and intra-class differences.

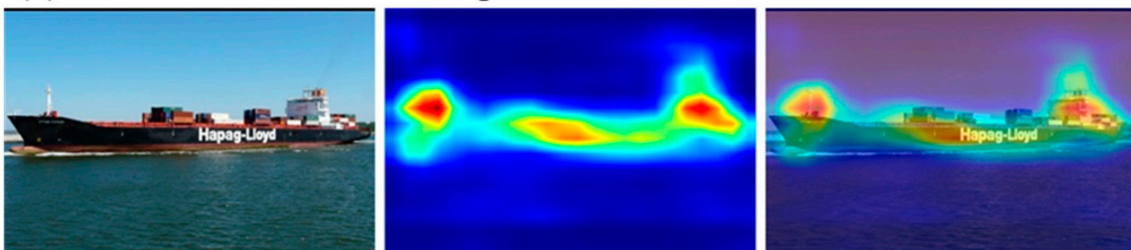
(a) ResNet50



(b) ResNet50+Non-local



(c) ResNet50+Gem Pooling



(d) FGFN

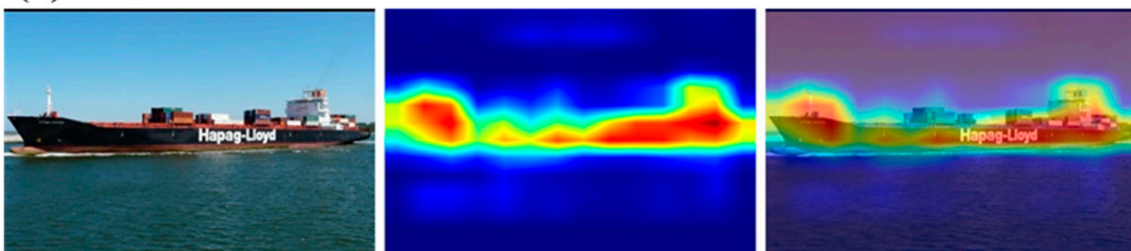


Figure 9. Feature response of ablation experiments.

Figure 12 shows the receiver operating characteristic (ROC) curves obtained from the baseline and proposed FGFN models. Each point on the ROC curve reflects the relationship between false positive (FP) and true positive (TP) corresponding to different thresholds. The horizontal coordinate is the false positive rate $FPR = FP / (FP + TN)$, and the vertical coordinate is the true positive rate $TPR = TP / (TP + FN)$. The red line is the resulting curve of random guessing, the purple line is the Baseline, and the green line is the ROC curve of the FGFN model, which obviously achieves a lower false positive rate of ReID.



Figure 10. Re-identification results on the VesselID-700 dataset.

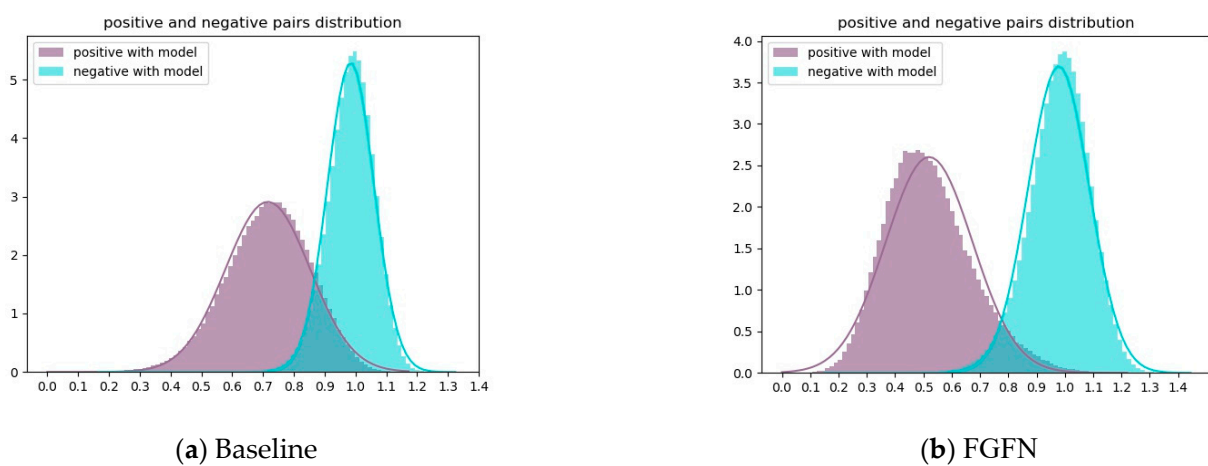


Figure 11. Cosine distance distribution of positive and negative sample pairs under Baseline and FGFN.

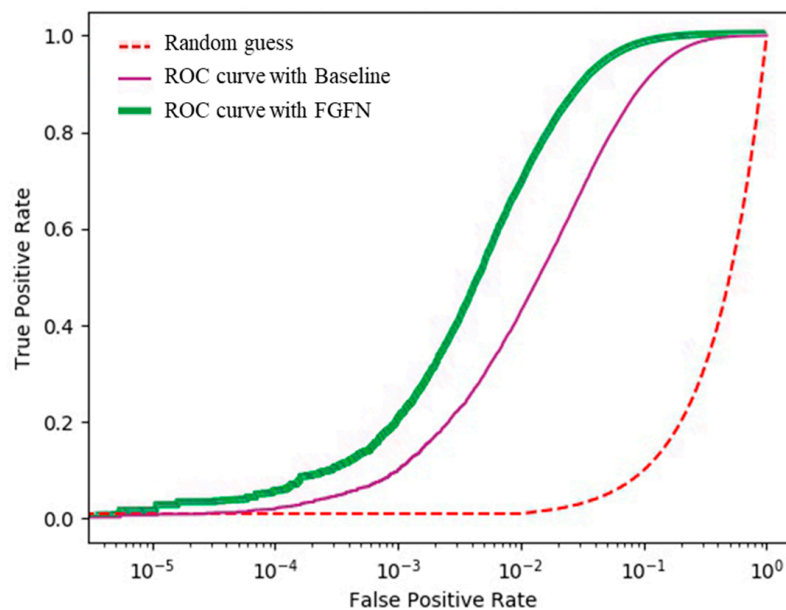


Figure 12. ROC curves of Baseline and FGFN.

3.3. Generalization Performance

In this section, we validate the generalization performance of FGFN on publicly available pedestrian re-identification and vehicle re-identification datasets to assess the model's ability to generalize to different rigid targets and multiple viewpoints for re-identification tasks. The generalization experiments aim to compensate for the limitation of the current ship re-recognition approach, which does not consider the UAV viewpoint. The vehicle target is also a rigid body and thus shares the characteristic of exhibiting significant differences in visual features under changes in the viewpoint.

FGFN can also achieve better results on other target public re-identification datasets, especially vehicle ReID datasets. As shown in Table 4, three datasets, Market1501 [54], VeRI-776 [55], and UAV-VeID [56], are used. The images of the pedestrian re-identification dataset Market1501 are from six fixed cameras on campus, including 32,643 images of 1501 pedestrians. The images of vehicle re-identification dataset VeRI-776 come from 17 public cameras on public roads in the city, recording a total of 51,035 images of 776 vehicles. The images of the vehicle ReID dataset UAV-VeID come from videos of urban public roads taken by UAVs, and a total of 58,767 images of 4601 vehicles are registered.

Table 4. Performance of FGFN on other target re-identification datasets.

Target	Dataset	Model	Rank-1 (%)	mAP (%)
Pedestrian	Market1501	FGFN	95.3	87.9
		Circle Loss [57]	96.1	87.4
		FGFN	96.0	78.3
Vehicle	VeRI-776	PRN [58]	94.3	74.3
		PGAN [59]	96.5	79.3
	UAV-VeID	FGFN	80.0	85.6
		VSCR [56]	70.6	--

In detail, the experiments on Market1501 focus on verifying the model's generalizability, and the test results show that FGFN has reached the mainstream level on the pedestrian ReID task. It could be seen that FGFN lags only 0.8% behind Rank-1 of the model corresponding to the SOTA level of Circle Loss on the pedestrian re-recognition task. The experiments of FGFN on vehicle ReID datasets, on the other hand, verify that the model can reach the SOTA level against various types of rigid targets. The accuracy performance

on the VeRI-776 dataset is comparable to that of the PGAN model based on the a priori knowledge space attention mechanism. The FGFN model performs well on the UAV-VeID dataset with UAV views, which indicates that FGFN can achieve the SOTA levels even in the case of overhead views with significant changes in views. The experimental data of vehicle ReID from the UAV view can also prove the effectiveness of the proposed model in rigid target re-identification from the UAV view. However, the UAV view ship image data in the real environment are currently unavailable.

3.4. Background Noise

As mentioned above, we used vanilla YOLOv3 to perform ship target bounding box cropping on the VesselID-700 dataset, and the results before and after bounding box cropping are shown in Figure 13. To demonstrate the importance of cropping and noise reduction, we conducted a comparison experiment under the Baseline and obtained the experimental results as shown in Table 5. The dataset with bounding box cropping brings a significant improvement of about 4% and 8.5% to the re-identification Rank-1 and mAP performance, respectively, and the removal of background noise ensures the efficiency of local features in their similarity.

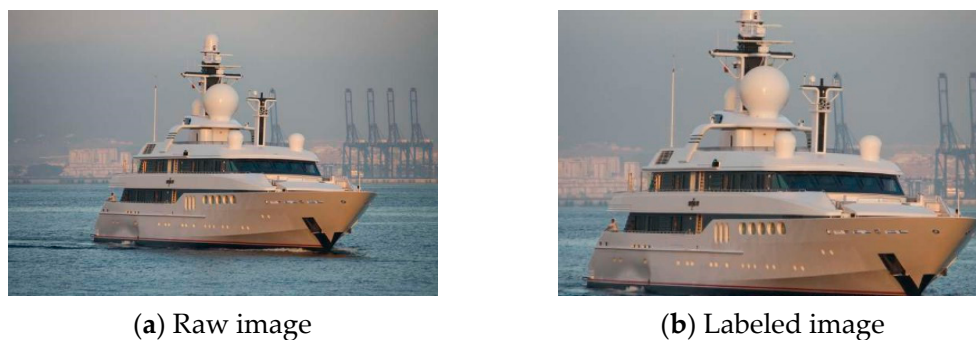


Figure 13. Target detection bounding box labeling effect.

Table 5. Impact of bounding box labeling on ReID performance.

Bounding Box Labeling	Rank-1 (%)	mAP (%)
False	79.01	33.80
True	83.10	42.33

3.5. Homologous and Heterologous Multi-View Fusion Retrieval Ranking Performance

N continuous angle query images are randomly taken from the small 3D ship model test set above during the homologous feature fusion experiments. The N 2048-dimensional feature vectors computed by the FGFN backbone network are fused into one 2048-dimensional feature vector under different approaches and used to conduct similarity ranking and ID query. Figure 14 shows the Rank-1 performance of the re-identification of homologous query images under different fusion methods. The horizontal coordinate is the number of fused features. The fusion number of 1 means no fusion is conducted, and the vertical coordinate is the query accuracy. The two solid lines represent the Max and Average Pooling fusion methods. The test results show that the Average Pooling of multiple homogenous views brings different performance improvements compared to a single image query. On the other hand, due to the difference in attention between multiple output features caused by viewpoint variations, pure maximum pooling is likely to cause fine detail loss, leading to a degradation of the overall fusion performance.

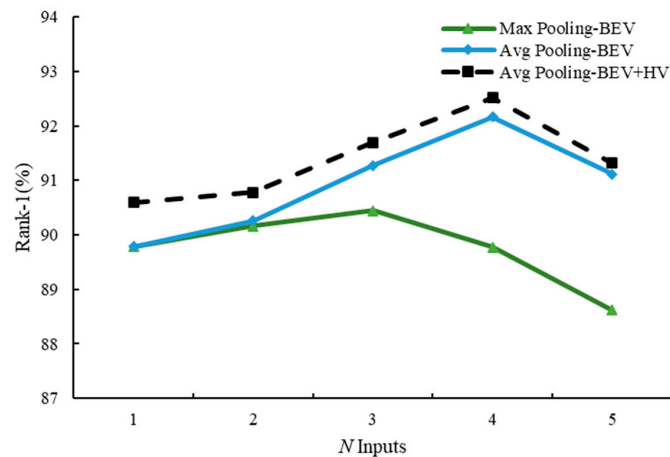


Figure 14. Homologous and heterologous fusion experiments.

For the heterogeneous multi-view fusion process of the USV and the UAV, we consider the Average Pooling fusion between the horizontal view of the boat and the bird's eye views of the UAV. The test results, Figure 14, show that Rank-1 of the heterogeneous multi-view achieves a slight improvement of about 3%. However, as the fusion scale increases, the multi-views bring saturation to the representation of ship features, and thus the gain generated by the flat view is no longer prominent.

3.6. Fusion Time Consumption

In this subsection, we test the feature fusion retrieval ranking elapsed time based on the USV-UAV collaborative architecture with independent ranking and average pooled fusion ranking for N consecutive query inputs under a fixed-size test set, respectively. The experiments measure the independent retrieval ranking time and the average pooled retrieval ranking time with the complete test set, where the fused retrieval ranking process includes the feature fusion and the fused retrieval ranking process. The test results show that as the fusion scale increases the query times decrease, which significantly reduces the frequency and time cost of similarity measures in re-identification. It means that the time cost required to process the same batch of data is decreasing, and it can save 73% of the retrieval ranking time when fusing four features. Although the time to compute the features of the query image is constant for any size of fusion retrieval, as shown in Figure 15, the complete ReID process also takes time to compute more image features when extending the fusion scale.

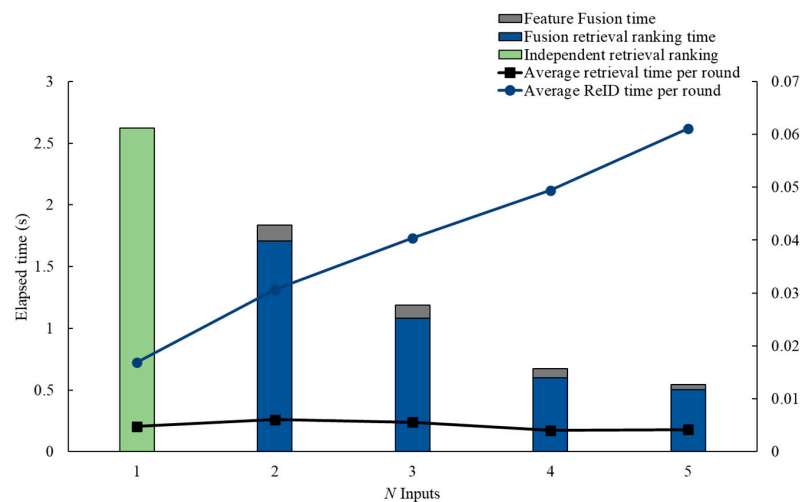


Figure 15. Fusion retrieval ranking elapsed time test.

4. Conclusions

Under the trend of safety supervision in the marine agriculture and transportation industry, this paper presents our research results in various aspects such as ship ReID datasets, feature extraction networks, metric loss, and UAV-USV-based ranking optimization. In addition, we designed a USV-UAV collaboration ReID architecture in conjunction with the ship ReID network. The conclusions of this study are as follows:

We collected and constructed the VesselID-700 ship ReID dataset, providing a foundational resource for optimizing the ship ReID algorithm. We employed angle-based grouping and optimization techniques utilizing HOG feature descriptors and K-means clustering. Additionally, we applied noise reduction to the dataset using vanilla YOLOv3. A comparative experiment confirmed the positive impact of background noise removal on enhancing ReID accuracy. A fine-grained feature network (FGFN) solves the challenge of small inter-class distance and large intra-class variation of samples in ship ReID. The network incorporates the self-attentive mechanism and the generalized mean pooling based on ResNet. It uses a classification loss and a TriHard Loss with difficult sample mining as multi-task loss functions. The ablation experiments show that the network could achieve an 89.78% Rank-1 accuracy and 65.72% mAP accuracy on VesselID-700, respectively, with 6.7% and 23.4% improvement compared to the base model Baseline. The generalization experiments show that the FGFN achieves a 96.0% Rank-1 on the vehicle ReID dataset VeRI-776. On the vehicle ReID dataset UAV-VeID from the UAV view, it substantially outperforms the VSCR method. The above experimental data show that the proposed FGFN model can better cope with the inter-class similarity and intra-class difference problems in ship re-identification. The feature fusion ranking optimization settles the information redundancy problem that the UAV continuous-angle query inputs in the USV-UAV collaboration platform. The retrieval elapsed time test shows that the feature fusion retrieval can reduce the time by 73% and improve the accuracy rate by about 3%.

The ship ReID technology based on the USV-UAV collaboration is the basis of future unmanned supervision systems in the marine economy field. We will focus more on task-oriented USV-UAV control and multi-source data fusion in the future. Like the perspective and computational advantages of USV-UAV collaboration for the ReID task, the task-inspired cooperative control of the UAV and the USV will also provide more diversified data information and opportunities for unmanned systems at sea.

Author Contributions: Conceptualization, W.D. and L.Z.; methodology, W.D. and L.Z.; software, L.Z.; validation, W.D., L.Z. and Y.W.; formal analysis, W.D. and L.Z.; investigation, L.Z.; resources, Y.W.; data curation, L.Z.; writing—original draft preparation, W.D. and L.Z.; writing—review and editing, W.D. and S.W.; visualization, W.D. and L.Z.; supervision, Y.W. and S.W.; project administration, Y.W.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Science and Technology Project of Shenzhen under Grant JCYJ20200109113424990; Marine Economy Development Project of Guangdong Province under Grant GDNRC [2020]014; CCF-Baidu Apollo Joint Development Project Fund.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, Z.; Ni, G.; Xu, Y. Trajectory prediction based on AIS and BP neural network. In Proceedings of the 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 11–13 December 2020; pp. 601–605.
2. Zhao, L.; Yang, J.; Shi, G. A Correction Method for Time of Ship Trajectories Based on AIS. In Proceedings of the 1st International Conference on Big Data Research, Osaka, Japan, 22–24 October 2017; pp. 83–88.
3. Zheng, Z.; Zheng, L.; Yang, Y. A discriminatively learned CNN embedding for person reidentification. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2017**, *14*, 1–20. [[CrossRef](#)]

4. Li, J.; Wang, J.; Tian, Q.; Gao, W.; Zhang, S. Global-local temporal representations for video person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3958–3967.
5. Wang, Y.; Peng, J.; Wang, H.; Wang, M. Progressive learning with multi-scale attention network for cross-domain vehicle re-identification. *Sci. China Inf. Sci.* **2022**, *65*, 160103. [[CrossRef](#)]
6. Lian, J.; Wang, D.; Zhu, S.; Wu, Y.; Li, C. Transformer-based attention network for vehicle re-identification. *Electronics* **2022**, *11*, 1016. [[CrossRef](#)]
7. Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; Tian, Q. Person re-identification in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1367–1376.
8. Suh, Y.; Wang, J.; Tang, S.; Mei, T.; Lee, K.M. Part-aligned bilinear representations for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 402–419.
9. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 480–496.
10. Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; Cristani, M. Person re-identification by symmetry-driven accumulation of local features. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2360–2367.
11. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; Yang, Y. Improving person re-identification by attribute and identity learning. *Pattern Recognit.* **2019**, *95*, 151–161. [[CrossRef](#)]
12. Matsukawa, T.; Suzuki, E. Person re-identification using CNN features learned from combination of attributes. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 2428–2433.
13. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3754–3762.
14. Sun, Y.; Zheng, L.; Deng, W.; Wang, S. Svdnet for pedestrian retrieval. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3800–3808.
15. Chen, D.; Xu, D.; Li, H.; Sebe, N.; Wang, X. Group consistent similarity learning via deep crf for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8649–8658.
16. Wang, Y.; Wang, L.; You, Y.; Zou, X.; Chen, V.; Li, S.; Huang, G.; Hariharan, B.; Weinberger, K.Q. Resource aware person re-identification across multiple resolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8042–8051.
17. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Mask-guided contrastive attention model for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1179–1188.
18. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
19. Variator, R.R.; Shuai, B.; Lu, J.; Xu, D.; Wang, G. A siamese long short-term memory architecture for human re-identification. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 135–153.
20. Ye, M.; Liang, C.; Wang, Z.; Leng, Q.; Chen, J. Ranking optimization for person re-identification via similarity and dissimilarity. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 1239–1242.
21. Zhong, Z.; Zheng, L.; Cao, D.; Li, S. Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1318–1327.
22. Zeng, G.; Wang, R.; Yu, W.; Lin, A.; Li, H.; Shang, Y. A transfer learning-based approach to maritime warships re-identification. *Eng. Appl. Artif. Intell.* **2023**, *125*, 106696. [[CrossRef](#)]
23. Qiao, D.; Liu, G.; Dong, F.; Jiang, S.-X.; Dai, L. Marine vessel re-identification: A large-scale dataset and global-and-local fusion-based discriminative feature learning. *IEEE Access* **2020**, *8*, 27744–27756. [[CrossRef](#)]
24. Ghahremani, A.; Alkanat, T.; Bondarev, E.; de With, P.H. Maritime vessel re-identification: Novel VR-VCA dataset and a multi-branch architecture MVR-net. *Mach. Vis. Appl.* **2021**, *32*, 1–14. [[CrossRef](#)]
25. Groot, H.G.; Zwemer, M.H.; Wijnhoven, R.G.; Bondarau, E. Vessel-speed enforcement system by multi-camera detection and re-identification. In Proceedings of the 15th International Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP2020), Valetta, Malta, 27–29 February 2020; pp. 268–277.
26. Ribeiro, R.; Cruz, G.; Matos, J.; Bernardino, A. A data set for airborne maritime surveillance environments. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *29*, 2720–2732. [[CrossRef](#)]
27. Ghahremani, A.; Kong, Y.; Bondarev, E.; de With, P.H. Towards parameter-optimized vessel re-identification based on IORnet. In Proceedings of the Computational Science–ICCS 2019: 19th International Conference, Faro, Portugal, 12–14 June 2019; pp. 125–136.
28. Li, W.; Ge, Y.; Guan, Z.; Ye, G. Synchronized Motion-Based UAV–USV Cooperative Autonomous Landing. *J. Mar. Sci. Eng.* **2022**, *10*, 1214. [[CrossRef](#)]
29. Shao, G.; Ma, Y.; Malekian, R.; Yan, X.; Li, Z. A novel cooperative platform design for coupled USV–UAV systems. *IEEE Trans. Ind. Inform.* **2019**, *15*, 4913–4922. [[CrossRef](#)]

30. Huang, T.; Chen, Z.; Gao, W.; Xue, Z.; Liu, Y. A USV-UAV Cooperative Trajectory Planning Algorithm with Hull Dynamic Constraints. *Sensors* **2023**, *23*, 1845. [[CrossRef](#)]
31. Li, W.; Ge, Y.; Guan, Z.; Gao, H.; Feng, H. NMPC-based UAV-USV cooperative tracking and landing. *J. Frankl. Inst.* **2023**, *360*, 7481–7500. [[CrossRef](#)]
32. Yao, P.; Gao, Z. UAV/USV Cooperative Trajectory Optimization Based on Reinforcement Learning. In Proceedings of the 2022 China Automation Congress (CAC), Xiamen, China, 25–27 November 2022; pp. 4711–4715.
33. Wei, W.; Wang, J.; Fang, Z.; Chen, J.; Ren, Y.; Dong, Y. 3U: Joint design of UAV-USV-UUV networks for cooperative target hunting. *IEEE Trans. Veh. Technol.* **2022**, *72*, 4085–4090. [[CrossRef](#)]
34. Lewicka, O.; Specht, M.; Stateczny, A.; Specht, C.; Dardanelli, G.; Brčić, D.; Szostak, B.; Halicki, A.; Stateczny, M.; Widźgowski, S. Integration data model of the bathymetric monitoring system for shallow waterbodies using UAV and USV platforms. *Remote Sens.* **2022**, *14*, 4075. [[CrossRef](#)]
35. Han, Y.; Ma, W. Automatic Monitoring of Water Pollution based on the Combination of UAV and USV. In Proceedings of the 2021 IEEE 4th International Conference on Electronic Information and Communication Technology (ICEICT), Xi'an, China, 18–20 August 2021; pp. 420–424.
36. Wang, Y.; Liu, W.; Liu, J.; Sun, C. Cooperative USV-UAV marine search and rescue with visual navigation and reinforcement learning-based control. *ISA Trans.* **2023**, *137*, 222–235. [[CrossRef](#)]
37. Wu, J.; Li, R.; Li, J.; Zou, M.; Huang, Z. Cooperative unmanned surface vehicles and unmanned aerial vehicles platform as a tool for coastal monitoring activities. *Ocean. Coast. Manag.* **2023**, *232*, 106421. [[CrossRef](#)]
38. Li, Y.; Li, S.; Zhang, Y.; Zhang, W.; Lu, H. Dynamic Route Planning for a USV-UAV Multi-Robot System in the Rendezvous Task with Obstacles. *J. Intell. Robot. Syst.* **2023**, *107*, 52. [[CrossRef](#)]
39. Yu, Y.; Rodríguez-Piñeiro, J.; Shunqin, X.; Tong, Y.; Zhang, J.; Yin, X. Measurement-based propagation channel modeling for communication between a UAV and a USV. In Proceedings of the 2022 16th European Conference on Antennas and Propagation (EuCAP), Madrid, Spain, 27 March–1 April 2022; pp. 01–05.
40. Shao, Z.; Wu, W.; Wang, Z.; Du, W.; Li, C. Seaships: A large-scale precisely annotated dataset for ship detection. *IEEE Trans. Multimed.* **2018**, *20*, 2593–2604. [[CrossRef](#)]
41. Prasad, D.K.; Rajan, D.; Rachmawati, L.; Rajabally, E.; Quek, C. Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 1993–2016. [[CrossRef](#)]
42. ShipSpotting. Available online: www.shipspotting.com (accessed on 19 July 2023).
43. Naderializadeh, N.; Orhan, O.; Nikopour, H.; Talwar, S. Ultra-dense networks in 5G: Interference management via non-orthogonal multiple access and treating interference as noise. In Proceedings of the 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), Toronto, ON, Canada, 24–27 September 2017; pp. 1–6.
44. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Society. Ser. C (Appl. Stat.)* **1979**, *28*, 100–108. [[CrossRef](#)]
45. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
47. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
48. Radenović, F.; Tolias, G.; Chum, O. Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1655–1668. [[CrossRef](#)]
49. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
50. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
51. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
52. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An empirical study of spatial attention mechanisms in deep networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6688–6697.
53. Gu, Y.; Li, C.; Xie, J. Attention-aware generalized mean pooling for image retrieval. *arXiv* **2018**, arXiv:1811.00202.
54. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1116–1124.
55. Liu, X.; Liu, W.; Ma, H.; Fu, H. Large-scale vehicle re-identification in urban surveillance videos. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
56. Teng, S.; Zhang, S.; Huang, Q.; Sebe, N. Viewpoint and scale consistency reinforcement for UAV vehicle re-identification. *Int. J. Comput. Vis.* **2021**, *129*, 719–735. [[CrossRef](#)]
57. Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; Wei, Y. Circle loss: A unified perspective of pair similarity optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6398–6407.

-
58. Chen, H.; Lagadec, B.; Bremond, F. Partition and reunion: A two-branch neural network for vehicle re-identification. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 184–192.
 59. Zhang, X.; Zhang, R.; Cao, J.; Gong, D.; You, M.; Shen, C. Part-guided attention learning for vehicle instance retrieval. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 3048–3060. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.