Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.050879

ARTICLE

Fine-Grained Ship Recognition Based on Visible and Near-Infrared Multimodal Remote Sensing Images: Dataset, Methodology and Evaluation

Shiwen Song, Rui Zhang, Min Hu* and Feiyao Huang

Department of Aerospace Science and Technology, Space Engineering University, Beijing, 101416, China

*Corresponding Author: Min Hu. Email: humin@hgd.edu.cn

Received: 21 February 2024 Accepted: 22 April 2024 Published: 20 June 2024

ABSTRACT

Fine-grained recognition of ships based on remote sensing images is crucial to safeguarding maritime rights and interests and maintaining national security. Currently, with the emergence of massive high-resolution multi-modality images, the use of multi-modality images for fine-grained recognition has become a promising technology. Fine-grained recognition of multi-modality images imposes higher requirements on the dataset samples. The key to the problem is how to extract and fuse the complementary features of multi-modality images to obtain more discriminative fusion features. The attention mechanism helps the model to pinpoint the key information in the image, resulting in a significant improvement in the model's performance. In this paper, a dataset for fine-grained recognition of ships based on visible and near-infrared multi-modality remote sensing images has been proposed first, named Dataset for Multimodal Fine-grained Recognition of Ships (DMFGRS). It includes 1,635 pairs of visible and near-infrared remote sensing images divided into 20 categories, collated from digital orthophotos model provided by commercial remote sensing satellites. DMFGRS provides two types of annotation format files, as well as segmentation mask images corresponding to the ship targets. Then, a Multimodal Information Cross-Enhancement Network (MICE-Net) fusing features of visible and near-infrared remote sensing images, has been proposed. In the network, a dual-branch feature extraction and fusion module has been designed to obtain more expressive features. The Feature Cross Enhancement Module (FCEM) achieves the fusion enhancement of the two modal features by making the channel attention and spatial attention work cross-functionally on the feature map. A benchmark is established by evaluating state-of-the-art object recognition algorithms on DMFGRS. MICE-Net conducted experiments on DMFGRS, and the precision, recall, mAP0.5 and mAP0.5:0.95 reached 87%, 77.1%, 83.8% and 63.9%, respectively. Extensive experiments demonstrate that the proposed MICE-Net has more excellent performance on DMFGRS. Built on lightweight network YOLO, the model has excellent generalizability, and thus has good potential for application in real-life scenarios.

KEYWORDS

Multi-modality dataset; ship recognition; fine-grained recognition; attention mechanism

1 Introduction

Ship detection based on remote sensing images refers to the use of modern technology to extract the ship's position from the image taken by a remote sensing satellite and determine its category, which can be used to manage ships in the sea area and improve the level of maritime traffic management. Based on ship detection, the fine-grained recognition of ships will be further refined to carry out a finer division into sub-classes. For remote sensing ship recognition tasks, based on the classical image detection framework, ship recognition methods based on manual features and deep learning have been proposed. In the early stages, low-level global geometric features such as scale, aspect ratio and shape were used as the basis for ship recognition algorithms [1–3]. With the application and excellent performance of deep learning in computer vision, ship recognition algorithms based on deep learning have gradually become a research hotspot for the detection of ships with the characteristics of small objects, rotating objects and complex backgrounds [4–8]. In addition, multi-scale is one of the key points of ship recognition [9,10].

The current mainstream ship recognition methods are mainly based on single-source remote sensing images, such as SAR (Synthetic Aperture Radar), visible images, etc. [11,12]. With the development of multimodal image fusion technology and the emergence of multimodal cameras, some studies have gradually applied them to the object recognition field [13–16]. Multi-spectral images can provide combined information to make object detection and recognition applications more reliable and robust in real-world scenarios [17]. In the application of the multimodal image-based object recognition framework, there are two main branches. First, the image fusion method is used to generate a fused image with higher image quality instead of the source image as the input of the object recognition framework. Based on the better visual effect and richer information of the fused image, a better recognition performance is obtained. The second is the combination of the visible and infrared image fusion algorithm with the object recognition algorithm. The image fusion algorithm and the object recognition algorithm are placed in one framework. After the feature extraction and fusion of visible and infrared images, the intermediate fusion information of the image fusion is directly utilized by the object recognition stage, omitting the step of reconstructing the fused image.

Fine-grained recognition of objects mainly solves the problem of fine-grained classification to distinguish different subclasses under the same category. Ship fine-grained recognition based on ship detection will further refine the class recognition of ships with finer subclasses. HRSC2016 classified the recognition of ships into three levels (L1–L3) [18]. Fine-grained image classification is very hard due to the small granularity of the classification and the subtle variation within the target class, so the detailed information contained in the image has a significant impact on the effectiveness of fine-grained recognition.

Remote sensors receive light reflected or emitted from ground objects and convert it into electrical signals, which are processed to produce digital images. Visible images contain red, green, and blue bands with a spectral range of 400–700 nm, while near-infrared (NIR) images are in the spectral range of 700–1,100 nm. The quality of visible images is greatly affected by weather conditions, such as haze, smog, and fog. These weather conditions make the light form scattering and attenuates the contrast of the captured image, causing the image to lose detailed information. The NIR light has a high penetration ability for fog, compared to visible light, in this case, the NIR image provides higher contrast with richer texture details. Based on the complementarity between NIR remote sensing images and visible remote sensing images, some object recognition algorithm models based on dual-modality remote sensing images have appeared to obtain better recognition results.

The most essential aspect of recognition methods based on multimodal remote sensing images is how to utilise the complementary information in the multimodal data in order to benefit as much as possible from each modality [19]. Since Bahdanau, Cho and Bengio in 2015 used attention mechanisms for deep learning tasks, a wide variety of variants of attention mechanisms have emerged, which greatly improve the efficiency of visual information processing and also optimise the performance of visual tasks [20]. In deep neural networks, the implementation of the attention mechanism can be seen as a dynamic weighting operation. The attention mechanism obtains weight parameters based on the input information, and in turn, applies the weight parameters to the input information to achieve the purpose of focusing on the key regions of the input information. Attention mechanisms can adjust the observation of more informative features based on their relative importance, allowing the algorithm to focus on the most relevant parts of the input, shifting from focusing on global features to focusing on key features, thus conserving resources and obtaining the most effective information rapidly [20].

A large number of remote sensing image datasets have been applied to various tasks in the field of object recognition, including NWPU VHR-10 [21] for geospatial object detection, DOTA [22] for object detection in aerial images, and VEDAI [23] for vehicle detection in aerial images. In the field of ship recognition, several public datasets for ship recognition in remote sensing images have been proposed successively, such as HRSC2016 [18], DOSR [6], FGSCR-42 [24], etc. HRSC2016 organizes the ship model into a tree structure, which is composed of three levels: (L1–L3), which is the first public remote sensing dataset for ship recognition. DOSR supports the research of ship recognition in four typical scenarios: Chaotic scene, dense scene, small scene and large-scale variance scene. FGSCR-42 is the first public dataset published specifically for fine-grained ship classification, containing 42 distinct categories from 10 major ship classes.

The requirements for more fine-grained recognition and multi-modality image-based interpretation put new requirements on the dataset:

(1) The unity of image data quality. To improve the generalization ability and robustness of the model, the remote sensing image used in the training sample should be as close to the original remote sensing image as possible in terms of image quality.
(2) The diversity of image data sources. The application of image fusion algorithms in the field of object recognition has gradually attracted attention. The research of multi-modality remote sensing image object recognition based on image fusion is also inseparable from the support of multi-modality image datasets.
(3) The richness of the object image features. The research of fine-grained recognition requires the dataset to have rich feature information of the object and provide the object image under different environmental backgrounds, multi-angle, multi-phase shooting, and different weather conditions.
(4) The fineness of sample labelling. Provide more fine-grained annotation information, select the rotation annotation method with the higher fit degree to the object, and provide pixel-level segmentation annotation of the object image.

Object detection and recognition based on visible and NIR images have been developed and applied, but there are fewer studies and applications in the field of fine-grained recognition of ships in remote sensing images, and essential datasets are lacking. On the other hand, how to extract the complementary features of multimodal images and integrate the information between different modalities, and how to design a more effective cross-modal fusion mechanism to obtain more discriminative fusion features and improve the recognition effect is also a key issue.

The main contributions of our work are given as follows:

(1) A dataset, DMFGRS, has been established for the task of fine-grained recognition of ships based on multi-modality remote sensing imagery. DMFGRS provides a total of 1,635 digital orthophotos map (DOM) from commercial remote sensing satellites, including 3,689 samples

of ship objects in 20 categories. DMFGRS provides true-color images with resolutions of 0.5, 0.75, and 0.8 m for each category as well as near-infrared (NIR) images that correspond to the true-color images before and after the fusion process. NIR images that correspond to the true-color images before and after processing. At the same time, DMFGRS provides two kinds of annotation format files and gives the corresponding segmentation mask images of the ship objects.

(2) A multimodal information cross-enhancement network for fine-grained recognition of ships, called MICE-Net, is proposed. MICE-Net enhances the effectiveness of ship recognition based on single-modality remote sensing images by fusing features and information from visible-NIR remote sensing images through a dual-branch feature extraction and fusion network.

(3) A Feature cross enhancement module is designed to achieve a more informative and critically focused feature map. Cross-acting on both visible and NIR modal features through the attention mechanism to achieve mutual guidance of feature focus information and obtain enhanced and fused features.

(4) The performance of popular object detection algorithms on the DMFGRS is evaluated to provide a benchmark for future research. The effectiveness of MICE-Net is verified and superior performance is achieved on the DMFGRS dataset with MICE-Net outperforming the benchmark.

2 Related Work

2.1 Dataset

Since the emergence of object recognition algorithms, datasets have played an increasingly important role in data-driven research. DOTA is a large-scale aerial image object dataset used to advance object recognition research in Earth vision, which is favoured by researchers because it has enough images, categories and object instances [22]. In addition, DOTA uses directional bounding boxes to mark objects, which can better surround items and distinguish crowded objects. LEVIR is a large remote sensing building change recognition dataset consisting of a large number of high-resolution Google Earth images of 0.2 to 1.0 m pixels, covering most types of ground features of the human inhabited environment [25]. NWPU VHR-10 includes 10 types of geospatial objects, which are characterized by a balanced sample size among the various categories [21]. DIOR is a large-scale optical remote sensing image dataset for object detection, which not only has a considerable number of object instances and image numbers but also has a wide range of object scale changes [26]. FAIR1M is currently the largest remote sensing images dataset for fine-grained recognition, containing more than 15,000 images with resolutions better than 1 m and sizes ranging from thousands to tens of thousands of pixels, with more than 1 million finely labelled, multi-angle distribution objects, covering hundreds of typical cities, towns, as well as commonly used airports, ports, etc. [27]. It also provides data for the same region and different phases, which is a set of multi-temporal, multi-resolution and multi-factor remote sensing image standardized sample sets. VEDAI is used for vehicle detection in aerial images and supports the research of recognition algorithms in unconstrained environments [23]. DroneVehicle is oriented to the visual task of vehicle detection and counting, and its shooting environment covers day and night [28]. It is worth mentioning that the VEDAI and DroneVehicle datasets not only provide visible remote sensing images but also provide infrared images after registration.

As shown in Table 1, the above commonly used datasets for remote sensing object detection mostly focus on common broad categories of objects, such as aircraft, ships, oil tanks, automobiles, etc., but do not carry out further detailed annotations within each category, which is not enough to support fine-grained recognition research. There are two types of remote-sensing object labeling formats: Oriented Bounding Box (OBB) and Horizontal Bounding Box (HBB). Although VEDAI and DroneVehicle focus on vehicle categories, the number of categories in the dataset is too small, so the universality and generalization ability of the dataset are weak. In addition, the above remote sensing image datasets are not aimed at remote sensing image ship objects, and ship recognition has different characteristics from other remote sensing object recognition. For example, ship objects are mostly axisymmetric structures, and the shooting environment is susceptible to weather and illumination. Therefore, these datasets are not enough to support the research of ship fine-grained recognition tasks in remote sensing images.

Table 1: Datasets for remote sensing image object detection

Dataset	Object	#Categories	Annotation	Image	#Instances
DOTA	Aircraft, ships, oil tanks, bridges, etc.	15	OBB	Visible	188,282
LEVIR	Aircraft, ships, oil tanks, etc.	3	HBB	Visible	11,028
NWPU VHR-10	Airplanes, ships, oil tanks, baseball fields, etc.	10	HBB	Visible	3,651
DIOR	Airplanes, airports, bridges, chimneys, etc.	20	HBB	Visible	23,476
FAIR1M	Aircraft, ships, vehicles, etc.	37	OBB	Visible	1.02M
VEDAI	Vehicles	3	OBB	Visible + infrared	1,268
DroneVehicle	Vehicles	5	OBB	Visible + infrared	953,087
DMFGRS (ours)	Ships	20	OBB	Visible + NIR	3,689

The fine-grained recognition of ship objects is mainly to solve the problem of fine-grained classification based on detection, and the purpose is to distinguish different subclasses under the same category. HRSC2016 has divided the recognition of ships into three levels (L1–L3). Due to the small granularity of classification, it is very difficult to classify fine-grained images. Therefore, the research of fine-grained recognition algorithms has more stringent requirements on datasets, and the lack of clarity or image information will seriously affect the recognition effect.

Nowadays, as shown in Table 2, many datasets have made contributions in the field of ship fine-grained recognition. The HRSC2016 was published in 2016 and includes 2,976 objects in 3 broad and 27 subcategories. FGSC-23 is designed to meet the research requirements of ship object fine-grained recognition tasks based on deep learning [29]. The open high-resolution Google Earth and GF-2 satellite water surface scene remote sensing images containing ship objects are collected, and a high-resolution optical remote sensing image ship object fine recognition data set is constructed. FGSCR-42 consists of 9,320 optical satellite images of different spatial resolutions, ranging in size from approximately 50 × 50 to approximately 1,500 × 1,500 pixels, containing a total of 9,320 ship instances in 42 different classes from 10 major ship classes [25]. DOSR contains 1,066 optical remote sensing images and 6,127 ship instances, with image sizes ranging from 600 to 1,300 pixels and resolutions ranging from 0.5 to 2.5 m. There are a variety of scenes in DOSR, including clutter scenes, dense scenes, small scenes and large-scale change scenes [6].

Table 2: Datasets for ship fine-grained recognition

Dataset	#Categories	Type	Image	#Images	Level
DSCR [30] (2019)	7	Classification	Visible	6,685	L2
FGSC-23 (2020)	23	Classification	Visible	4,080	L2
FGSCR-42 (2021)	42	Classification	Visible	9,320	L3
DOSR (2022)	20	Object detection	Visible	1,066	L2
HRSC2016 (2016)	3/27	Object detection	Visible	1,061	L3
DMFGRS (2023) (ours)	20	Object detection, fine-grained recognition, segmentation, image fusion	Visible + NIR	1,635	L3

2.2 *Object Recognition Algorithms Based on Multi-Modality Image*

Reference [31] proposed a differential maximum loss function for extracting complementary features in visible and infrared images. The loss function directs the learning direction of the two basic neural networks and maximizes the difference between the features of the two basic neurons to extract complementary and diverse features. Reference [28] constructed a large-scale unmanned aerial vehicle (UAV-based) visible infrared vehicle recognition dataset, called DroneVehicle, which collected 28,439 visible infrared image pairs covering urban roads, residential areas, car parks and other scenes. An uncertainty-aware cross-modal vehicle recognition (UACMDet) framework is also proposed to extract complementary information from cross-modal images to significantly improve the recognition performance under low-light conditions. The framework includes an Uncertainty Awareness Module (UAM) to quantify the uncertainty weight of each modality, which is calculated from the modal cross IoU (Intersection over Union) and visible illumination values, in addition to a light-aware cross-modal. Non-Maximum Suppression algorithm is designed for better integration of modality-specific information in the inference stage. Some other researchers have conducted in-depth studies on object recognition based on the fusion of visible and infrared images [15,32]. Most of the existing methods for solving Red-Green-Blue (RGB) and Thermal (T) salient object detection (SOD) try to integrate multimodal information through various fusion strategies or reduce modal differences through unidirectional or undifferentiated bidirectional interactions, but in some challenging scenarios, these methods have little success, therefore, Xie et al. [33] proposed a new RGB-T network that includes an interaction branch to indirectly connect visible and thermal modalities, uses a double bidirectional interaction (DBI) module consisting of a forward interaction block (FIB) and a backward interaction block (BIB) to reduce cross-modal disparities, and introduces a multi-scale feature enhancement and fusion (MSFEF) module that fuses the multimodal features taking into account the internal gaps of different modalities. Finally, the cascaded decoder and a cross-level feature enhancement (CLFE) module are used to generate high-quality saliency maps.

Research on object recognition based on multimodal remote sensing images mainly focuses on visible and infrared images. Reference [34] proposed YOLOrs for real-time object recognition in multimodal remote sensing imagery. YOLOrs can detect objects at multiple scales, utilize a small receptive field to identify small objects, and predict object orientation. In addition, YOLOrs introduces a novel mid-level fusion architecture that makes it suitable for multimodal aerial images. Reference [35] proposed a new, lightweight multispectral feature fusion method based on the idea of preserving and enhancing modality-specific features and selecting modality-shared features from visible and thermal infrared modalities with common and differential modal focus, called Cross-Modality Attentive

Feature Fusion (CMAFF), the common selection sub-module and the differential enhancement submodule are the two unique parts of enhancement and selection of features in CMAFF, enabling the detector to achieve significant performance improvements while keeping the overhead small.

3 Dataset for Multimodal Fine-Grained Recognition of Ships (DMFGRS)

As shown in Table 2, most of the existing datasets for ship fine-grained recognition are oriented to classification tasks rather than detection tasks. They only provide the category information of ship objects, and the location information of ships is not given. However, in practical application, classification and location are equally important tasks for remote sensing image ship detection and recognition in sea areas and near shore. Therefore, it is very important and necessary to establish the dataset of remote sensing image ship fine-grained recognition objects. At the same time, most of the current datasets for the ship fine-grained recognition task only provide the visible light format, which brings obstacles to the research of multi-modality image fusion applications in object recognition.

Compared with the above datasets, DMFGRS provides visible and near-infrared images and supports the research of object recognition tasks of a single source (based on true color images or near-infrared images) and multi-source remote sensing images. The labelling level of each category reaches the L3 level, which supports the task of ship fine-grained recognition. At the same time, the segmentation mask image of the object instance is provided to support segmentation task research. Visible light and near-infrared complete registration processing, DMFGRS also supports image fusion research and its application in recognition tasks. Fig. 1 shows the difference in annotation information between DMFGRS and the other datasets. (a) The dataset for remote sensing object detection only distinguishes between ship and non-ship categories; (b) the ship-oriented fine-grained classification dataset only provides specific ship category information but does not label the object location; (c) the ship-oriented fine-grained recognition dataset, namely DMFGRS, not only provides fine-grained category information of ship objects but also accurately marks the location of object instances.

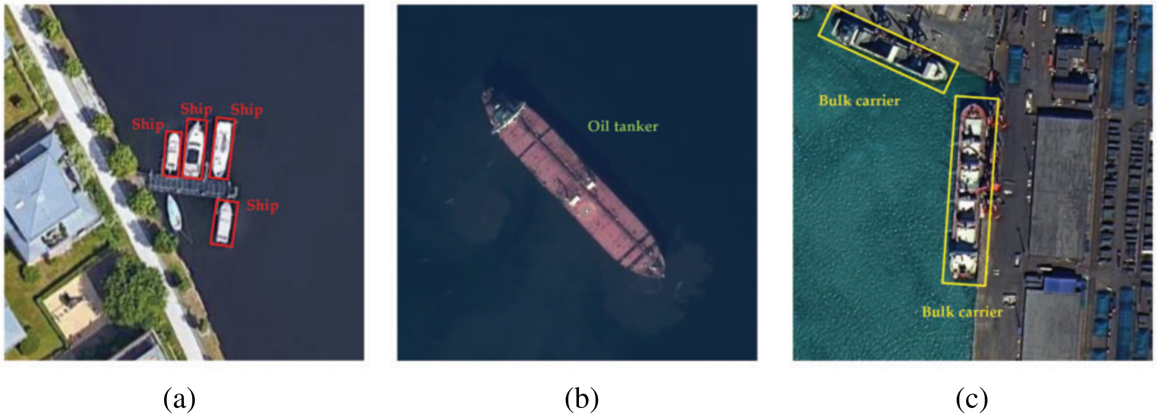


Figure 1: Comparison of sample labeling. (a) Dataset for remote sensing object detection. (b) Dataset for fine-grained ship classification. (c) Dataset for ship fine-grained recognition (ours)

3.1 Establishment of DMFGRS

3.1.1 Images Collection

In order to ensure the authenticity and reliability of the images in the dataset, as well as to increase the diversity of image data, the DMFGRS ship satellite image is derived from the orthographic remote sensing image (DOM) product and collected from commercial satellites, including two images in the visible (red, green and blue) band and near-infrared band. The image format is TIFF. The image of the same location has the difference between the time of shooting, the angle of shooting and the source satellite.

The DMFGRS images are all cut from the original satellite images, and the information contained in them is closer to the real scene. Due to the small difference between training and reasoning images, DMFGRS, as a network model trained by the training dataset, will relatively reduce the impact of post-processing when applied to the recognition of real remote sensing images. The network can have better migration ability and generalization ability, which is more suitable for the practical application scenarios of intelligent interpretation of on-orbit images.

3.1.2 Category Selection

DMFGRS selected 20 classic ship objects, and the image examples of some categories are shown in Fig. 2, where each class shows two visible light images with different resolutions. According to the ship recognition level assigned by HRSC2016, all categories belong to the L3 level, which meets the requirements of fine-grained recognition. Examples of visible and near-infrared images with different resolutions are shown in Fig. 3, where 0.5_RGB means a visible image with a resolution of 0.5 m. The aspect ratio of each category in DMFGRS is shown in the Fig. 4. It can be seen that the aspect ratio of the ship object is relatively large, and the minimum is also above 3. Most of them are concentrated in the range of 4 to 7, and the maximum aspect ratio is close to 8.

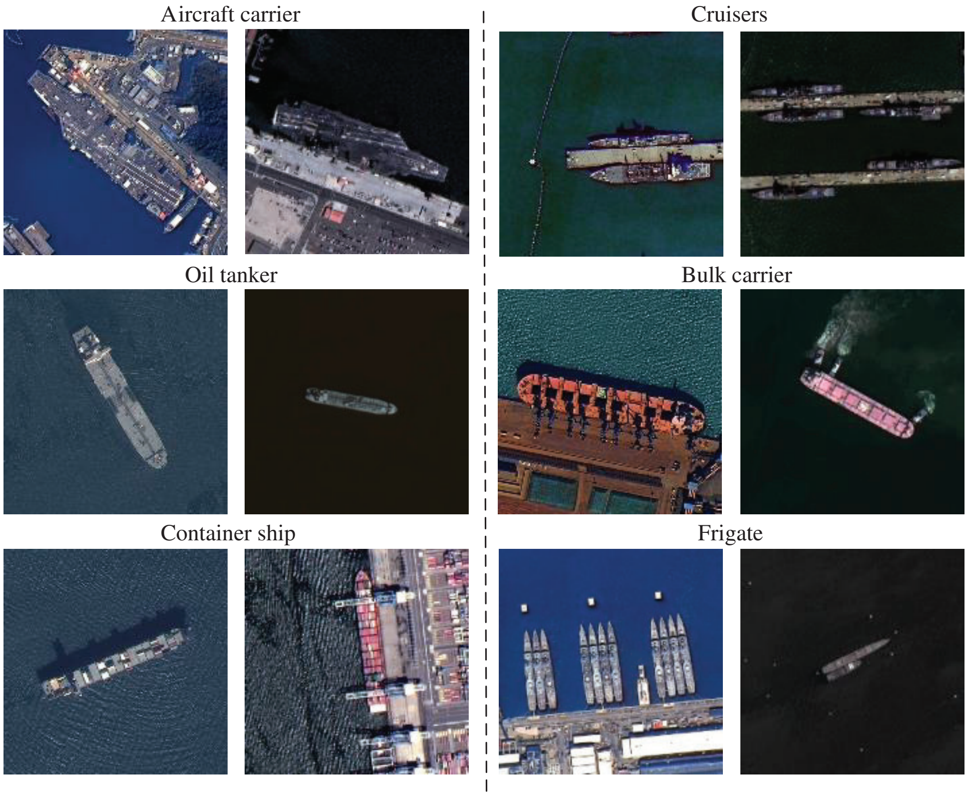
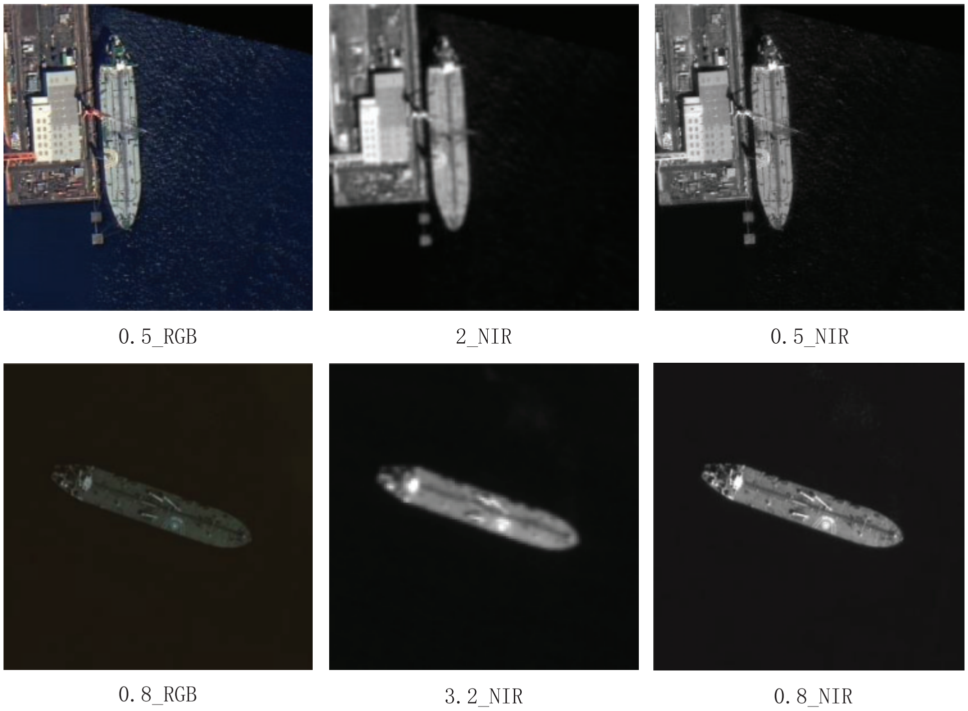


Figure 2: Examples of visible images (two visible images of different resolution)



0. 5_RGB

2_NIR

0. 5_NIR

0. 8_RGB

3. 2_NIR

0. 8_NIR

Figure 3: Examples of visible and NIR images

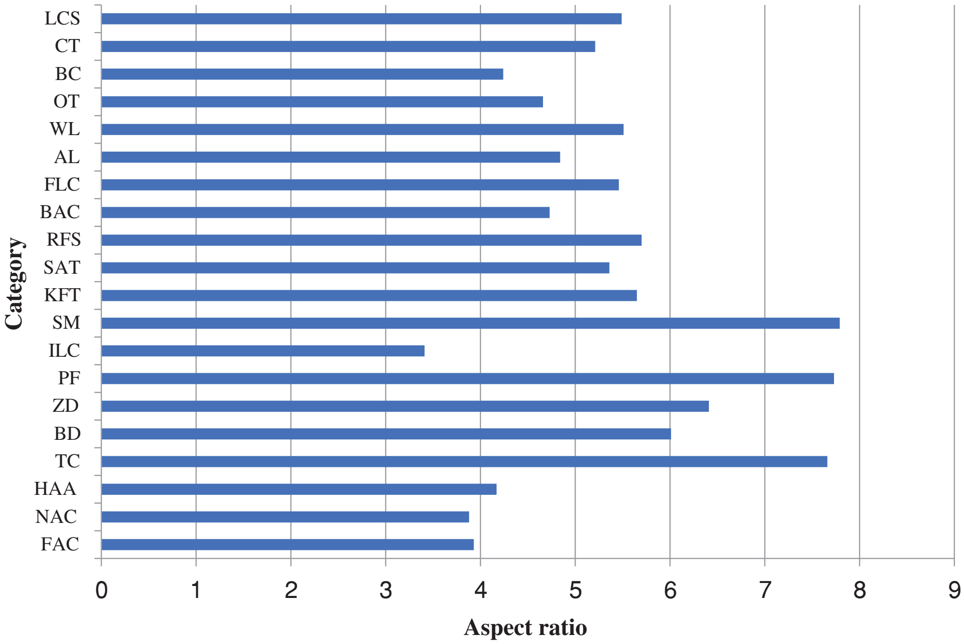


Figure 4: Aspect ratio for each category

3.1.3 Annotation Method

Different from the flat view shooting angle of objects in natural scenes, the shooting angle of remote sensing image is overlooking, so its object does not have a fixed direction like the object in natural scenes but has non-orientation. In addition, ships in remote sensing images are usually densely arranged on the port shore. If horizontal boundary boxes is used, there will be partial overlap between boundary boxes, which will undoubtedly mislead the object detection model and reduce its ability to distinguish object boundaries. However, rotating boundary boxes can well avoid this situation. As shown in Fig. 5, Fig. 5a shows the use of horizontal box annotations, due to the dense layout of the object, the overlap between adjacent boxes is high; Fig. 5b shows the use of rotating box annotations significantly mitigated the overlap problem, but only labelled the ship category; Fig. 5c shows the further annotating the class as a frigate based on (b), and (d) not only annotating the position with a rotating box but also refining the ship class to level L3 (ours).

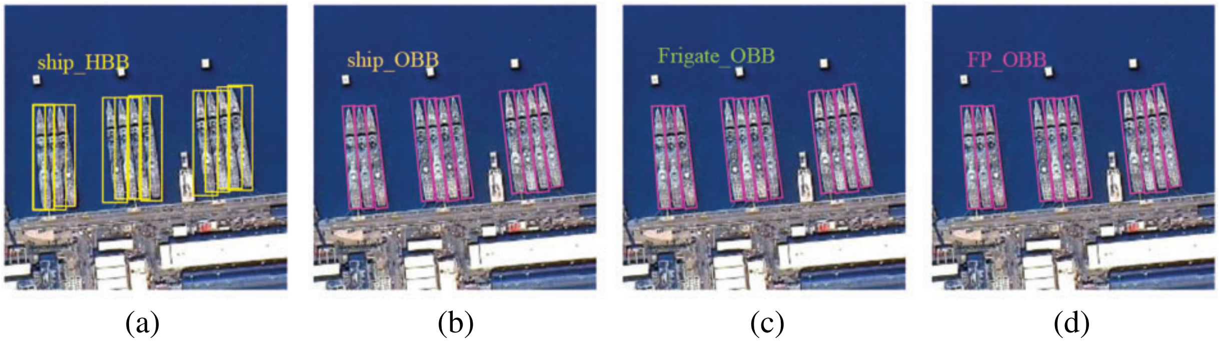
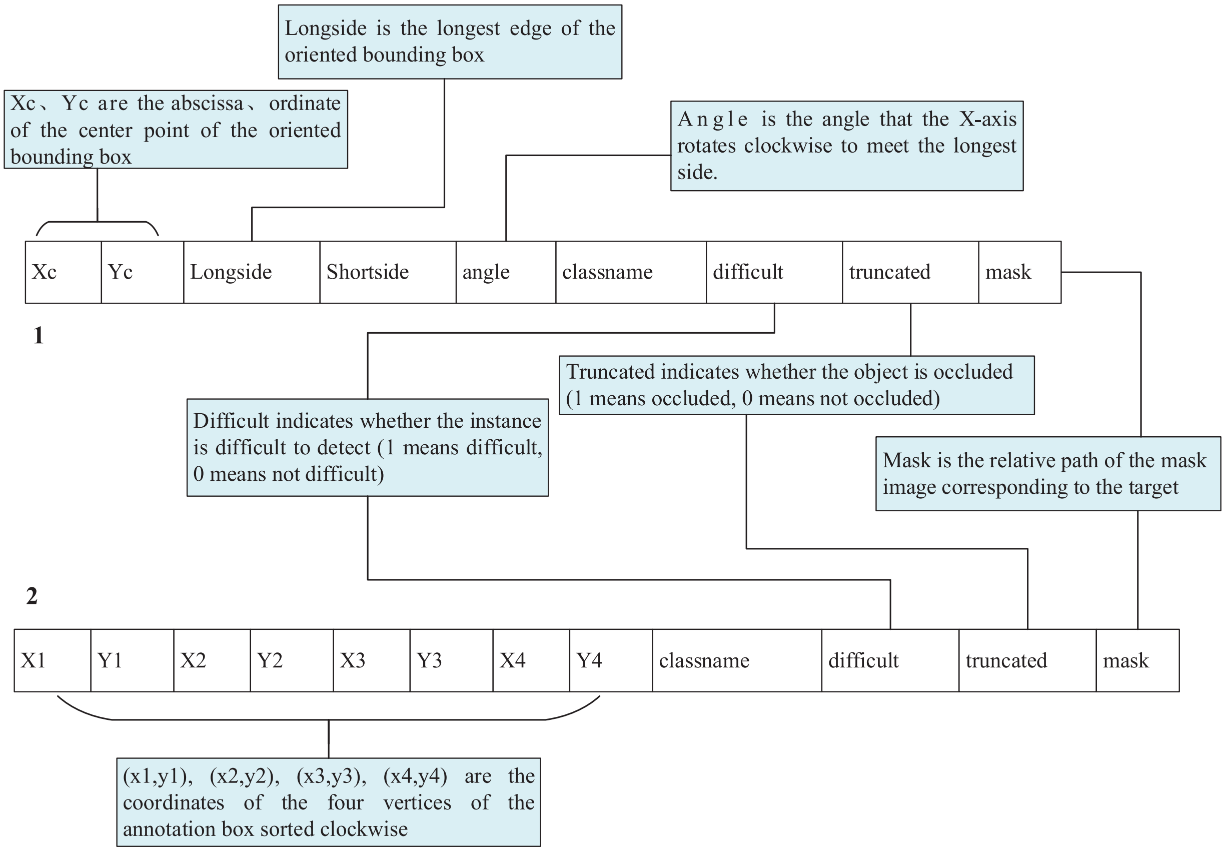


Figure 5: Comparison of annotation methods. (a) Horizontal box annotations. (b) Rotating box annotations. (c) Further annotating the class. (d) Refining the ship class to L3 (ours)

In order to ensure that the bounding box is better fitted with the object instance, avoid the overlap problem as much as possible, and provide more accurate information for the research of object recognition, the dataset adopts the method of rotating annotation and provides two formats of annotation files as shown in Fig. 6. In the DOTA annotation format, as shown in Fig. 7, (X1, Y1) represents the first vertex position of the upper left corner of the object instance, that is, the vertex of the position on the left side of the ship's bow. Therefore, DMFGRS not only gives the exact position of the ship instance but also indicates the direction of the bow.

3.1.4 Segmentation Mask Image Production

DMFGRS uses the polygon annotation box to draw the minimum envelope of the ship instance in the image to generate a binary mask gray pattern. In order to distinguish different types of ship objects, different gray values are used to correspond to different categories, as shown in Table 3. The illustration of the binary mask image is shown in Fig. 8.



Longside is the longest edge of the oriented bounding box

Xc、Yc are the abscissa、ordinate of the center point of the oriented bounding box

Angle is the angle that the X-axis rotates clockwise to meet the longest side.

1

Xc	Yc	Longside	Shortside	angle	classname	difficult	truncated	mask

Truncated indicates whether the object is occluded (1 means occluded, 0 means not occluded)

Difficult indicates whether the instance is difficult to detect (1 means difficult, 0 means not difficult)

Mask is the relative path of the mask image corresponding to the target

2

X1	Y1	X2	Y2	X3	Y3	X4	Y4	classname	difficult	truncated	mask

(x1,y1), (x2,y2), (x3,y3), (x4,y4) are the coordinates of the four vertices of the annotation box sorted clockwise

Figure 6: Annotation information in two formats. 1 is DOTA format, 2 is longsize format

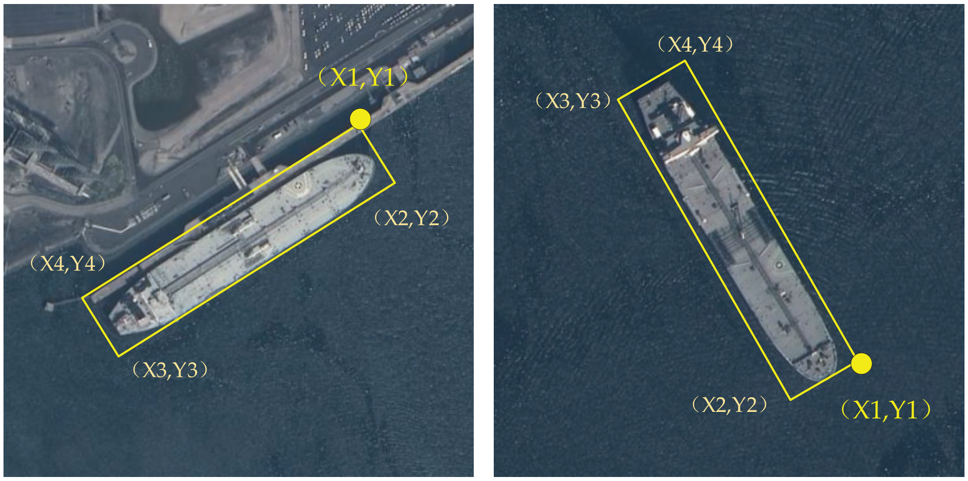


Figure 7: The first vertex position

Table 3: Greyscale values for each category of mask

Class	FAC	NAC	HAA	TC	BD	ZD	PF	ILC	SM	KFT
Gray	10	20	30	40	50	60	70	80	90	100
Class	SAT	RFS	BAC	FLC	AL	WL	OT	BC	CT	LCS
Gray	110	120	130	140	150	160	170	180	190	200

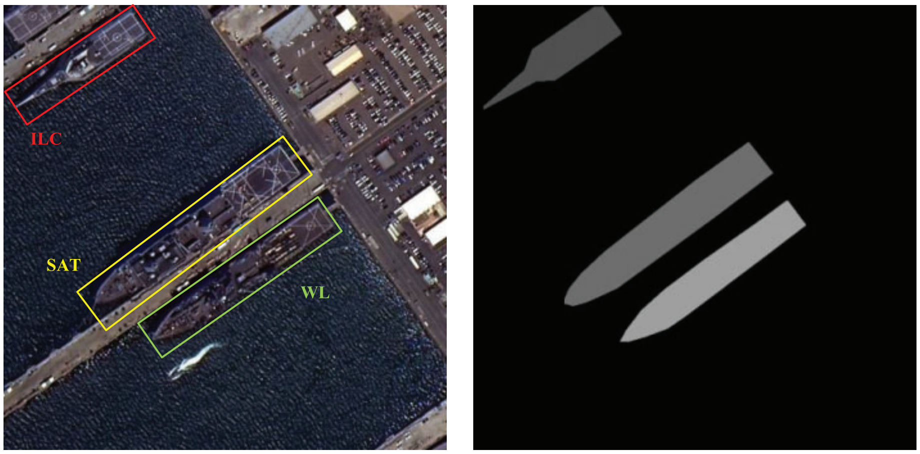


Figure 8: Examples of mask images

3.2 Properties of DMFGRS

3.2.1 Image Size

In order to ensure the integrity of large ship objects in images, and considering the size requirements for fine-grained recognition of objects in slice images, the slice sizes of this data set are designed as follows:

1) 0.5 m resolution visible and near-infrared slice size of 768 × 768.
2) 0.8, 0.75 m resolution visible and near-infrared slice size of 512 × 512.
3) Raw 2 m resolution and 3.2 m resolution near-infrared slices of 192 × 192 and 128 × 128.

It can be seen that the resolution and size of the original NIR slice is a quarter of that of the visible image, and similarly, the coordinates of the labelled object frame are a quarter of that of the visible image.

3.2.2 Spatial Resolution Information

The spatial resolution of the image has an important effect on the object recognition task, especially in the fine-grained recognition task. The difference between subcategories of fine-grained recognition tasks is subtle, and the detailed information has an important impact on the recognition effect. The image with high resolution is more conducive to the feature extraction network to obtain more detailed information. In addition, the use of training data with different resolutions can improve the robustness of the model to identify the same class of objects. The same ship objects will occupy different proportions in images with different resolutions, which also greatly improves the diversity of the datasets. Therefore, DMFGRS collects images in three different resolutions: 0.5 m resolution, 0.75 m resolution and 0.8 m resolution. Three images with different resolutions not only meet the requirements of fine-grained recognition, but also enhance the diversity and robustness of the dataset. DMFGRS also provides raw NIR images corresponding to 0.5, 0.75 and 0.8 m resolutions in quadruple relation, i.e., 2, 3 and 3.2 m resolutions, with the image size and labelled coordinates simultaneously reduced to the corresponding quadruples.

3.2.3 Instance Information

The distribution of the number of instances for each category is shown in Fig. 9. It is clear from the figure that the DMFGRS has an imbalance in the distribution of instances in each category. This imbalance is mainly due to the difference in the number of ships of each category used in reality. Since the number of ships in different categories varies in reality, the number is also affected when collecting ship objects.

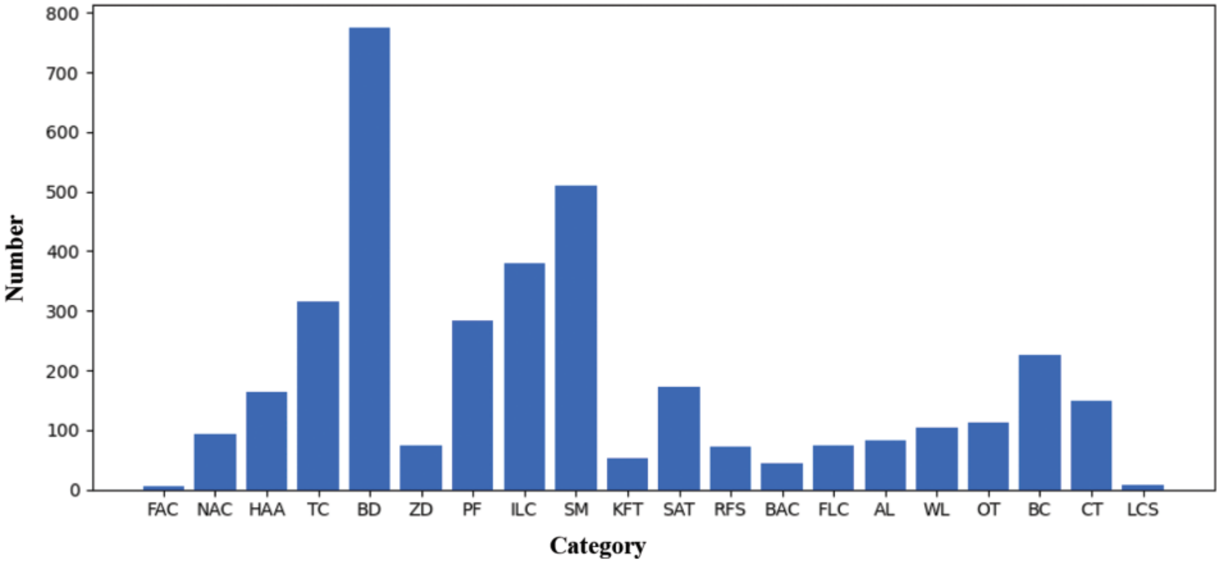


Figure 9: Number of instances per category

Because of the particularity of fine-grained recognition with little difference between classes and big differences within classes, the details of the object have a great impact on the results of fine-grained classification. If the object size in the image is too small, the object detection or fine-grained recognition model will not be able to capture the details in the image, which will affect the final recognition effect. The detection and recognition of small objects have the problems that their visual characteristics are not obvious and the available information is less, which is difficult at the ordinary recognition level, let alone the more demanding precision recognition task.

Compared with medium and large objects, the feature extraction of small objects is also difficult, but the quality of feature extraction will directly affect the recognition effect. Therefore, the existence of small objects undoubtedly brings negative effects on the research of ship fine-grained recognition. To better support the fine-grained recognition research of ships, the ship objects in DMFGRS are screened into medium and large targets according to the provisions of the MS-COCO dataset, and the interference of small objects to the research is eliminated [36]. As shown in Table 4 and Fig. 10, the instance size of each category in DMFGRS is in the medium and large object range.

Table 4: MS-COCO on large, medium and small object size requirements

	Min rectangle area	Max rectangle area
Small object	0×0	32×32
Medium object	32×32	96×96
Large object	96×96	$\infty \times \infty$

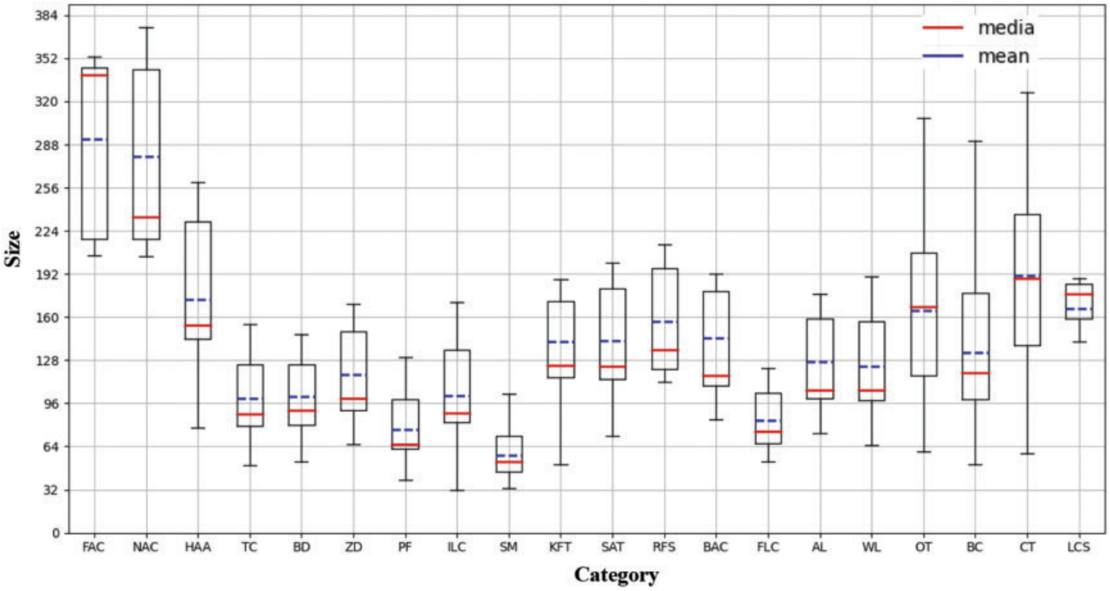


Figure 10: Instance size distribution of each category

Most of the existing public datasets for ship recognition only annotate the complete instance or use negative values to annotate the ship parts beyond the image. These negative labels will be filtered out on the grounds of illegal data in the data processing program of the object recognition algorithm. However, in the actual case of remote sensing ship recognition and recognition, the remote sensing images taken by the satellite do not fully guarantee that the ships at the edge are completely captured and presented, but it does not mean that these ships are not important, or in the actual scene, they cannot be regarded as illegal data and directly filtered out. Therefore, it is necessary to consider the detection and recognition of incomplete ship objects in the real scene.

Datasets that ignore incomplete instances cannot provide a learning experience for the recognition of incomplete ships in real situations. DMFGRS annotates the incomplete ship instances at the edge of the image and supports the object recognition algorithm to reason with the incomplete features of the ship in the learning and training stage, so that the detection and recognition model trained on this dataset has the ability to detect incomplete ship objects, which is more suitable for the actual scene, and also improves the universality of the model.

4 Multi-Modality Information Cross-Enhancement Network (MICE-Net)

4.1 Framework Overview

The overall architecture of the multi-modality information cross-enhancement network (MICE-Net) is shown in Fig. 11. MICE-Net is modified on the YOLO detection network, including three parts: Dual-modality information extraction and fusion module (DMIEF), Neck and Head. DMIEF module extracts features from visible and NIR images, respectively, on the features of two modes at different depths. The feature cross enhancement module (FCEM) is used for feature fusion to obtain more representative intermediate features, which are sent to the subsequent network. In the Neck, the feature pyramid structure is used to fuse high-level features with low-level features to enhance the expression ability of features and improve the performance of subsequent recognition categories and locations. The feature mapping is obtained on the feature layers of two different scales, and the feature mapping is input into the recognition head to obtain the recognition result. The input of the network is the registered visible and NIR remote sensing images, and the output is the category, location and confidence of the detected object.

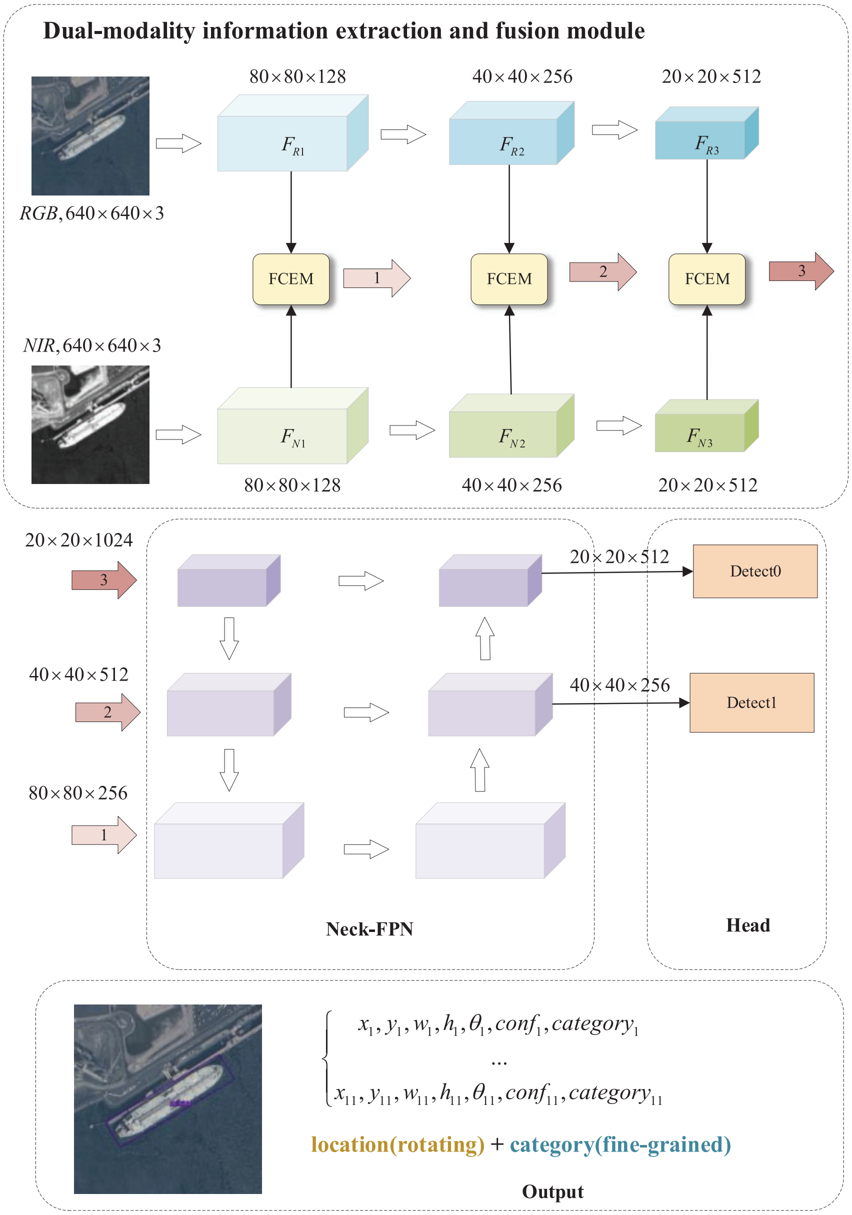


Figure 11: Framework of MICE-Net

In the DMFGRS proposed in the previous section, the two corresponding visible and NIR images have the same size and resolution, the image scene is unobscured, and the two images are registered, so the labels of the two modes are completely matched with each other. In addition, DMFGRS uses the parallelogram labelling method to label each object as a rotating object. Therefore, the loss function of the network also needs to consider the rotation angle. Thus, the loss function of the proposed MICE-Net is designed as follows:

$$L_{\text{recognition}} = \alpha_1 \times L_\theta + \alpha_2 \times L_{\text{class}} + \alpha_3 \times L_{\text{obj}} + \alpha_4 \times L_{\text{box}} \tag{1}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are the weights of each component loss; $L_{recognition}$, L_θ, L_{class}, L_{obj} and L_{box} are the overall loss, angle classification losses, object classification losses, confidence losses and bounding box regression loss.

The CIOU [37] was used for the calculation of the L_{box} L_θ, L_{class} and L_{obj} are calculated in the form of cross-entropy, and the respective calculations are given in the following equation:

$$L_{box} = 1 - IoU + \frac{\rho^2_{(b,b^{gt})}}{c^2} + \alpha v, \alpha = \frac{v}{(1 - IoU) + v}, v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{2}$$

$$L_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log\left(\hat{y}_{ic}\right) + (1 - y_{ic})\log\left(1 - \hat{y}_{ic}\right) \tag{3}$$

$$L_{obj} = -\frac{1}{NG}\sum_{i=1}^{NG}\left[y_i\log\left(\hat{y}_i\right) + (1 - y_i)\log\left(1 - \hat{y}_i\right)\right] \tag{4}$$

$$L_\theta = -\frac{1}{NT}\sum_{i=1}^{NT}\sum_{t=1}^{T} y_{it}\log\left(\hat{y}_{it}\right) + (1 - y_{it})\log\left(1 - \hat{y}_{it}\right) \tag{5}$$

where IoU is the Intersection over Union (IoU) between the prediction bounding box b and the ground truth bounding box b^{gt}, $\rho^2_{(b,b^{gt})}$ is the square of the Euclidean distance between the centroid of b and the centroid of b^{gt}. c is the diagonal length of the smallest box that can contain both b and b^{gt}. α is a weighting factor to balance the effects of different losses. v used to measure the consistency of the aspect ratio of b and b^{gt}. w and h are the width and height of b, while w^{gt} and h^{gt} are the width and height of b^{gt}. In Eq. (3), N is the number of detected objects, y_{ic} is the true label, $\hat{y}_{ic}$ is the probability that object i belongs to category c as predicted by the model. In Eq. (4), NG is the total number of bounding boxes of all grid cells, y_i is the true label, $y_i = 1$ if an object exists within bounding box i, otherwise $y_i = 0$. $\hat{y}_i$ is the confidence score of the bounding box i predicted by the model. L_θ is essentially a category loss, which is calculated by dividing the angles into 180 categories and then proceeding to perform a cross-entropy loss calculation.

4.2 Dual-Modality Information Extraction and Fusion Module (DMIEF)

The framework of dual-modality information extraction and fusion module is shown in the above figure. The basic structure of the DMIEF module is a dual-branch network, and each branch extracts features from shallow to deep layers from visible and NIR remote sensing images. Feature cross enhancement module (FCEM) is used to enhance the fusion of visible and NIR features between the feature maps of two modes with three scales of (80 × 80 × 128), (40 × 40 × 256) and (20 × 20 × 512).

In order to imitate the methods of human visual and cognitive systems, so that the model can focus on the key areas in the input data like humans, the attention mechanism is gradually applied to image processing tasks. The attention mechanism assigns different weights to different positions of the middle layer features so that the neural network can automatically pay attention to and learn those key information, thereby improving the performance and generalization ability of the model. In the object recognition task, the current popular attention mechanisms include the self-attention mechanism, spatial attention mechanism and channel attention mechanism. The application of these attention mechanisms greatly improves the sensitivity of the model to the object location and category, and effectively improves the accuracy of the model.

The attention mechanism is actually a weighting mechanism. It performs different weighting processing on different parts of the input data with different weights. In practical applications, it is usually achieved by calculating the weight vector. For single-modality input images, the attention mechanism acts on the feature map and selects high-value and more critical object regions from a large number of unrelated background regions. Due to the different imaging modes of visible and NIR, the wavelength of near-infrared is larger than that of visible light. Therefore, the attention mechanism will generate different weight vectors on visible and NIR images. The FCEM designed in this paper is different from the previous attention mechanism. As shown in Fig. 12, the attention weights of the two modes not only act on themselves, but also act on each other to achieve the purpose of enhancing the ability of feature expression, so it is called the feature cross enhancement module.

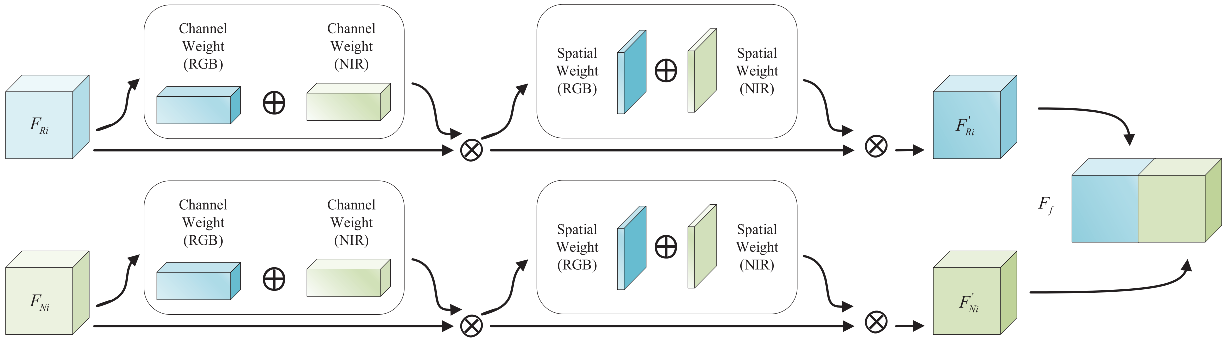


Figure 12: Structural diagram of the proposed feature cross enhancement module

The input of the FCEM is the feature map of visible and NIR modes at three different depths obtained by the Dual-modality information extraction and fusion module in the feature extraction stage: $F_{R\mathrm{i}}$ and $F_{N\mathrm{i}}(i \in 1, 2, 3)$. In order to make the attention mechanism not only guide the feature aggregation and enhancement of visible and NIR feature maps themselves but also guide each others. As shown in Fig. 12, the channel weights of the two modal feature maps are added to obtain the cross-channel weights, and the spatial weights of the two modal feature maps are added to obtain the cross-spatial weights. $F_{R\mathrm{i}}$ is multiplied by the cross-channel weight and the cross-spatial weight in turn, and $F_{N\mathrm{i}}$ repeats the same operation. Finally, the weighted feature maps of the two modes are concat to obtain the fusion result. The feature cross enhancement module is calculated as follows:

$$F'_{R_i} = F_{R_i} \times \left(w_C^N + w_C^R\right) \times \left(w_S^N + w_S^R\right) \tag{6}$$

$$F'_{N_i} = F_{N_i} \times \left(w_C^N + w_C^R\right) \times \left(w_S^N + w_S^R\right) \tag{7}$$

where $F_{R\mathrm{i}}$ and $F_{N\mathrm{i}}(i \in 1, 2, 3)$ are the respective feature maps of visible and NIR; w_C^N and w_C^R are the channel weights of $F_{R\mathrm{i}}$ and $F_{N\mathrm{i}}$; w_S^N and w_S^R are the spatial weights of $F_{R\mathrm{i}}$ and $F_{N\mathrm{i}}$;

The channel weight and spatial weight of visible and NIR feature maps are calculated by channel attention mechanism and spatial attention mechanism, respectively. The calculation process is as follows:

$$w_C^R = f_{CA}\left(F_{R_i}\right) \tag{8}$$

$$w_S^R = f_{SA}\left(F_{R_i} \times w_C^R\right) \tag{9}$$

$$w_C^N = f_{CA}\left(F_{N_i}\right) \tag{10}$$

$$w_S^N = f_{SA}\left(F_{N_i} \times w_C^N\right) \tag{11}$$

where $f_{CA}(\cdot)$ indicates the channel attention mechanism; $f_{SA}(\cdot)$ indicates the spatial attention mechanism.

The channel attention module and the spatial attention module refer to CBAM, the process of which is shown in Figs. 13 and 14. The computational function of the channel attention module is given in Eq. (12). The computational function of the spatial attention module is shown in Eq. (13):

$$f_{CA} = sigmoid((MLP(AvgPool(F_C) + (MLP(MaxPool(F_C)) \tag{12}$$

$$f_{SA} = sigmoid\left(Conv2d\left(Concat\left(AvgPool\left(F_S\right), MaxPool\left(F_S\right)\right)\right)\right) \tag{13}$$

where F_C refers to the features of the input channel attention module, F_S refers to the features of the input spatial attention module, *AvgPool* denotes the average pooling operation, *MaxPool* denotes the maximum pooling operation, *MLP* contains a multilayer perceptual machine (MLP) model with two linear layers and an activation function ReLU, and *sigmoid* denotes a sigmoid activation function. *Concat* denotes stacking the two vectors, and *Conv2d* denotes a 2D convolutional layer.

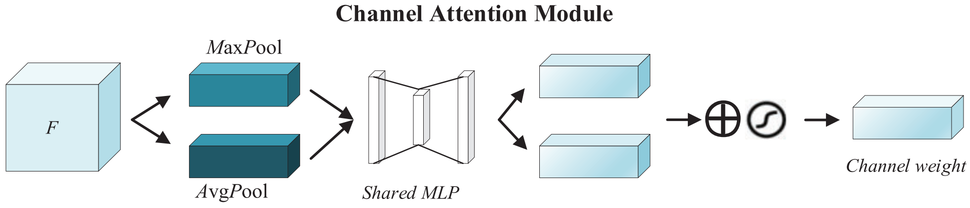


Figure 13: Channel attention module

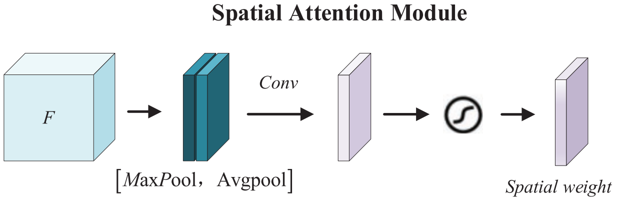


Figure 14: Spatial attention module

5 Experiments and Results

5.1 Implementation Details

5.1.1 Training Detail

The experiment is based on the 64-bit operating system Ubuntu 18.04 installed on a NVIDIA GeForce RTX 3080 and 10,015 MiB memory. We use the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.01, a final learning rate of 0.002, a weight attenuation of 0.0005, and a momentum of 0.937. The total epoch is set to 800 (if the model converges in the middle of training, the model can end the training early), and the batch size is 4 (the single-modality model reads four visible images or NIR images at one time, and the dual-modality model reads four visible images and four NIR images at one time), using the Mosaic [38] data augmentation method.

5.1.2 Evaluation Metrics

The evaluation metrics are an important basis for measuring the performance of remote-sensing object recognition algorithms. In order to evaluate the practicality of the proposed dataset and the effectiveness of the proposed algorithm, The evaluation metrics we use are shown in Table 5.

Table 5: Evaluation metrics for object recognition

Evaluation metrics	Description
Precision	Reflect the proportion of real positive samples in the test results
Recall	Reflects the proportion of correctly predicted positive samples in all positive samples to be detected
AP	The test results of each category are good or bad
mAP0.5	The average value of all types of APs (IoU = 0.5)
mAP0.5:0.95	The average value of all types of APs (IoU from 0.5 to 0.95, step size 0.05)

Hence, the precision, recall indicators are formulated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

where TP (True positives) refers to the number of samples correctly predicted as positive samples; FP (False positives) refers to the number of samples wrongly predicted as positive samples, and FN (False negatives) refers to the number of samples wrongly predicted as negative samples.

For a certain category, different thresholds are selected to obtain the corresponding accuracy and recall rate. It is drawn as a curve in the coordinate system with Recall in the abscissa and Precision in the ordinate, which is the P-R (Precision-Recall) curve. The area under the PR curve is the AP of this class.

5.2 Evaluation of Advanced Object Recognition Algorithms on DMFGRS

We evaluate the state-of-the-art object recognition algorithms on DMFGRS, including YOLOv5s+CSL [39], R-CNN [40], Faster-RCNN [41], RetinaNet [42], etc. Because DMFGRS adopts rotation annotation method, the verified model comes from the modified rotation detection algorithm on the MMrotate platform. MMRotate is a toolkit that provides a unified training and evaluation framework for rotating object detection methods and supports three rotating frame definition methods: OpenCV definition method, 135° long side definition method and 90° long side definition method. The OpenCV definition method is used for the test models in this section. MMrotate can specify the angle prediction methods, including CSL and KLD [43].

The benchmark test results of the selected baseline model are shown in Table 6. The YOLOv5s outperformed R-CNN, Faster R-CNN, and RetinaNet. The worst performance is the model RetinaNet. The performance of RetinaNet is greatly affected by the values of the hyperparameters α and γ. It takes continuous experiments to determine the optimal α and γ parameters. In addition, category LCS and category FAC performed poorly, which may be related to their small sample size. From Fig. 4, it can be seen that the aspect ratio of category ILC is the smallest, and the aspect ratio of adjacent category SM and category PF is almost doubled. Category ILC is among the top in each model, and the comparison with category SM and category PF is more obvious. Therefore, the large aspect ratio is an important factor to be considered in the optimization of fine-grained recognition algorithms for ship objects. On the RetinaNet model, we tested the two angle prediction methods, CSL and KLD, respectively, and it can be seen that the KLD method is 7.6 percentage points higher than the CSL method in mAP, which shows that for the RetinaNet model, the boxes predicted by the KLD method are more accurate.

Table 6: Benchmark results of baseline models

Model	Config	FAC	NAC	HAA	TC	BD	ZD	PF	ILC	SM	KFT	mAP
		SAT	RFS	BAC	FLC	AL	WL	OT	BC	CT	LCS	
YOLOv5+CSL	s	0	92.8	89.9	93.5	97.7	98.8	95.1	98.2	88.9	71.2	**80.9**
		88.5	82.2	99.5	80.8	99.5	89.4	91.3	78.6	81.3	0	
R-CNN	R50-FPN	0	97.2	98.2	84.3	89.3	95.2	88.0	99.2	80.0	75.1	**78.92**
		82.0	97.3	89.3	69.8	99.1	72.8	90.4	86.9	78.8	5.5	
Faster R-CNN	R50-FPN	33.3	83.5	79.5	62.5	81.2	78.4	49.3	99.2	75.4	62.0	**70.6**
		74.8	92.9	59.1	73.4	64.1	55.6	66.6	83.4	61.0	77.3	
RetinaNet	R50-FPN	40.0	78.8	66.9	56.3	59.3	60.0	44.9	84.7	32.2	16.1	**46.4**
		26.6	80.4	33.6	20.8	57.5	23.5	42.4	58.5	45.1	0	
	R50-FPN-kld	100	87.4	56.1	59.1	60.9	72.6	42.4	81.4	44.2	26.8	**53.0**
		48.0	71.9	58.7	25.0	49.9	29.1	46.0	57.0	44.2	0	

From Table 6, it can be seen that the classical object detection framework with the best validation effect on DMFGRS is the YOLOv5s model, which combines both detection accuracy and detection efficiency, and also has generality for the detected object, and the number of parameters is relatively small, which saves space resources. Moreover, the entire YOLOv5s model architecture is highly modular and suitable for making alterations. Therefore, we choose YOLOv5s as the base framework of our model.

5.3 Evaluation of Multimodal Information Fusion Network on DMFGRS

5.3.1 Comparison with Single-Modality Image Recognition

We validate the proposed ship fine-grained recognition network MICE-Net based on multimodal information cross-enhancement on the established DMFGRS, because the DMFGRS recognition process refers to the idea of YOLOv5s, and as shown in Table 7, the YOLOv5s outperforms the other several models in terms of speed and recognition effect, so this section mainly compares with the recognition effect of YOLOv5s.

Table 7: Evaluation results of MICE-Net on DMFGRS

Modality	Method	Precision	Recall	mAP0.5	mAP0.5:0.95
Visible	YOLOV5s	77.3	76.6	80.9	58.8
NIR	YOLOV5s	81.6	75.3	79.6	56.5
Visible + NIR	MICE-Net	**87**	**77.1**	**83.8**	**63.9**

As shown in Table 7, the performance of MICE-Net is significantly better than YOLOv5s in all the metrics. MICE-Net uses visible-NIR dual-modal remote sensing images for fine-grained recognition of ships on DMFGRS, and it achieves 87%, 77.1%, 83.8%, and 63.9% in the metrics of precision, recall, and mAP0.5 and mAP0.5:0.95. Compared to YOLOv5s for single-modality visible remote sensing images, MICE-Net is 9.7% higher on precision, 0.5% higher on recall, 2.9% higher on mAP0.5, and 5.1% higher on mAP0.5:0.95. Compared to YOLOv5s for single-modality NIR remote sensing images, MICE-Net is 5.4% higher on precision, 1.8% higher on recall, 4.2% higher on mAP0.5, and 7.4% higher on mAP0.5:0.95.

Fig. 15 shows the changes in the four metrics precision, recall, and mAP from the start of training to the convergence of the model. As can be seen from the figure, within 300 epochs, the purpleline representing MICE-Net is always located above the other two lines, i.e., during the training process, the convergence speed of MICE-Net using visible-NIR bimodal remote sensing images for ship recognition is significantly better than that of using YOLOv5s for ship recognition based on the single-modality remote sensing images in the pre-training period. In particular, from the trend plot representing the change in mAP0.5:0.95, it can be seen that throughout the training period including the final test phase shown in the table, the bimodal training was superior to the two unimodal ones. While mAP0.5:0.95 indicates that when the IOU threshold changes from 0.5 to 0.95, the average value of mAP corresponding to each threshold is taken. Compared with mAP0.5, mAP0.5:0.95 can evaluate the model performance more comprehensively and accurately by considering the average accuracy under multiple IOU thresholds at the same time. The mAP0.5:0.95 of MICE-Net far exceeds the single-modality case, indicating that MICE-Net has a wide coverage and can be applied to different scenarios and application requirements.

Compared to the training effect of YOLOv5s for single-modality visible images, the training effect is about the same or even slightly better than that of MICE-Net in the late convergence stage though. However, when the trained model is applied to a test dataset with no training added at all, the recognition results of MICE-Net, including precison, recall, mAP0.5 and mAP0.5:0.95, are significantly better than those of YOLOv5s for unimodal visible or NIR images. Therefore, MICE-Net has better transferability and generalizability.

Table 8 shows the metrics comparison between YOLOv5s and MICE-Net on DMFGRS specific to each category, where P represents Precision, R represents Recall, 0.5 represents mAP0.5, and 0.95 represents mAP0.5:0.95. As can be seen from the table, the performance of all models is poor in the FAC category due to the extremely low number of instances, but the performance of MICE-Net on dual-modality images exceeds the performance of YOLOv5s on single-modality images in most of the categories. In particular, the performance of the mAP0.5:0.95 index, except for the categories ZD, KFT and OT, the performance of MICE-Net is significantly superior. Combined with Table 7, although there are still categories where MICE-Net fails to achieve the top performance in terms of Precision, Recall and mAP0.5, the overall performance is best in all categories, and therefore MICE-Net is better able to balance the differences between categories using the feature information of the dual-modality images, and has a better generalisability. However, we should admit that MICE-Net still has space for improvement and progress.

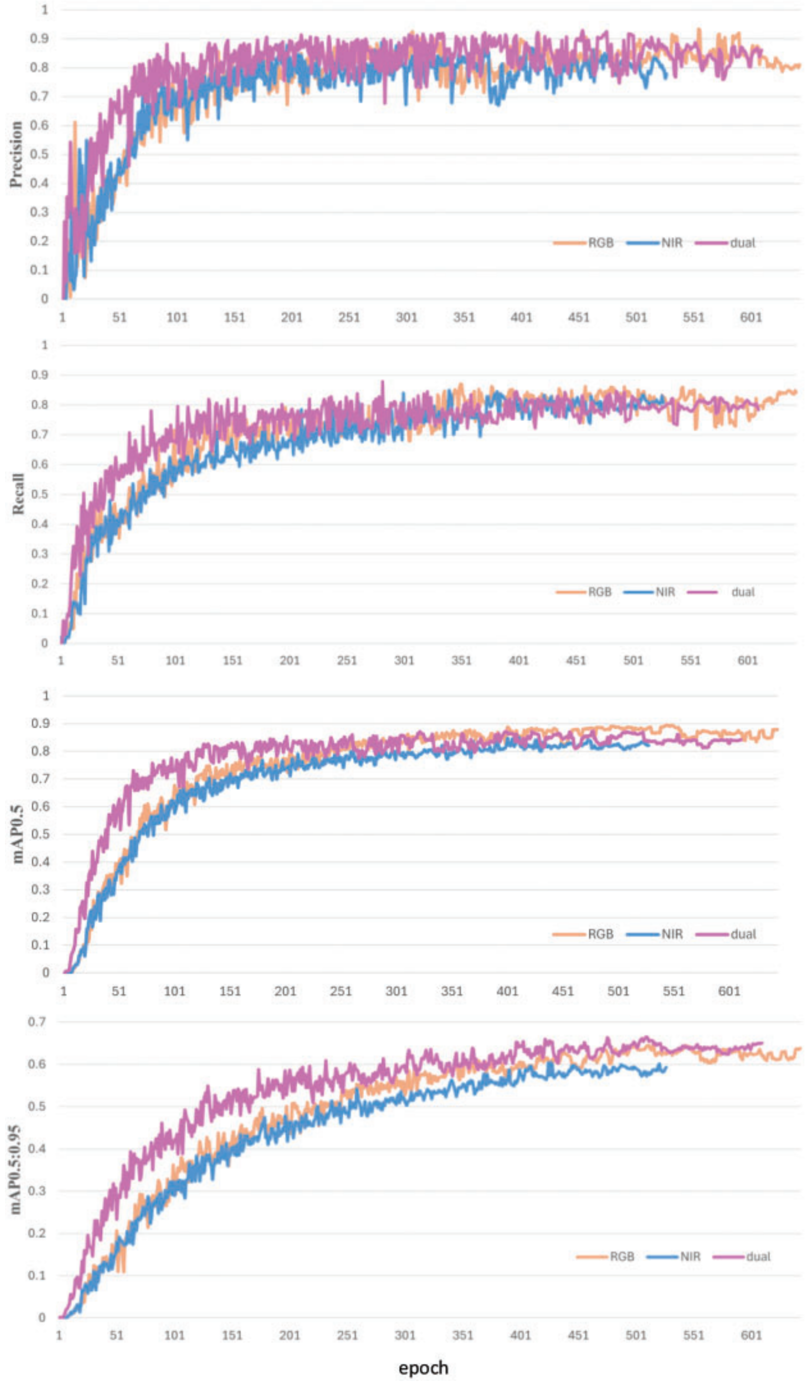


Figure 15: Comparison of metrics during training

Table 8: Comparison of performance in different categories

Class	Modality	P	R	0.5	0.95	Class	Modality	P	R	0.5	0.95
FAC	Visible	0	0	0	0	SAT	Visible	76	81.1	88.5	73.5
	NIR	0	0	0	0		NIR	74.5	81.1	83.8	70.9
	Dual	0	0	0	0		Dual	**91.6**	**83.8**	**91.9**	**80.8**
NAC	Visible	90	75	92.8	72.9	RFS	Visible	**91.4**	78.9	82.2	47.8
	NIR	89.9	74.1	94.2	71.6		NIR	81.2	78.9	81.5	53.5
	Dual	**100**	**89.9**	**97.5**	**75**		Dual	86.1	**84.2**	**92.7**	**58.1**
HAA	Visible	86.6	**82.2**	89.9	67.5	BAC	Visible	90.9	**100**	99.5	89.1
	NIR	84.1	**82.2**	87.3	61.4		NIR	92	**100**	99.5	82.5
	Dual	**97.3**	81.2	**92.8**	**70.2**		Dual	**95.3**	92.3	**99**	**89.5**
TC	Visible	86.4	91.4	**93.5**	69.1	FLC	Visible	82.6	80	80.8	59.3
	NIR	**87.1**	93.1	93.3	65.6		NIR	**85.6**	79.2	**89.2**	67.2
	Dual	83.9	84.5	90.7	**73.5**		Dual	84.4	**86.7**	89.1	**72.8**
BD	Visible	90.2	**94.1**	**97.7**	74.3	AL	Visible	**100**	99.8	**99.5**	72.8
	NIR	87.7	90.1	93.8	71.5		NIR	93.3	92.8	94.5	72.4
	Dual	**93.3**	90.9	95.8	**76.6**		Dual	95.3	**100**	**99.5**	**76.9**
ZD	Visible	93.2	**100**	98.8	**77**	WL	Visible	76.1	**79.2**	**89.4**	72.1
	NIR	89.1	87.5	91.8	73.4		NIR	85.8	75	80.7	65.3
	Dual	**97.6**	**100**	**99.5**	75.8		Dual	**86.3**	78.5	85.8	**72.5**
PF	Visible	89.2	92.6	95.1	58.4	OT	Visible	**95.3**	**85.2**	**91.3**	**69.7**
	NIR	91	93.3	95.8	53.1		NIR	85	84.2	86.5	57.6
	Dual	**92.5**	**96.3**	**98.4**	**64.1**		Dual	87.7	79.3	90.3	69.2
ILC	Visible	**97.8**	93.1	98.2	83.2	BC	Visible	72.2	**71**	78.6	55.8
	NIR	**97.8**	**93.2**	**99.1**	83.2		NIR	67.8	70.5	72.4	55.1
	Dual	97.3	92.8	96.9	**85.7**		Dual	**90.9**	68.1	**80.1**	**67.7**
SM	Visible	78.9	**88.6**	88.9	41.1	CT	Visible	82.6	76.2	81.3	41.5
	NIR	81	83.8	87.1	41.7		NIR	80.2	**84**	**91.2**	43.6
	Dual	**93.4**	85.4	**92.2**	**53.3**		Dual	**90.4**	76	86.7	**51.9**
KFT	Visible	66.1	63.6	71.2	**50.5**	LCS	Visible	0	0	0	0
	NIR	**78.8**	63.6	70.5	39.8		NIR	**100**	0	0	0
	Dual	75.8	**72.7**	**71.5**	49.8		Dual	**100**	0	**24.9**	**14.9**

Table 9 shows the model complexity of MICE-Net, which mainly includes the number of parameters and the time consumed. The time is calculated when the input image size of the network is (4, 3, 768, 768). Because MICE-Net is changed from a single-branch network to a two-branch network, and from one Backbone to two Backbones, the parameters of the network will inevitably increase.

Compared to YOLOV5s, the number of parameters of MICE-Net has increased from 7,549,525 to 14,164,660, which is a total increase of 6,615,135. Each detection time consists of three parts: Pre-process, inference and NMS. MICE-Net needs to process two images for each detection, so the duration of both pre-process and inference is increased. However, since MICE-Net only has two detections for medium and large objects, so that the duration of NMS decreases. Total time increased 4 ms from 5.8 to 9.8 ms.

Table 9: Comparison of MICE-Net model complexity

Model	Parameters	Time (ms)			
		Pre-process	Inference	NMS	Total
YOLOV5s	7,549,525	0.2	4.3	1.3	5.8
MICE-Net	14,164,660 (↑6,615,135)	0.5 (↑0.3)	8.2 (↑3.9)	1.1 (↓0.2)	9.8 (↑4)

In summary, MICE-Net can indeed achieve significant improvement in detection by fusing and enhancing dual-modality image feature information, but it also results in an increase in model complexity, including a rise in the number of parameters as well as an increase in inference time. Therefore, in the future, we will further process the model, such as lightweight, to improve the detection efficiency of the model.

5.3.2 Ablation Studies

In this ablation study, we analyzed in detail the effects of the effectiveness of FCEM, attention mechanism structure and DMIEF structure on the performance of MICE-Net.

Effectiveness of FCEM FCEM obtains more critical and representative fusion vectors by making the attentional mechanism work crosswise on the feature vectors of the visible and NIR images so that the focus of the fused features refers to the salient regions of the information of both modalities at the same time. To verify the effectiveness of FCEM, we replaced the FCEM in MICE-Net with the most basic Concat module, allowing the feature vectors of the two modalities to be stacked directly. We then compared the experimental results of this configuration with those obtained when using FCEM. The experimental results are shown in Table 10. When we used the simple Concat module, which directly stacks the information from the two modalities as the fusion step, it can be seen that the recognition performance was far inferior to using FCEM. In terms of the four metrics of Precision, Recall, mAP0.5, and mAP0.5:0.95, the use of FCEM achieved significant improvements of 1.4%, 3.5%, 3.8%, and 5.9% respectively compared to the use of Concat. Combining Tables 7 and 10, it can be seen that the fusion method using simple stacking is even less effective than directly using visible images for recognition. This may be because the model did not selectively combine all the information from the two modalities. Some redundant information not only failed to play a positive role but also reduced the model's sensitivity to the object, resulting in unsatisfactory final recognition performance. Therefore, it can be concluded that the FCEM we designed is very necessary and effective.

Table 10: Performance comparison on the effects of FCEM

Module	Method	Precision	Recall	mAP0.5	mAP0.5:0.95
Visible + NIR	Concat	85.6	73.6	80	**58**
Visible + NIR	FCEM	87 (↑1.4)	77.1 (↑3.5)	83.8 (↑3.8)	**63.9 (↑5.9)**

Attention mechanism structure As described in Section 4.2, FCEM utilizes temporal and spatial attention mechanisms to act on the feature maps of the two modalities separately. Based on the derived weights of the two attention mechanisms, cross-guidance is performed on the feature vectors before fusion. The function of the attention mechanism is inseparable from pooling operations. Therefore, to explore a more effective FCEM architecture, we have designed three attention implementation methods. Maximum pooling is theoretically more capable of extracting salient features of the object from the background, so we experimented with using maximum pooling instead of the average pooling in the spatial attention module and channel attention module. Table 11 compares the recognition performance of different attention implementation methods on DMFGRS, where max-CA indicates that the average pooling in channel attention is replaced with maximum pooling, and max-SA indicates that the average pooling in spatial attention is replaced with maximum pooling. It can be seen that compared to the other two methods, the first method has a better overall effect, with higher scores on both mAP0.5 and mAP0.5:0.95. However, in terms of the Recall metric, the second method performs better, achieving 77.8%.

Table 11: Performance comparison on the attention mechanism structure

Structure	Precision	Recall	mAP0.5	mAP0.5:0.95
SA + CA	**87**	77.1	**83.8**	**63.9**
SA + max-CA	78.9	**77.8**	81.4	60.5
Max-SA + CA	85.9	76.1	81.7	60.4

DMIEF structure The depth of the network is closely related to the dimensionality of the feature vectors and the expressive ability of modality information. The position of the feature fusion module within the network is closely related to the fusion effect and directly impacts the final recognition performance. Therefore, we designed to use FCEM at different positions within the network, resulting in three DMIEF structures. These three scenarios are illustrated in Fig. 16. Table 12 compares the recognition performance of different DMIEF structures on DMFGRS. It can be seen that structure (A) has the best recognition effect, thus making it our final design choice. Worth mentioning is that structure (C) performs the best in terms of accuracy, with a precision score of 88.8%.

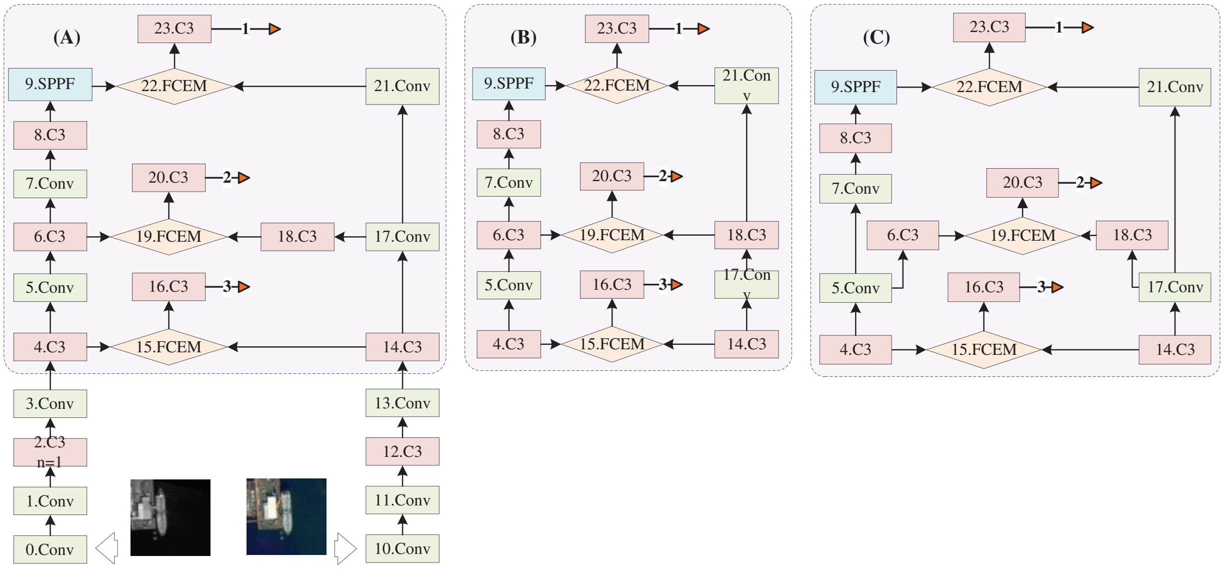


Figure 16: An illustration of different DMIEF structures

Table 12: Performance comparison on the DMIEF structure

Structure	Precision	Recall	mAP0.5	mAP0.5:0.95
A	87	**77.1**	**83.8**	**63.9**
B	79.9	76	81	58.2
C	**88.8**	71.2	80.9	60.2

6 Conclusion

In this paper, a multi-modality image dataset DMFGRS for ship fine-grained recognition research and a ship fine-grained recognition model MICE-Net based on visible and NIR remote sensing images are proposed. The remote sensing images of DMFGRS are all derived from digital orthophoto maps (DOM) provided by commercial remote sensing satellites. The authenticity of the image sources makes the trained model closer to real-world scenarios and more suitable for future on-orbit applications. DMFGRS provides high-resolution, finely annotated visible/near-infrared multimodal images as well as object segmentation mask images, supporting research on single/dual modality object detection, fine-grained ship recognition and segmentation tasks. MICE-Net is an end-to-end single-modality fine-grained ship recognition model that is modified from a single-modality recognition model that balances performance and speed. It utilizes an attention mechanism to perform cross-guidance on the information from the two modalities, achieving multimodal feature enhancement and fusion, and thus significantly improving recognition performance. The effectiveness and usability of DMFGRS have been validated by conducting experiments based on commonly used object recognition algorithms. Through comparisons of experimental data, the superior performance of MICE-Net demonstrates its good portability and generalizability.

However, DMFGRS has limitations in terms of the number of categories and a lack of instances for certain types of ships. Additionally, due to the addition of an extra feature extraction backbone network, the number of parameters in MICE-Net increases significantly, resulting in slower computation speed. In future work, the number of categories and images in DNGFRS will continue to expand, and the number of instances for each category will be balanced and increased. For MICE-Net, we will further optimize and lightweight to obtain better recognition and faster recognition rate, and gradually carry out specific research tasks such as on-orbit applications.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Shiwen Song; data collection: Shiwen Song, Rui Zhang, Min Hu; analysis and interpretation of results: Shiwen Song, Rui Zhang; draft manuscript preparation: Shiwen Song, Rui Zhang, Min Hu, Feiyao Huang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Lin, S. Song, and J. Yang, “Ship classification based on MSHOG feature and task-driven dictionary learning with structured incoherent constraints in SAR images,” *Remote Sens.*, vol. 10, no. 2, pp. 190, Jan. 2018. doi: 10.3390/rs10020190.

[2] Q. Oliveau and H. Sahbi, “Learning attribute representations for remote sensing ship category classification,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, no. 6, pp. 2830–2840, Jun. 2017. doi: 10.1109/JSTARS.2017.2665346.

[3] L. Huang, W. Li, C. Chen, F. Zhang, and H. Lang, “Multiple features learning for ship classification in optical imagery,” *Multimed. Tools Appl.*, vol. 77, no. 11, pp. 13363–13389, Jun. 2018. doi: 10.1007/s11042-017-4952-y.

[4] Z. Lu, P. Wang, Y. Li, and B. Ding, “A new deep neural network based on SwinT-FRM-ShipNet for sar ship detection in complex near-shore and offshore environments,” *Remote Sens.*, vol. 15, no. 24, pp. 5780, Dec. 2023. doi: 10.3390/rs15245780.

[5] S. Liu, P. Chen, and Y. Zhang, “A multiscale feature pyramid SAR ship detection network with robust background interference,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 9904–9915, Oct. 2023. doi: 10.1109/JSTARS.2023.3325376.

[6] Y. Han, X. Yang, T. Pu, and Z. Peng, “Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, Oct. 2022. doi: 10.1109/TGRS.2021.3123666.

[7] J. Chen and Y. Qian, “Hierarchical multilabel ship classification in remote sensing images using label relation graphs,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 1, pp. 1–13, Sep. 2022. doi: 10.1109/TGRS.2021.3111117.

[8] W. Wu, X. Li, Z. Hu, and X. Liu, “Ship detection and recognition based on improved YOLOv7,” *Comput. Mater. Contin.*, vol. 76, no. 1, pp. 489–498, Jun. 2023. doi: 10.32604/cmc.2023.039929.

[9] S. Basalamah, S. D. Khan, and H. Ullah, "Scale driven convolutional neural network model for people counting and localization in crowd scenes," *IEEE Access*, vol. 7, pp. 71576–71584, May 2019. doi: 10.1109/ACCESS.2019.2918650.

[10] S. D. Khan, L. Alarabi, and S. Basalamah, "A unified deep learning framework of multi-scale detectors for geo-spatial object detection in high-resolution satellite images," *Arab J. Sci. Eng.*, vol. 47, no. 8, pp. 9489–9504, Aug. 2022. doi: 10.1007/s13369-021-06288-x.

[11] Y. Zhang, D. Han, and P. Chen, "Swin-PAFF: A SAR ship detection network with contextual cross-information fusion," *Comput. Mater. Contin.*, vol. 77, no. 2, pp. 2657–2675, Nov. 2023. doi: 10.32604/cmc.2023.042311.

[12] K. Liu, S. Yu, and S. Liu, "An improved inceptionV3 network for obscured ship classification in remote sensing images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 4738–4747, Aug. 2020. doi: 10.1109/JSTARS.2020.3017676.

[13] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li and Q. Du, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, Mar. 2023. doi: 10.1109/TGRS.2023.3258666.

[14] H. Bi, R. Wu, Z. Liu, H. Zhu, C. Zhang and T. Z. Xiang, "Cross-modal hierarchical interaction network for RGB-D salient object detection," *Pattern Recognit.*, vol. 136, no. 5, pp. 109194, Apr. 2023. doi: 10.1016/j.patcog.2022.109194.

[15] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu, "LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-Thermal images," *IEEE Trans. Image Process.*, vol. 32, pp. 1329–1340, Feb. 2023. doi: 10.1109/TIP.2023.3242775.

[16] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognit.*, vol. 85, no. 4, pp. 161–171, Jan. 2019. doi: 10.1016/j.patcog.2018.08.005.

[17] Q. Fang, D. Han, and Z. Wang, "Cross-modality fusion transformer for multispectral object detection," *SSRN Electron. J.*, Oct. 2022. doi: 10.2139/ssrn.4227745.

[18] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods (ICPRAM)*, Porto, Distrito do Porto, Portugal, 2017, pp. 324–331.

[19] S. Lee, J. Park, and J. Park, "CrossFormer: Cross-guided attention for multi-modal object detection," *Pattern Recognit. Lett.*, vol. 179, no. 11, pp. 144–150, Mar. 2024. doi: 10.1016/j.patrec.2024.02.012.

[20] S. Lu, M. Liu, L. Yin, Z. Yin, X. Liu and W. Zheng, "The multi-modal fusion in visual question answering: A review of attention mechanisms," *PeerJ Comput. Sci.*, vol. 9, no. 18, pp. e1400, May 2023. doi: 10.7717/peerj-cs.1400.

[21] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016. doi: 10.1109/TGRS.2016.2601622.

[22] G. S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, Utah, USA, 2018, pp. 3974–3983.

[23] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, no. 10, pp. 187–203, Jan. 2016. doi: 10.1016/j.jvcir.2015.11.002.

[24] Y. Di, Z. Jiang, and H. Zhang, "A public dataset for fine-grained ship classification in optical remote sensing images," *Remote Sens.*, vol. 13, no. 4, pp. 747, Feb. 2021. doi: 10.3390/rs13040747.

[25] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1100–1111, Mar. 2018. doi: 10.1109/TIP.2017.2773199.

[26] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020. doi: 10.1016/j.isprsjprs.2019.11.023.

[27] X. Sun *et al.*, “FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery,” *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 116–130, Feb. 2022. doi: 10.1016/j.isprsjprs.2021.12.004.

[28] Y. Sun, B. Cao, P. Zhu, and Q. Hu, “Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6700–6713, Oct. 2022. doi: 10.1109/TCSVT.2022.3168279.

[29] L. Yao, X. Zhang, Y. Lyu, W. Sun, and M. Li, “FGSC-23: A large-scale dataset of high-resolution optical remote sensing image for deep learning-based fine-grained ship recognition,” *J. Image Graph.*, vol. 26, no. 10, pp. 2337–2345, Oct. 2021. doi: 10.11834/jig.200261.

[30] Y. Di *et al.*, “A public dataset for ship classification in remote sensing images,” in *Proc. Img. Sig. Process. Remote Sens. XXV*, Strasbourg, Grand Est, France, 2019, pp. 515–521.

[31] X. Xiao *et al.*, “Infrared and visible image object detection via focused feature enhancement and cascaded semantic extension,” *Remote Sens.*, vol. 13, no. 13, pp. 2538, Jun. 2021. doi: 10.3390/rs13132538.

[32] J. U. Kim, S. Park, and Y. M. Ro, “Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1510–1523, Mar. 2022. doi: 10.1109/TCSVT.2021.3076466.

[33] Z. Xie *et al.*, “Cross-modality double bidirectional interaction and fusion network for RGB-T salient object detection,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4149–4163, Aug. 2023. doi: 10.1109/TCSVT.2023.3241196.

[34] M. Sharma *et al.*, “YOLOrs: Object detection in multimodal remote sensing imagery,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 1497–1508, Nov. 2021. doi: 10.1109/JSTARS.2020.3041316.

[35] Q. Y. Fang and Z. K. Wang, “Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery,” *Pattern Recognit.*, vol. 130, no. 3, pp. 108786, Oct. 2022. doi: 10.1016/j.patcog.2022.108786.

[36] T. Lin *et al.*, “Microsoft coco: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Canton of Zurich, Switzerland, 2014, pp. 740–755.

[37] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye and D. Ren, “Distance-IoU loss: Faster and better learning for bounding box regression,” in *Proc. AAAI Conf. Artif. Intell.*, New York, NY, USA, Apr. 2020, pp. 12993–13000.

[38] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, “YOLOv4: Optimal speed and accuracy of object detection,” arXiv Preprint arXiv:2004.10934, Apr. 2020.

[39] X. Yang and J. Yan, “On the arbitrary-oriented object detection,” *Classification Based Approaches Revisited, Int. J. Comput. Vis.*, vol. 130, no. 5, pp. 1340–1365, May 2022. doi: 10.1007/s11263-022-01593-w.

[40] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Ithaca, NY, USA, 2014, pp. 580–587.

[41] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017. doi: 10.1109/TPAMI.2016.2577031.

[42] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020. doi: 10.1109/TPAMI.2018.2858826.

[43] X. Yang, X. Yang, J. Yang, Q. Ming, and J. Yan, “Learning high-precision bounding box for rotated object detection via kullback-leibler divergence,” *Adv. Neural Inf. Process. Syst.*, vol. 22, pp. 18381–18394, Jun. 2021.