



Research Article

Dual-Band Maritime Imagery Ship Classification Based on Multilayer Convolutional Feature Fusion

Xiaohua Qiu ^{1,2}, Min Li ¹, Lin Dong,² Guangmang Deng,² and Liqiong Zhang^{1,2}

¹*Xi'an Research Institute of Hi-Tech, Xi'an 710025, China*

²*School of Information Engineering, Engineering University of PAP, Xi'an 710086, China*

Correspondence should be addressed to Xiaohua Qiu; qxh_1025@163.com and Min Li; proflimin@163.com

Received 18 July 2020; Revised 11 October 2020; Accepted 3 November 2020; Published 2 December 2020

Academic Editor: Everardo Vargas-Rodriguez

Copyright © 2020 Xiaohua Qiu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Addressing to the problems of few annotated samples and low-quality fused feature in visible and infrared dual-band maritime ship classification, this paper leverages hierarchical features of deep convolutional neural network to propose a dual-band maritime ship classification method based on multilayer convolutional feature fusion. Firstly, the VGGNet model pretrained on the ImageNet dataset is fine-tuned to capture semantic information of the specific dual-band ship dataset. Secondly, the pretrained and fine-tuned VGGNet models are used to extract low-level, middle-level, and high-level convolutional features of each band image, and a number of improved recursive neural networks with random weights are exploited to reduce feature dimension and learn feature representation. Thirdly, to improve the quality of feature fusion, multilevel and multilayer convolutional features of dual-band images are concatenated to fuse hierarchical information and spectral information. Finally, the fused feature vector is fed into a linear support vector machine for dual-band maritime ship category recognition. Experimental results on the public dual-band maritime ship dataset show that multilayer convolution feature fusion outperforms single-layer convolution feature by about 2% mean per-class classification accuracy for single-band image, dual-band images perform better than single-band image by about 2.3%, and the proposed method achieves the best accuracy of 89.4%, which is higher than the state-of-the-art method by 1.2%.

1. Introduction

Object classification is a fundamental problem with numerous applications in computer vision and has been extensively studied for visible (VIS) image in the past decades. Because infrared (IR) image provides additional information of the same scene, it helps address various challenges in VIS image, such as variation illumination and occluded appearances. Thus, dual-band data consisting of VIS and IR images has been successfully applied to face recognition [1–3]. Many recent works in object classification [4], person reidentification [5], and pedestrian detection [6] show that dual-band data can improve performance and offer competitive advantages over single band.

After the breakthrough research in image classification by Krizhevsky et al. [7], deep convolutional neural network (CNN) has achieved remarkable success on the ImageNet challenge [8] and produced a number of excellent CNN

models like AlexNet [7], VGGNet [9], GoogleNet [10], and ResNet [11]. Researchers found that features learned from CNN are hierarchical in the whole network [12]; that is, the low-level layer features are similar to Gabor filters and color blobs, the middle-level layer features include fine visual details and semantic information, and the high-level layer features are distinctive semantic features. Furthermore, they also demonstrated the generality and specificity of convolutional feature [13]; that is, first-layer features are general to many datasets and tasks, and last-layer features are specific to a particular dataset or task. However, large-scale datasets like ImageNet are expensive or difficult to collect and time-consuming to train in practical maritime applications. Thus, in order to improve performance for various practical tasks, such as ship classification, the well-known pretrained CNN models like AlexNet and VGGNet have been widely used to fine-tune on ship image [14–16] and extract meaningful ship features [17, 18].

Shi et al. [19] combined low-level features obtained by Gabor filter and multiscale completed local binary patterns (MS-CLBP) with high-level features extracted from the pretrained CNN model with fine-tuning and classified ship categories on VIS images. Shi et al. [20] also proposed a classification framework, which consists of a multifeature ensemble based on convolutional neural network (ME-CNN), and improved the classification accuracy of VIS images. Zhang et al. [21] combined the pretrained VGG-16 model with gnostic fields to improve dual-band maritime ship classification performance. Santos and Bhanu [22] extracted features from the 5th convolutional layer of the pretrained VGG-19 model [9] for both VIS and IR images and proposed a decision level fusion of convolutional networks using a probabilistic model. Due to being limited by high dimension of each layer, most of these methods extracted feature from only one convolutional layer or one fully connected layer. Zhang et al. [4] exploited linear discriminant analysis (LDA) to reduce feature dimension of a convolutional layer, then presented a multifeature fusion method, which combines structure fusion with spectral regression discriminant analysis (SF-SRDA) to learn structure information of convolutional feature, and achieved a promising result. However, features of single layer cannot provide sufficient information. Besides, LDA is a supervised dimensionality reduction technology, and thus, it requires the additional class labels. Although the combination of multilayer features provides richer information, it produces higher-dimensional data and requires more calculations. To address the above problems, recursive neural networks (RNNs) [23] provide one possible solution through a systematic feature learning strategy.

RNN comprises a class of architecture in which the same set of weights is recursively applied within a structural setting and, in particular, on directed acyclic graphs [24]. The main idea of RNN is to learn distributed feature representation by exploiting the same neural network recursively in a tree structure, and it is suitable for processing structured data such as natural language processing [25]. In order to process feature extracted from CNN, a fixed-tree RNN with blocks was presented for multiclass object classification tasks in [23]. The RNN uses nonoverlapping receptive fields instead of overlapping receptive fields in CNN. Besides, it not only reduces the dimension of convolutional feature but also learns feature representation to improve classification performance. Thus, the RNN allows us to transfer information from multiple layers effectively [23]. This characteristic is particularly helpful in feature fusion of multiple layers. Recently, it is also extended to object classification [25, 26] and image super-resolution [27, 28].

In this paper, we present a multilayer convolutional feature fusion method for dual-band maritime ship classification by taking advantage of CNN and RNN. The pretrained and fine-tuned VGGNet models are used to extract convolutional feature of each band image. A number of RNNs with random weights are applied to reduce feature dimension and learn feature representation. The concatenation of low-dimensional hierarchical convolutional features provides abundant information; thus, the proposed method has the

potential to significantly improve classification performance while speeding up the network adaptation process. The main contribution can be concluded as follows:

- (1) A multilayer convolutional feature fusion method is proposed for dual-band maritime ship classification, and three combinations of two feature extractors are investigated
- (2) A number of improved RNNs with random weights are exploited to reduce convolutional feature dimension and learn feature representation
- (3) Multilayer convolutional features of the pretrained and fine-tuned VGGNet models are fused to improve classification performance. The proposed dual-band feature fusion method achieves the best classification accuracy of 89.4% and outperforms state-of-the-art method by 1.2%

The remainder of the paper is organized as follows: the next section introduces the proposed method and improved RNN in details, Section 3 shows and analyzes the experimental results, and Section 4 draws the conclusions.

2. Proposed Method

In our work, we explore the effectiveness of using CNN together with RNNs to recognize maritime ship categories of dual-band data. Especially, the pretrained VGG-f model [29] is applied to extract raw convolutional feature, and the multiple improved RNNs are used to learn feature representation. The proposed framework is illustrated in Figure 1. As is known to all, due to over-fitting, fine-tuning directly the pretrained CNN model in small-scale dataset may not achieve the well classification performance [21]. However, fine-tuning the CNN model on specific dataset can learn specific semantic information of middle and high layers [12]. Therefore, we also take the pretrained VGG-f model with fine-tuning as feature extractor. A classification architecture forwards through five steps, as shown in Figure 1. Firstly, dual-band data including VIS and IR image is taken as the inputs. Secondly, multilevel features of each band image are extracted from the pretrained VGG-f models. Thirdly, a number of improved RNNs without training are employed to learn feature representation, which are hierarchically concatenated for each band image, respectively. Fourthly, the final feature representation of VIS and IR images is fused in the way of concatenation or summation and fed into a linear support vector machine (SVM) classifier in the last step.

2.1. Convolutional Feature Extraction. The pretrained VGG-f model is used to extract image feature in our work. VGG-f network consists 8 layers, 5 of which are convolutional layers (namely, C1, C2, C3, C4, and C5 in Figure 1), and the last 3 are fully connected layers (namely, F6, F7, and F8 in Figure 1). The network architecture was trained on VIS images with 224×224 size and three channels from ImageNet dataset. The first and second layers learn general features similar to Gabor filters and color blobs, which are

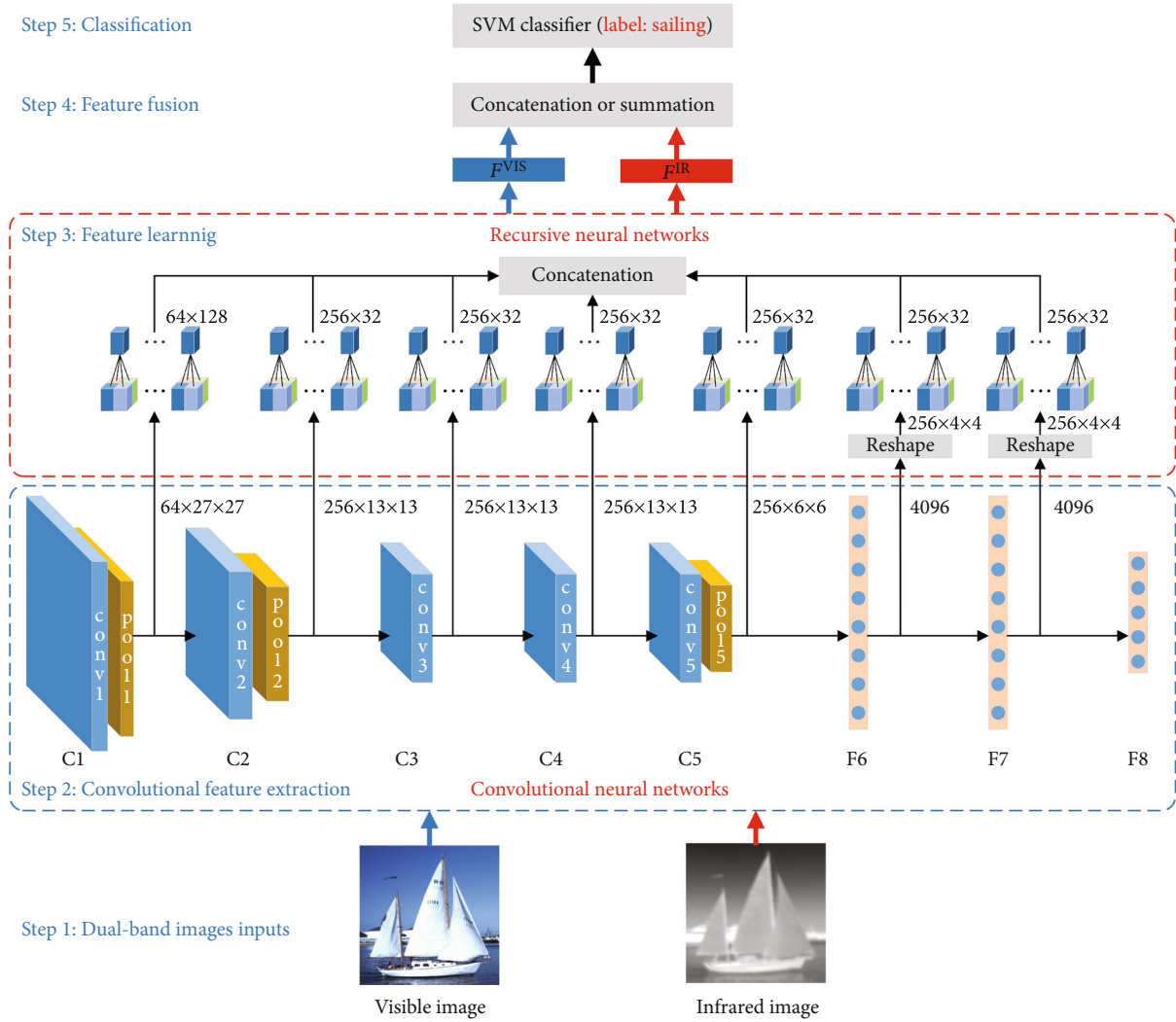


FIGURE 1: An illustration of overview pipeline for dual-band maritime ship classification.

suitable for many datasets and tasks. Then, as the depth of CNN architecture increases, features transition from general to specific [12]. At the last fully connected layer, features are finally specific to a particular dataset and task, such as 1000 classes of ImageNet. Unlike using single middle-level feature [25] and combined middle-level features [26], we exploit the general and specific features extracted from the low-level, middle-level, and high-level layers of the VGG-f network for each band image, such as C2, C5, and F6 layers. Meanwhile, in order to capture the semantic feature of ship, the pretrained VGG-f model with fine-tuning on each band training images of the VAIS dataset [21] is also used as a feature extractor.

2.2. Dimensionality Reduction and Feature Learning. Convolutional features has high dimensions, especially in low-level and middle-level convolutional layers. To exploit the features of different levels, we adopt the improved RNN to reduce the dimension of feature space and learn feature representation. Figure 2 shows an example of two improved RNN architectures.

2.2.1. Multilayer Block RNN. RNN is firstly introduced to learn distributed representation of structured data such as logical terms in [24] and then extended to construct a binary tree in a bottom-up fashion for natural language processing [29]. Although the binary-tree RNN allows the input for more flexibility, the search over optimal trees slows down the architecture. Besides, it was not necessary to obtain high performance for task based on convolutional feature. Therefore, a fixed-tree RNN architecture named Multilayer Block RNN (MB-RNN) is proposed for object classification based on CNN [23]. MB-RNN learns feature representation from convolutional feature and generalized this architecture to allow each layer to merge blocks of adjacent vectors instead of only paired vectors of binary-tree RNN, then improved the performance of classification. An example of MB-RNN is shown in Figure 2(a), with details as follows.

Assume a given convolutional feature is a 3D matrix $X = x_1; \dots; x_{r^2}$, ($X \in R^{K \times r \times r}$), in which K is the filter bank size and $r \times r$ is the size of feature maps. A square block with the size of $K \times b \times b$ is defined as a list of adjacent column vector, which are merged into a parent vector $p_i \in R^K$. Thus, there

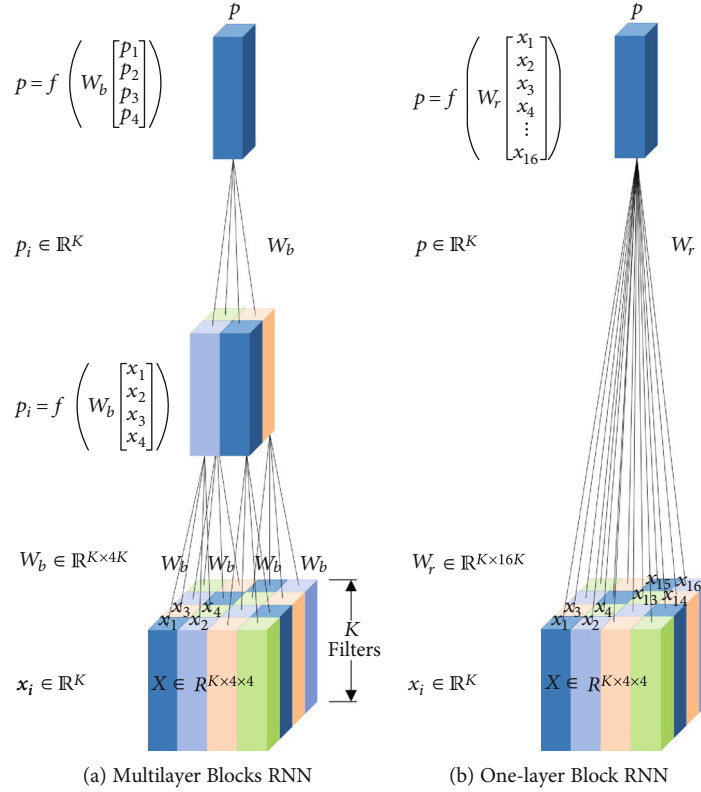


FIGURE 2: An example of two improved RNN architectures. (a) Multilayer Block RNN merges convolutional feature X into a parent vector p through multiple layers and blocks of $b \times b$ children in the end. (b) One-layer Block RNN merges convolutional feature X into a parent vector p through one layer and one block of $r \times r$ children. Note that $r = 16$ and $b = 4$.

are b^2K -dimensional vectors in each block. The parent vector p_i is computed by

$$p_i = f \left(W_b \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{b^2} \end{bmatrix} \right), \quad (1)$$

where $f = \tanh$, the weight parameter matrix $W_b \in R^{(K \times b^2 K)}$, and $i = 1, \dots, (m/b)^2$, where m should be a multiple of b . Equation (1) will be applied to all blocks of vectors in X with the same weights W_b . In general, due to using nonoverlapping receptive field in RNN, $(r/b)^2$ parent vectors form a new matrix $P_1 = p_1, \dots, p_i$. The vectors in P_1 will again be merged in blocks just as those in matrix X using Equation (1) with the same tied weights resulting in matrix P_2 . This recursive procedure continues until only one parent vector p remains.

2.2.2. One-Layer Block RNN. MB-RNN needs a suitable size of convolutional feature because of its architecture. Thus, MB-RNN is not suitable to the size of feature extracted from C2 to C5 layers of the pretrained VGG-f model. Addressing to this problem, one-layer block RNN architecture (OB-

RNN) is proposed to improve MB-RNN. OB-RNN takes the convolutional feature X as a big block of adjacent vectors, that is, $b = r$ in Equation (1), and directly merges X into a parent vector p through one layer. Then, the parent vector p is passed through a nonlinear squash function. An example of OB-RNN is shown in Figure 2(b). Therefore, a feature $X \in R^{K \times r \times r}$ is fed into an OB-RNN and resulted in a K -dimensional vector. Feature dimension is reduced from $K \times r \times r$ to K . Besides, an OB-RNN with weight W_r learns a kind of feature representation, and N OB-RNNs with different weights learn N kinds of feature representation and produce a NK -dimensional feature vector. The larger the N , the higher the feature dimension. Therefore, the number of OB-RNNs N is critical and will be discussed in Section 3.3.2.

Additionally, due to the fact that dimensions of features extracted from convolutional layers have the form of $K \times r \times r$ except fully connected layers, we reshape the features of F6 and F7 by fixing the number of filter bank sizes to 256. Thus, the outputs of F6 and F7 layers are formed into $256 \times 4 \times 4$ dimensions.

2.2.3. One-Layer Block RNN with Random Weights. Generally, training RNN and learning weights require back-propagation through the structure [24]. However, even with random weights, RNN architectures can be inherently frequency selective and translation invariant [30]. In addition,

the RNNs with random weights can also produce high-quality feature vectors for multiclass object classification task [23]. Therefore, the OB-RNN with randomly initialized weights is used to produce feature representation in our work. We forward propagate through all of N OB-RNNs and concatenate their outputs to produce a NK -dimensional feature vector, which is then given into the following feature fusion. The above procedure is applied to each layer of the VGG-f model for both VIS and IR images.

The role of OB-RNN in the process is twofold. First, like MB-RNN architecture, it transforms feature into lower dimension space and improves classification performance and is random-weight-based architecture without requiring back-propagation. Second, because it uses one layer instead of multiple layers, it allows more flexibility for the features extracted from the pretrained CNN model and runs faster than MB-RNN. Meanwhile, OB-RNN would not degenerate the performance of MB-RNN.

2.2.4. Feature Fusion and Classification. The learned feature representation at different layers is fused by *concatenation* for each band image, and the final concatenated feature vector of VIS and IR images is fused by *concatenation* or *summation*. *Concatenation* and *summation* are the two most common vector fusion methods and are often used to fuse the features of multimodal or dual-band data [31]. The fusion goal is to integrate two feature vectors F^{VIS} and F^{IR} to a fused feature vector F^F , where $F^{\text{VIS}}, F^{\text{IR}} \in \mathbb{R}^D$ denote the feature vector of VIS and IR images, respectively, and D is the dimension of a feature vector.

Concatenation is to directly concatenate two feature vectors, which can be defined as

$$\begin{aligned} F^F &= f^{\text{concat}}(F^{\text{VIS}}, F^{\text{IR}}), \\ F_d^F &= F_d^{\text{VIS}}, \\ F_{D+d}^F &= F_d^{\text{IR}}, \end{aligned} \quad (2)$$

where $F_d^{\text{VIS}}, F_d^{\text{IR}}$, and F_d^F represent the d^{th} value of $F^{\text{VIS}}, F^{\text{IR}}$, and F^F , respectively. F_{D+d}^F is the $(D+d)^{\text{th}}$ value of F^F . $1 \leq d \leq D$ and $F^F \in \mathbb{R}^{2D}$. This fusion method concatenates the dimensions of the two input feature vectors.

Summation is a simple addition of the corresponding dimensions of two feature vectors, which can be defined as

$$\begin{aligned} F^F &= f^{\text{sum}}(F^{\text{VIS}}, F^{\text{IR}}), \\ F_d^F &= F_d^{\text{VIS}} + F_d^{\text{IR}}, \end{aligned} \quad (3)$$

where $F_d^{\text{VIS}}, F_d^{\text{IR}}$, and F_d^F represent the d^{th} value of $F^{\text{VIS}}, F^{\text{IR}}$, and F^F , respectively. $1 \leq d \leq D$ and $F^{\text{VIS}}, F^{\text{IR}}, F^F \in \mathbb{R}^D$. The dimension of the fused feature vector is the same as that of the input feature vector.

Concatenation combines two feature vectors with any dimension but generates a feature vector with twice the dimension than *summation* in the case of two feature vectors with the same dimension. In our work, the input feature vectors of *concatenation* have the same dimension for either

single-band or dual-band images. After feature fusion, the final feature representation of the original input dual-band ship data is given to a linear SVM classifier for achieving ship classification task.

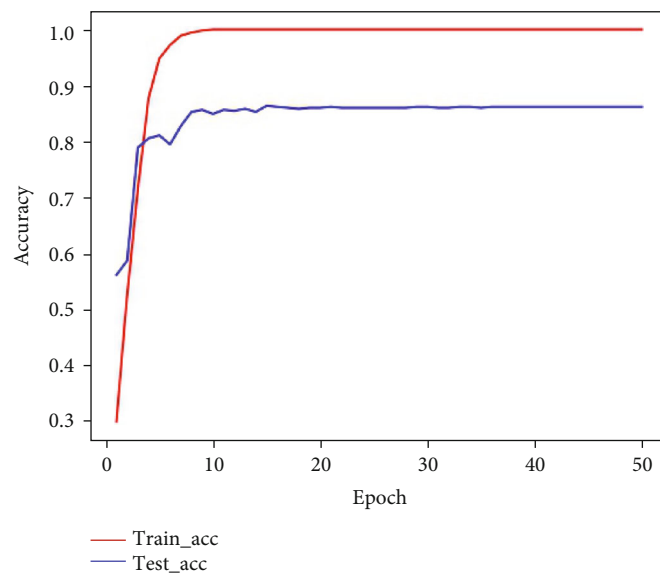
3. Results

3.1. Dataset. We investigate nine fusion models in the proposed fusion architecture on the publicly available VAIS [21] dataset, which is the only existing public database of paired VIS and IR ship imagery. The dataset contains 2865 images (1623 VIS images and 1242 IR images), of which there are 1088 “VIS-IR” unregistered image pairs, and includes 6 categories: cargo ships, medium-other ships, passenger ships, sailing ships, small boats, and tug boats. However, the images are captured at different distance and various times of day, including dusk and dawn. Therefore, some images are high-resolution while a part of images may appear dim and hard to recognize even with manual inspection. In the dataset, the paired VIS/IR image set is partitioned into 539 image pairs for training and 549 image pairs for testing. A sample pair from VAIS is illustrated in Figure 1. Following the baseline method [21], the same train data and test data are used.

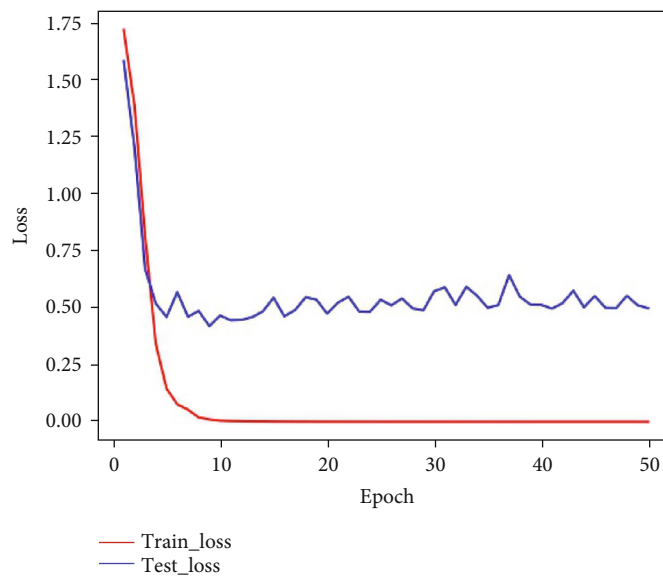
3.2. Implementation Platform and Details. Our processing platform is a standard personal computer with Ubuntu 16.04, with a simple CPU (4.20 GHz) of an Intel Core i7-7770K with 16 GB of random access memory and NVIDIA GTX1080Ti Graphics PU. The computation environment is MATLAB R2017a with MatConvNet [32] toolbox for CNN computation and Liblinear [33] toolbox for classification. Additionally, as the pretrained VGG-f model expects 224×224 three channels VIS image as input, we simply duplicated IR image into three channels. Meanwhile, both VIS and IR images are resized to 224×224 using the nearest interpolation. Besides, the pretrained VGG-f model is fine-tuned on training images of the VAIS dataset, with stochastic gradient descent. Epoch is set to 50 for VIS image and 100 for IR image, learning rate is set to 0.001, and batch size is set to 32. To avoid over-fitting, a dropout layer is applied after the second fully connected layer and its rate is set to 0.5. In addition, due to OB-RNNs with random weights, there are slight fluctuations of classification accuracy in each time for the same procedure. Therefore, we take the mean per-class classification accuracy as evaluation for each time and run the same procedure 50 times for more accurate evaluation, then take the average accuracy together with standard deviation among 50 times as the final evaluation.

3.3. Experimental Results and Analysis

3.3.1. Performance Analysis of Feature Extractors. Firstly, we fine-tune the pretrained VGG-f model on VIS and IR training images of VAIS, respectively. In our previous experiments, data argument and dropout regularization techniques are used to avoid over-fitting during fine-tuning VGG-f model. However, data argument cannot achieve well performance, even if together with dropout. Fortunately, just only using dropout to fine-tune model gets satisfied results on VIS images, but not always good performance on IR

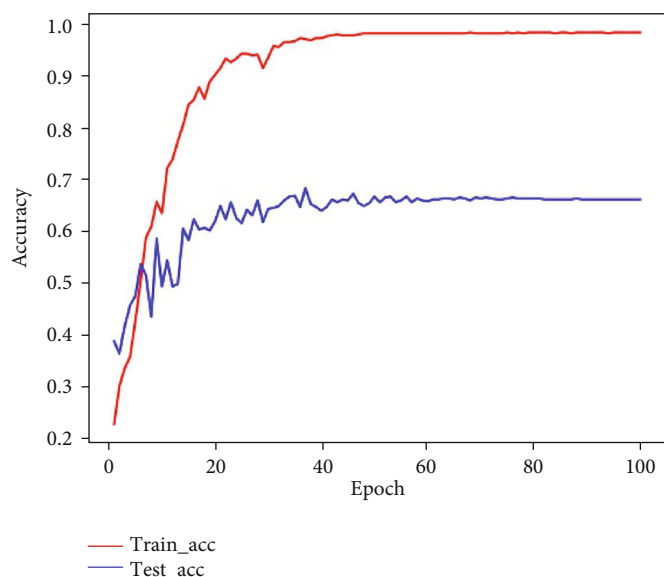


(a) The accuracy curves on VIS images

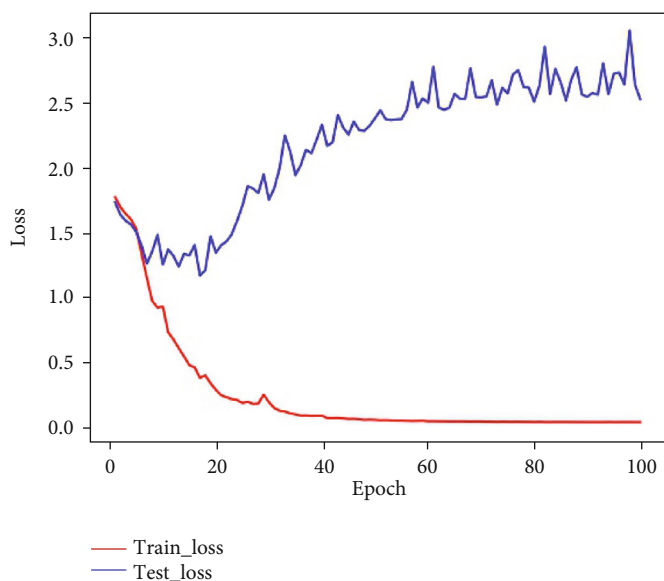


(b) The loss curves on VIS images

FIGURE 3: Continued.



(c) The accuracy curves on IR images



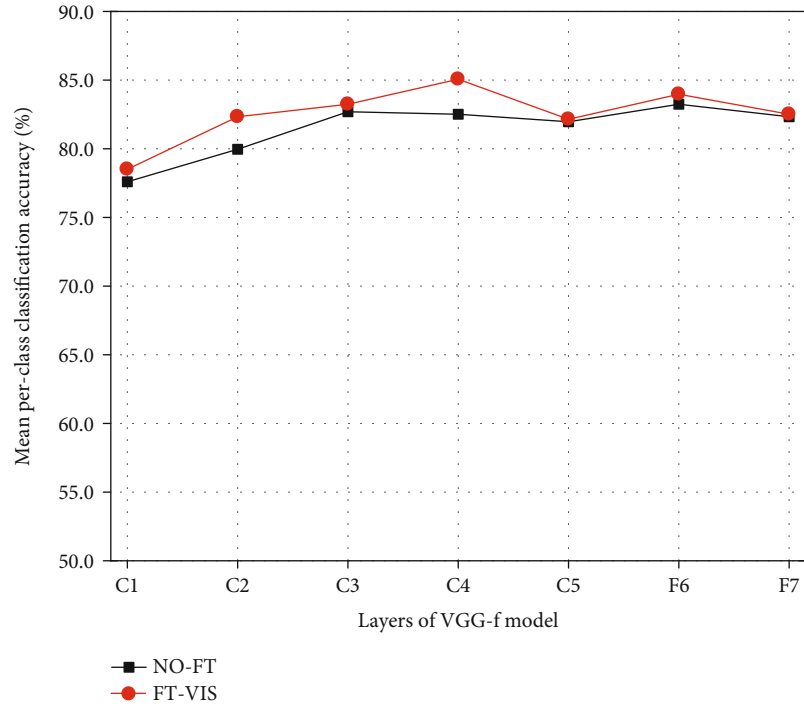
(d) The loss curves on IR images

FIGURE 3: The accuracy and loss curves of the pretrained VGG-f model with fine-tuning on VIS and IR images.

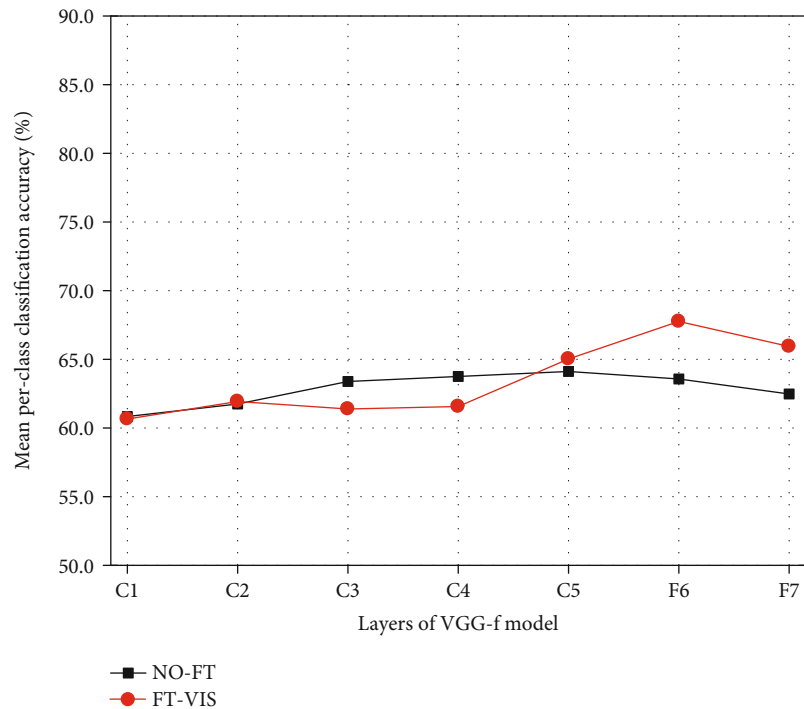
images. Figure 3 shows the accuracy and loss curves of fine-tuning VGG-f model. Train accuracy and loss curves perform well for VIS and IR images, but test accuracy and loss curves show different performance. As shown in Figures 3(a) and 3(b), the test accuracy and loss curves are stable after 20 epochs on VIS image. However, for the results of IR images shown in Figure 3(d), the test loss decreases before epoch 20s, but increases between epochs 20 and 60, then is gradually stable. The test accuracy increases until it stabilizes after epoch 60. Comparing the train and test loss curves, we can find that fine-tuning on IR images has the over-fitting problem. The main reason may be that IR images have low resolution, and some of them are too blur. Meanwhile, the VGG-f model is trained on ImageNet dataset, in which all of images are VIS images. After fine-tuning model

several times, we observed that test accuracy of VIS and IR is about 85.0% and 63.0%, respectively.

Secondly, we take the pretrained and fine-tuned VGG-f models as the feature extractors of dual-band images and investigate the influence of original features produced by the feature extractors for each band on classification performance. Due to the over-fitting problem of fine-tuning the VGG-f model on IR training images, the fine-tuned model cannot be taken as a feature extractor. For convenience, the pretrained VGG-f models without fine-tuning and with fine-tuning on VIS training images are abbreviated as NO-FT and FT-VIS, respectively. As shown in Figure 4(a), the results of FT-VIS are better than those of NO-FT for VIS testing images. From Figure 4(b), we find a great change by comparing the classification accuracy of the two feature



(a) On VIS images



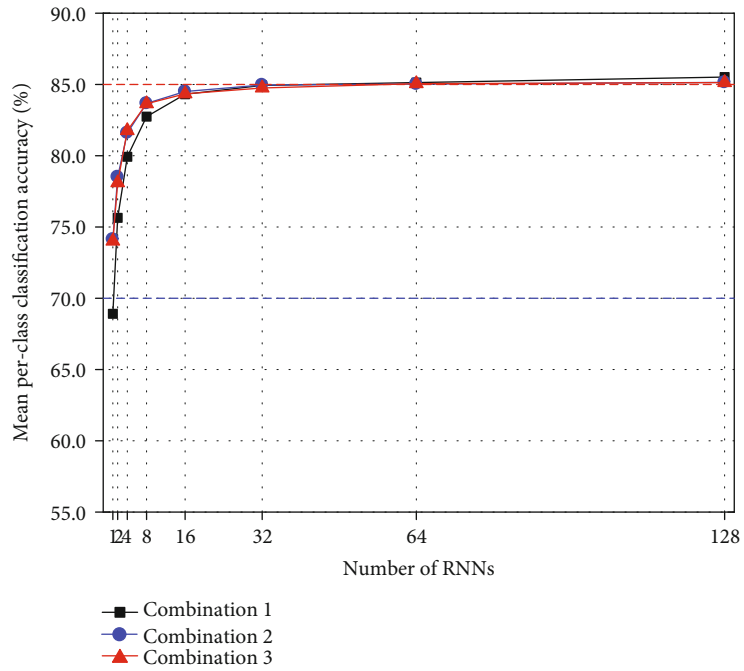
(b) On IR images

FIGURE 4: Influence of original features produced by two feature extractors for each band image on classification performance.

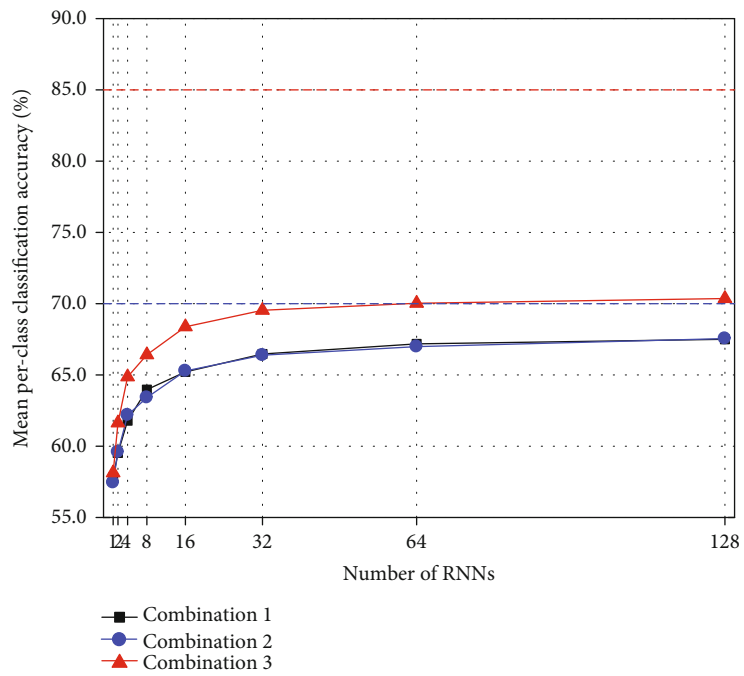
extractors on each layer for IR testing images. Features extracted by FT-VIS significantly increase the accuracy of the last three layers but decrease that of C3 and C4 layers. Generally, compared with VIS image, IR image has low resolution, high contrast and more object contour while less details. It is the reason that FT-VIS cannot greatly improve

TABLE 1: The three combinations of two feature extractors for dual-band maritime ship classification.

Images	Combination 1	Combination 2	Combination 3
VIS	NO-FT	FT-VIS	FT-VIS
IR	NO-FT	NO-FT	FT-VIS

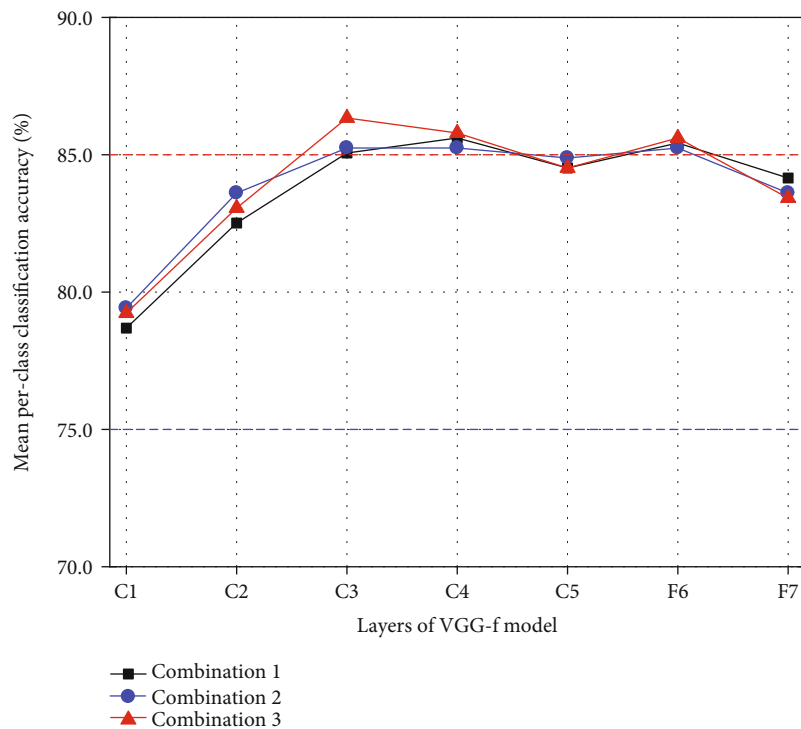


(a) On VIS images

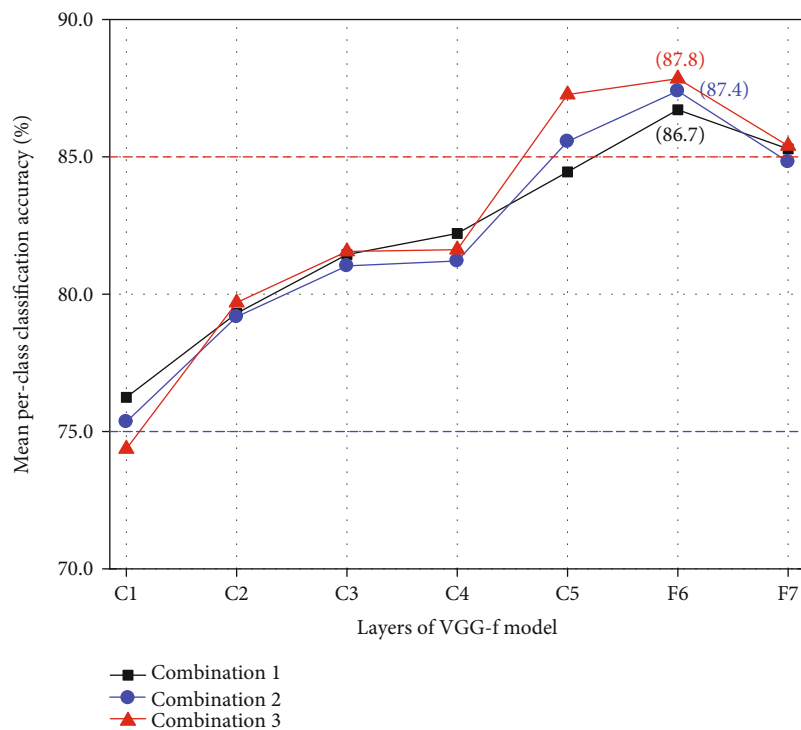


(b) On IR images

FIGURE 5: Continued.



(a) Without RNNs



(b) With RNNs

FIGURE 6: Influence of original features produced by three pretrained VGG-f models for each band image on classification performance of three combinations.

the classification performance on the low-level layers capturing feature such as color, corners, and line segments, but obviously improves the accuracy of the last three layers. According to the above analysis, fine-tuning the pretrained

VGG-f model on VIS training images can learn ship semantic information. Therefore, we investigate three combinations of the two feature extractors in our feature fusion architecture, as shown in Table 1.

TABLE 3: The classification accuracy (%) of two combinations on two and three layers of the VGG-f model.

2*layers	Combination 2				Combination 3			
	VIS	IR	CON	SUM	VIS	IR	CON	SUM
F6	84.9 ± 0.7	66.4 ± 1.3	87.4 ± 0.8	87.4 ± 0.7	84.9 ± 0.7	69.3 ± 1.1	87.8 ± 0.6	87.6 ± 0.7
C2F6	86.5 ± 0.6	71.6 ± 0.8	88.9 ± 0.6	88.7 ± 0.7	86.6 ± 0.7	69.6 ± 1.2	89.0 ± 0.5	88.8 ± 0.7
C3F6	86.9 ± 0.8	71.2 ± 0.9	89.1 ± 0.6	88.9 ± 0.8	86.8 ± 0.7	70.1 ± 1.0	89.1 ± 0.5	89.0 ± 0.6
C4F6	86.5 ± 0.7	71.4 ± 1.0	88.8 ± 0.6	88.5 ± 0.6	86.6 ± 0.7	69.7 ± 1.0	89.1 ± 0.5	88.8 ± 0.5
C5F6	85.9 ± 0.7	71.1 ± 0.8	88.9 ± 0.6	88.8 ± 0.7	86.0 ± 0.6	70.5 ± 0.9	88.6 ± 0.6	88.4 ± 0.7
C2C5F6	86.6 ± 0.7	71.7 ± 0.6	89.0 ± 0.6	88.8 ± 0.7	86.8 ± 0.7	70.2 ± 0.9	89.4 ± 0.5	89.1 ± 0.5
C3C5F6	87.2 ± 0.7	71.4 ± 0.5	89.4 ± 0.5	89.1 ± 0.8	87.2 ± 0.5	70.0 ± 0.7	89.3 ± 0.4	89.1 ± 0.5
C4C5F6	87.0 ± 0.9	72.2 ± 0.9	89.3 ± 0.4	89.0 ± 0.7	86.8 ± 0.5	70.5 ± 0.7	89.1 ± 0.6	88.9 ± 0.5

Accuracy evaluation using the average accuracy together with standard deviation in 50 times. CON and SUM represent *concatenation* and *summation* feature fusion methods, respectively. Abbreviated symbol C2F6 represents that C2 layer and F6 layer features for each band image are concatenated, the same as to others. Bold denotes that the average accuracy is the best one in the corresponding column of the table.

TABLE 4: Comparison of classification accuracy (%) with other state-of-the-arts on VAIS dataset.

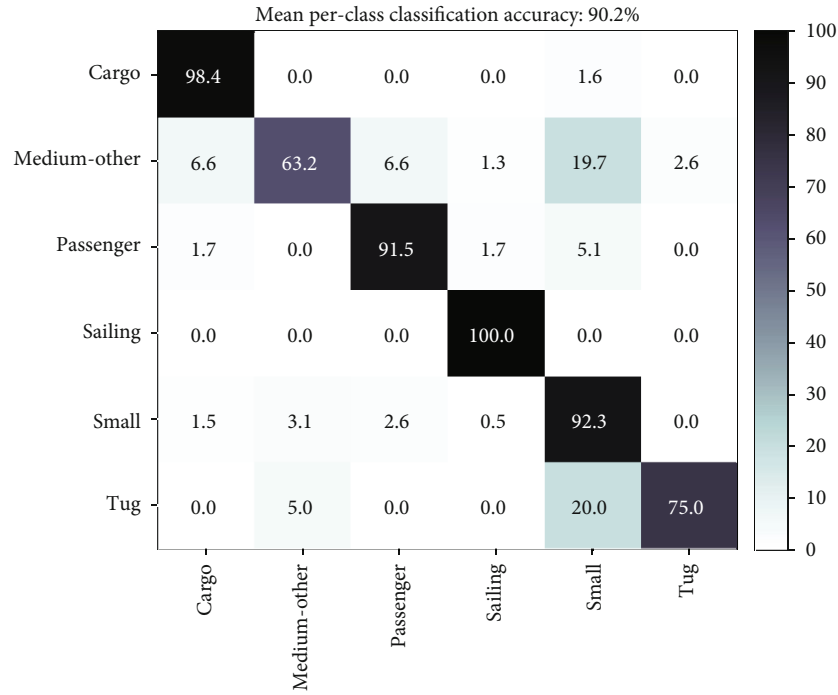
Methods	VIS	IR	VIS + IR
CNN [21]	81.9	54.0	82.1
Gnostic field [21]	82.4	58.7	82.4
CNN + gnostic field [21]	81.0	56.8	87.4
ME-CNN [20]	87.3	—	—
MFL (feature-level) + ELM [34]	87.6	—	—
CNN + Gabor + MS-CLBP [19]	88.0	—	—
Multimodal CNN [15]	—	—	86.7
DyFusion [22]	—	—	88.2 ± 0.2
SF-SRDA [4]	87.6	74.7	88.0
Proposed Combination 3-SUM (C2C5F6)	86.8 ± 0.7	70.2 ± 0.9	89.1 ± 0.5
Proposed Combination 3-SUM (C3C5F6)	87.2 ± 0.5	70.7 ± 0.7	89.1 ± 0.5
Proposed Combination 2-CON (C3C5F6)	87.2 ± 0.7	71.4 ± 0.5	89.4 ± 0.5
Proposed Combination 3-CON (C2C5F6)	86.8 ± 0.7	70.2 ± 0.9	89.4 ± 0.5

CON and SUM represent *concatenation* and *summation* feature fusion methods, respectively. Abbreviated symbol C2C5F6 represents that C2 layer, C5 layer, and F6 layer features for each band image are concatenated, the same as to others. Bold indicates the best one.

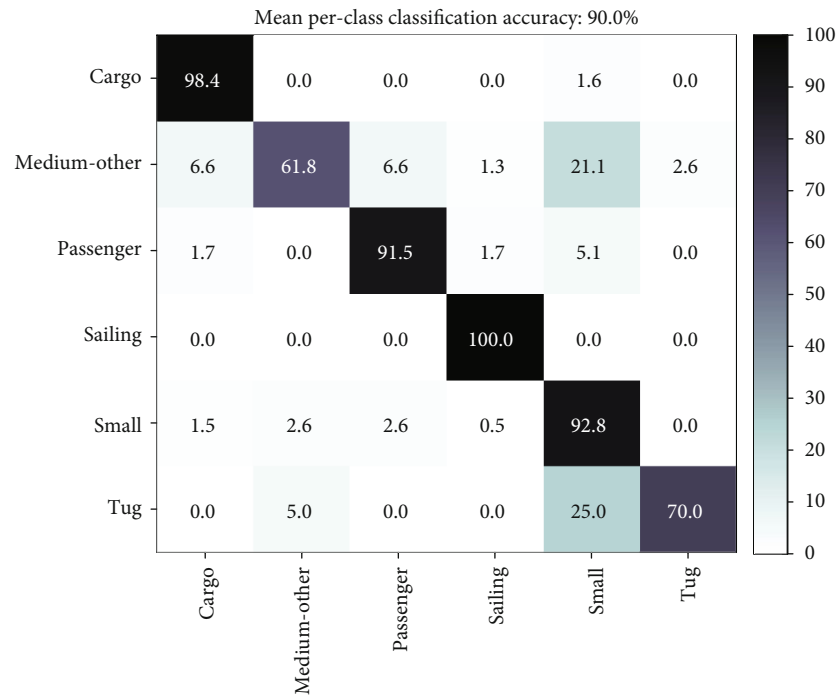
3.3.2. Classification Accuracy Evaluation of Single-Layer Feature Fusion. In this section, we analyze the above three combinations on the single layer of the VGG-f model in two aspects. Firstly, we analyze the effect of the number of RNNs. Figure 5 shows the influence of the number of RNNs on the classification accuracy of the F6 layer. The number ranges from 1 to 128. For whether single band image [see Figures 5(a) and 5(b)] or dual-band images [see Figures 5(c) and 5(d)], increasing the number of RNNs improves the classification accuracy, and it levels off at around 32. However, the larger the number of RNNs, the more time it takes to learn. Therefore, according to the same size of feature for each layer, the number of RNNs of each layer is set to 32 except 128 for the C1 layer. In addition, the influence of RNNs on the feature extractors from Figures 5(a) and 5(b), especially for IR test images, is also found. The results of Combination 2 and Combination 3 on IR images [see lines with circle and triangle in Figure 5(b)], which use the FT-VIS feature extractor, are bet-

ter than that of Combination 1 using the NO-FT feature extractor. Furthermore, comparing the four figures in Figure 5 by using red and blue dotted lines as reference, we observe that feature fusion with RNNs of dual-band images improves the classification performance of each band image no matter how the number changes. Meanwhile, the classification accuracies of Combination 2 and Combination 3 on dual-band images are higher than that of Combination 1.

Secondly, we evaluate the classification performance of three combinations without RNNs and with RNNs on a single layer. Table 2 gives the feature size of each layer without and with RNNs. Feature size affects classification accuracy and efficiency. The smaller the size of the feature, the faster the SVM classifier processes. Figure 6 shows the classification accuracy of three combinations without and with RNNs on each layer. Comparing Figures 6(a) and 6(b), it can be found that the classification accuracies of three combinations with RNNs are better than those of without RNNs on the last three



(a) Fusion with *concatenation*



(b) Fusion with *summation*

FIGURE 7: Confusion matrices of fusion results for Combination 3-CON (C2C5F6) in one time. Note the 90.2% accuracy for the *concatenation* method and 90.0% accuracy for the *summation* method.

layers (C5, F6, and F7), and all of them achieve the best on the F6 layer. However, the accuracy of the first four layers (C1, C2, C3, and C4) is reduced. According to the results in Figures 5(c) and 5(d), if we increase the number of RNNs, it will improve the accuracy but consume much time. In our work, we mainly focus on multilayer feature fusion not just single-layer feature fusion.

3.3.3. *Classification Accuracy Evaluation of Multilayer Feature Fusion.* The above analysis indicates that RNN improves classification performance for dual-band images and also shows that Combination 2 and Combination 3 outperform Combination 1 on single-layer feature fusion. Therefore, this subsection investigates the classification performance of Combination 2 and Combination 3 on multiple

layer feature fusion with RNN. Table 3 shows the classification accuracy of three combinations on two and three layers of the VGG-f model, and the results of single F6 layer feature fusion are shown for comparison. As shown in Table 3, we found three points. Firstly, multilayer feature fusion improves the classification performance of VIS image, IR images, and dual-band images and especially outperforms single layer by 1.1%~2.3% for VIS images and by 0.8%~2.0% for dual-band images. Secondly, the accuracy of dual-band images is higher than that of VIS image by about 2.3%, and three-layer feature fusion performs better than two-layer feature fusion by about 0.3%. Thirdly, the results of the *concatenation* feature fusion method are almost higher than those of *summation* by 0.2%~0.3%. However, the feature size of *concatenation* is twice that of *summation*. Therefore, as the number of combination layers increases, the *summation* fusion method runs faster than *concatenation*.

3.3.4. Comparison with Other State-of-the-Arts. We compare the best of our fusion method with seven methods on the VAIS dataset: (1) the baseline method (CNN + gnostic field) [21], (2) Multimodal CNN [15], (3) DyFusion [22], (4) SF-SRDA [4], (5) MFL (feature-level) + ELM [34], (6) CNN + Gabor + MS-CLBP [19], and (7) ME-CNN [20]. The first four methods are on paired images, and the last three methods are on VIS images of the paired images. Table 4 shows the experimental results. As it is shown, CNN + Gabor + MS-CLBP obtains the best classification performance on VIS images, and SF-SRDA achieves the highest classification accuracy on IR images. Obviously, the proposed method performs better than the other methods on dual-band images and achieves 89.4% of the best classification accuracy, outperforming the current best method (DyFusion) by 1.2%. Therefore, it also shows that the proposed method is more suitable for dual-band ship classification than single band. Figure 7 shows the confusion matrices of classification result on Combination 3 for one time. In the experiments of Combination 2 and Combination 3, all categories except for medium-other and tug are above 90% accuracy, and sailing ship is sometimes 100% accuracy. However, classification accuracy of medium-other ship and tug boat are always less than 80%. As we found, medium-other ship and tub boat are often confused with small ship.

4. Discussions

The proposed method exploits a pretrained or fine-tuned VGG-f model to extract image features, and it is suitable for small-scale datasets with few data samples. The OB-RNN is flexible for layer convolutional features produced by most of pretrained well-known CNN models. The OB-RNNs reduce the dimension of convolutional feature to avoid the “curse of dimensionality” caused by the fusion of low-level, middle-level, and high-level convolutional features. The feature of multilayer convolutional features fusion includes richer information and stronger feature representation ability than any single-layer convolutional feature. Moreover, there is a great potential for further improvement of the proposed method. One potential factor is that the

VGG-f model we used can be replaced by the pretrained well-known CNN models such as VGG-16, ResNet, and GoogleNet. Besides, training OB-RNNs can also further improve the feature representation ability and classification accuracy. In our method, the simple fusion strategy *concatenation* and *summation* are used to fuse the features of dual-band images. Therefore, putting the features of dual-band images into a feature space to learn a common feature representation is also a future direction.

5. Conclusions

According to few annotated dual-band samples, we propose a multilayer convolutional feature fusion method to recognize maritime ship category. Fine-tuning the pretrained VGG-f model on VIS images captures specific ship information and improves classification accuracy. The improved RNN with random weights reduces convolutional feature dimension and learns more feature representation as the number of RNNs increases. The low-level, middle-level, and high-level convolutional features are concatenated for producing complementary information and improving classification performance. Experimental results on the public VAIS dataset demonstrate that the best multilayer feature fusion performs better than other existed methods and confirm that our method is more suitable for dual-band ship classification than single band. We will focus on the decision level fusion in the future.

Data Availability

The Excel data used to support the findings of this study are included within the supplementary information file (available here).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China (Grant No. 61102170) and the National Social Science Foundation of China (Grant No. 15GJ003-243).

Supplementary Materials

(1) The worksheet “Fig.4” shows classification accuracy of each layer of original CNN feature for each band image, which is produced by two feature extractors named NO-FT and FT-VIS. Therefore, “Figure 4” in our manuscript is formed by these values in the worksheet “Fig.4.” (2) The worksheet “Fig.5” shows classification accuracy of F6 layer with different RNN numbers (1, 2, 4, 8, 16, 32, 64, and 128) for each band image and dual-band images. “Figure 5” in our manuscript is formed by these “Mean” values in the worksheet “Fig.5.” (3) The worksheet “Fig.6(a)” shows classification accuracy of original CNN feature fusion (that is Without RNNs) on single layer of three combinations in

VIS and IR images. “Figure 6(a)” in our manuscript is formed by these “Mean” values for “CON” in the worksheet “Fig. 6(a).” (4) The worksheet “Fig.6(b)” shows classification accuracy of CNN feature fusion with RNNs (that is With RNNs) on single layer of three combinations in VIS and IR images. “Figure 6(b)” in our manuscript is formed by these “Mean” values for “CON” in the worksheet “Fig. 6(b).” (5) The worksheet “Table3” shows classification accuracy of two/three layers feature fusion with RNNs of three combinations in VIS and IR images. “Table 3” in our manuscript is based on these “Mean + Std” values for “Combination 2” and “Combination 3” in the worksheet “Table3.” (6) The worksheet “FeatureFusion-all” shows classification accuracy of single, two, and three layers feature fusion with RNNs of three combinations in VIS and IR images. (*Supplementary Materials*)

References

- [1] K. Guo, S. Wu, and Y. Xu, “Face recognition using both visible light image and near-infrared image and a deep network,” *CAA Transactions on Intelligence Technology*, vol. 2, no. 1, pp. 39–47, 2017.
- [2] G. Hermosilla, M. Rojas, J. Mendoza et al., “Particle swarm optimization for the fusion of thermal and visible descriptors in face recognition systems,” *IEEE Access*, vol. 6, pp. 42800–42811, 2018.
- [3] C. Peng, N. Wang, J. Li, and X. Gao, “DLFace: deep local descriptor for cross-modality face recognition,” *Pattern Recognition*, vol. 90, pp. 161–171, 2019.
- [4] E. Zhang, K. Wang, and G. Lin, “Classification of marine vessels with multi-feature structure fusion,” *Applied Sciences*, vol. 9, no. 10, p. 2153, 2019.
- [5] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, “Visible thermal person re-identification via dual-constrained top-ranking,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 1092–1099, Stockholm, Sweden, 2018, International Joint Conferences on Artificial Intelligence Organization.
- [6] L. Zhang, Z. Liu, S. Zhang et al., “Cross-modality interactive attention network for multispectral pedestrian detection,” *Information Fusion*, vol. 50, pp. 20–29, 2019.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, pp. 1097–1105, arXiv, 2012.
- [8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: a large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, FL, USA, 2009, IEEE.
- [9] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” <https://arxiv.org/abs/1409.1556>.
- [10] C. Szegedy, V. Vanhoucke, S. Ioe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, Las Vegas, NV, USA, 2016, IEEE.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016, IEEE.
- [12] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS’14*, pp. 3320–3328, Montreal, Canada, 2014, MIT Press.
- [13] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision-ECCV 2014*, pp. 818–833, Springer International Publishing, Cham, 2014.
- [14] B. Solmaz, E. Gundogdu, V. Yucesoy, and A. Koc, “Generic and attribute-specific deep representations for maritime vessels,” *IPSA Transactions on Computer Vision and Applications*, vol. 9, no. 1, p. 22, 2017.
- [15] K. Aziz and F. Bouchara, “Multimodal deep learning for robust recognizing maritime imagery in the visible and infrared spectrums,” in *Image Analysis and Recognition*, A. Campilho, F. Karray, and B. Haar Romeny, Eds., pp. 235–244, Springer International Publishing, 2018.
- [16] M. Milicevic, K. Zubrinic, I. Obradovic, and T. Sjekavica, “Application of transfer learning for fine-grained vessel classification using a limited dataset,” in *Applied Physics, System Science and Computers III*, pp. 125–131, Springer International Publishing, Cham, 2019.
- [17] F. Bousetouane and B. Morris, “Off-the-shelf CNN features for fine-grained classification of vessels in a maritime environment,” in *Advances in Visual Computing*, pp. 379–388, Springer International Publishing, Cham, 2015.
- [18] E. Gundogdu, B. Solmaz, V. Yucesoy, and A. Koc, “MARVEL: a large-scale image dataset for maritime vessels,” in *Computer Vision - ACCV 2016*, pp. 165–180, Springer International Publishing, Cham, 2017.
- [19] Q. Shi, W. Li, F. Zhang, W. Hu, X. Sun, and L. Gao, “Deep CNN with multi-scale rotation invariance features for ship classification,” *IEEE Access*, vol. 6, pp. 38656–38668, 2018.
- [20] Q. Shi, W. Li, R. Tao, X. Sun, and L. Gao, “Ship classification based on multifeature ensemble with convolutional neural network,” *Remote Sensing*, vol. 11, no. 4, p. 419, 2019.
- [21] M. M. Zhang, J. Choi, K. Daniilidis, M. T. Wolf, and C. Kanan, “VAIS: a dataset for recognizing maritime imagery in the visible and infrared spectrums,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 10–16, Boston, MA, USA, 2015, IEEE Computer Society.
- [22] C. E. Santos and B. Bhanu, “Dyfusion: dynamic IR/RGB fusion for maritime vessel recognition,” in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1328–1332, Athens, Greece, 2018, IEEE.
- [23] R. Socher, B. Huval, B. Bhat, C. D. Manning, and A. Y. Ng, “Convolutional-recursive deep learning for 3d object classification,” in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS’12*, pp. 656–664, Lake Tahoe, NV, USA, 2012, Curran Associates Inc..
- [24] C. Goller and A. Kuchler, “Learning task-dependent distributed representations by backpropagation through structure,” in *Proceedings of International Conference on Neural Networks (ICNN’96)*, pp. 347–352, Washington, DC, USA, 1996, IEEE.
- [25] H. M. Bui, M. Lech, E. Cheng, K. Neville, and I. S. Burnett, “Object recognition using deep convolutional features transformed by a recursive network structure,” *IEEE Access*, vol. 4, pp. 10059–10066, 2016.

- [26] A. Caglayan and A. B. Can, "Exploiting multi-layer features using a CNN-RNN approach for RGB-D object recognition," in *Computer Vision - ECCV 2018 Workshops*, pp. 675–688, Springer International Publishing, Cham, 2019.
- [27] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1637–1645, Las Vegas, NV, USA, 2016, IEEE.
- [28] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2790–2798, Honolulu, HI, USA, 2017, IEEE.
- [29] R. Socher, C. C.-Y. Lin, A. Y. Ng, and C. D. Manning, "Parsing natural scenes and natural language with recursive neural networks," in *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11*, pp. 129–136, Bellevue, WA, USA, 2011, Omnipress.
- [30] A. M. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Y. Ng, "On random weights and unsupervised feature learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11*, pp. 1089–1096, Bellevue, WA, USA, 2011, Omnipress.
- [31] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 243–250, Honolulu, HI, USA, 2017, IEEE.
- [32] A. Vedaldi and K. Lenc, "MatConvNet: convolutional neural networks for MATLAB," in *Proceedings of the 23rd ACM International Conference on Multimedia - MM '15*, pp. 689–692, Brisbane, Australia, 2015, ACM Press.
- [33] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: a library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.
- [34] L. Huang, W. Li, C. Chen, F. Zhang, and H. Lang, "Multiple features learning for ship classification in optical imagery," *Multimedia Tools and Applications*, vol. 77, no. 11, pp. 13363–13389, 2018.