A Novel Multiscale Contrastive Learning Network for Fine-Grained Ocean Ship Classification

Shaokang Dong, Jiangfan Feng, and Dongxu Fang

***Abstract*—Fine-grained ocean ship classification plays a crucial role in maritime military surveillance, traffic management, and antismuggling operations. However, the complex backgrounds of remote sensing images (RSIs), as well as significant interclass similarities and intraclass differences, result in poor classification performance. Hence, we propose MSCL-Net, a multiscale contrastive learning network for fine-grained ship classification (FGSC). First, we introduce ResNet50 as the backbone network and extract the multilayer features by using the FPN for FGSC. Second, a channel spatial attention module (CSAM) is proposed to extract the similarity (contrastive) feature of the same class, strengthening the representation learning ability for addressing issues caused by significant interclass similarity and intraclass difference. Third, a region cropping and enlargement module is proposed to extract the fine-grained features of local discriminant regions in RSIs to overcome the challenge of background complexity. Finally, we used the CSAM to fuse the features of the original image and the cropped region image for FGSC. In addition, we introduce a combined loss based on center loss and PolyLoss to enhance the discrimination ability of features and make it more suitable for the imbalance dataset compared with cross-entropy. We use a public FGSC dataset, FGSC-23, and our FGSC-41 to evaluate the performance of MSCL-Net. The experimental results show superior performance compared to other state-of-the-art methods, highlighting the effectiveness of MSCL-Net in addressing the challenges associated with FGSC. Ablation experiments also suggest the effectiveness of our design.**

***Index Terms*—Contrastive learning, fine-grained, multiscale learning, PolyLoss, ship classification.**

I. Introduction

Accurately classifying ocean ships at a fine-grained level is crucial for diverse applications, including port supervision, resource allocation, target classification, and civilian watercraft protection. In military contexts, precise ship identification is indispensable for devising intricate combat strategies, enhancing sea target surveillance, and ensuring national defense security. As a result, achieving accurate and efficient ocean ship identification holds great promise and significance, driving ongoing research efforts in this field.

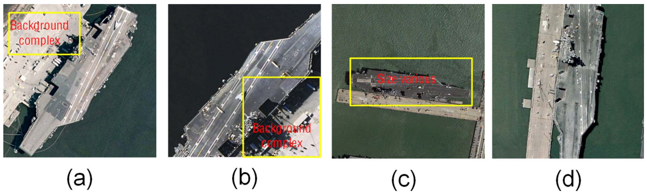


Fig. 1. (a) and (b) show the optical RSIs of two Kitty Hawk aircraft carriers, and (c) and (d) show the optical RSIs of two Nimitz aircraft carriers.

Limited by the difficulty of data acquisition and the low quality of data [1], the study of ship classification initially focused on the target level and coarse granularity. These methods primarily classify the ship target and backgrounds [2] or distinguish a few types of ships, such as fishing boats, cargo ships, and warships [3]. With advances in optical remote sensing technology, images with enormous high resolution, interpretability, and semantic richness can be obtained. These advancements have revolutionized ocean ship classification and provided a solid foundation for numerous research studies focused on fine-grained ship classification (FGSC).

FGSC was developed from the fine-grained classification of natural images [4], [5], [6]. However, the backgrounds and sizes of ship targets in RSIs are complex and varied, and there is significant interclass similarity as well as intraclass difference. For instance, Fig. 1(a) and (b) displays optical RSIs of two Kitty Hawk aircraft carriers, while Fig. 1(c) and (d) depicts optical RSIs of two Nimitz aircraft carriers. Fig. 1(a) shows that the background of the ship target is complex, and Fig. 1(c) shows that the size of the ship target is different from that of other images. In addition, we can discover that ships of different classes have similar outlines, but significant differences exist within the same class. Furthermore, due to the difficulty of collecting optical RSIs of naval ships, labeled datasets are usually lacking and imbalanced.

These issues make FGSC based on optical RSIs difficult, which leads to plenty of methods have been proposed for FGSC. Some methods focus on few-shot learning [7], [8], [9], [10], generative adversarial network [11], [12], and semisupervised learning [13], [14]. Although these methods address the lack of labeled data, it is difficult to extract the similarity features of the same class image and the fine-grained features of local discriminant regions in the context of significant interclass similarity and intraclass difference.

Therefore, some researchers focus on extracting the multiscale features for FGSC, which can improve the accuracy of

Manuscript received 19 February 2024; revised 17 April 2024 and 30 April 2024; accepted 7 May 2024. Date of publication 10 May 2024; date of current version 30 May 2024. This work was supported by the National Natural Science Foundation of China under Grant 41971365. *(Corresponding author: Jiangfan Feng.)*

Shaokang Dong and Jiangfan Feng are with the School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: d220201004@stu.cqupt.edu.cn; fengjf@cqupt.edu.cn).

Dongxu Fang is with the China Mobile Group Chongqing Company, Ltd., Chongqing 400065, China (e-mail: fangdongxu@cq.chinamobile.com).

Digital Object Identifier 10.1109/JSTARS.2024.3399310

the model with significant interclass similarity and intraclass difference [15], [16], [17], [18], [19], [20], [21], [22]. With the progress of contrastive learning, some methods were proposed to make the features of the same class closer by loss function [1], [23]. However, multiscale-based methods either extract multiscale features from the network, which can be influenced by complex backgrounds, or extract local features of the most attention regions of the feature map, which can be influenced by noise. The method based on contrastive learning may fail when ship images of the different classes are highly similar.

In addition, several object detection methods were proposed for FGSC in the case of background complexity [24], [25], [26]. The method based on object detection can accurately locate the position of the ship target and increase the classification accuracy with complex backgrounds, but the label data of the bounding box are lacking. In conclusion, the significant interclass similarity, intraclass difference, and background complexity are still the essential challenges of FGSC.

To address these challenges, we propose a multiscale contrastive learning network (MSCL-Net) for FGSC, including four parts: multiscale feature learning, contrastive learning, feature fusion, and combined loss. The multiscale feature consists of the multilayer feature of ResNet50 and the fine-grained feature. The FPN is used to extract the multilayer features of ResNet50. Unlike the existing methods of fine-grained feature extraction, we propose a region cropping and enlargement module (RCEM) to extract the discriminant of local key regions. RCEM selects the maximum connected area of the mask matrix to cut the original images rather than the region with the highest response of the feature map, which avoids the influence caused by noise. In contrastive learning, we propose a channel spatial attention module (CSAM) aimed at extracting the most similar features of the same class, which can overcome the challenge of significant interclass similarity and intraclass difference. In addition, we use CSAM to fuse the features of the cut images with those of the original images. The fusion features provide more information for predicting the ship class. Finally, we design a combined loss consisting of center loss and PolyLoss. The center loss encourages feature vectors from the same classes to be closer and enhances feature distinguishability [27]. PolyLoss is an improvement of cross-entropy and focal loss, which can significantly improve the performance of classification methods in the context of imbalanced datasets [28]. The main contributions of this study are as follows.

1) We pioneered a state-of-the-art MSCL-Net for FGSC, which can extracts contrastive features of the same class and multiscale features. This approach enhances classification performance, especially in significant interclass similarity, intraclass difference, and complex backgrounds.
2) We propose RCEM to crop the maximum connected area of the mask matrix and as the input of ResNet50 to extract the fine-grained feature for FGSC. In contrastive learning, we propose CSAM to extract the most similar channel features and use spatial similarity to enhance it, which overcomes the challenge of the significant interclass similarity and intraclass difference.
3) We created a new dataset, namely FGSC-41, to evaluate the performance of the MSCL-Net. The effectiveness of our method has been confirmed through detailed experiments using two different datasets with superior accuracy and efficiency over existing methods. Additional ablation experiments further confirmed the efficacy of each part.

II. Related Work

FGSC methods have undergone significant developments. In this section, we first briefly review and discuss the fine-grained ship classification methods based on optical RSIs. In addition, the design of MSCL-Net refers to contrastive learning and multiscale feature learning. Therefore, we also discuss the progress of these two parts in computing vision.

A. FGSC in Optical Remote Sensing Images

With the success of remote sensing technology and deep learning, plenty of methods were proposed to classify the ship images at the fine-grained level. Early, due to the limited label data, some methods based on semisupervised, few-shot learning, and GAN were used to improve the accuracy of ship classification. Li et al. [7] introduced a novel rotation-invariant module to extract the rotation-invariant features for reconstruction. This method enhances the suitability for few-shot FGSC. Shi et al. [8] proposed a metric-based few-shot method that can obtain the novel class representation using the nearest neighbor prototype to improve the accuracy of FGSC in the case of small training samples. Oliveau et al. [13] proposed a semisupervised deep attribute network to extract the discriminative image features for FGSC when the labeled data are lacking. These methods based on few-shot learning fully solve the lack of label data. However, it may lead to poor generalization performance and overfitting since the small sample data may not adequately represent the diversity and complexity.

Hence, Luu et al. [29] employed three data augmentation steps, including random rotation and flipping, to obtain additional remote sensing ship images for FGSC. Liu et al. [12] proposed the local-aware CycleGAN to complete the image translation of the background and foreground, which can improve the reality of synthetic images. Moon et al. [11] used generative adversarial networks to construct a dataset without safety problems and proved that data augmentation is useful for FGSC. Kim et al. [14] used simulation programs to generate synthetic naval ship images for FGSC. Yi et al. [21] designed an essential feature mining network (EFM-Net) based on deep CNN to extract the most discriminative features. These methods were more robust, generalizable representations of the data distribution and reduced the risk of overfitting by generating realistic synthetic samples or extracting key features. Moreover, benefiting from two large-scale FGSC datasets, called FGSC-23 [20] and FGSCR-42 [30], further address the issue of insufficient labeled data. These studies also further make the method of FGSC more focused on other challenges, such as the significant interclass similarity and intraclass difference, as well as background complexity.

To enhance representation learning of the network, some methods based on contrastive learning were proposed for closing the feature vectors of images of the same class. Chen et al. [23] proposed a push-and-pull network that includes push and pull two stages for FGSC. P^2Net makes the features stay away from different classes and aggregate the features of the same class. Pan et al. [1] developed a contrastive learning network (C2Net) for FGSC, which applies counterfactual causal reasoning to make decisions at the logical level and enhances attention to local details. Zhang et al. [31] introduced a part assignment module and proposed a similarity learning by ranking contrastive learning framework for FGSC. This framework aims to capture the subtle differences between ship instances by ranking similarity. By incorporating the part assignment module, the approach can effectively handle the FGSC of RSIs.

Furthermore, to overcome the influence of complex backgrounds, Huang et al. [18] combined CNN and Swin transformers to extract multiscale features for FGSC. The method leverages the advantages of the Swin module to capture long-range dependencies across the entire image, enabling a better understanding of the relationships between different regions in the image. Huang et al. [15] proposed a method that fuses low-layer local features and high-layer global features of CNN for FGSC. Chen et al. [19] proposed a method to extract three-scale ship features from three-scale images. The three-scale image includes the original image, ship targets by Grad-CAM obtained and affine transformation. Song et al. [22] introduced an attention classification reduction network that utilized local and global features for FGSC. Zhang et al. [20] proposed an attribute-guided multilevel enhanced feature representation network (AMEFRN) for FGSC. AMEFRN used the multilevel feature and attribute feature to improve the ability to fine-grained classify ships. Meng et al. [17] proposed a global-to-local progressive learning module (GLPM), which uses the global and local features for FGSC. These methods can extract the multiscale feature from different layers and image regions. However, the extraction method based on the region image usually cuts the most attention region to extract the features, which leads the network to lose some suboptimal information. It is essential to extract the maximum connected region of the mask matrix, which includes more information and is more robust to noise, which is better for FGSC.

In addition, Han et al. [25] proposed a novel efficient information reuse network, which can maximize using multiscale information, suppress noise, and highlight targets. Liu et al. [26] first applied the oriented RPN for FGSC. Ma et al. [32] proposed a multiscale deep learning training model based on Fast-R-CNN and used guided filtering to remove fog.

With the success of interpretability methods and large models in the semantic segment, image classification, anomaly detection [33], [34], and object detection, such as Li et al. [33] presented a new interpretable network called LRR-Net for anomaly detection, which leverages the alternating direction method of multipliers optimizer to solve the LRR model efficiently and incorporates the solution as prior knowledge into the deep network to guide the optimization of parameters. Moreover, LRR-Net transforms the regularized parameters into trainable parameters of the deep neural network, thus alleviating the need for manual parameter tuning. Some methods were proposed for FGSC. Xiong et al. [35] focused on interpretability and proposed an interpretable attention network for FGSC. Xiong et al. [36] proposed a cognitive network, an inherently interpretable model tailored for FGSC. Hong et al. [37] created a universal RS foundation model named SpectralGPT for the first time. A large number of RSIs are used to train the SpectralGPT, which can enhance the generalizability of the network. In addition, SpectralGPT is widely applied not only for FGSC but also for semantic segmentation and object detection.

Our proposed method is based on contrastive learning and multiscale feature learning, which aims to improve the existing method and address the challenge of background complexity, significant interclass similarity, and intraclass difference. In our study, we propose two modules: CSAM and RCEM. CSAM is used to enhance the similarity feature of the same class image in contrastive learning and enhance the key region features of original images in feature fusion. RCEM is used to extract fine-grained features.

B. *Multiscale Feature Learning*

Multiscale feature extraction broadly contains two aspects: the multilayer or multibranch features of the network and the different region features of cut original images. For the multilayer feature extraction, which is mainly used for object detection, the typical network is the feature pyramid network (FPN), which fuses the low-layer feature and higher layer feature to improve the accuracy of object detection [38]. Li et al. [39] also applied the multiscale network for building footprint extraction of RSIs. For the multibranch network, Hong et al. [40] proposed a high-resolution domain adaptation network, HighDAN, for semantic segments, which is capable of capturing multiscaled image representations from parallel high-to-low-resolution subnetworks, yielding repetitive information exchange across different resolutions in a highly efficient manner. Wu et al. [41] embed a tiny U-Net into a larger U-Net backbone, enabling the multilevel and multiscale representation learning of objects, which can enhance the global and local features to improve the object detection performance. In addition, some methods are also used for fine-grained visual classification (FGVC). In particular, Qian et al. [42] proposed a multiscale covariance pooling network that can capture and better fuse features at different scales to generate more representative features for FGVC. Liu et al. [43] proposed a scale-consistent attention part network, which can be learned in an end-to-end way for FGVC.

The different region features of cut images are also used for FGVC. Du et al. [4] utilized a random jigsaw patch generator to obtain many cut images of multiscale, encouraging the network to learn features at specific granularities and enhancing the classification accuracy. Further, some network fusion multiscale features of two aspects for FGVC. Zhang et al. [6] proposed a multibranch and multiscale attention learning network (MMAL-Net) for FGVC. Feng et al. [5] proposed a progressive region-focused network for fine-grained human behavior recognition.

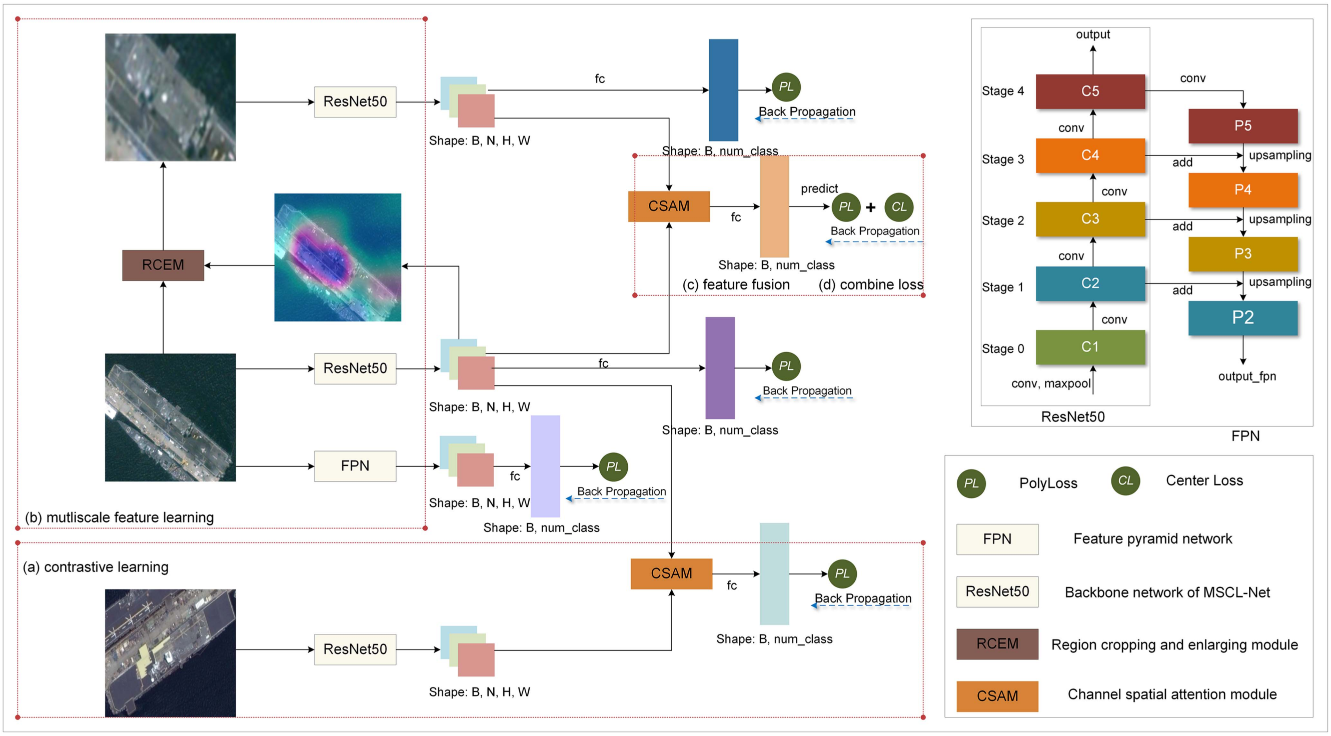


Fig. 2. Overall framework of the MSCL-Net.

Although these methods extract rich features for FGVC, the region feature extraction process usually randomly cuts original images or cuts the most attention region of the feature map. The method can obtain key features, ignore the suboptimal feature, and can be influenced by noise. Unlike these existing methods, we used the FPN to learn the multilayer feature of ResNet50 and proposed RCEM to extract fine-grained features. The design extracts multiscale features from two aspects and cuts original images by the maximum connection area of the mask matrix to extract the fine-grained feature, ensuring that the MSCL-Net can learn more rich and key features for FGSC.

C. *Contrastive Learning*

Contrastive learning can enhance the representation learning of images of the same class by the triplet, max-margin, and N-pairs loss. The method based on contrastive learning aims to pull tighter the anchor and positive samples and push apart the anchor from many negative samples in feature space through contrastive training data [44], [45], [46]. In recent years, plenty of methods have utilized contrastive learning loss to improve the performance of FGVC methods.

Wei et al. [47] designed a noise-tolerant supervised contrastive learning loss that incorporates a weight-aware mechanism for noisy label correction and selectively updating momentum queue lists, which can reduce the influence of noise. Bukchin et al. [48] used angular contrastive learning with coarse labels for FGVC. Zhang et al. [49] proposed a progressive co-attention network (PCA-Net), which can extract the similarity features of the channel to improve the accuracy of classification. Breiki et al. [50] used contrastive learning based on the SimCLR model for FGVC.

These methods that use contrastive loss aim to enhance the representation learning of global features in the feature space, and the purpose is to learn the similarity or otherness of training samples. These methods may not be suitable when the intraclass sample is highly similar, which leads to classification errors. Instead, PCA-Net extracting the similarity feature from the feature map can avoid these issues. However, the method only extracts the channel similarity feature for FGVC, ignoring the spatial similarity. We propose a module focusing on the channel and spatial similarity features, aiming to learn the most discriminant features for FGSC.

III. Methodology

The fine-grained ocean ship classification task presents challenges, including significant interclass similarity and intraclass difference, and background complexity. To address these challenges, we design an MSCL-Net, and Fig. 2 shows the overall framework of MSCL-Net. The backbone of MSCL-Net is ResNet50 and mainly includes four parts: a) multiscale feature learning, b) contrastive learning, c) feature fusion, and d) combined loss.

The multiscale feature learning consists of FPN and RCEM. FPN aims to extract multilayer features of ResNet50, which can fuse the lower layer detailed features for FGSC. RCEM is used to extract the features of the maximum connected region in the mask matrix, that is, fine-grained features, which can overcome the challenge of background complexity. In contrastive learning, we propose a CSAM to extract the similarity (contrastive) feature of the same class, strengthening the representation learning ability of the network. Unlike existing methods, CSAM obtains the channel similarity features and uses spatial similarity to enhance them, which can improve the performance of FGSC with the challenge of significant interclass similarity and intraclass difference.

Only utilizing the feature of the original images extracted by ResNet50 to predict the ship class can reduce the accuracy due to background complexity. Moreover, only utilizing the fine-grained features may cause some information to be lost. Hence, we use CSAM to fuse the features of the cut images with those of the original images. The fusion features provide more information for predicting the ship class.

In addition, in the case of datasets with significant interclass similarity and imbalance, the performance would reduce when using cross-entropy loss. Therefore, we introduce the combined loss to optimize the weight of MSCL-Net. The combined loss includes center loss and PolyLoss, which can ensure that the samples of the same class gather more closely in the feature space, which makes the feature representation more discriminating. At the same time, PolyLoss can improve the accuracy of MSCL-Net in the context of imbalanced datasets.

A. Multiscale Features Extraction

The pioneering work by Karen Simonyan and others underscores the profound influence of deep networks on image classification tasks [51]. With the rapid advancements in CNN and evolving requirements, various network architectures have emerged, including ResNet50, ConvNext, AlexNet, and DenseNet. ResNet50, renowned for its effectiveness in image classification tasks, is widely used as a backbone network for feature extraction. ResNet50 is also commonly employed for FGSC to extract image features automatically. Hence, we utilize ResNet50 as the backbone of MSCL-Net. ResNet50 usually uses the last-layer feature that includes rich semantic information for predicting the ship class, ignoring the detailed features of the low layer. To solve this issue, the FPN is usually used to extract the multilayer feature for classification. Likewise, we use it to improve the classification performance. In addition, we propose RCEM to crop and enlarge the original image and further as the input of ResNet50 to extract the fine-grained features.

1) Feature Pyramid Network: The framework of FPN is shown on the right of Fig. 2. The FPN consists of bottom-up and top-down pathways. The part of the bottom-up path is ResNet50. Specifically, we divided ResNet50 into five stages, namely, {stage0, stage1, stage2, stage3, and stage4}. These stages produce outputs {C1, C2, C3, C4, C5} and the corresponding feature map dimensions {(64, W/2, H/2),(256, W/4, H/4), (512, W/8, H/8), (1024, W/32, H/32), (2048, W/64, H/64)}, where W and H represent the input image size.

In the top-down pathways, a 1 × 1 convolutional layer is usually used to obtain P5, and the channel dimension of the feature map is 256. Then, through the operation of upsample and add, we can obtain P4, P3, and P2. Finally, the P2 fusion the high-layer rich semantic information and the low-layer rich detail, textural, and shape features. We utilize the P2 train the MSCL-Net to improve the accuracy.

2) Fine-Grained Feature Extraction: Following the classification way of person, the fine-grained feature of the key region is useful for FGSC. In this section, we introduce a novel module called the RCEM, which can locate the maximum attention area of the mask matrix, further map the position of the feature map to the original image and crop and enlarge the area to the size of the original image. Finally, we can use the cut image as the input of ResNet50 to extract the fine-grained features for FGSC. Fig. 3 shows the process of RCEM.

The feature map $F \in \mathbb{R}^{N\times H\times W}$, where N represents the channel, H and W denote the height and width, respectively. Initially, for each pixel of F, we compute the mean value of all channels by (1). The computed result is a matrix A with the shape H × W

$$A = \sum_{i=1}^{N}(F_i) \tag{1}$$

where F_i represents the feature of the ith channel. Then, (2) is used to compute the mean value $\bar{a}$ of all values in A

$$\bar{a} = \frac{\sum\limits_{x=1}^{W-1}\sum\limits_{y=1}^{H-1}(A(x,y))}{H\times W}. \tag{2}$$

Furthermore, we use (3) to acquire the mask matrix M. If $A(x,y) > \bar{a}$, we set the value to 1 in the corresponding position, otherwise, the value is set to 0

$$M(x,y) = \begin{cases} 0, & \text{if } A(x,y) > \bar{a} \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

The existing methods usually used the areas with the maximum value of A as the crop region, which can lead to classification errors when the RSIs include isolated noise. We applied the maximum connected graph of M as the crop area, which is conducive to obtaining the key region for FGSC and can avoid the influence of isolated noise. We defined that the coordinates of the minimum circumscribed matrix of the maximum connected graph are (x_1, y_1), (x_2, y_2), (x_3, y_3), and (x_4, y_4). The new coordinates in the original image can be computed by

$$\begin{aligned} x_n &= W' \times x_i/W \\ y_n &= H' \times y_i/H \end{aligned} \tag{4}$$

where W' and H' are the width and height of the original images, respectively. x_i and y_i represent the coordinate values on M. x_n and y_n denote the coordinate values on the original image.

Finally, we use the four new coordinates to crop the original image and enlarge the size of the original image. The new image is used as the input of ResNet50 to extract the fine-grained features of RSI, which can optimize the results of MSCL-Net.

B. Contrastive Learning

In our study, contrastive learning mainly extracts the channel similarity features of RSIs of the same class and uses spatial similarity to enhance them for FGSC, which can be conducive to

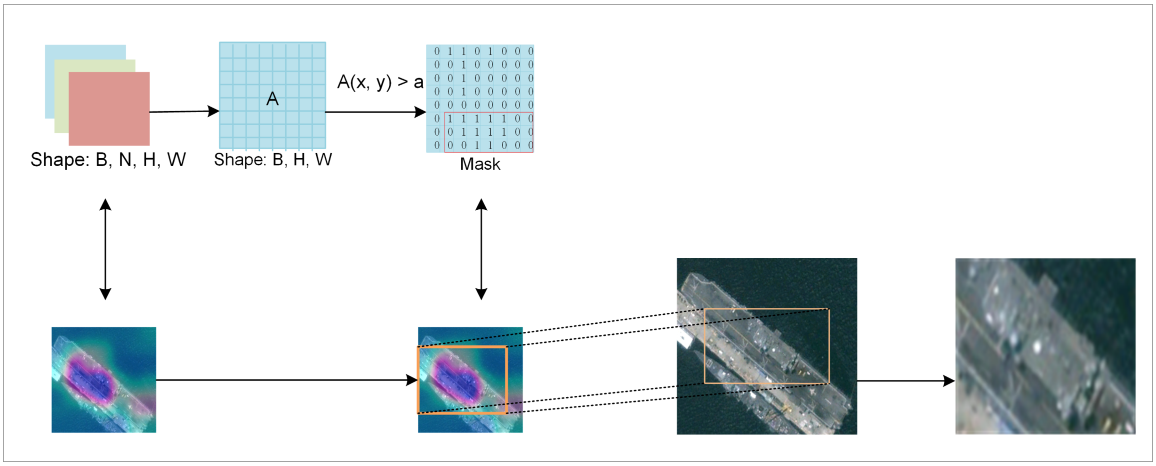


Fig. 3. Steps of the RCEM.

the MSCL-Net to learn the most similarity features and address the challenges of significant interclass similarity and intraclass difference. Presently, Zhang et al. [49] proposed a co-attention module (CA-Module) to extract the interaction features among channels within two images belonging to the same class. However, the CA-Module extracts only the similarity feature with the channel, ignoring the spatial similarity. Hence, we propose a CSAM based on cosine similarity, which can extract the most similar (contrastive) feature to enhance the feature of the original image. The details are introduced as follows.

Define F_o denotes the feature map of original images, F_c denotes the feature map of contrastive images, and the shape of F_o and F_c is (B, N, W, H). First, to compute the similarity of the channel, we reshape the F_o and F_c to (B, N, H*W). Then, we compute the corresponding channel similarity by

$$\text{CA} = \frac{\sum_{i=1}^{N} (F_{oi} \times F_{ci})}{\sqrt{\sum_{i=1}^{N} (F_{oi})^2} \times \sqrt{\sum_{i=1}^{N} (F_{ci})^2}} \tag{5}$$

where N represents the number of channels, F_{oi} and F_{ci} denote the ith channel vector.

Likewise, for spatial similarity, we compute each pixel similarly in the feature map by

$$\text{SA} = \frac{\sum_{m=1}^{H} \sum_{n=1}^{W} (F_{omn} \times F_{cmn})}{\sqrt{\sum_{m=1}^{H} \sum_{n=1}^{W} (F_{omn})^2} \times \sqrt{\sum_{m=1}^{H} \sum_{n=1}^{W} (F_{cmn})^2}}. \tag{6}$$

Finally, we utilize the CA and SA to extract the most similar channel and spatial features. Meanwhile, the suboptimal information is useful for FGSC. Hence, we use the channel similarity feature map F_{CA} to enhance the feature of the direct use channel and spatial attention obtained, which can be computed by

$$\begin{aligned} F_{\text{CA}} &= \text{CA} \bigotimes F_o \\ F_{\text{CSAM}} &= \text{SA} \bigotimes F_{\text{CA}} \bigoplus F_{\text{CA}} \end{aligned} \tag{7}$$

where $\bigoplus$ represents an addition and $\bigotimes$ represents a matrix multiplication.

C. Feature Fusion and Inference Process

In general, the fine-grained feature or the feature of the original image is used to predict the ship class in RSI. However, considering that the mask matrix M can consist of multiple attention areas, the fine-grained features of crop images can lose some vital region features. In addition, the direct use of the features of the original image can decrease the classification accuracy. Therefore, we use the CSAM to fuse the fine-grained feature with the original feature, which can retain the global information and enhance the local region feature.

In addition, this section introduces the training and testing processes of MSCL-Net. As illustrated in Fig. 2, MSCL-Net comprises five branches, all of which share weight parameters. The five branches utilize different features to train MSCL-Net, including the features extracted from the original image using ResNet50 and FPN, the contrastive features, the fine-grained features, and the fusion features. During training, these branches work collectively to update the parameters of the network. In the testing process, we use only the fusion feature to predict the ship class. This design allows the MSCL-Net to use multitype features during the training and test process, which improves performance and generalizability.

D. Combined Loss

Fig. 4 shows the distribution of annotated instances per class for FGSC-23 and FGSC-41. Figs. 1 and 4 show the RSIs are usually significant interclass similarity and intraclass difference, and the dataset is imbalanced. With significant advancements in deep learning applied in natural image classification, various loss functions have been proposed to enhance the performance of classification methods. For the challenge of significant interclass similarity and intraclass difference, the center loss is used for FGVC [49]. Likewise, focal loss [52] or PolyLoss [28] are used to address the issues caused by data imbalance.

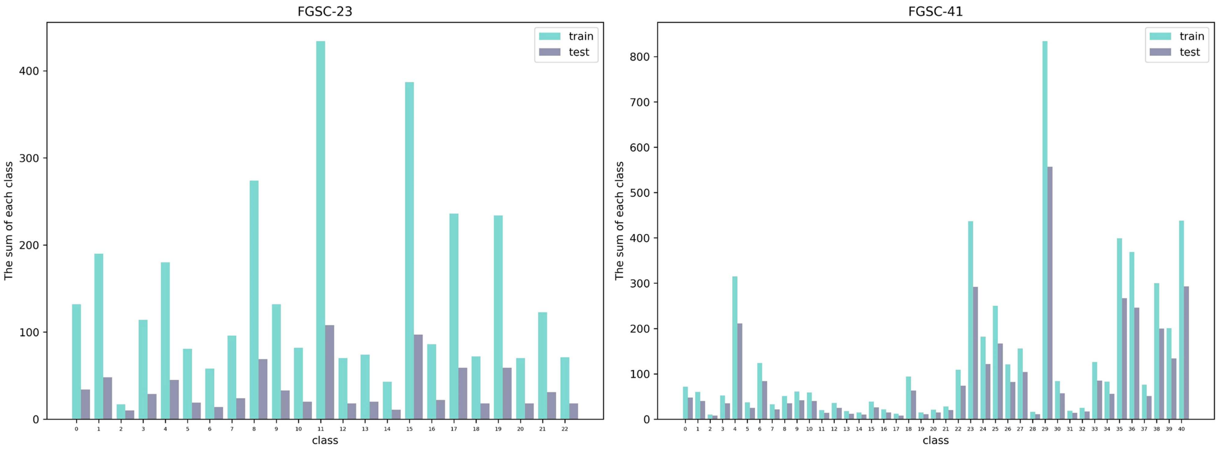


Fig. 4. Distribution of annotated instances per class for FGSC-23 and FGSC-41.

The primary purpose of the center loss is to minimize the distance between each sample and its class center, thereby enhancing the discriminative power of the classification network. The center loss can be computed by

$$L_{\text{CL}} = \frac{1}{2}\sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2 \tag{8}$$

where x_i and c_{y_i} represent the samples in the feature space and the category centers associated with class y_i. $x_i - c_{y_i}$ denotes the euclidean distance from sample x_i to its corresponding class center c_{y_i}.

PolyLoss aimed to address the limitations of cross-entropy loss for imbalanced datasets. This enables the network to apply a stronger penalty to hard samples, further improving its classification performance for imbalanced datasets. The calculation equation is as follows:

$$\begin{aligned} L_{\text{PL}} &= \alpha_1(1-P_t) + \alpha_2(1-P_t)^2 \cdots + \alpha_N(1-P_t)^N \\ &= \sum_{j=1}^{\infty}(\alpha_j(1-P_t)^j). \end{aligned} \tag{9}$$

Improving the loss to address the challenge reduces the complexity of network and computational resources. Hence, we propose replacing the common cross-entropy loss with a combined loss to improve the accuracy of MSCL-Net. This combined loss includes center loss and PolyLoss, as expressed in

$$L = \frac{1}{m}\sum_{i=1}^{m} L_{\text{PL}}(F_i) + L_{\text{CL}} \tag{10}$$

where L_{CL} and L_{PL} denote the center loss and PolyLoss, respectively. F_i represents different types of features extracted from MSCL-Net.

IV. Experimental Results

A. Dataset

The fine-grained ocean ship classification datasets based on optical RSIs primarily include FGSCR-42 [30] and FGSC-23 [20]. Currently, many researchers employ these two datasets to evaluate their methods. Nevertheless, the test set and training set of FGSCR-42 contain numerous repeated images, and even some images from different classes are repeated, as shown in Fig. 5. Consequently, we used FGSC-23 to verify our proposed method. Meanwhile, to avoid the result occasionally, we created a new dataset named FGSC-41 to further evaluate MSCL-Net. The images of FGSC-41 were collected from publicly available optical remote sensing datasets, Google Earth resources, and ship RSIs from the internet. It encompasses 41 distinct classes of ships, including military and commercial ships. The details of FGSC-23 and FGSC-41 are as follows.

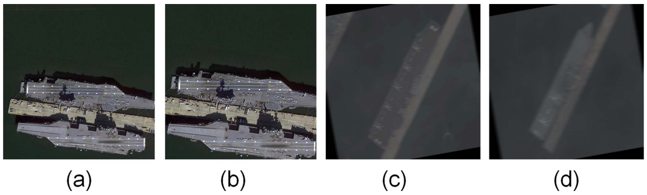


Fig. 5. (a) is from the training set of the Nimitz-class aircraft carrier category, (b) is from the testing set of the same category, (c) is from the Lzumo-class helicopter destroyer category, and (d) is from the Osumi-class landing ship category.

1) FGSC-23: This dataset consists of high-resolution RSIs and is usually used for FGSC. These images are primarily sourced from Google Earth and Gaofen-1 (GF-1), which offers diverse image scenes, intricate ship classification, and comprehensive labeling. It comprises a total of 4080 images distributed across 23 different ship classes. In addition, FGSC-23 is partitioned to the training set with 3256 samples and the test set with 824 samples.

2) FGSC-41: This is a self-built dataset constructed by cropping, resizing, eliminating duplicates, and labeling the images sourced from DOTA [53], HRSC2016 [3], ShipRSImageNet [54], FGSCR-42, and NWPUVHR-10 [55]. FGSC-41 includes 25 types of military ships (aircraft carriers, cruisers, transport ships, etc.) and 16 types of civilian ships (cargo ships, barge ships, oil tanker ships, fishing boats, yachts, motorboats, etc.). The detailed information about FGSC-41 is presented in Table I, which includes class details and the number of samples.

TABLE I
DETAILED INFORMATION OF FGSC-41

Id	Class name (level1)	Class name (level2)	Class name (level3)	Class numbers	Train set	Test set
1	Warship	Aircraft carrier	Nimitz-class aircraft carrier	100	60	40
2			Midway-class aircraft carrier	18	10	8
3			Enterprise-class aircraft carrier	87	52	35
4		Destroyer	Arleigh Burke-class destroyer	526	315	211
5			Asagiri-class destroyer	62	37	25
6			Atago-class destroyer	208	124	84
7			Hatsuyuki-class destroyer	55	33	22
8			Hyuga-class destroyer	86	51	35
9		Amphibious ship	Austin-class amphibious transport dock	103	61	42
10			Whidbey Island-class amphibious transport dock	99	59	40
11			Osumi-class amphibious transport dock	34	20	14
12			San Antonio-class amphibious transport dock	61	36	25
13			YuDao-class amphibious transport dock	30	18	12
14			YuDeng-class amphibious transport dock	25	15	10
15			YuTing-class amphibious transport dock	65	39	26
26			Yuzhao-class amphibious transport dock	37	22	15
17			Wasp-class amphibious assault ship	20	12	8
18			America-class amphibious assault ship	157	94	63
19		Perry-class frigate	Perry-class frigate	500	300	200
20		Ticonderoga-class cruiser	Ticonderoga-class cruiser	335	201	134
21		Auxiliary ship	Masyuu-class auxiliary ship	36	21	15
22		Oiler ship	Oiler ship	48	28	20
23		Command ship	Command ship	120	72	48
24		Submarine	Submarine	731	438	293
25		Expeditionary fast transport	Expeditionary fast transport	26	15	11
26	Civil ship	Barge ship	Barge ship	183	109	74
27		Cargo ship	Other cargo ship	729	437	292
28			Container ship	304	182	122
29		Fishing vessel	Fishing vessel	417	250	167
30		Motorboat	Motorboat	1391	834	557
31		Oil tanker	Oil tanker	141	84	57
32		Test ship	Test ship	33	19	14
33		Training ship	Training ship	42	25	17
34		Tugboat	Tugboat	211	126	85
35		Patrol ship	Patrol ship	139	83	56
36		Sailboat	Sailboat	666	399	267
37		Yacht	Yacht	615	369	246
38		RoRo ship	RoRo ship	127	76	51
39		Hovercraft	Hovercraft	260	156	104
40		Medical ship	Medical ship	27	16	11
41		Ferry	Ferry	203	121	82

FGSC-41 consists of a total of 9057 ship images. We partitioned the dataset into training and testing sets at a 6:4 ratio, resulting in 5419 samples as the training set and 3638 samples as the testing set.

B. Evaluation Metrics

Considering the sample imbalance in the FGSC datasets, the overall accuracy (OA), average accuracy (AA), weighted mean precision (MP), and P-R curve were selected as the main metrics for evaluating the performance of the model. OA mainly evaluates the model from the overall perspective. In short, it is the proportion of correctly predicted samples from the total samples and can be computed by

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{11}$$

where true positive (TP) denotes the correct classification of positive samples, true negative (TN) denotes the correct classification of negative samples, false positive (FP) represents the error classification of positive samples, and false negative (FN) indicates the error classification of negative samples.

TABLE II
COMPARISON RESULTS OF DIFFERENT METHODS ON FGSC-23, INCLUDING OVERALL ACCURACY (OA%), MEAN PRECISION (MP%), AVERAGE ACCURACY (AA%), AND THE ACCURACY OF EACH CLASS

Method	MMAL-Net (2021)	GLPM (2022)	P^2 Net (2022)	PRF-Net (2023)	EFM-Net (2023)	MSCL-Net (ours)	Δ
class0	88.24	**94.12**	85.29	91.18	**94.12**	91.18	–2.94
class1	93.52	95.83	85.42	93.75	93.75	**97.22**	+1.39
class2	90.91	90.00	70.00	**100.0**	**100.0**	95.45	–4.55
class3	93.22	93.10	82.76	93.10	96.55	**98.31**	+1.76
class4	72.22	62.22	51.11	66.67	64.44	**83.33**	+11.11
class5	83.05	84.21	68.42	84.21	73.68	**88.14**	+3.93
class6	83.33	92.86	71.43	92.86	**100.0**	94.44	–5.56
class7	93.55	**100.0**	95.83	**100.0**	**100.0**	**100.0**	0.00
class8	61.11	73.91	66.67	75.36	84.06	**88.89**	+4.84
class9	89.58	81.82	48.48	87.88	87.88	**97.92**	+8.34
class10	**100.0**	45.00	35.00	60.00	45.00	**100.0**	0.00
class11	72.41	88.89	93.52	84.26	96.30	**96.55**	+0.25
class12	**82.22**	55.56	66.67	72.22	72.22	75.56	–6.66
class13	89.47	**100.0**	**100.0**	95.00	**100.0**	84.21	–15.79
class14	92.86	90.91	**100.0**	90.91	**100.0**	**100.0**	0.00
class15	95.83	82.47	78.35	86.60	81.44	**100.0**	+4.17
class16	72.46	90.91	50.00	86.36	**95.45**	84.06	–11.39
class17	75.76	**96.61**	89.83	89.83	88.14	93.94	–2.67
class18	60.00	**94.44**	66.67	72.22	88.89	70.00	–24.44
class19	61.11	79.66	**83.05**	79.66	83.05	72.22	–10.83
class20	95.00	94.44	66.67	66.67	**100.0**	**100.0**	0.00
class21	90.91	96.77	77.42	93.55	**100.0**	**100.0**	0.00
class22	89.69	83.33	55.56	88.89	88.89	**90.72**	+1.03
OA	85.07	85.07	76.46	84.34	87.62	**91.50**	+3.88
MP	86.53	85.11	76.79	84.59	87.94	**91.63**	+3.69
AA	83.76	85.53	73.40	84.83	88.43	**91.40**	+2.97

The bold values represent the optimal results.

AA denotes the average accuracy of each class, which can be used to evaluate the performance of each class. The AA metric is defined as

$$\mathrm{AA} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{TP}_i + \mathrm{TN}_i}{\mathrm{TP}_i + \mathrm{TN}_i + \mathrm{FPI} + \mathrm{FN}_i} \tag{12}$$

where n represents the number of classes. TP_i denotes ith class correct classification of positive samples.

Precision is usually used as the evaluation metric for imbalanced datasets and is defined in (13). In our study, the dataset is also imbalanced and includes multiple classes. We use the weighted MP to evaluate the performance and compute it by (14)

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP + FP}} \tag{13}$$

$$\mathrm{MP} = \sum_{i=1}^{n}(P_i \times W_i) \tag{14}$$

where P_i and W_i represent the precision and weight of the ith classes, respectively. The P-R curve represents the relationship between recall and precision and can show the classification performance more intuitively and clearly.

In addition, floating point operations (FLOPs) and model parameters (Params) are adopted to illustrate the computational complexity of the network in the inference phase.

C. Implementation Details

All the experiments were implemented on an NVIDIA GeForce RTX 2080ti using PyTorch 1.9.1. We did not use auxiliary information in the experiment, and the class label of the image is the only label used for training. During the preprocessing, the images in datasets are resized to 224 × 224. Moreover, the random gradient descent optimizer (SGD) is used. The initial learning rate was set at 0.01 and gradually decreased with increasing epochs.

D. Comparison With State-of-the-art Methods

With the advance of fine-grained classification, plenty of state-of-the-art methods have been proposed for FGSC or FGVC, such as PRF-Net [5], MMAL-Net [6], EFM-Net [21], P^2Net [23], and GLPM [17]. We evaluate the performance of MSCL-Net by comparing with these methods on FGSC-23 and FGSC-41. The detailed comparison results are as follows.

TABLE III
COMPARISON RESULTS OF DIFFERENT METHODS ON FGSC-41, INCLUDING OVERALL ACCURACY (OA%), MEAN PRECISION (MP%), AVERAGE ACCURACY (AA%), AND THE ACCURACY OF EACH CLASS

Method	MMAL-Net (2021)	GLPM (2022)	P^2 Net (2022)	PRF-Net (2023)	EFM-Net (2023)	MSCL-Net (ours)	Δ
class0	79.17	77.08	66.67	72.92	81.25	**91.67**	+10.42
class1	85.00	87.50	70.00	87.50	**97.50**	**97.50**	0
class2	87.50	87.50	**100.0**	50.00	**100.0**	50.00	–50.00
class3	**91.43**	82.86	68.57	85.71	88.57	88.57	–2.86
class4	95.73	91.94	88.15	90.52	94.31	**98.58**	+2.85
class5	68.00	56.00	**72.00**	52.00	44.00	**72.00**	0
class6	91.67	83.33	89.29	96.43	92.86	**96.43**	0
class7	**59.09**	40.91	27.27	36.36	45.45	50.00	–9.09
class8	91.43	94.29	94.29	94.29	94.29	**94.29**	0
class9	90.48	80.95	71.43	83.33	**92.86**	88.10	–4.76
class10	80.00	67.50	55.00	72.50	87.50	**90.00**	+2.50
class11	**100.0**	**100.0**	92.86	92.86	**100.0**	**100.0**	0
class12	84.00	88.00	52.00	76.00	**92.00**	**92.00**	0
class13	**100.0**	**100.0**	83.33	91.67	**100.0**	**100.0**	0
class14	**80.00**	60.00	20.00	70.00	60.00	70.00	–10.00
class15	**92.31**	**92.31**	80.77	88.46	**92.31**	88.46	–3.85
class16	**86.67**	**86.67**	73.33	**86.67**	**86.67**	80.00	–6.67
class17	87.50	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	0
class18	90.48	95.24	74.60	93.65	**96.83**	95.24	–1.59
class19	**100.0**	72.73	36.36	72.73	81.82	81.82	–18.18
class20	86.67	86.67	**100.0**	80.00	86.67	**100.0**	0
class21	**65.00**	50.00	60.00	50.00	55.00	**65.00**	0
class22	52.70	48.65	32.43	54.05	52.70	**58.11**	+4.06
class23	74.66	78.08	66.10	83.90	83.90	**85.96**	+2.06
class24	73.77	73.77	35.25	70.49	**82.79**	78.69	–4.10
class25	46.71	64.07	38.32	58.08	58.68	**67.07**	+3.00
class26	29.27	48.78	14.63	43.90	**51.22**	50.00	–1.22
class27	93.27	86.54	81.73	**98.08**	97.12	97.12	–0.96
class28	90.91	**100.0**	27.27	63.64	**100.0**	**100.0**	0
class29	77.20	76.66	78.99	75.40	78.82	**82.23**	+3.24
class30	45.61	56.14	26.32	47.37	66.67	**68.42**	+1.75
class31	64.29	78.57	50.00	71.43	57.14	**92.86**	+14.29
class32	**88.24**	82.35	82.35	70.59	82.35	76.47	–11.77
class33	47.06	49.41	36.47	47.06	48.24	**55.29**	+5.88
class34	80.36	82.14	75.00	73.21	78.57	**91.07**	+8.93
class35	75.28	77.90	62.92	76.78	**85.39**	83.90	–1.49
class36	59.35	**82.93**	63.41	80.08	76.83	82.11	–0.82
class37	92.16	94.12	74.51	80.39	**96.08**	94.12	–1.96
class38	**96.50**	92.00	86.50	89.00	94.50	94.50	–2.00
class39	88.81	82.84	75.37	90.30	91.04	**96.27**	+5.23
class40	97.27	96.93	95.56	98.29	99.32	**100.0**	+0.68
OA	77.71	79.49	68.91	78.89	82.60	**85.18**	+2.58
MP	78.42	79.57	68.27	79.03	82.44	**85.21**	+2.77
AA	79.65	78.86	65.34	75.50	81.74	**84.00**	+2.26

The bold values represent the optimal results.

1) Results of OA, MP, AA, FLOPs and Params: Table II shows the comparison results of different methods on FGSC-23, including OA, MP, AA, and the accuracy of each class. The OA, MP, and AA of MSCL-Net are 91.50, 91.63, and 91.40, respectively, which are the highest compared with those of other state-of-the-art methods. Specifically, OA is 3.88 higher, MP is 3.69 higher, and AA is 2.97 higher. By analyzing the accuracy of each class, the accuracy of most classes is higher than that of other state-of-the-art methods. Notably, the accuracy of each class is very balanced, unlike other methods, which have very low accuracy for some particular classes. This phenomenon proves that the generalization ability and stability of MSCL-Net for different classes of images are better.

Table III shows the comparison results of different methods on FGSC-41. FGSC-41 is a dataset that consists of more classes than FGSC-23, which can be used to evaluate better the performance of MSCL-Net. The OA, MP, and AA of MSCL-Net are 85.18, 85.21, and 84.00 on FGSC-41. OA is 2.58 higher

TABLE IV
COMPARISON OF MODEL PARAMETERS AND FLOPS

Method	FLOPs (G)	Params (M)
MMAL-Net (2021)	57.84	**23.59**
P^2 Net (2022)	17.01	35.32
GLPM (2022)	119.91	99.40
PRF-Net (2023)	21.44	74.01
EFM-Net (2023)	15.37	91.76
MSCL-Net (ours)	**13.55**	24.67

The bold values represent the optimal results.

than the state-of-art methods, MP is 2.77 higher, and AA is 2.26 higher. In addition, the accuracy of each class on FGSC-41 is also better balanced, which is consistent with the results on FGSC-23. These results mean the performance of MSCL-Net is also the best for the more complex datasets.

From Tables II and III, we can discover that the performance of EFM-Net surpasses that of other comparative methods, indicating the importance of essential features for FGSC. In addition, the results of MMAL-Net, GLPM-Net, and PRF-Net are almost consistent on FGSC-23, while PRF-Net outperforms MMAL-Net and GLPM-Net on FGSC-41. These methods all utilize local and global features for FGSC, and PRF-Net incorporates additional attention mechanisms and region cropping networks. This represents that attention mechanisms and region-focused features are useful for FGSC, particularly in the case with significant interclass similarity and intraclass differences.

Table IV shows the compare result of model parameters and FLOPs. The floating point operations and parameters of MSCL-Net are 13.55 G and 24.67 M. The FLOPs are the lowest compared to those of the other methods, which means that the computational complexity of MSCL-Net is lower. The number of parameters of MMAL-Net is the lowest, which is 1.08 M lower than that of MSCL-Net. The reason can be that MSCL-Net adds the FPN for FGSC compared with MMAL-Net. However, although the design of MSCL-Net has increased parameters, it has also improved accuracy steeply. The MSCL-Net balances the computational efficiency and parameters compared with other methods, ensuring that the computational complexity can be kept low even when the parameters are slightly higher.

We further explore Table III and observe the accuracy of class 2, class 7, class 26, and class 33 are close to 50. By analyzing the original images of FGSC-41, we discovered that the images of these classes are fuzzy and should be seriously influenced by various degradation, noise effects, or variabilities. Our method can locate the similarity feature of the same image and the fine-grained feature in RSI, which can address the challenge of complex backgrounds. However, spectral variability and environmental variability can lead to the same class of ship exhibiting different visual features, which increases the difficulty of FGSC. At present, some methods are proposed to address the challenge caused by spectral variability. Specifically, Hong et al. [56] proposed an augmented linear mixing model (ALMM) to address spectral variability for hyperspectral unmixing. The ALMM utilizes not only the principal scaling factor but also introduces the spectral variability dictionary to expand the scalability of the endmember dictionary. The results of ALMM are better than those of methods that do not consider spectral variability. These impact factors imply designing the methods based on the process of the noise data or considering spectral variability may be efficient in future work.

2) Visual Analysis: To more clearly analyze the results, we further provided some visual analysis. Fig. 6 shows the precision-recall (P-R) curve of MSCL-Net on FGSC-23 and FGSC-41, which provides a clear visual representation. Generally, the P-R curve approaching the top-right corner demonstrates that the model achieves high precision while maintaining a high recall rate. Likewise, in Fig. 6, the curve of MSCL-Net approaches the top-right corner, which also indicates that the performance of MSCL-Net is best.

We also observe that the performance of MMAL-Net is lower than PRF-Net on FGSC-41, while it is higher on FGSC-23. In addition, PRF-Net incorporates features from multiple stages of ResNet50 and attention mechanisms for fine-grained classification, unlike MMAL-Net. In FGSC-41, the size of ship targets in RSIs varies, further indicating that multilayer features of ResNet50 are useful for FGSC.

E. Parameter Sensitivity Analysis

In the section on implementation details, we introduce two important parameters when training the network: the learning rate and batch size. The learning rate controls the step size during optimization. A learning rate that is too small may unnecessarily lengthen the training process and lead to a suboptimal solution. A value that is too large can cause instability in the learning process. To evaluate the learning rate sensitivity, we trained MSCL-Net with learning rates of 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, and 0.1. The batch size defines the number of training samples considered for each calculation of weight update. Setting this hyperparameter too high can result in high memory requirements. A value that is too low can cause the model to bounce back and forth without converging. We evaluated the performance of MSCL-Net with batch sizes of 8,16, 32, and 64.

Table VI shows the results with different learning rates and batch sizes on FGSC-23. The higher results center on the learning rates of 0.01–0.001 and batch sizes of 8 and 16, consistent with the well-known learning rate and batch size setting. These results show it is not essential to set a particular learning rate and batch size for MSCL-Net during training, which means our network is minimally impacted by the parameters that we set.

V. DISCUSSION

MSCL-Net mainly includes four parts: multiscale feature learning, contrastive learning, feature fusion, and combined loss. The multiscale feature learning can be further divided into multilayer feature extract of network and fine-grained feature extract. Likewise, the combined loss includes two aspects of improvement. Specifically, the PolyLoss is used to replace the cross-entropy, and the center loss increases the distance between different class features in the feature space. To evaluate the

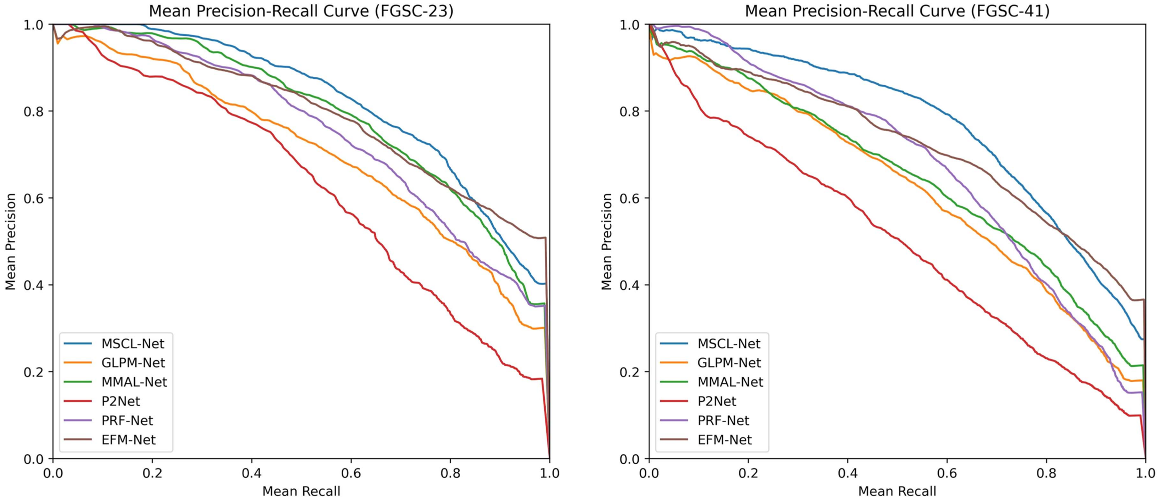


Fig. 6. P-R curves of MSCL-Net on FGSC-23 and FGSC-41.

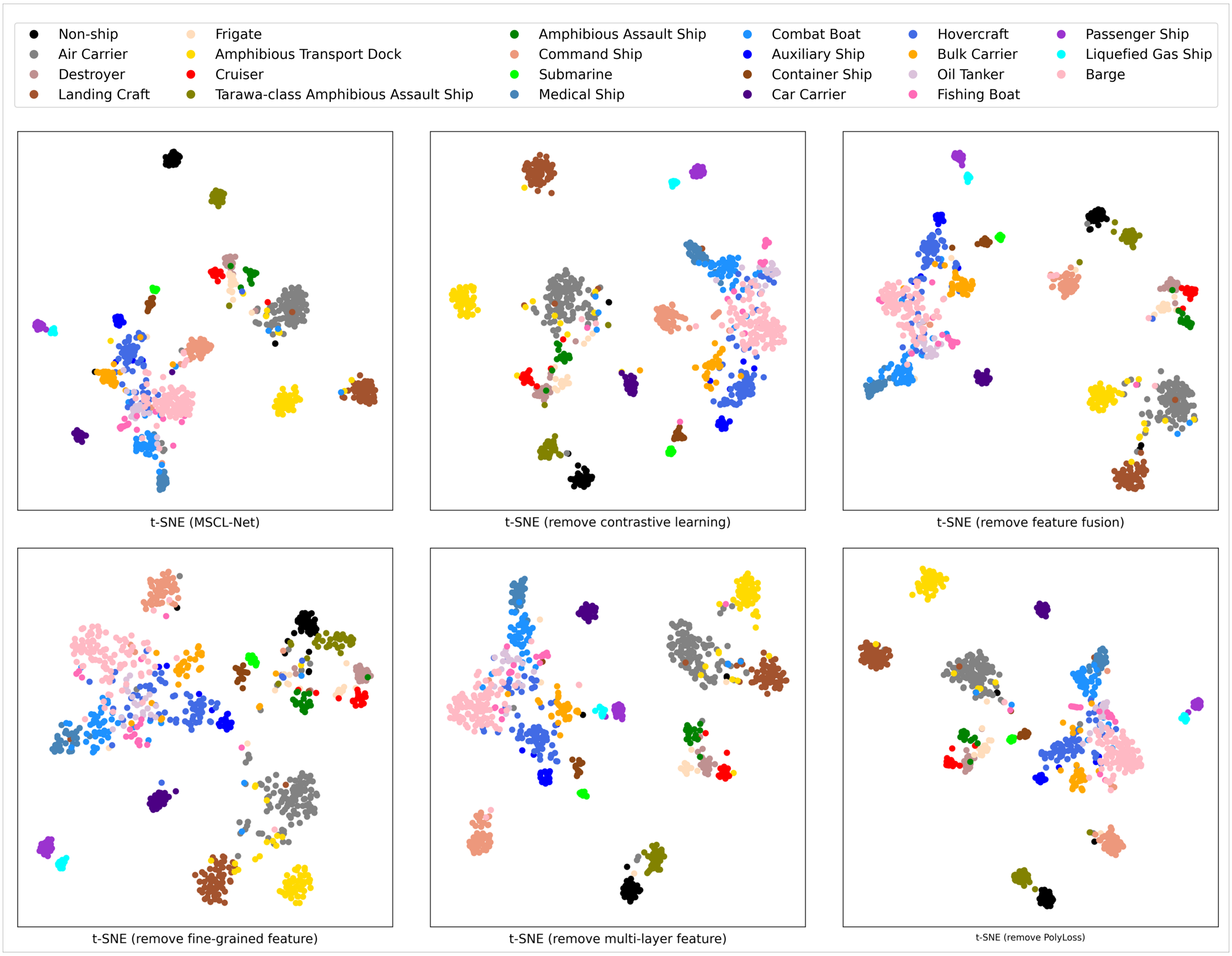


Fig. 7. Feature distribution of the predicted features on FGSC-23.

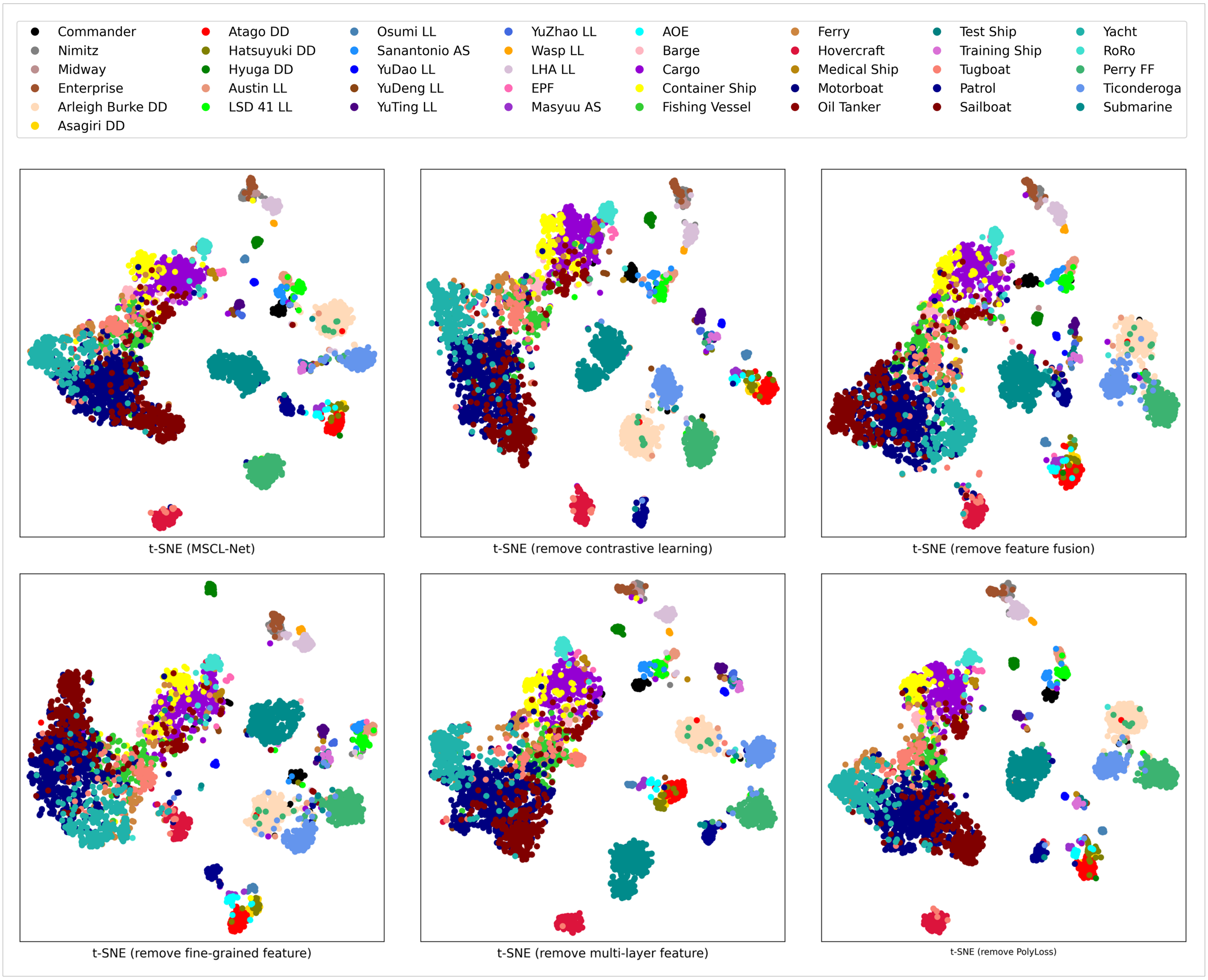


Fig. 8. Feature distribution of the predicted features on FGSC-41.

contributions of each part of MSCL-Net, we conducted a series of ablation studies.

We investigate the ablation studies of different parts in two ways. First, we explore the influence of each part by removing one part. Second, to analyze the relationships among the different parts, we designed the ablation studies by randomly removing two main parts. Specifically, we focus on studying the relationships among multilayer feature extraction, fine-grained feature extraction, and contrastive learning, that is, exploring whether the fusion of these parts contributes to improving the accuracy of FGSC.

Table V shows the ablation results on FGSC-23 and FGSC-41. The result of randomly removing one or two parts is lower than that of MSCL-Net, which implies that these parts contribute to improving the performance of the network. Furthermore, the OA, MP, and AA are 87.14, 87.77, and 88.22 when removing the part of the fine-grained feature extraction on FGSC-23. Likewise, the OA, MP, and AA are 79.80, 79.67, and 77.25 on FGSC-41. The results of removing the fine-grained feature extraction are the lowest, representing the part with the greatest contribution to FGSC. In addition, we can discover the accuracy of removing center loss is closest to MSCL-Net, which represents the parts that can be difficult to distinguish the features in the feature space due to the challenge of significant interclass similarity and intraclass difference. The AA of replacing PolyLoss with cross-entropy on FGSC-23 is closer to the accuracy of MSCL-Net than the result on FGSC-41. The reason is that the dataset of FGSC-41 is more imbalanced than FGSC-23, which means that PolyLoss is efficient in improving the accuracy of the network with the imbalanced dataset.

The results of removing two parts of contrastive learning and fine-grained feature extraction are 86.77, 87.11, and 86.36 on FGSC-23. The OA, MP, and AA are 4.73, 4.52, and 5.04 lower than the results of MSCL-Net on FGSC-23. The results of remove contrastive learning are 1.09, 0.83, and 0.69 lower than the results of MSCL-Net. The results of removing fine-grained feature extract are 4.36, 3.86, and 3.18 lower than the results of MSCL-Net. Generally, the results of removing

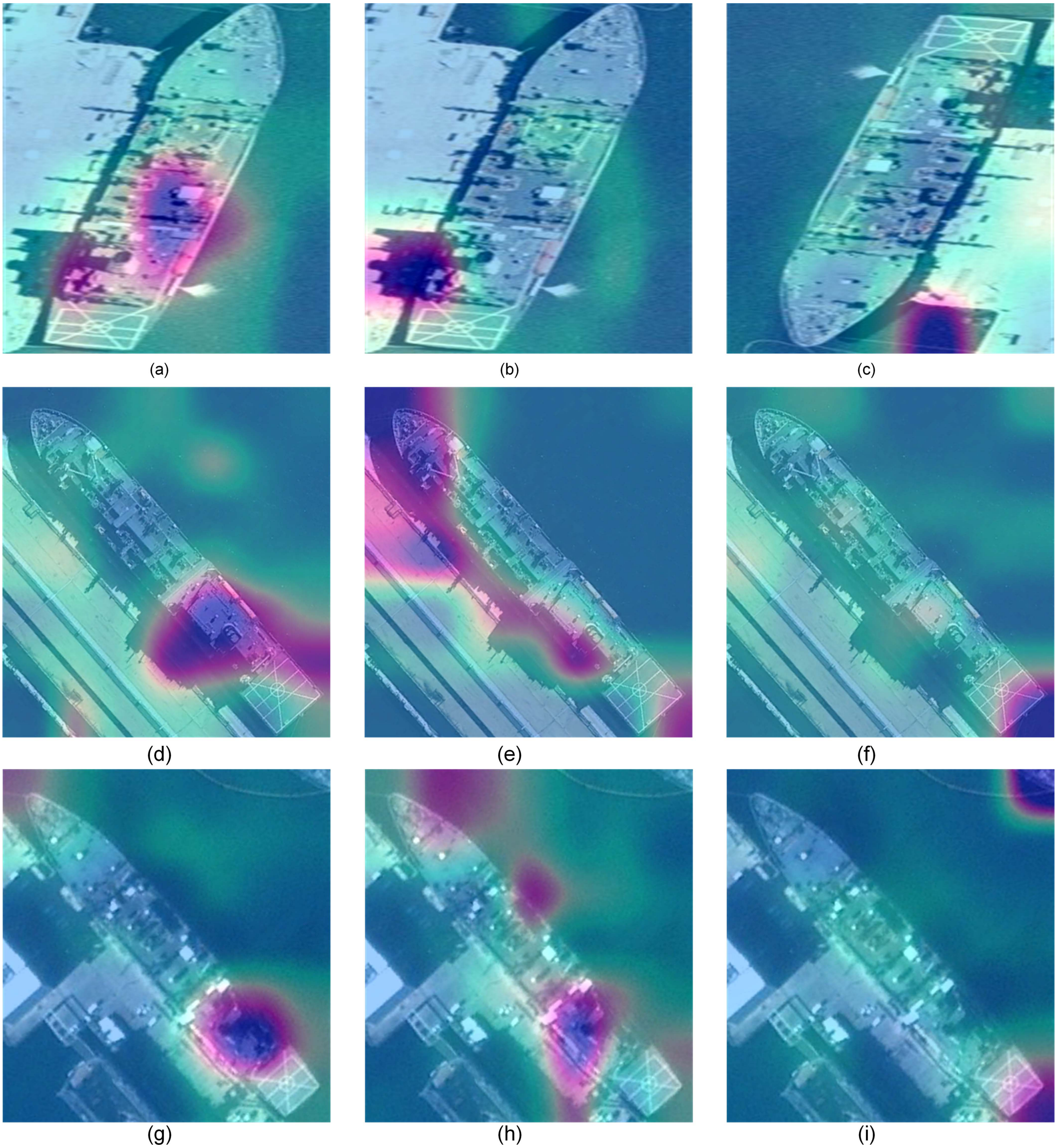


Fig. 9. Heatmap of the features of three images with the same class. Figures (a), (d), and (g) are the heatmap using the predicted feature of MSCL-Net. Figures (b), (e), and (h) are the heatmap using the predicted feature of removing the contrastive learning part. Figures (c), (f), and (i) are the heatmap using the predicted feature of removing the fusion feature part.

these two parts should be 5.45, 4.69, and 3.87 lower than the results of MSCL-Net. However, OA and MP are higher, and AA is lower, representing that combining contrastive learning and fine-grained feature extraction can improve the total classification accuracy and balance the network performance. Similarly, removing contrastive learning and multilayer feature extraction results are 1.69, 1.75, and 1.26 lower than the result of MSCL-Net on FGSC-23. The results of MP and AA are higher, and OA is lower than the mean of independent removal of the two parts, which represents these two parts as being more focused on the classification of each class. The results for removing the multilayer feature and fine-grained feature

TABLE V
RESULTS OF THE ABLATION EXPERIMENT ON FGSC-23 AND FGSC-41

Fine-grained	Contrastive	Multilayer	Feature	Center	Poly	FGSC-23						FGSC-41					
feature	learning	feature	fusion	loss	Loss	OA	Δ	MP	Δ	AA	Δ	OA	Δ	MP	Δ	AA	Δ
–	+	+	+	+	+	87.14	–4.36	87.77	–3.86	88.22	–3.18	79.80	–5.38	79.67	–5.54	77.25	–6.75
+	–	+	+	+	+	90.41	–1.09	90.80	–0.83	90.71	–0.69	83.48	–1.70	83.95	–1.26	80.87	–3.13
+	+	–	+	+	+	89.93	–1.57	90.06	–1.57	89.76	–1.64	83.51	–1.67	83.85	–1.36	81.21	–2.79
+	+	+	–	+	+	88.96	–2.54	89.02	–2.61	88.90	–2.50	80.40	–4.78	80.63	–4.58	78.33	–5.67
+	+	+	+	–	+	90.78	–0.72	90.92	–0.71	90.20	–1.20	84.69	–0.49	84.79	–0.42	82.72	–1.28
+	+	+	+	+	–	90.78	–0.72	91.01	–0.62	90.95	–0.45	84.91	–0.27	84.90	–0.31	82.28	–1.72
–	–	+	+	+	+	86.77	–4.73	87.11	–4.52	86.36	–5.04	78.50	–6.68	79.23	–5.98	73.94	–10.06
+	–	–	+	+	+	89.81	–1.69	89.88	–1.75	90.14	–1.26	84.61	–0.57	84.92	–0.29	81.57	–2.43
–	+	–	+	+	+	88.83	–2.67	89.08	–2.55	88.37	–3.03	79.33	–5.85	79.93	–5.28	76.02	–7.98
+	+	+	+	+	+	**91.50**	/	**91.63**	/	**91.40**	/	**85.18**	/	**85.21**	/	**84.00**	/

The bold values represent the optimal results.

TABLE VI
RESULTS WITH DIFFERENT LEARNING RATES AND BATCH SIZES ON FGSC-23

Learning rate	batch = 8			batch = 16			batch = 32			batch = 64		
	OA	MP	AA	OA	MP	AA	OA	MP	AA	OA	MP	AA
LR = 0.1	77.18	77.68	76.22	80.10	80.46	78.82	81.19	81.36	81.25	69.78	69.98	67.96
LR = 0.05	83.13	84.05	83.82	83.74	83.94	83.58	87.62	87.92	88.09	90.53	90.74	**90.78**
LR = 0.01	90.41	90.57	89.86	91.50	91.63	91.40	**91.26**	91.24	90.55	**90.41**	**90.34**	89.06
LR = 0.005	91.63	91.72	91.22	**91.99**	**92.05**	**91.88**	**91.26**	**91.26**	**90.85**	89.32	89.45	89.90
LR = 0.001	**91.99**	**92.25**	**92.12**	90.41	90.48	90.29	90.53	90.65	90.76	87.14	87.54	83.74
LR = 0.0005	90.90	91.15	92.04	90.78	90.79	90.75	87.99	87.35	84.74	80.70	79.89	72.76
LR = 0.0001	85.80	85.30	79.76	72.09	69.23	58.15	46.84	42.79	30.33	27.91	21.11	11.74

The bold values represent the optimal results.

extraction are higher than the mean for independently removing the two parts, demonstrating that multiscale feature learning contributes to improving the performance of the network for FGSC.

Likewise, the results on FGSC-41 of removing the contrastive learning and fine-grained feature extraction, as well as removing the two parts of multiscale, are consistent with the results on FGSC-23. However, different from the results on FGSC-23, the removed contrastive learning and multilayer feature extract results are 0.57, 0.29, and 2.43 lower than the result of MSCL-Net on FGSC-41. The results of OA, MP, and AA are higher than the mean of independent removing the two parts. The reason can be the ship target size on FGSC-41 is various, and numerous images of ship targets exist with small sizes.

T-distributed stochastic neighbor embedding (t-SNE) is often used for 2-D visualization of high-dimensional features. To further explore the contributions of each part, we use the t-SNE to visualize the feature distribution of the predicted features. The vision image is shown in Figs. 7 and 8. The result images show the distribution of removing the fine-grained feature extraction is the most chaotic, which represents the part of the fine-grained feature extraction that is more important for FGSC. In addition, for the MSCL-Net, we can see that classes are much farther away from each other, while images of the same class are much closer.

In addition, the most important parts of MCSL-Net are contrastive learning and feature fusion. These two parts ensure that the network can locate the most similar area of the same class and avoid the influence of complex backgrounds. To explore the mechanism, we use the predicted feature of MSCL-Net, the predicted feature of removing the contrastive learning part, and the predicted feature of removing the fusion feature part to draw a heatmap. The results are shown in Fig. 9. Fig. 9(a), (d), and (g) is the heatmap using the predicted feature of MSCL-Net. Fig. 9(b), (e), and (h) is the heatmap using the predicted feature of removing the contrastive learning part. Fig. 9(c), (f), and (i) is the heatmap using the predicted feature of removing the fusion feature part. By observing the left of Fig. 9, we can easily discover that the attention area of MSCL-Net extracted is the same for two images of the same class, and the attention area is the ship rather than the background.

The center part of Fig. 9 displays the area attention is different. The result implies that contrastive learning can make the network notice the similarity area of the same class image. The right part of Fig. 9 shows the result of the predicted feature after removing the feature fusion. We can see that the attention area of the predicted feature focuses on the background area, which means that the part of fine-grained feature extraction and feature fusion contributes to addressing the challenge of the complex background in RSI. These results prove it is correct for our motivation.

VI. CONCLUSION

Two challenges still exist in FGSC: significant interclass similarity and intraclass difference and background complexity. To challenge these challenges, we proposed a novel MSCL-Net.

In contrastive learning, the CSAM is proposed to extract the channel similarity feature and use spatial similarity to enhance the channel similarity feature. In multiscale feature learning, we use FPN to extract the multilayer features of ResNet50, which include rich semantic and detailed information. Meanwhile, we propose an RCEM to cut the original images, which aims to crop and enlarge the maximum attention area of the mask matrix and as the input of ResNet50. In feature fusion, we utilize the CSAM to fuse the fine-grained feature and the feature of the original image, which can extract the most similar channel feature and enhance the spatial similarity feature. Moreover, we propose a combined loss, including PolyLoss and center loss, to improve the performance of MSCL-Net. We evaluate MSCL-Net on FGSC-23 and FGSC-41 and design a series of ablation experiments to verify the efficiency of each part. The results show that MSCL-Net is superior to other state-of-the-art methods.

In addition, spectral variability and environmental variability can lead to the same class of ship exhibiting different visual features, which increases the difficulty of FGSC. In future work, we will design the FGSC network based on the process of the noise data or consider spectral variability to address the influence caused by spectral variability and environmental variability.

Data availability

The data will be made available upon request.

Acknowledgment

The authors would like to express their sincere thanks to the editors and anonymous reviewers for their valuable comments and contributions.

References

[1] C. Pan, R. Li, Q. Hu, C. Niu, W. Liu, and W. Lu, “Contrastive learning network based on causal attention for fine-grained ship classification in remote sensing scenarios,” *Remote Sens.*, vol. 15, no. 13, 2023, Art. no. 3393.

[2] Y. You, B. Ran, G. Meng, Z. Li, F. Liu, and Z. Li, “OPD-Net: Prow detection based on feature enhancement and improved regression model in optical remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6121–6137, Jul. 2021.

[3] 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*, 2017, vol. 2, pp. 324–331.

[4] R. Du et al., “Fine-grained visual classification via progressive multi-granularity training of Jigsaw patches,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 153–168.

[5] J. Feng and M. Gou, “A progressive region-focused network for fine-grained human behavior recognition,” *Hum.-Centric Comput. Inf. Sci.*, vol. 13, 2023, Art. no. 106099.

[6] F. Zhang, M. Li, G. Zhai, and Y. Liu, “Multi-branch and multi-scale attention learning for fine-grained visual categorization,” in *Proc. 27th Int. Conf. MultiMedia Model.*, 2021, pp. 136–147.

[7] Y. Li, L. Chen, W. Li, and N. Wang, “Few-shot fine-grained classification with rotation-invariant feature map complementary reconstruction network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5608312.

[8] J. Shi, Z. Jiang, and H. Zhang, “Few-shot ship classification in optical remote sensing images using nearest neighbor prototype representation,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3581–3590, 2021.

[9] Y. Li and C. Bian, “Few-shot fine-grained ship classification with a foreground-aware feature map reconstruction network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622812.

[10] Y. Li, C. Bian, and H. Chen, “Generalized ridge regression-based channelwise feature map weighted reconstruction network for fine-grained few-shot ship classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600910.

[11] S. Moon, J. Lee, J. Lee, A. R. Oh, D. Nam, and W. Yoo, “A study on the improvement of fine-grained ship classification through data augmentation using generative adversarial networks,” in *Proc. Int. Conf. Inf. Commun. Technol. Convergence*, 2021, pp. 1230–1232.

[12] B. Liu, L. Li, Q. Xiao, W. Ni, and Z. Yang, “Remote sensing fine-grained ship data augmentation pipeline with local-aware progressive image-to-image translation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5631716.

[13] Q. Oliveau and H. Sahbi, “Semi-supervised deep attribute networks for fine-grained ship category recognition,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 6871–6874.

[14] Y. Kim, H. Jang, S. Park, J. Lee, and C. Kim, “Semi-supervised synthetic-to-real domain adaptation for fine-grained naval ship image classification,” 2020. [Online]. Available: https://easychair.org/publications/preprint_open/LFmr

[15] S. Huang, H. Xu, X. Xia, F. Yang, and F. Zou, “Multi-feature fusion of convolutional neural networks for fine-grained ship classification,” *J. Intell. Fuzzy Syst.*, vol. 37, no. 1, pp. 125–135, 2019.

[16] Z. Zhang et al., “Fine-grained ship image recognition based on BCNN with inception and AM-softmax,” *Comput. Mater. Contin.*, vol. 73, no. 1, pp. 1527–1539, 2022.

[17] H. Meng, Y. Tian, Y. Ling, and T. Li, “Fine-grained ship recognition for complex background based on global to local and progressive learning,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 3511905.

[18] L. Huang, F. Wang, Y. Zhang, and Q. Xu, “Fine-grained ship classification by combining CNN and swin transformer,” *Remote Sens.*, vol. 14, no. 13, 2022, Art. no. 3087.

[19] Y. Chen, Z. Zhang, Z. Chen, Y. Zhang, and J. Wang, “Fine-grained classification of optical remote sensing ship images based on deep convolution neural network,” *Remote Sens.*, vol. 14, no. 18, 2022, Art. no. 4566.

[20] X. Zhang, Y. Lv, L. Yao, W. Xiong, and C. Fu, “A new benchmark and an attribute-guided multilevel feature representation network for fine-grained ship classification in optical remote sensing images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1271–1285, 2020.

[21] Y. Yi, Y. You, C. Li, and W. Zhou, “EFM-Net: An essential feature mining network for target fine-grained classification in optical remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5606416.

[22] Y. Song et al., “An attention cut classification network for fine-grained ship classification in remote sensing images,” *Remote Sens. Lett.*, vol. 13, no. 4, pp. 418–427, 2022.

[23] J. Chen, K. Chen, H. Chen, W. Li, Z. Zou, and Z. Shi, “Contrastive learning for fine-grained ship classification in remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4707916.

[24] L. Shuxin and Z. Hua, “Ship target detection and fine-class recognition based on course-to-fine cascade neural networks,” in *Proc. 2nd Int. Conf. Robot., Intell. Control Artif. Intell.*, 2020, pp. 370–374.

[25] 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, 2022, Art. no. 5612318.

[26] Y. Liu, J. Liu, Z. Yu, and Z. Wu, “SFINet: An oriented fine-grained ship identification network based on remote sensing image,” in *Proc. Int. Conf. Image, Vis. Intell. Syst.*, 2023, pp. 206–215.

[27] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

[28] Z. Leng et al., “Polyloss: A polynomial expansion perspective of classification loss functions,” 2022, *arXiv:2204.12511*.

[29] V. H. Luu, V. K. Dinh, N. H. H. Luong, Q. H. Bui, and T. N. T. Nguyen, “Improving the bag-of-words model with spatial pyramid matching using data augmentation for fine-grained arbitrary-oriented ship classification,” *Remote Sens. Lett.*, vol. 10, no. 9, pp. 826–834, 2019.

[30] 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, 2021, Art. no. 747.

[31] Z. Zhang, T. Zhang, Z. Liu, and Y. Li, “Contrastive learning with part assignment for fine-grained ship image recognition,” in *Proc. Int. Conf. Pattern Recognit., Mach. Vis. Intell. Algorithms*, 2023, pp. 260–265.

[32] J. Ma, J. Yu, H. Yang, H. Jiang, and W. Li, “Fine-grained recognition of ships under complex sea conditions,” *Int. J. Adv. Netw., Monit. Controls*, vol. 7, no. 4, pp. 39–46.

[33] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, “LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513412.

[34] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, “Low-rank representations meets deep unfolding: A generalized and interpretable network for hyperspectral anomaly detection,” 2024, *arXiv:2402.15335*.

[35] W. Xiong, Z. Xiong, and Y. Cui, “An explainable attention network for fine-grained ship classification using remote-sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620314.

[36] W. Xiong, Z. Xiong, L. Yao, and Y. Cui, “Cog-Net: A cognitive network for fine-grained ship classification and retrieval in remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5608217.

[37] D. Hong et al., “SpectralGPT: Spectral remote sensing foundation model,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024, to be published, doi: 10.1109/TPAMI.2024.3362475.

[38] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[39] Y. Li, D. Hong, C. Li, J. Yao, and J. Chanussot, “HD-Net: High-resolution decoupled network for building footprint extraction via deeply supervised body and boundary decomposition,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 209, pp. 51–65, 2024.

[40] D. Hong et al., “Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks,” *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[41] X. Wu, D. Hong, and J. Chanussot, “UIU-Net: U-Net in U-Net for infrared small object detection,” *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2023.

[42] L. Qian, T. Yu, and J. Yang, “Multi-scale feature fusion of covariance pooling networks for fine-grained visual recognition,” *Sensors*, vol. 23, no. 8, 2023, Art. no. 3970.

[43] H. Liu, J. Li, D. Li, J. See, and W. Lin, “Learning scale-consistent attention part network for fine-grained image recognition,” *IEEE Trans. Multimedia*, vol. 24, pp. 2902–2913, 2022.

[44] P. Khosla et al., “Supervised contrastive learning,” in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 18661–18673.

[45] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2018, *arXiv:1807.03748*.

[46] O. Henaff, “Data-efficient image recognition with contrastive predictive coding,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4182–4192.

[47] Q. Wei, L. Feng, H. Sun, R. Wang, C. Guo, and Y. Yin, “Fine-grained classification with noisy labels,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11651–11660.

[48] G. Bukchin et al., “Fine-grained angular contrastive learning with coarse labels,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8726–8736.

[49] T. Zhang, D. Chang, Z. Ma, and J. Guo, “Progressive co-attention network for fine-grained visual classification,” in *Proc. Int. Conf. Vis. Commun. Image Process.*, 2021, pp. 1–5.

[50] F. A. Breiki, M. Ridzuan, and R. Grandhe, “Self-supervised learning for fine-grained image classification,” 2021, *arXiv:2107.13973*.

[51] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.

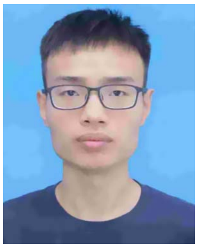
[52] J. Feng, E. Tang, M. Zeng, Z. Gu, P. Kou, and W. Zheng, “Improving visual question answering for remote sensing via alternate-guided attention and combined loss,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, 2023, Art. no. 103427.

[53] G.-S. Xia et al., “DOTA: A large-scale dataset for object detection in aerial images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[54] Z. Zhang, L. Zhang, Y. Wang, P. Feng, and R. He, “ShipRSImageNet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8458–8472, 2021.

[55] 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.

[56] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, “An augmented linear mixing model to address spectral variability for hyperspectral unmixing,” *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.



Shaokang Dong received the master’s degree in computer technology from the Qingdao University of Science and Technology, Qingdao, China, in 2022. He is currently working toward the Ph.D. degree in computer science and technology with the Chongqing University of Posts and Telecommunications, Chongqing, China.

His research interests include fine-grained ship classification, remote sensing, and oil detection.

Jiangfan Feng received the Ph.D. degree in cartography and geographical information system from Nanjing Normal University, Nanjing, China, in 2007.

He is currently a Professor with the Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include GIS, artificial intelligence, remote sensing, and VQA.

Dongxu Fang received the master’s degree in communication and information systems from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2007.

He is currently a Professor Engineering with China Mobile Group Chongqing Company, Ltd., Chongqing, China. His research interests include mobile communication networks, artificial intelligence, and data mining.