A New Benchmark and an Attribute-Guided Multilevel Feature Representation Network for Fine-Grained Ship Classification in Optical Remote Sensing Images

Xiaohan Zhang, Yafei Lv, Libo Yao, Wei Xiong, and Chunlong Fu

***Abstract*—Maritime activities are essential aspects of human society. Accurate classification of ships is vital for maritime surveillance and meaningful to numerous civil and military applications. However, most studies conducted are limited to the coarse-grained ship classification. Few studies on fine-grained ship classification have been undertaken despite its accuracy and practicability. In this study, we construct a new benchmark for fine-grained ship classification which consists of 23 fine-grained categories of ships. Besides the category label, the benchmark contains several other attribute information. To solve the problem of interclass similarity, an attribute-guided multilevel enhanced feature representation network (AMEFRN) is proposed. Concretely, a multilevel enhanced visual feature representation is designed to fuse the reweighted regional features in order to focus more on the silent region and suppress the other regions. Further to this, considering the complementary role of attribute information in ship identification, an attribute-guided feature extraction branch is proposed, which extracts the auxiliary attribute features by utilizing the attribute information as supervision. Finally, the attribute features and the enhanced visual features jointly function as a feature representation for classification. Compared to other existing classification models, AMEFRN has better performance with an overall accuracy rate of 93.58% on the established fine-grained ship classification dataset. Moreover, it can be easily embedded into most CNN models as well as trained end-to-end.**

***Index Terms*—Attribute information, fine-grained classification, multilevel features, optical remote sensing image, ship classification.**

I. Introduction

MARITIME activities, for instance, maritime transportation, commercial trades, maritime security, and anti-illegal activities, are important to the human society, as they impact economic and social development. Daily, numerous ships of different types cruise the sea and the classification of ships through optical remote sensing images constitutes one of the basic technologies for marine surveillance [1], [2]. Hence, this technology has numerous civil and military applications [3].

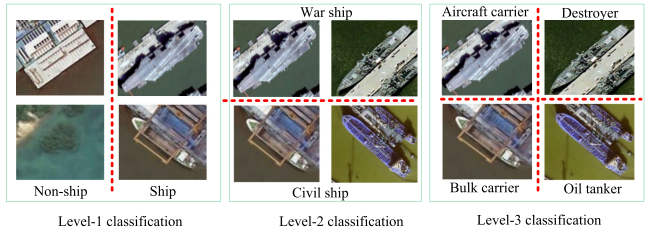


Fig. 1. Three-level ship classification task.

Based on practical demands, there are three levels of ship classification [4]. In level-1 classification, the meta-classification, ships and nonship objects are separated. In level-2 classification, the coarse-grained classification, ships are classified according to some criteria into coarse categories, i.e., warship or civilian ship. Level-3 classification refers to the fine-grained classification where ships are distinguished into their precise categories. The differences in the three-level classification are shown in Fig. 1. The complexity in classification increases from one level to the next. A significant number of studies conducted mainly address the first two levels of classification, with only a few methods and datasets proposed for the fine-grained ship classification. More precise and detailed classification in this level can be more practical and valuable compared to the other two levels of classifications in many applications [5].

There is growing attention on ship classification in the remote sensing field with numerous methods proposed to solve this task. The existing techniques are primarily categorized into handcrafted feature-based and deep learning-based approaches. The handcrafted feature-based methods are further divided into global feature-based methods and local feature-based methods. Previously, low-level global features such as geometric features, i.e., scale, aspect ratio, and shape, aided in ship classification [6]. However, these features were only used in simple cases. Explorations of some local features such as the location of the mast were found to be more discriminative in ship classification

Manuscript received January 8, 2020; revised February 26, 2020 and March 9, 2020; accepted March 13, 2020. Date of publication March 23, 2020; date of current version April 16, 2020. This work was supported by the National Natural Science Foundation of China under Grant 91538201, 61790554, 61531020, and 61971432. *(Corresponding author: Libo Yao.)*

Xiaohan Zhang, Yafei Lv, Libo Yao, and Wei Xiong are with the Research Institute of information Fusion, Naval Aviation University, Yantai 264001, China (e-mail: xhan_zhang@163.com; yfei_lv@163.com; ylb_rs@126.com; xiongwei@csif.org.cn).

Chunlong Fu is with the Troops 90139 of PLA, Beijing 100001, China (e-mail: 13718013639@163.com).

Digital Object Identifier 10.1109/JSTARS.2020.2981686

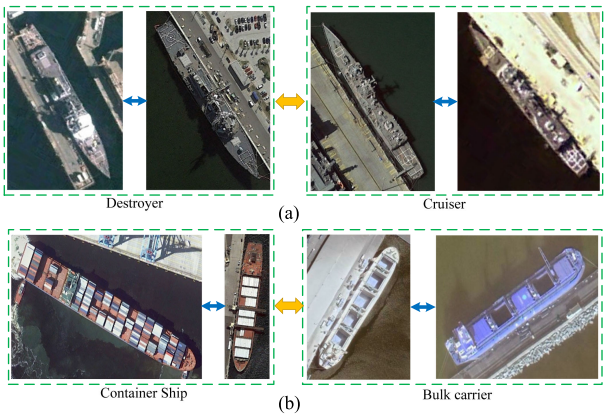


Fig. 2. Two groups of ship samples. (a) Destroyer versus cruiser. (b) Container ship versus cargo.

[7], [8]. Further, other extracted local features, including scale-invariant feature transform (SIFT) [9], local binary patterns (LBP) [10], and hierarchical multiscale LBP (HMLBP) [11], effectively improve the classification performances at varying degrees. Both the global and local features have been combined in multiple feature frameworks, such as the Gabor-based multi-scale completed LBP (MS-CLBP) [12], the joint feature-based model [13], and MSHOG feature-based task-driven dictionary learning [14]. Recently, deep learning-based methods provided impressive results in multiple tasks of computer vision and object classification. In the task of ship classification in remote sensing images, Shi *et al.* [15] used two-branch CNN to extract features based on two-dimensional (2-D) discrete fractional Fourier transform (2D-DFrFT). Multiple traditional features obtained by Gabor filter, LBP (CLBP), 2D-DFrFT, and deep CNN were adopted and applied to extract high-level abstract features automatically [16]. This method achieved state-of-the-art performance on a four-category ship classification dataset [9]. Compared with the handcrafted feature-based methods, deep-learning-based features provide more discriminative high-level visual features. These features more effectively bridge the semantic gap existing between the hand-crafted feature representation and the content of remote sensing images.

There still exist some limitations in the use of deep learning-based features despite their strong discriminative ability in natural image classification and high potential in fine-grained ship classification.

1) Lack of annotated data. Few datasets are constructed with the fine-grained ship category labels due to challenges in remote sensing images' collection and data annotation. The existing datasets [9], [17] are mainly constructed for level-1 and level-2 classification tasks.
2) Interclass similarity. As is shown in Fig. 2, ships of different types may have similar appearance features and ships of the same type may appear differently under different conditions which causes difficulties in discriminating both similar and different ships.
3) The particularity of ship classification. Discriminative visual features are usually used decisively in the classification of common objects, especially from the dominant performances acquired by deep CNNs.

Regarding the ship identification, which belongs to rigid body target, some attributes can be simple but effective features in the differentiation of similar subcategories. For instance, a destroyer can be easily classified with a fishing ship and an oil tanker following its larger aspect ratio without considering the visual features' differences. Therefore, attributes which usually describe some unique features of ships can provide adequate auxiliary information for ship classification.

In this article, the challenging level-3 ship classification task is explored. First, a new benchmark consisting of high-resolution optical remote sensing images for fine-grained ship classification is established. After that, an attribute-guided multilevel enhanced feature representation network (AMEFRN) is proposed for fine-grained ship classification. Specifically, a CNN with a multilevel enhanced local feature representation module is adopted to obtain the discriminative visual features, while a novel attribute-guided branch is designed to obtain auxiliary attribute features. Finally, the enhanced visual features and attribute features are fused up as the feature representation for the fine-grained classification. Comprehensive experiments on the new benchmark validate its usefulness and the effectiveness of AMEFRN. In summary, this study's contributions entail the following.

1) The challenging task of fine-grained ship classification in remote sensing image is explored. A fine-grained ship classification dataset containing 23 categories with high-resolution optical remote sensing images is established. Besides the labels of ship category, the attributes of ship aspect ratio and angles are also annotated in this dataset.
2) A novel AMEFRN is proposed for the fine-grained ship classification. Based on the generic CNN model, multilevel enhanced visual features are extracted, where an RNN-based attention mechanism is used to reweigh the importance of features in different regions. Moreover, a new branch for attribute feature learning supervised by the attribute information is designed to enhance normal category supervised learning. These two schemes can be easily embedded into most CNN frameworks and be trained end-to-end.
3) Numerous experiments are conducted on the new benchmark and the effectiveness of the proposed method is fully verified. Compared with other methods, this framework shows state-of-the-art performance on the constructed dataset. Besides, some baseline models are evaluated on the benchmark.

The rest of this article is organized as follows. Section II gives a detailed introduction to the FGSC-23. Section III contains the proposed classification framework. Section IV gives details of the experiments and result analysis while Section V points out the conclusions.

II. New Benchmark for Fine-Grained Ship Classification

A. Ship Classification Datasets in Previous Works

Several object detection and classification datasets [4], [9], [18], [19] in remote sensing have been established in previous works. The information of some datasets related to ship

TABLE I
COMPARING OUR BENCHMARK WITH OTHER SHIP CLASSIFICATION DATASETS

Dataset	Data source	Image type	Image amount	Categories	Classification task level	Category balance	Ship orientation
MASATI	Microsoft Bing maps	Color image	6212	2/7	Level-1	Unbalance	Arbitrary-oriented
BT1000	Electro-Optical (EO) satellite	Greyscale image	2000	2	Level-2	Balanced	Arbitrary-oriented
CCT250	Electro-Optical (EO) satellite	Greyscale image	750	3	Level-2	Balanced	Arbitrary-oriented
BCCT200	Electro-Optical (EO) satellite	Greyscale image	800	4	Level-2	Balanced	Arbitrary-oriented
BCCT200-resize	Electro-Optical (EO) satellite	Greyscale image	800	4	Level-2	Balanced	Fixed-oriented
HRSC2016	Google Earth	Color image	1161/2976	4/25	Level-1,2,3	Unbalance	Arbitrary-oriented
FGSC-23(ours)	Google Earth, GF-1	Color image	4080	23	Level-3	Unbalance	Arbitrary-oriented

classification using optical remote sensing images is listed in Table I. Concretely, the Maritime Satellite Imagery (MASATI) created by Gallego *et al.* [19] had two main classes, i.e., ship and nonship, and the subclasses of ship category included ship, detail, multi, and coast and ship. Thus, it only met with level-1 classification task. BT1000 [9] divided 2000 ships into bulk carriers and tankers; CCT250 [9] divided 750 ships into cargo ships, container ships, and tankers while BCCT200 [9] added a barge category based on CCT250. All the three datasets had an equal number of images per category and met the level-2 classification task. The ships in these datasets were arbitrary oriented and the size of the images was different. In the BCCT200-resize dataset [15], preprocessing was accomplished based on BCCT200 by fixing the orientations of the ships and resizing the images to 300×150. Although these datasets were used in previous works [15], [16], [20], the ship categories were not abundant enough for the fine-grained ship classification task. HRSC2016 [4] could be regarded as the first public fine-grained high-resolution ship detection dataset, which contained 1161 images, and 2976 ships were labeled with locations and fine-grained or coarse-grained categories. Regarding the level-3 classification task, 2285 samples were labeled with 25 fine-grained categories. However, the number of samples in 13 categories were less than 100, while seven categories were with less than ten ships. Besides, as a ship detection dataset, the fine-grained classification of ships could only be done based on the correct detection of ships. Although the dataset was an inspiring work for the fine-grained ship classification, it was not adequate enough. Thus, in this study, a collection of high-resolution optical remote sensing ship images was made and it could be used for fine-grained ship classification.

B. Collection and Annotations of FGSC-23

FGSC-23 has a total of 22 categories of ships and 4080 chips. Some negative samples which look like ships are labeled as "nonship" category. Therefore, 23 categories constitute this dataset. All the ship chips are obtained from Google Earth public images and GF-1 satellite. The sizes of images are not fixed, ranging from 40 to 800 pixels. Approximately 1600 chips are taken from the HRSC2016 dataset. All the ship categories are labeled by human interpretation. Typical samples of each ship category are shown in Fig. 3, and the number of each category is listed in Table II.

Except for the category labels, the attributes of the ship aspect ratio and the angle between the ship's central axis and the image's horizontal axis are also annotated. The illustrations of the two attribute labels are shown in Fig. 4.

C. Properties of FGSC-23

The FGSC-23 has the following properties.

1) Category diversity: FGSC-23 divides ships into fine-grained categories. For example, for a coarse category of cargo ship, it is divided into container ship, bulk carrier, car carrier, oil tanker, and liquefied gas ship fine-grained categories.
2) Image diversity: As is shown in Fig. 3, ships under different illumination conditions, with onshore or offshore backgrounds and with arbitrary-oriented distributions are involved in this dataset. Moreover, the resolutions of images are not fixed, ranging from 0.4 to 2 m. On the one hand, the diversity of data places higher requirements on classification algorithms. Still, on the other hand, it is beneficial to train a model with stronger learning capability and better generalization.
3) Label diversity: Each sample of the dataset is annotated with three labels—ship category, ship aspect ratio, and the distribution direction. Therefore, this dataset can also be used for other tasks such as ratio and direction estimation.
4) Category imbalance: There exists an issue of category imbalance in FGSC-23, as is shown in Table II: In the real world, specific ships, such as the medical ship, are much less than others. Therefore, category imbalance is unavoidable to some degree. This necessitates further improvements to provide a solution to this issue.

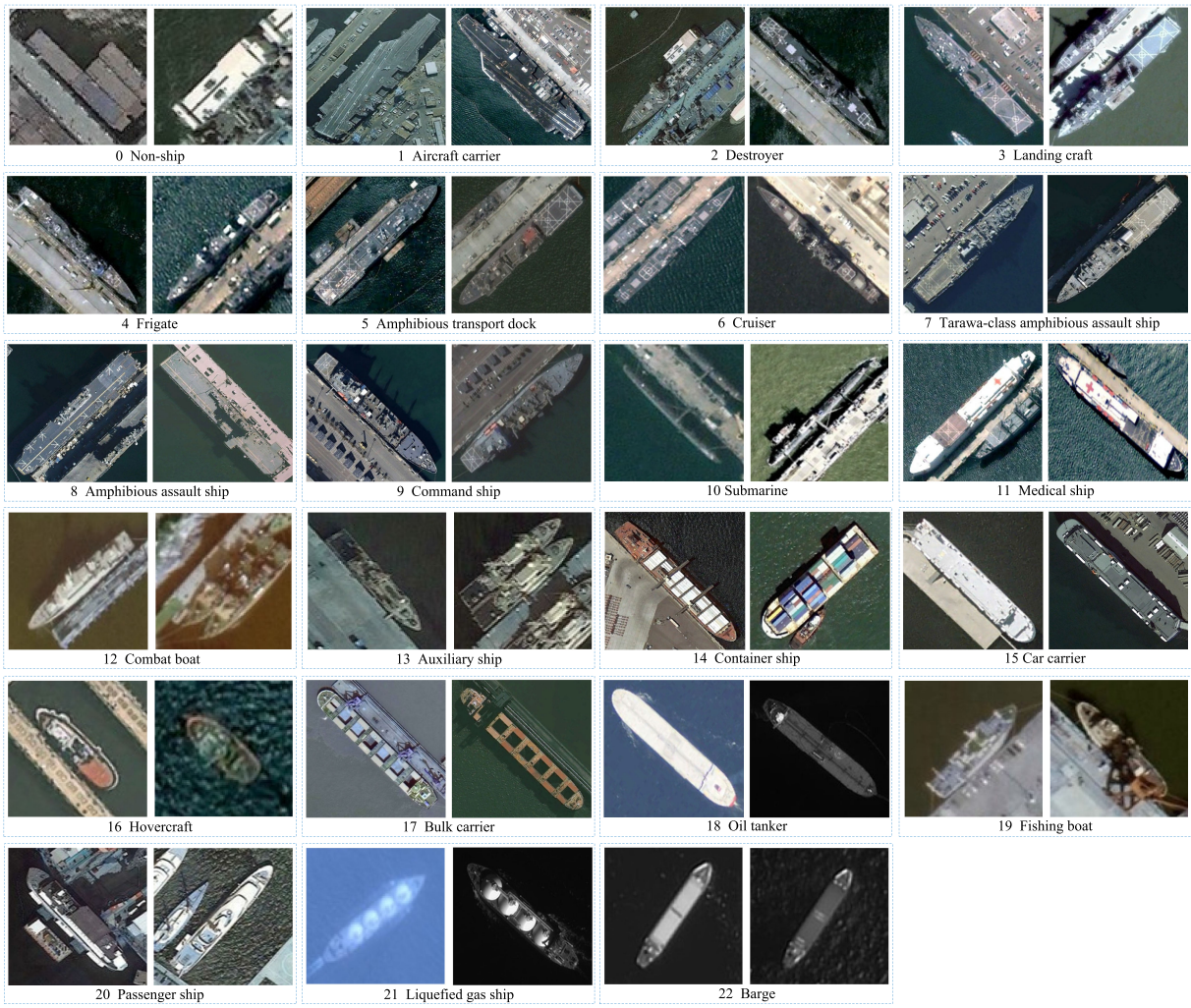


Fig. 3. Fine-grained categories and ship slice samples of each category in FGSC-23.

TABLE II
NUMBER OF EACH CATEGORY IN FGSC-23

Category ID	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Amount	484	165	542	108	295	90	293	88	154	89	238	27
Category ID	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**Total**
Amount	143	225	101	72	120	343	165	102	88	94	54	4080

The correspondence between category ID and category can be found in Fig. 3.

5) Public availability: The dataset is publicly available for free for scientific research, and the link to the current version is[1] (extraction code: n8ra).

III. PROPOSED METHOD

A. Architecture of the Proposed Method

Based on the interclass similarity issue, a discriminative feature representation is highly required for a fine-grained ship classification task [21], [22]. In this study, an AMEFRN is designed. Two schemes, i.e., multilevel visual feature representation and attribute feature representation are proposed to optimize the feature representation of general CNN. The overall architecture of the proposed method is shown in Fig. 5 and the classical network of VGG16 [23] is taken as an example for illustration. Multilevel convolutional visual features are extracted from VGG16 and the local features are weighed by an RNN-based attention mechanism to get an enhanced visual feature representation. To obtain the attribute features, a novel attribute-guided branch trained by additional attribute supervision information is designed. Consequently, the visual features and attribute features are associated together as classification features and fed into the classifier. Details and analysis of the proposed schemes are introduced in the following sections.

[1]https://pan.baidu.com/s/1h_F7c-btLqhOxLT20XHWBg

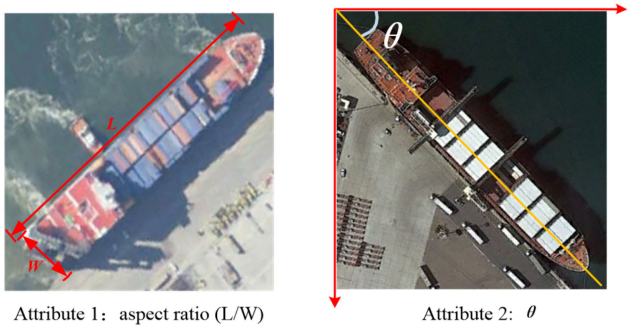


Fig. 4. Illustration of the two attribute labels in FGSC-23. (a) Aspect ratio r ($r = \lfloor L/W \rfloor$). (b) Angle $\theta \in [0, 180°]$.

B. Multilevel Enhanced Feature Representation

Local features play an important role in distinguishing the fine-grained categories of ships. Some specific local features might become a "symbol" of a particular category of the ships. For instance, as is shown in Fig. 6, a symbol of a red cross on a ship signifies a medical ship, while a special shape of the bow, aircrafts on the deck, and a flight runway illustrate an aircraft carrier. The difference in cargos is an essential element to distinguish a container ship from a bulk carrier. Compared with features extracted from other regions, these specific local features are crucial for better classification performance and require more attention. However, since the focus areas in different images are varied, the classification network should be guided to weigh the features of multilevel areas automatically.

Generally, in CNN frameworks, the high-level convolutional features from the last convolutional block are usually translated to a feature vector by flattening or pooling operation. The feature vector merges all areas of the image; hence, it can be considered as a global feature vector with senior semantic information. However, the global feature vector fails to distinguish the features of different areas in the image. Thus, in this framework, a multilevel feature representation is proposed to make up for the deficiencies. Classical VGG16 is taken as an example for illustration. The input of the network is resized to $224 \times 224 \times 3$, and after a series of pooling operations, the last convolutional feature map drops to $7 \times 7 \times 512$. It is treated as a collection of local feature vectors: $F_{L1} = \{v_1^{512}, v_2^{512}, \ldots, v_{7\times7}^{512}\}$, where v_i^{512} denotes the features of the corresponding area in the original image. Then, an average pooling operation with a kernel size 2×2 is conducted and the second-level feature collection of enlarged local areas is obtained by $F_{L2} = \{w_1^{512}, w_2^{512}, \ldots, w_{3\times3}^{512}\}$. Further to this, a global average pooling operation with kernel size 7×7 is done to translate features into the third-level global feature vector $F_{L3} = f_G^{512}$. The illustration of the three-level visual feature representation and the visualization effects is shown in Fig. 7. The number of feature levels is not fixed and can be adjusted according to actual requirements.

The feature vectors in F_{L1} and F_{L2} which denote features of different local areas are accorded with different weights. When we scan an image, our eyes move in turns, and we will pay more attention to some salient parts which is acquired through contextual comparison. Based on this fact, these feature vectors are regarded as a set of ordered sequences, and spatially adjacent vectors usually share more semantic associations, which inspires us to use the RNN network in learning their importance weights. One of our previous works, an RNN-based attention module [24], is adopted here to weigh these feature vectors. For the ith level feature map $F_{Li} = \{f_1^{512}, f_2^{512}, \ldots, f_k^{512}\}$, the k regional feature vectors $\{f_j^{512}\}(j = 1, 2, \ldots, k)$are entered into the gated recurrent units (GRU) [25] in turns, which outputs sequences $H = \{h_1, h_2, \ldots, h_k\}$. H is treated as a learned revised regional feature representation considering context information. After that, two fully connected (FC) layers are connected and the attention weight map of the regional features $A_i = \{a_1, a_2, \ldots, a_k\}$ is acquired. F_{Li} is revised by weight map A: $F'_{Li} = \{f'_j\} = \{a_1 \times f_1^{512}, a_2 \times f_2^{512}, \ldots, a_k \times f_k^{512}\}$. The weighted feature vectors in F'_{L1} and F'_{L2} are summed up, respectively, to obtain the local feature vector f_{L1} and f_{L2} as follows:

$$f_{Li}^{512} = \sum_{j=1}^{k} f'_j = \sum_{j=1}^{k} a_j \times f_j^{512}. \tag{1}$$

Together with the global feature vector f_G^{512}, the three feature vectors $\{f_G^{512}, f_{L1}^{512}, f_{L2}^{512}\}$ form a three-level visual feature representation.

C. Attribute-Guided Feature Extraction Branch

As discussed in Section I, some inherent attributes of ships, i.e., scale and aspect ratio, are effective auxiliary information helpful for fine-grained ship classification. Inspired by the self-supervised learning [26], [27] where the feature representation learning is achieved by predicting the image rotations, the attribute information of ships might be used as the supervision information as well, which can enhance the feature representation learning. The learned feature supervised by the attribute information is regarded as an attribute feature and is the auxiliary feature to visual features extracted in Section III-B. The attribute of the ship's aspect ratio is not affected by the image resolution; besides, it is easy to acquire as the ship detection technology with oriented bounding boxes is quite mature [28], [29]. Thus, it is adopted as the supervision information to construct an attribute-guided feature extraction branch for the attribute feature representation.

The attribute-guided feature extraction branch is built based on the general CNN structure and its architecture is shown in Fig. 5. The input of the original network is fed into this branch, and after five blocks of convolution-pooling layers, features of this branch are fed into two FC layers. A *ReLU* function is then used in predicting the aspect ratio attribute. The backbone of this branch is similar to the structure of VGG16 and its detailed composition is shown in Fig. 8.

To guide this branch learning the attribute of aspect ratio automatically, a supervised $L1$ loss is adopted which is shown as follows:

$$L_{\text{attr}} = |\hat{p}_a - g_a| \tag{2}$$

where $\hat{p}_a$ is the attribute prediction of the network and g_a is the ground truth.

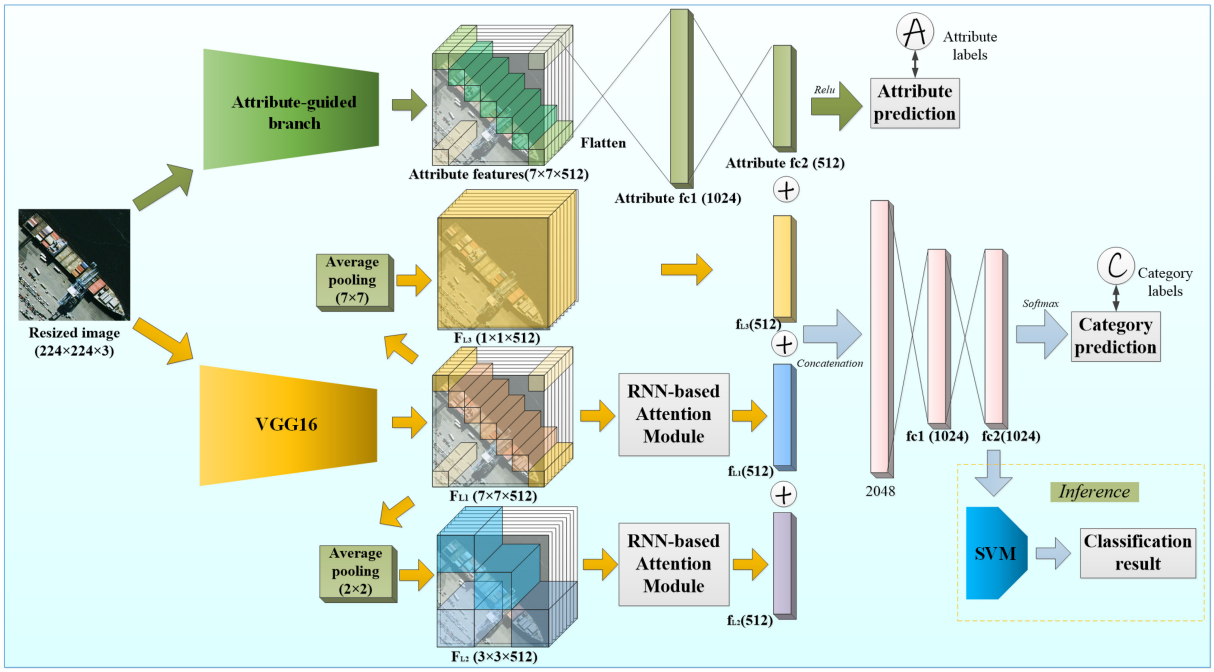


Fig. 5. Framework of the proposed network.

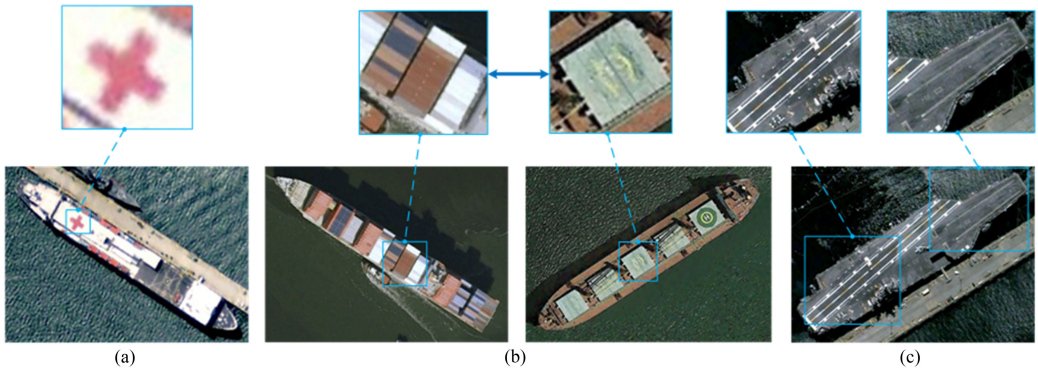


Fig. 6. Samples of salient regions for fine-grained ship classification. (a) Important symbol of medical ship. (b) Local features to distinguish container ship and bulk carrier. (c) Special local information of aircraft carrier.

The output of the last FC layer is a 512-dimesional feature vector f_{attr}^{512}, which is directly used for attribute prediction. Naturally, it holds much geometric attribute information, which is beneficial for category classification. Therefore, it is fused with the multilevel visual features extracted for subsequent category classification. Hence, the whole network obtains attribute supervision information and category supervision information, and the attribute features are extracted as auxiliary features to multilevel version features. This enhances the learning capability of the network for a fine-grained classification task.

D. Training and Inference

Using the framework mentioned above, attribute feature f_{attr}^{512} and multilevel visual features $\{f_G^{512}, f_{L1}^{512}, f_{L2}^{512}\}$ are acquired. They are concatenated together and then, two FC layers follow to reduce the dimension and further modify the feature. The classification feature vector f_{class}^{1024} achieved in this way is then entered to *Softmax* classifier for output prediction. Category labels are used for supervised learning and cross-entropy loss described in formula (2) is adopted for training

$$L_{\text{cat}} = -\sum_{c=1}^{M} y_c \log(\hat{p}_c) \tag{3}$$

where M is the categorynumber, $\hat{p}_c$ is the category prediction, and y_c is the ground truth. The overall training loss of the whole network constitutes the summary of L_{attr} and L_{cat}

$$L = L_{\text{cat}} + L_{\text{attr}}. \tag{4}$$

During the end-to-end training, *Softmax* function is adopted for classification. In inference, following [24], classifier of linear

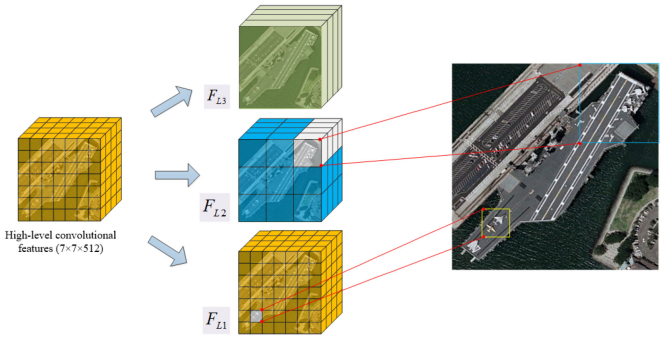


Fig. 7. Illustration of multilevel feature representation. The case of three-level feature representation is visualized.

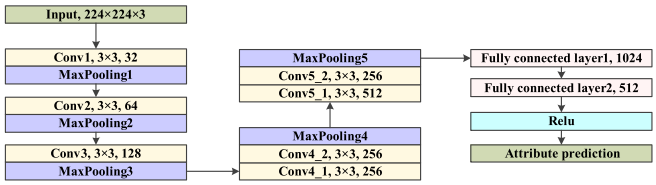


Fig. 8. Architecture of attribute-guided branch.

SVM is adopted to replace the *Softmax* function due to its powerful classification ability. The SVM is trained with classification vector f_{class}^{1024} as input and category of the target as output.

IV. Experiments and Result Analysis

In this section, extensive experiments are conducted on the FGSC-23 dataset to validate the effectiveness of the proposed method.

A. Experimental Setup

In this section, some details about the experiment, including the dataset setting, image preprocess method, evaluation metrics, and experiment environment, are presented as follows.

1) Dataset Setting: The FGSC-23 is separated into training set and testing set. From each category, 20% images are randomly selected for testing and the rest for training. Due to the issue of sample imbalance, data augmentation is done to enlarge the number of images of some categories in the training set, and the methods include changing the image lightness in the range of [0.5, 1.5], image scaling in the range of [0.8, 1.2], image flip, and random cropping. For the categories with less than 200 ships in the training set, the images are selected randomly and the augmentation means mentioned above are conducted to enlarge the training images to 200. Therefore, there are 825 chips in testing set, and the 3255-sample training set is enlarged to 5165 samples.

2) Image Preprocess: The CNN models with FC layers for the classification tasks require the inputs with a fixed size. Therefore, images of different shapes need resizing. However, the general operation [30] of resizing an image to a fixed size by interpolation or downsampling changes the aspect ratios of objects in the nonsquare images, which has a negative influence in the training of the attribution-aware branch. To solve this issue, an image resizing operation of zero padding is proposed to maintain the aspect ratios of objects. The inputs require resizing to $224 \times 224 \times 3$. The longer side of the image is first upsampled or downsampled to 224. Upsampling or downsampling on the shorter side of the image with the same ratio follows. The rest of the image area is then padded with zeros. This maintains the attribute of the ship's aspect ratio. The resizing results of the general operation and zero-padding-based operation are compared in Fig. 9. Besides, experiments are conducted to examine the influences of the two image preprocessing means on detection.

3) Evaluation Metrics: Three indicators, i.e., accuracy rate (AR) of each category, overall accuracy (OA), and the confusion matrix (CM) [24], are adopted to evaluate the classification results in the experiments. The AR measures the ratio of correctly classified images and total testing images among a category, while the OA measures the ratio of correctly classified images and total testing images regardless of categories. The CM is the visualization of the classification matrix, which records the detailed classification results for every category. Each element a_{mn} in CM denotes the proportion of the chips predicted to be the nth category while it actually belongs to the mth category.

4) Experiment Environment: The experiments are conducted using Keras framework on a 64-b computer under Ubuntu 16.06 with one NVIDIA GTX 1080Ti GPU for acceleration. During training, a batch size of 32 is set and an initial learning rate of 0.0001. All the models are trained for 100 epochs.

B. Establishment of Baselines

The proposed schemes, i.e., multilevel enhanced feature representation and the attribute-guided branch in AMEFRN, can be conveniently embedded into most CNN models. Thus, two representative models, VGG16 and ResNet50 [31], are adopted as baseline feature extractors to test the effects of the proposed method in the experiments. Due to the relatively small scale of our dataset, pretrained parameters on ImageNet [32] are used to initialize the baseline models. Here, comparisons are made between classifiers of *Softmax* and the linear SVM, as well as image preprocessing methods of general resizing operation and the proposed zero-padding method. VGG16 and ResNet50 are trained with resized data and zero-padded data, respectively. *Softmax* function is used in the end-to-end training of both VGG16 and ResNet50. But during testing, *Softmax* function and linear SVM are adopted in turn on the trained VGG16 and ResNet50 to compare their classification performances. The classification feature vector f_{class}^{1024} extracted by trained VGG16 and ResNet50 is used as an input to SVM, and all the images in the training set are used to train the SVM. The OAs of these eight models are listed in Table III, and the visualizations of respective CMs are shown in Fig. 10.

The above-mentioned results reveal that under the same conditions, SVM provides a better classification performance on both VGG16 and ResNet50. The effects of the zero-padding method are validated as it behaves better than the general image resizing method. Thus, in the subsequent experiments, the zero-padding method is used for image preprocessing. VGG16 with

Fig. 9. Comparisons of our resizing method and general resizing method.

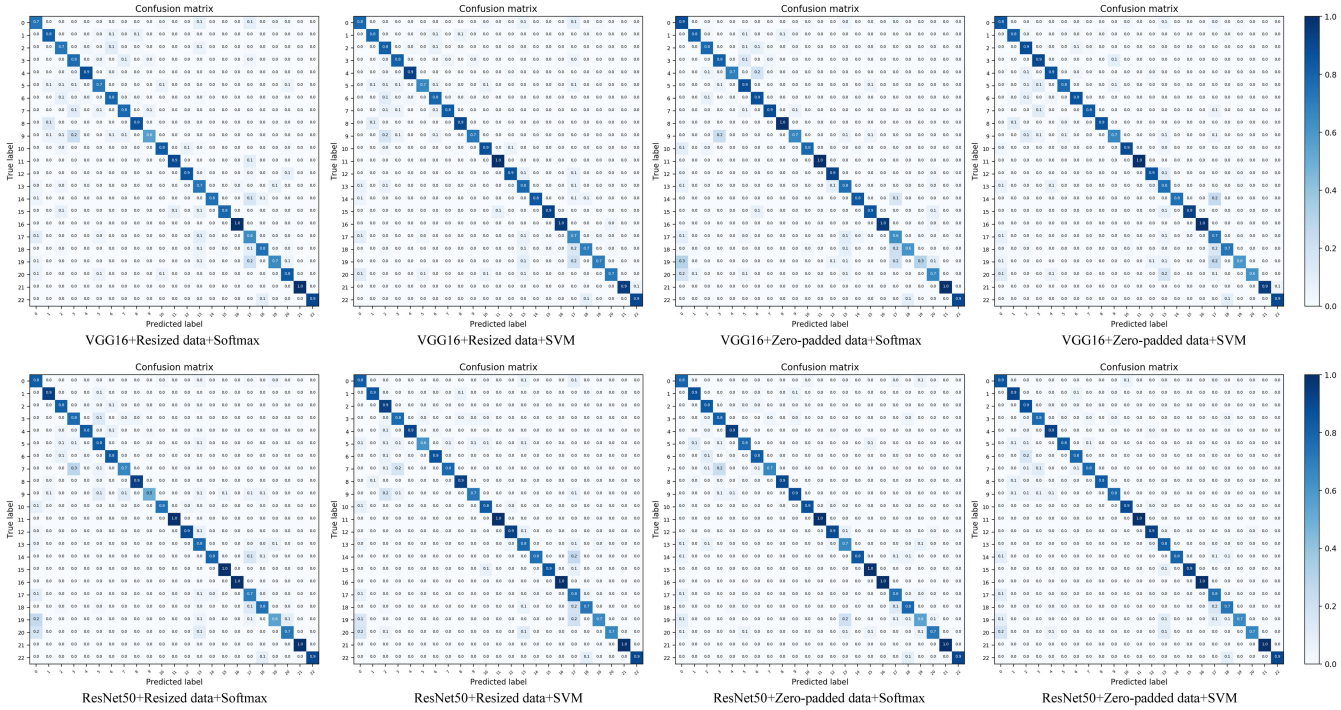


Fig. 10. CM of each model in baseline establishment.

SVM and ResNet50 with SVM are adopted as baselines in our experiments, namely, baseline1 and baseline2, respectively.

C. Ablation Study

In this section, ablation studies are conducted to verify the effects of the proposed AMEFRN. The experiments are divided into three parts. First, the effect of the attribute-guided branch used alone on the two baselines is validated. Then, different local feature levels are tested to find the best setting of the proposed multilevel enhanced feature representation on the baselines. Finally, the two schemes are applied jointly to test the compound influence to the two baselines.

1) Effects of the Attribute-Guided Branch: The proposed attribute-guided branch (named scheme1) is applied to baseline1 and baseline2, respectively. The proposed zero-padding method is used to modify the input images. Each category's classification result, as well as the OA of the testing set, is shown in Table IV.

From the results above, we can see that the attribute-guided branch added to baseline models definitely improves the OA of both VGG16 and ResNet50, improving by 1.93% and 0.86%, respectively. Here, ResNet50-based models show better performance than VGG16-based models, implying that ResNet50 has a better feature representation capability compared to VGG16. However, the addition of our attribute-guided branch narrows the gap between the two baseline models to some degree.

2) Effects of Multilevel Enhanced Feature Representation: In this section, the effects of multilevel enhanced feature representation (named scheme2) are verified on both baseline1 and baseline2. Specially, different feature representation levels for ship classification are tested. Four conditions are considered, i.e., using 1-level features, 2-level features, 3-level features, and 4-level features, respectively, in the classification. Concretely, the 1-level feature representation has only global feature vector (f_G^{512} described in Section III-B); the 2-level feature representation contains the global feature vector and enhanced

TABLE III
OA (%) OF EACH MODEL IN BASELINE ESTABLISHMENT

Models	Resized data	Zero-padded data	Softmax	SVM	OA
VGG 16	√	-	√	-	77.82
	√	-	-	√	81.09
	-	√	√	-	79.64
	-	√	-	√	**81.95**
ResNet 50	√	-	√	-	81.21
	√	-	-	√	84.00
	-	√	√	-	83.03
	-	√	-	√	**84.60**

The portions in bold represent the best performance.

TABLE IV
AP (%) AND OA (%) OF THE BASELINES WITH SCHEME1

Result	Baseline1	Baseline1 +Scheme1	Baseline2	Baseline2 +Scheme1
AR_0	84.54	84.54	85.57	**88.66**
AR_1	90.74	85.29	**91.18**	85.29
AR_2	83.33	**93.52**	87.96	**93.52**
AR_3	**90.91**	86.36	77.27	86.36
AR_4	91.53	84.75	**93.22**	91.53
AR_5	66.67	72.22	**83.33**	77.78
AR_6	81.36	**83.05**	79.66	**83.05**
AR_7	**88.89**	**88.89**	77.78	**88.89**
AR_8	87.10	87.10	83.87	**96.77**
AR_9	**72.22**	66.67	**77.78**	61.11
AR_{10}	87.50	87.50	**89.58**	**89.58**
AR_{11}	90.00	**100.00**	**100.00**	**100.00**
AR_{12}	82.76	**93.10**	**93.10**	86.21
AR_{13}	77.78	77.78	**82.22**	80.00
AR_{14}	80.00	**90.00**	80.00	**90.00**
AR_{15}	85.71	78.57	**92.86**	**92.86**
AR_{16}	**100**	95.83	**100**	**100**
AR_{17}	71.01	72.46	**75.36**	72.46
AR_{18}	60.61	**78.79**	72.73	**78.79**
AR_{19}	60.00	55.00	**65.00**	55.00
AR_{20}	55.56	**66.67**	**66.67**	61.11
AR_{21}	90.00	95.00	**100**	95.00
AR_{22}	**90.91**	**90.91**	**90.91**	**90.91**
OA	81.95	83.88	84.60	**85.46**

The portions in bold represent the best performance.

local 3 × 3-subregion feature vector ($f_G^{512}$and f_{L2}^{512} described in Section III-B); the 3-level feature representation contains global feature vector, enhanced local 3 × 3-sub-region feature vector and local 7 × 7-subregion feature vector (f_G^{512},f_{L1}^{512}, and f_{L2}^{512} described in Section III-B); and for 4-level feature representation, except for these feature vectors, an additional 14 × 14-subregion feature vector is contained, which is achieved from the feature map of the second-last convolutional block. The four feature representations are embedded into baseline1 and baseline2, respectively, and the classification results are shown in Table V. The visualizations of CMs for the models are shown in Fig. 11, which provide a more intuitive reflection of the classification results.

Here, the 1-level feature representation is achieved by global pooling operation to the last convolutional feature layer of the CNN model, which is similar to that of general CNNs. The classification results of 1-level feature representation are also similar to their baselines. Further, the multilevel enhanced feature representations share significant improvements to classification performance than the baselines. Among them, 3-level feature representation achieves the best performance on both baseline1 and baseline2. The 4-level features do not perform optimally like the 3-level features, which is due to lower-level feature maps being used in 4-level feature representation to achieve higher resolution regional features, which is not semantic enough for the classification task. Therefore, 3-level feature representation is the best setting for FGSC-23. In subsequent experiments, this setting is adopted.

Besides, we find out that the proposed scheme2 significantly improves the performance of baseline1 and narrows the OA gaps between baseline1 and baseline2. The best performance in this group of experiments is achieved by baseline1 with 3-level feature representation. More classification details are reflected in the CM of each model.

3) Compound Effects of Proposed Schemes: In this section, the proposed scheme1 and scheme2 are used jointly to test their compound influence on the classification performance of the two baselines. A 3-level feature representation is used for scheme2. For a more intuitive comparison, in Table VI, we present the classification performances of the baselines with the schemes used alone. Besides, the CMs for each model in this group are visualized and shown in Fig. 12.

From the results, it can be seen that scheme1 improves the baseline1 by 1.93%, scheme2 improves it by 9.2%, and scheme1 and scheme2 jointly improve it by 11.63%. For baseline2, these schemes improve it by 0.86%, 5.7%, and 8.49%, respectively. The improvements caused by a combination of the two schemes are even larger than the sum of the improvements achieved by individual schemes, indicating that the two schemes are more powerful when used jointly in classification models. When it comes to the detailed performances in each category, the schemes also improve the classification accuracy in most categories. Notably, the combination of the two schemes yields the best ARs of the two baselines in 19 categories out of 23 categories. Evidently, our proposed method effectively improves the accuracy of CNN models in fine-grained ship classification task.

Besides, we find that our schemes seem to have a larger influence on VGG16 than ResNet50, especially scheme2.

The VGG16 has fewer layers and simpler structure compared with ResNet50, and ResNet50 is proved to have stronger feature representation capability in Section IV-B. The output of the last convolutional layer of VGG16 is a feature map of 7 × 7× 512, while the output of the last convolutional layer of ResNet50 is a 7 × 7× 2048 feature map. Thus, the 3-level feature representation of scheme2 forms a 6144-dimensional concatenated visual feature vector in ResNet50, much larger than the 1356-dimensional concatenated visual feature vector in

TABLE V
AP (%) AND OA (%) OF MODELS WITH MULTILEVEL FEATURE REPRESENTATION

Result	Baseline1 +1-level features	Baseline1 +2-level features	Baseline1 +3-level features	Baseline1 +4-level features	Baseline2 +1-level features	Baseline2 +2-level features	Baseline2 +3-level features	Baseline2 +4-level features
AR_0	77.32	87.63	90.72	91.75	86.60	91.75	**92.78**	**92.78**
AR_1	91.18	**94.12**	88.24	94.12	88.24	91.18	**94.12**	**94.12**
AR_2	89.81	95.37	96.30	94.44	87.96	96.30	**99.07**	95.37
AR_3	**90.91**	81.82	86.36	81.82	77.27	77.27	86.36	77.27
AR_4	86.44	94.912	**96.61**	93.22	89.83	**96.61**	**96.61**	94.92
AR_5	**88.89**	83.33	83.33	72.22	66.67	83.33	72.22	77.78
AR_6	86.44	**89.83**	88.13	84.75	81.36	88.13	88.13	**89.83**
AR_7	83.33	83.33	83.33	**88.88**	77.78	**88.89**	83.33	83.33
AR_8	90.32	90.32	93.55	93.55	93.55	90.32	**96.77**	**96.77**
AR_9	72.22	72.22	**83.33**	66.67	72.22	77.78	77.78	66.67
AR_{10}	87.50	93.75	93.75	93.75	91.67	93.75	**95.83**	**95.83**
AR_{11}	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
AR_{12}	93.10	93.10	**96.55**	**96.55**	89.66	86.21	**96.55**	**96.55**
AR_{13}	73.3	80.00	**91.11**	77.78	71.11	80.00	86.67	77.78
AR_{14}	80.00	90.00	**95.00**	90.00	85.00	90.00	80.00	90.00
AR_{15}	92.86	92.86	92.86	**100.00**	92.86	**100.00**	92.86	**100.00**
AR_{16}	95.83	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
AR_{17}	71.01	78.26	**86.96**	81.16	71.01	79.71	82.61	82.61
AR_{18}	69.70	78.79	84.85	78.79	75.76	**87.88**	78.79	78.79
AR_{19}	55.00	40.00	75.00	70.00	55.00	55.00	65.00	**80.00**
AR_{20}	66.67	77.78	**83.33**	77.78	72.22	77. 78	77.78	**83.33**
AR_{21}	95.00	**100.00**	**95.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
AR_{22}	**100.00**	90.91	**100.00**	90.91	90.91	90.91	90.91	90.91
OA	83.15	87.64	**91.15**	88.48	83.52	88.97	90.30	89.82

The portions in bold represent the best performance.

VGG16, which means that there are more parameters to train in the following FC layers in baseline2. However, the FGSC-23 is still a small-scale dataset, and cannot train well a model which is too complex. Thus, using VGG16 with the two schemes seems to be more suitable for ship classification task in FGSC-23.

4) Efficiency of the Proposed Schemes: The results and analysis presented above highlight the classification effects of different models on FGSC-23. Here, a further discussion is made to analyze the efficiency of the proposed method.

The two schemes unavoidably reduce the processing speed as they add extra layers and parameters to the original CNN model. To explore the efficiency of our method, the training and testing time of the eight models in Section IV-B is recorded as shown in Table VII. The aspects of time included are the total training time per epoch, average training time on a single image, total testing time, and the average testing time per image. For each model, the testing time is measured ten times and the average testing time is recorded.

The data shown in the table reveal that compared with the baseline1, the training time of our AMEFRN increases by about 77% and the testing time increases by about 39%; compared with baseline2, the training time and testing time of our AMEFRN increase by 58% and 18%, respectively. Besides, scheme1 seems more time-consuming than scheme2. The average testing time per image is about 7.1 ms on baseline1 and 7.43 ms on baseline2. In general, the absolute time of our method can be accepted in real applications.

D. Comparison With Other Classification Methods

In this section, our proposed model is compared with other classification models on the FGSC-23. Studies have shown that models based on deep CNN-based features are more effective than most traditional handcrafted feature-based methods for object classification [33], [34]. On this basis, we compare our method only with deep CNN-based methods. They involve representative CNN models proposed in recent years including Inception-v3 [35], DenseNet121 [36], MobileNet [37], and Xception [38], fine-grained classification models for remote sensing images including ME-CNN [16], FDN [15], LGFFE [24], as well as fine-grained object classification models for natural scene images, i.e., B-CNN [39] and DCN [40]. Among them, for ME-CNN, FDN, LGFFE, and DCN, the settings with the best performance reported in their works are adopted. The parameters of the rest five models are tuned for times and the best classification performances are reported. Best performances

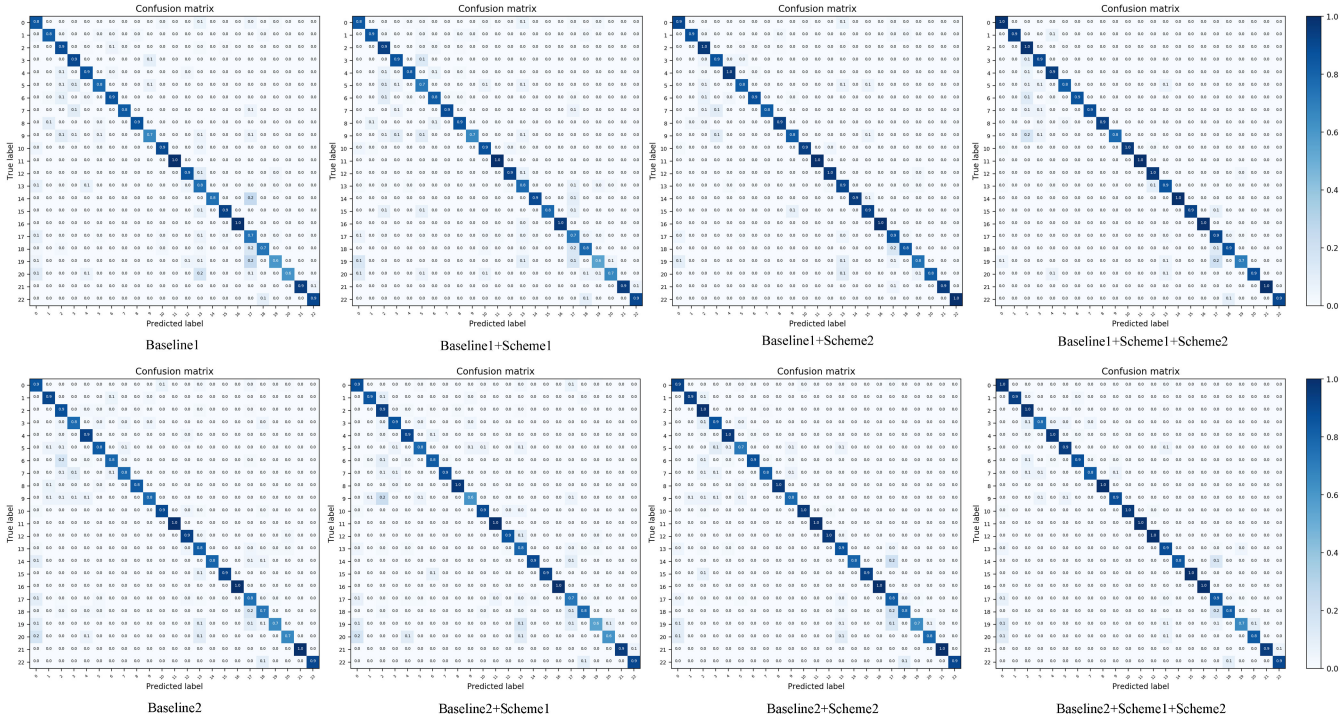


Fig. 11. CMs of models with multilevel feature representation.

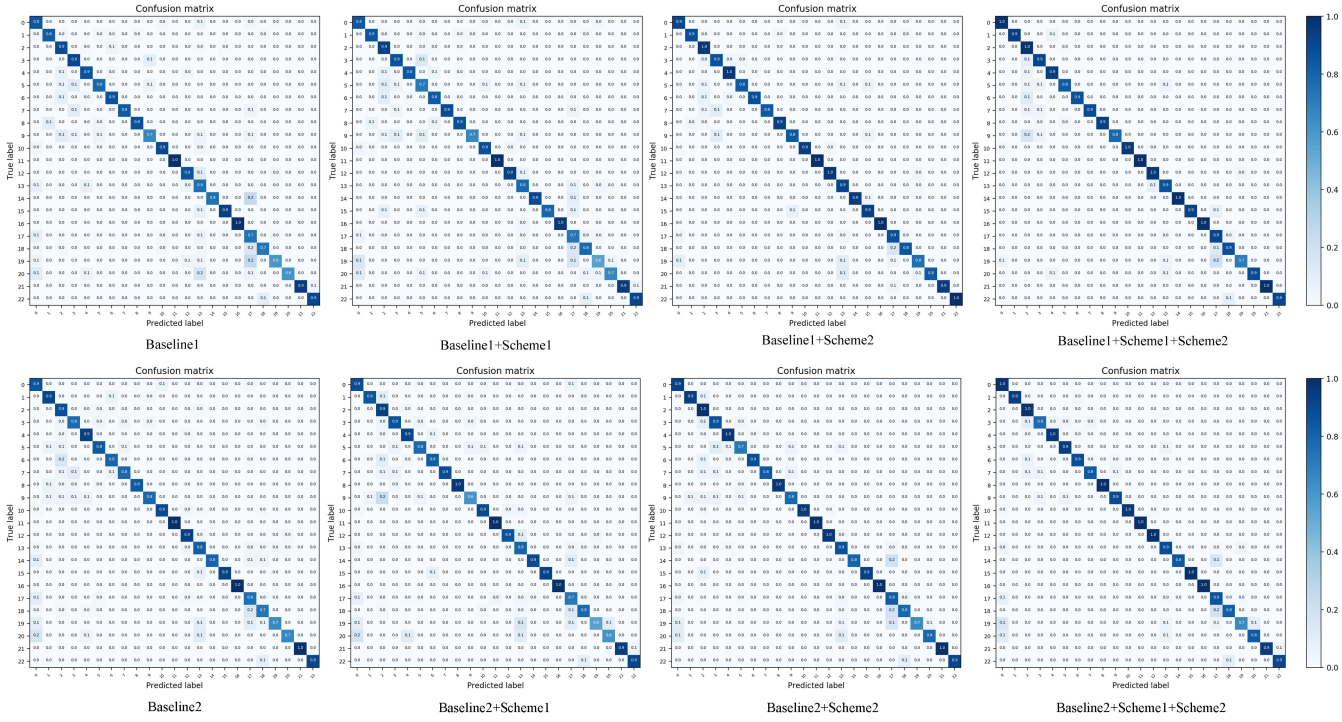


Fig. 12. CMs of models in ablation study.

obtained from these models are then compared with the best performance achieved by our method. The AP and OA percentages of these models as well as the average training and testing time per image are listed in Table VIII, and the CM of each model is shown in Fig. 13.

These results show that among the four deep CNN baselines, the Xception model, which integrates the advantages of ResNet series and Inception series, performs better than the other three models, while the MobileNet runs the fastest. FDN and ME-CNN are two models designed for ship classification in satellite

TABLE VI
AP (%) AND OA (%) OF ABLATION STUDY

Result	Baseline1	Baseline1+ scheme1	Baseline1+ scheme2	Baseline1+ scheme1+ scheme2	Baseline2	Baseline2+ scheme1	Baseline2+ scheme2	Baseline2+ scheme1+ scheme2
AR_0	84.54	84.54	90.72	**98.97**	85.57	88.66	92.78	97.94
AR_1	90.74	85.29	88.24	**94.12**	91.18	85.29	**94.12**	94.12
AR_2	83.33	93.52	96.30	**99.07**	87.96	93.52	**99.07**	98.15
AR_3	90.91	86.36	86.36	**90.91**	77.27	86.36	86.36	77.27
AR_4	91.53	84.75	**96.61**	94.92	93.22	91.53	**96.61**	**96.61**
AR_5	66.67	72.22	83.33	83.33	83.33	77.78	72.22	**94.44**
AR_6	81.36	83.05	88.13	88.98	79.66	83.05	88.13	**89.83**
AR_7	**88.89**	**88.89**	83.33	**88.89**	77.78	**88.89**	83.33	83.33
AR_8	87.10	87.10	93.55	93.55	83.87	96.77	96.77	**100.00**
AR_9	72.22	66.67	83.33	83.33	77.78	61.11	77.78	**88.89**
AR_{10}	87.50	87.50	93.75	**95.83**	89.58	89.58	**95.83**	**95.83**
AR_{11}	90.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
AR_{12}	82.76	93.10	96.55	96.55	93.10	86.21	96.55	**100.00**
AR_{13}	77.78	77.78	**91.11**	86.67	82.22	80.00	86.67	**91.11**
AR_{14}	80.00	90.00	95.00	**100.00**	80.00	90.00	80.00	85.00
AR_{15}	85.71	78.57	92.86	92.86	92.86	92.86	92.86	**100.00**
AR_{16}	**100.00**	95.83	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
AR_{17}	71.01	72.46	86.96	**91.30**	75.36	72.46	82.61	89.86
AR_{18}	60.61	78.79	84.85	**87.88**	72.73	78.79	78.79	84.85
AR_{19}	60.00	55.00	**75.00**	70.00	65.00	55.00	65.00	70.00
AR_{20}	55.56	66.67	83.33	**88.89**	66.67	61.11	77.78	83.33
AR_{21}	90.00	95.00	95.00	**100.00**	**100**	95.00	**100.00**	95.00
AR_{22}	90.91	90.91	**100.00**	90.91	90.91	90.91	90.91	90.91
OA	81.95	83.88	91.15	**93.58**	84.60	85.46	90.30	93.09

The portions in bold represent the best performance.

TABLE VII
TRAINING AND TESTING TIME OF EACH MODEL

Model	Total training time(*ms/epoch*)	Training time per image(*ms*)	Total testing time(*ms*)	Testing time per image(*ms*)
Baseline1	31474.22	6.08	4156.35	5.09
Baseline1+scheme1	53406.17	10.34	4813.35	5.90
Baseline1+scheme2	34056.72	6.59	4781.14	5.86
Baseline1+scheme1+ scheme2	5576.31	10.79	5791.28	7.10
Baseline2	35530.31	6.88	5122.36	6.28
Baseline2+scheme1	55007.25	10.65	5685.98	6.97
Baseline2+scheme2	36870.86	7.13	5620.42	6.89
Baseline2+scheme1+ scheme2	56169.38	10.88	6060.37	7.43

images. They differ in that FDN extracts features from Fourier domain of images, while ME-CNN extracts image features by Gabor filter, LBP operator, and 2-D DFrFT and uses CNN to further extract classification information. ME-CNN has a better performance than FDN while it runs slower than FDN. The LGFFE is also an effective feature representation model and has shown state-of-the-art performance in many classification tasks with remote sensing images. These three models are all proposed for remote sensing classification tasks, and notably, LGFFE achieves the best performance among them. B-CNN and DCN are all designed for fine-grained object recognition task, and the DCN outperforms B-CNN in classification performance for FCSC-23. Although not the fastest algorithm, our method exhibits the best performance, exceeding Xception by 5.82%, LGFFE by 4.13%, and DCN by 2.92% for OA. Besides, our method yields the best AR percentages in 15 out of 23 categories. Evidently, our method displays a state-of-the-art performance in the fine-grained ship classification task of FGSC-23.

TABLE VIII
AP (%) AND OA (%) OF DIFFERENT CLASSIFICATION MODELS

Models	Inception-v3	DenseNet 121	MobileNet	Xception	FDN	ME-CNN	LGFFE	B-CNN	DCN	VGG16+ AMEFRN
AR_0	87.63	86.60	84.54	89.69	85.57	93.81	92.78	84.54	93.81	**98.97**
AR_1	91.18	82.35	88.24	88.24	88.24	91.18	**94.12**	91.18	**94.12**	**94.12**
AR_2	89.81	89.81	88.89	91.67	85.19	87.04	95.37	86.11	97.22	**99.07**
AR_3	77.27	72.73	86.36	81.82	81.82	63.64	81.82	**90.91**	86.36	**90.91**
AR_4	91.53	84.75	88.16	**96.61**	89.83	86.44	**96.61**	89.83	94.92	94.92
AR_5	83.33	77.78	77.78	**88.89**	77.78	77.78	83.33	83.33	83.33	83.33
AR_6	76.27	72.88	86.44	86.44	81.36	76.27	83.05	76.27	86.44	**88.98**
AR_7	**94.44**	77.78	83.33	83.33	72.22	66.67	72.22	83.33	83.33	88.89
AR_8	**100.00**	**100.00**	**100.00**	90.32	77.42	83.87	96.77	96.77	**100.00**	93.55
AR_9	55.56	77.78	77.78	**94.44**	66.67	83.33	83.33	77.78	88.89	83.33
AR_{10}	89.58	93.75	91.67	93.75	87.50	**100.00**	97.92	91.67	93.75	95.83
AR_{11}	**100.00**	**100.00**	90.00	**100.00**	**90.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
AR_{12}	93.10	93.10	93.10	82.76	93.10	93.10	**96.55**	93.10	**96.55**	**96.55**
AR_{13}	68.89	75.56	75.56	77.78	77.78	82.22	77.78	73.33	84.44	**86.67**
AR_{14}	80.00	85.00	80.00	85.00	75.00	85.00	80.00	80.00	90.00	**100.00**
AR_{15}	92.86	**100.00**	92.86	92.86	85.71	**100.00**	92.86	**100.00**	**100.00**	92.86
AR_{16}	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
AR_{17}	76.81	75.36	75.36	76.81	73.91	78.26	85.51	72.46	81.16	**91.30**
AR_{18}	66.67	78.79	69.70	81.82	72.73	81.82	81.82	75.76	**87.88**	**87.88**
AR_{19}	40.00	55.00	45.00	**70.00**	45.00	60.00	65.00	55.00	55.00	**70.00**
AR_{20}	72.22	66.67	55.56	77.78	77.78	66.67	77.78	61.11	83.33	**88.89**
AR_{21}	90.91	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	95.00	**100.00**
AR_{22}	87.63	90.91	90.91	90.91	90.91	**100.00**	90.91	90.91	90.91	90.91
OA	83.88	84.00	84.24	87.76	82.30	85.58	89.45	84.00	90.66	93.58
Training time(ms)	7.51	7.89	**6.08**	10.75	10.62	23.96	7.29	12.53	7.91	10.79
Testing time(ms)	6.89	7.05	**5.92**	9.17	7.92	19.68	6.85	11.17	7.01	7.10

The portions in bold represent the best performance.

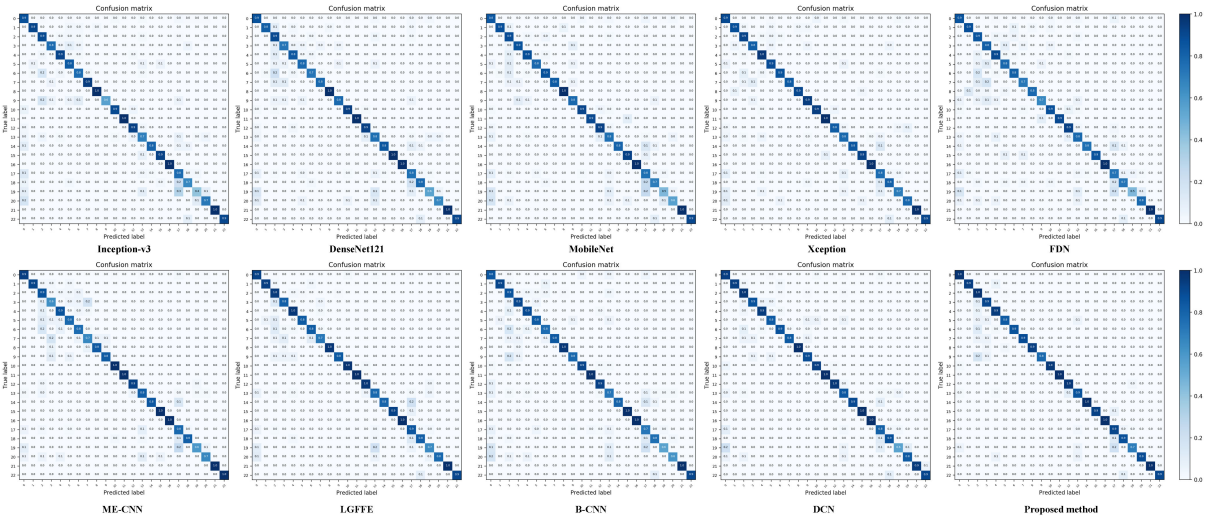


Fig. 13. CMs of models for comparison.

V. Conclusion

In this article, the task of fine-grained ship classification in optical remote sensing images is explored. A set of solutions are proposed to address the challenges of this task. A 23-category fine-grained ship classification dataset called FGSC-23 is established for this investigation, which compensates the lack of relevant data. To the best of our knowledge, it is the second public dataset with fine-grained ship categories after HRSC2016 and with the characters of data diversity, label diversity, and category diversity. A novel attribute-guided classification framework with multilevel enhanced feature representation is proposed for fine-grained ship classification in remote sensing images. We attempt to solve the classification task from two perspectives—enhancing the multilevel visual feature representation of CNN and adding additional attribute supervision information to the CNN framework. Concretely, multilevel local and global features are extracted and the local features are weighed using RNN-based attention module to guide the network focus on silent areas and suppress unimportant areas. Attribute information is added to an attribute-aware branch to extract attribute features, which is auxiliary to the enhanced visual features. The extra supervision information based on the ship's attribute effectively improves the learning capability of classification models. The two schemes proposed in this study can be easily embedded into most CNN models and can be trained end-to-end. Experiments have proven that the two schemes optimize the classification performance and their benefits are compounded when they are used jointly. Our AMEFRN presents state-of-the-art performance on the FGSC-23 dataset, exceeding that of other baselines and classification models.

Despite the high performance of our AMEFRN, we plan to further modify the dataset to resolve the issue of category imbalance and explore the fine-grained ship classification algorithms in other remote sensing images in future work.

Acknowledgment

The authors would like to thank the anonymous reviewers for their hard work. The authors would also like to thank L. Zikun *et al.* for their help with data and the interpretation of part of ships.

References

[1] J. Versteegen *et al.*, "On payload spatial and spectral resolutions for automatic ship detection in satellite images," in *Proc. 5th Int. Workshop Earth Observ. Remote Sens. Appl.*, Xi'an, China, 2018, 1–5.

[2] X. Yang *et al.*, "Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 132.

[3] J. Lan and L. Wan, "Automatic ship target classification based on aerial images," *Proc. SPIE*, vol. 7156, 2008, Art. no. 715612.

[4] Z. Liu *et al.*, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 324–331.

[5] Z. L. Szpak and J. R. Tapamo, "Maritime surveillance: Tracking ships inside a dynamic background using a fast level-set," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 6669–6680, 2011.

[6] L. J. W. Min and M. Dongping, "Extract ship targets from high spatial resolution remote sensed imagery with shape feature," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 30, no. 8, pp. 685–688, 2005.

[7] H. U. Jun-Hua *et al.*, "Detection of ships in harbor in remote sensing image based on local self-similarity," *J. Image Graph.*, vol. 14, no. 8, pp. 591–597, 2009.

[8] D. X. Zhang *et al.*, "Ship targets detection method based on multi-scale fractal feature," *Laser Infrared*, vol. 39, pp. 315–318, 2009.

[9] K. Rainey and J. Stastny, "Object recognition in ocean imagery using feature selection and compressive sensing," in *Proc. IEEE Appl. Imagery Pattern Recognit. Workshop*, 2011, pp. 1–6.

[10] S. Zhina, S. Haigang, and W. Yujie, "Automatic ship detection for optical satellite images based on visual attention model and LBP," in *Proc. IEEE Workshop Electron., Comput. Appl.*, May 2014, pp. 722–725.

[11] Z. Guo, L. Zhang, D. Zhang, and X. Mou, "Hierarchical multiscale LBP for face and palmprint recognition," in *Proc. IEEE Int. Conf. Image Process.*, Hong Kong, Sep. 2010, pp. 4521–4524.

[12] L. Huang, W. Li, C. Chen, F. Zhang, and H. Lang, "Multiple features learning for ship classification in optical imagery," *Multimedia Tools Appl.*, vol. 77, pp. 13363–13389, 2018.

[13] H. Lang, J. Zhang, X. Zhang, J. Zhang, X. Zhang, and J. Meng, "Ship classification in SAR image by joint feature and classifier selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 212–216, Feb. 2016.

[14] 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, 2018, Art. no. 190.

[15] Q. Shi, W. Li, and R. Tao, "2D-DFrFT based deep network for ship classification in remote sensing imagery," in *Proc. 10th IAPR Workshop Pattern Recognit. Remote Sens.*, Beijing, China, Aug. 2018, pp. 1–5.

[16] Q. Shi *et al.*, "Ship classification based on multifeature ensemble with convolutional neural network," *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 419.

[17] K. Li *et al.*, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, 2020.

[18] 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.

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

[20] C. M. Ward, J. Harguess, and C. Hilton, "Ship classification from overhead imagery using synthetic data and domain adaptation," in *Proc. OCEANS MTS/IEEE*, Charleston, SC, USA, 2018, pp. 1–5.

[21] T. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, pp. 1449–1457.

[22] J. Yu, M. Tan, H. Zhang, D. Tao, and Y. Rui, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, doi: 10.1109/TPAMI.2019.2932058.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc of 2015 International Conference on Learning Representations(ICLR)*, San Diego, CA, USA, May, 2015.

[24] Y. Lv *et al.*, "An end-to-end local-global-fusion feature extraction network for remote sensing image scene classification," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 3006.

[25] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. IEEE Conf. Comput. Lang.*, 2014, pp. 1724–1734.

[26] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. of 6th International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, April, 2018.

[27] K. Han *et al.*, "Attribute-aware attention model for fine-grained representation learning," in *Proc. ACM Multimedia Conf.*, 2018, pp. 2040–2048.

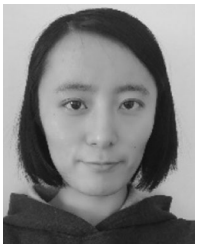
[28] W. Liu, L. Ma, and H. Chen, "Arbitrary-oriented ship detection framework in optical remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 6, pp. 937–941, Jun. 2018.

[29] J. Ding, N. Xue, Y. Long, G. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.

[30] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.

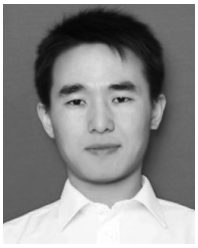
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[32] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vis. Pattern Recognit.*, 2009, pp. 248–255.
[33] G. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[34] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
[36] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vis. Pattern Recognit.*, 2007, pp. 2261–2269.
[37] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.
[38] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1800–1807.
[39] T. Lin, R. Chowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.
[40] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 5157–5166.



Xiaohan Zhang received the B.S. and M.S. degrees in 2014 and 2017, respectively. She is currently working toward the Ph.D. degree in information and communication engineering with Naval Aviation University, Yantai, China.

Her research interests include target recognition and detection in remote sensing and deep learning.

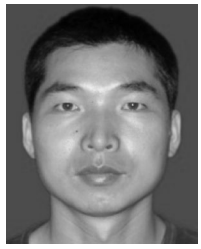


Yafei Lv received the B.S. and M.S. degrees in 2014 and 2017, respectively. He is currently working toward the Ph.D. degree in information and communication engineering with Naval Aviation University, Yantai, China.

His research interests include image retrieval and target association in remote sensing and deep learning.

Libo Yao received the B.S. and M.S. degrees in 2006 and 2019, respectively.

He is currently an Associate Professor with Naval Aviation University, Yantai, China. His research interests include satellite remote sensing information fusion and military big data.



Wei Xiong received the B.S., M.S., and Ph.D. degrees from Naval Aviation University, Yantai, China, in 1998, 2001, and 2005, respectively.

From 2007 to 2009, he was a Postdoctoral Researcher with the Department of Electronic Information Engineering, Tsinghua University, Beijing, China. He is currently a Full Professor with the Naval Aviation University, where he teaches random signal processing and information fusion. He is one of the Founders and the Directors of the Research Institute of information Fusion, Naval Aviation University. He is the Member and Director General of the Information Fusion Branch of the Chinese Society of Aeronautics and Astronautics. His research interests include pattern recognition, remote sensing, and multisensor information fusion.

Chunlong Fu received the B.S. degree from the Wuhan University of Surveying and Mapping, Wuhan, China, in 2000.

He is currently a Senior Engineer with Troops 90139 of PLA, Beijing, China. His research interests include geographic information systems.