applied sciences

MDPI

Article

Multi-Feature Fusion with Convolutional Neural Network for Ship Classification in Optical Images

Yongmei Ren [1,2], Jie Yang [1,*], Qingnian Zhang [3] and Zhiqiang Guo [1]

1 Hubei Key Laboratory of Broadband Wireless Communication and Sensor Networks, School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China; renyongmei@whut.edu.cn (Y.R.); guozhiqiang@whut.edu.cn (Z.G.)
2 School of Electrical and Information Engineering, Hunan Institute of Technology, Hengyang 421002, China
3 School of Transportation, Wuhan University of Technology, Wuhan 430070, China; zhangqn@whut.edu.cn
* Correspondence: jieyang@whut.edu.cn

Received: 15 September 2019; Accepted: 27 September 2019; Published: 9 October 2019

Abstract: The appearance of ships is easily affected by external factors—illumination, weather conditions, and sea state—that make ship classification a challenging task. To facilitate realization of enhanced ship-classification performance, this study proposes a ship classification method based on multi-feature fusion with a convolutional neural network (CNN). First, an improved CNN characterized by shallow layers and few parameters is proposed to learn high-level features and capture structural information. Second, handcrafted features of the histogram of oriented gradients (HOG) and local binary patterns (LBP) are combined with high-level features extracted by the improved CNN in the last fully connected layer to obtain discriminative feature representation. The handcrafted features supplement the edge information and spatial texture information of the ship images. Then, the Softmax function is used to classify different types of ships in the output layer. Effectiveness of the proposed method is evaluated based on its application to two datasets—one self-built and the other publicly available, called visible and infrared spectrums (VAIS). As observed, the proposed method demonstrated attainment of average classification accuracies equal to 97.50% and 93.60%, respectively, when applied to these datasets. Additionally, results obtained in terms of the F1-score and confusion matrix demonstrate the proposed method to be superior to some state-of-the-art methods.

Keywords: ship classification; convolution neural network; multi-feature fusion; HOG; LBP; VAIS

1. Introduction

In accordance with global economic expansion and increasing foreign trade, maritime traffic has witnessed tremendous growth, and this has resulted in an obvious increase in the number of ships taking to the oceans. Ship classification assumes great importance with regard to several aspects, such as maritime safety, traffic monitoring, and maritime-domain awareness [1,2]. In addition, with the recent advancement in artificial intelligence, the concept of intelligent ships [3] is being projected as the next big thing, as regards to the future of the maritime industry. Intelligent ships are expected to be characterized by safety, reliability, energy-saving potential, environmental friendliness, and economic efficiency. The information-perception technology—a key aspect of intelligent ships—has been developed to ensure that ships can obtain accurate information regarding themselves as well as their surrounding environment, including neighboring ships, video-surveillance information, and obstacle information.

The purpose of ship classification is to identify accurately ship classes within ship datasets. Since visual sensors are often susceptible to weather changes and illumination conditions, ship classes

can, at times, be difficult to identify. Moreover, intra-class variations within certain ship classes make ship classification more complex and challenging [4,5]. Generated from different sensors, ship images can be classified into synthetic aperture radar (SAR) images, infrared images, and visible optical images respectively. SAR images can be generated at all day/night time and under weather conditions (clear or cloudy), but SAR images have some limitations. SAR images mainly obtained from a radar, which is expensive and vulnerable to other electromagnetic interference. SAR images with low resolution do not facilitate the subsequent target tracking, and the object detection methods based on SAR images can only distinguish large vessels from background, but ignoring the boats with a long distance. In addition, using radar would give away the vessel's position, which precludes its use in some military applications. Infrared images are monitored by thermal imaging instead of natural light. However, infrared images have high sensitivity to the thermal radiation of the target, but not sensitive to the brightness change. Moreover, with low spatial resolution, infrared images cannot provide texture details. Dealing with targets with long distance, infrared images have a poor signal-to-noise ratio, seriously affecting the efficiency. By contrast, visible optical images contain grayscale information of multiple bands and have characteristics of high resolution and detailed texture. Therefore, visible optical images can provide abundant visual information and have better discrimination of targets. Using an appropriate algorithm, more ship features can be extracted to facilitate further ship classification and detection. Even more noteworthy is that the camera used to obtain the visible optical images has a low cost and power consumption, in addition, it can be implemented easily thanks to the small size.

Owing to the high importance associated with ship recognition and classification over the past few decades, several investigations have been performed in this regard by a number of researchers, and numerous approaches for ship classification have been proposed [6,7].

The traditional classification method is based on handcrafted features, and it involves use of a combination of target features, such as contour, texture, area, and color, for class recognition. Several feature-extraction techniques associated with face recognition have also been employed to facilitate class recognition within certain sample ship datasets [8]. Among traditional feature-description methods, the ones commonly used include the histogram of oriented gradients (HOG) [9], scale-invariant feature transform (SIFT) [10], local binary patterns (LBP) [11], etc. Many research works have been done on ship classification. Rainey et al. [12] proposed several image classification and feature extraction algorithms on ship imagery, which obtained good results. Arguedas [13] developed the local binary patterns (LBP) operator for vessel classification. Parameswaran et al. [14] used the bag of visual words (BOVW) in vessel classification. These methods, however, mainly focus on extraction of low-level visual features. Image backgrounds in actual scenarios can be complex. Additionally, different illumination intensities and viewing angles affect feature extraction. The color and texture information corresponding to the same object may differ in different images, thereby resulting in inaccurate feature extraction, which in turn, leads to lower classification accuracy. Methods based on extraction of handcrafted features, therefore, are greatly limited in terms of their ability to express images accurately. Consequently, they are only suitable for use in specific applications and rely on expert knowledge.

In recent years, deep learning methods that integrate feature extraction and classifier training to realize end-to-end machine learning have witnessed rapid development. The convolution operation automatically acquires structural information by combining low-level features to form high-level features that are more abstract. Since the emergence of convolutional neural networks at the ImageNet Challenge, deep learning methods have come to be recognized as a favorable means for solving target recognition problems, such as classification, positioning, and detection. In addition, such methods have been successfully employed in such applications as speech recognition [15], behavior detection [16], image classification [17], traffic-sign recognition [18], and other tasks. Since the labeling process of visible ship images is difficult and expensive, the number of ship images is often very small. Hence, at present, most research concerning ship classification is based on satellite and synthetic aperture radar (SAR) imaging. Only very few studies have been performed concerning class recognition of

ships based on images captured by a camera. Rainey et al. [19] designed a convolutional neural network (CNN) to facilitate ship-class recognition based on satellite imaging, thereby demonstrating attainment of high classification efficiency. Bentes et al. [20] also proposed a CNN model capable of operating on multi-resolution input, and evaluated its classification performance by providing as input TerraSAR-X images comprising five maritime classes—cargo, tanker, windmill, platform, and harbor structure. Providing images containing a combination of different resolutions as input helps improve classification accuracy. However, further investigation needs to be performed to understand how changes in image resolution affects internal activations within CNN. Khellal et al. [21] proposed a CNN-based method involving extreme learning machine for recognition of infrared ship images. This method was applicable to infrared-based recognition systems, and required an extreme learning machine based ensemble for image classification post learning of CNN features. Consequently, the algorithm on which this method was based was rather complicated. With technological advancements, deep-classification models, such as AlexNet [22], very deep convolutional networks (VGGNet) [23], and ResNet [24] have also been developed. Shi et al. [25] used deep CNN with multi-scale rotation invariant features to facilitate ship classification. The classification accuracy achieved equaled 98.33% when operating on the BCCT200-RESIZE (barges, cargo ships, container ships, and tankers, 200 images per class) dataset [26]; however, when operating on the visible and infrared spectrums (VAIS) dataset [27], the classification accuracy recorded equaled only 88%. Liu et al. [28] proposed a method for ship detection and classification based on remote sensing images with an improved residual network. However, the dataset was small, thereby resulting in overfitting. Zhao et al. [29] introduced a method based on CNN to extract features from ship images. Their method combined the HOG and hue, saturation, and value (HSV) algorithms to extract edge and color features, respectively, from images, and demonstrated attainment of a classification accuracy of 93.55% when applied to the visible ship dataset. Cao et al. [30] proposed a ship recognition method combined with image segmentation and deep learning feature extraction in video surveillance, the method can effectively identify three types of ships, with an average detection accuracy of 87%. Zhang et al. [31] designed a multi-feature structure fusion method based on spectral regression discriminant analysis (SF-SRDA). Shi et al. [32] proposed a classification framework consists of a multi-feature ensemble based on the convolutional neural network (ME-CNN) for optical remote sensing images.

According to the current relevant researches, the feature characterization ability and classification accuracy of the ship classification method needs to be further improved. The proposed study aims at attainment of high performance and ship-classification accuracies. Although the CNN can automatically capture structural information and has performed encouragingly, it still has some limitations. Firstly, in the CNN model, low-level features extracted from the first convolutional layer are fed into the top layer via layer-by-layer propagation to generate high-level features. However, the low-level features obtained in the first convolution layer may cause some important low-level information lost, such as edges and contours. Secondly, CNN requires large amounts of labeled data to train models. Network architectures, such as the VGGNet and ResNet, are relatively complex, and their direct application to ship datasets containing few samples may result in significant overfitting, thereby affecting algorithm performance.

In view of this, a multi-feature fusion with CNN has been proposed for ship classification. First, an improved shallow layers CNN-based method has been proposed to facilitate the learning of more useful, robust CNN features. Additionally, because there are many mature handcrafted feature extraction methods, such as HOG that can extract the edge of an image, and LBP can extract the local structure and texture information of an image [13]. Therefore, HOG and LBP are selected to describe the ship edge features and texture features more accurately. Subsequently, these three types of features are fused, and the Softmax function is used to classify different types of ship images in the output layer.

Major contributions of this paper can be summarized as follows. (1) An improved CNN with shallow layers and relatively few convolution kernels has been proposed to facilitate ship classification. The proposed method automatically extracts and learns features from ship images, thereby avoiding

occurrence of overfitting problems caused by a lack of training samples. Few network parameters can improve the efficiency of the algorithm. (2) The HOG and LBP features are fused with the high-level features extracted by the improved CNN network to further supplement the edge, profile information and spatial texture information of the ship images. The fused features also take the advantages of both high-level and handcrafted features to obtain more comprehensive ship features and improve the ability to describe and identify ships. Moreover, handcrafted features are not affected by the number of labeled samples. Namely, they can be extracted even when the amount of dataset is very small. (3) To facilitate comparison of the proposed method against other deep CNN models, such as AlexNet, VGG-16, and ResNet-18, to validate its effectiveness, the said three networks are separately fine-tune into network architectures suitable for an image size of 64 pizels × 64 pixels. (4) We construct a self-built ship dataset, including 9000 images. The classification performance of the proposed method is compared with the existing methods on two datasets—the self-built ship dataset and the publicly available yet challenging VAIS dataset.

The remainder of this paper is organized as follows. Section 2 provides a basic understanding of CNNs. Section 3 describes the proposed classification method in detail. Section 4 introduces the two experimental datasets employed in this study and discusses experimental results obtained and their subsequent analysis. Lastly, Section 5 lists major conclusions drawn from this study along with a brief discussion of future endeavors.

2. CNN Structure

CNNs are deep neural networks that are good at mining local features of input images. The "local weight sharing" and "down-sampling" characteristics of CNNs ensures invariance in the shifting, scaling, and rotation of images up to a certain extent. Local connections serve to reduce parameters that need to be trained within the network, thereby speeding up the training process. CNNs are multilayered learning networks comprising an input layer, convolution layer, pooling layer (i.e., down-sampling layer), fully connected layer, and output layer. The basic structure of a CNN is depicted in Figure 1. In a typical CNN, the first few layers of the network usually comprise alternating convolution and pooling layers, whereas the last few layers near the output layer usually represent fully connected networks [33].

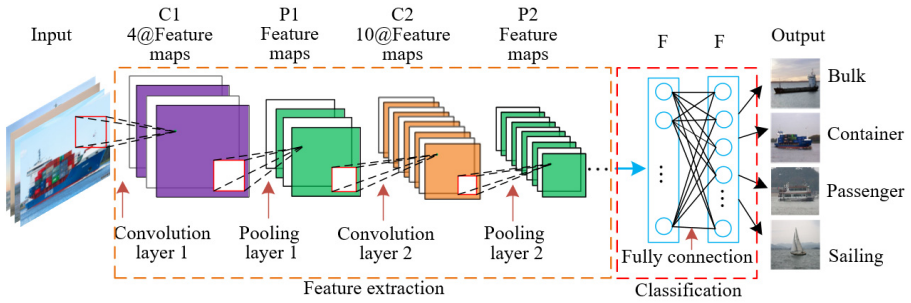


Figure 1. Typical convolutional neural network (CNN) structure.

Once an input image has been processed by the convolution layer, multiple feature maps of the image are obtained, and each feature map represents the extraction of a particular feature. During the extraction process, neurons of the same feature map share a set of weights (i.e., convolution kernel), and n feature maps can be obtained using n convolution kernels.

The pooling layer reduces the dimensionality of the convolution result to facilitate reduction in computation. Average pooling retains more image-background information, whereas max pooling retains more information concerning image texture. In view of this, max pooling was favored in this study to ensure retention of more texture information to facilitate accurate image classification.

The fully connected layer connects all features extracted from the upper layer, reduces these features to their one-dimensional forms, and sends the output value to the output layer for classification.

The output layer solves the multiclass classification problem. The Softmax function used for solving such problems is good at approximating complex nonlinear relationships whilst offering advantages of high training speed and high classification accuracy. Since the number of ship classes within the two datasets considered in this study equaled 4 and 6, respectively, the Softmax function was considered for label prediction of ship images. The output layer (Softmax layer) was placed after the last fully connected layer within the network architecture.

3. Proposed Ship Classification Method

The key to enhancing ship-classification accuracy lies in selection of appropriate features that characterize ship-image properties. Deep neural networks can automatically capture structural information. Compared to their low-level counterparts, features captured by deep neural networks are more abstract, robust, and discriminative when dealing with in-class differences and inter-class similarities. Thus, deep neural networks demonstrate good feature-extraction and classification abilities. At present, deep learning is considered the most advanced method, and it has tasted great success in the field of image classification. In scene classification, the purpose is to separate different kinds of objects in the same picture, and the between-class scatter is relatively large. However, there is small difference between classes for ship classification. CNN features are based on low-level features obtained in the first convolution layer, which may not fully capture all local features and result in the loss of some important information, such as edges and contours. HOG has excellent capacity to describe the contour of objects. LBP can capture spatial texture information and local structure. Thus, in this study, we proposed an effective multi-feature fusion learning framework for ship classification, which combines the handcrafted features with high-level features obtained by CNN. Figure 2 depicts the flowchart of proposed ship classification framework. Firstly, the improved CNN in this paper is used to capture high-level features. Then the HOG and LBP features are used to extract handcrafted features to supplement global information of ship images. High-level features and handcrafted features are concatenated together to obtain a more discriminating representation. Finally, the Softmax function is used to classify different types of ship images in the output layer.

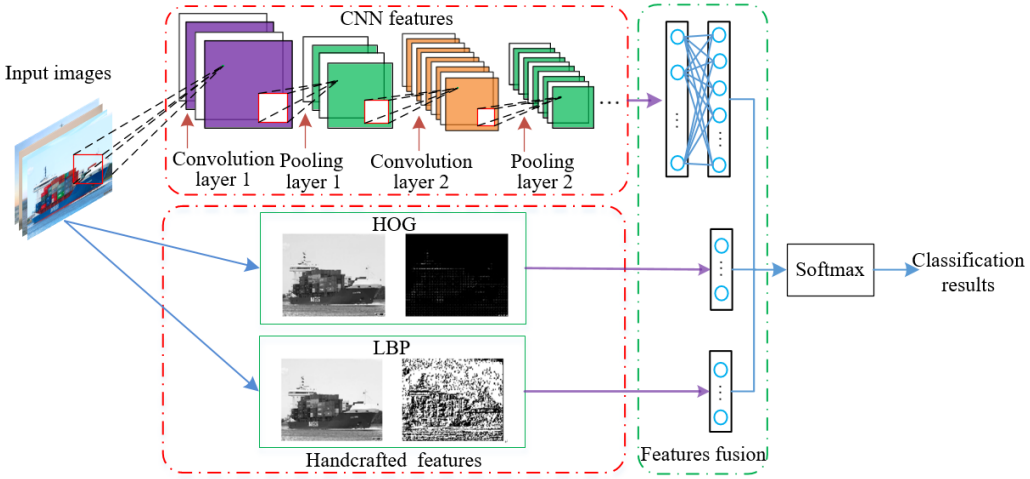


Figure 2. The flowchart of the proposed ship classification framework.

3.1. Feature Extraction Based on Improved CNN

3.1.1. Motivation for Development of Improved CNN

Although images within ship datasets can be easily affected by external factors, such as weather conditions, illumination, and viewing angles, each ship class possesses its own unique shape different from other classes. The key to correct ship classification, therefore, lies in the establishment of a CNN model capable of reliably identifying in-class differences and inter-class similarities. In addition, it is

also very important to select the appropriate network structure according to the number of the labeled samples. In this study, experiments were performed using the LeNet-5 network comprising few layers followed by fine-tuning of deep CNN networks (refer Section 3.4) to obtain the fine-AlexNet, fine-VGG-16, and fine-ResNet-18 versions for application to the self-built ship dataset (refer Section 4.1). The observed classification accuracy equaled only 89.99% when employing the LeNet-5 network while that obtained using the fine-AlexNet CNN was the highest (i.e., 93.22%). This indicates that feature-extraction characteristics of LeNet-5 network were not of the highest order, and consequently the corresponding ship-classification accuracy was not satisfactory. That said, although classification accuracy of the fine-AlexNet network equaled 93.22%, it comprises deeper layers and is easily susceptible to the occurrence of overfitting in a dataset with relatively few samples. Additionally, the number of parameters involved therein is relatively large, which in turn, affects the efficiency of the algorithm. Thus, classification accuracy of the said CNN can be further enhanced.

In this study, changes have been effected in the AlexNet network structure to develop an improved ship-classification CNN that comprises shallow layers and few convolution kernels. The number of convolution and pooling layers has been reduced from five to four and three to two, respectively, to facilitate enhanced performance and reduced computational complexity. Additionally, the size of an image has been considered to be directly related to the convolution kernel being selected. Likewise, selection of the convolution-kernel size is related to whether image features can be effectively extracted. The size of the ship images considered in this study equaled 64 pixels × 64 pixels, and after several experiments, convolution kernel sizes of 5 × 5 and 3 × 3 were considered for the extraction of ship features. The specific analysis can be described as follows. The convolution kernel size of typical classification networks, such as the LeNet-5, AlexNet, VGGNet, and ResNet, were considered to obtain the different combinations of kernel sizes listed in Table 1 (all other parameters are the same), wherein numbers 1–4 represent the four combination cases, respectively. Figure 3 depicts the classification accuracy for each ship class (refer Section 4.1) corresponding to the said four cases.

Table 1. Combination of different convolution kernels.

Layer	Kernel Size			
	1	**2**	**3**	**4**
Conv1	7 × 7	5 × 5	5 × 5	3 × 3
Conv2	3 × 3	3 × 3	3 × 3	3 × 3
Conv3	3 × 3	3 × 3	3 × 3	3 × 3
Conv4	3 × 3	3 × 3	5 × 5	3 × 3

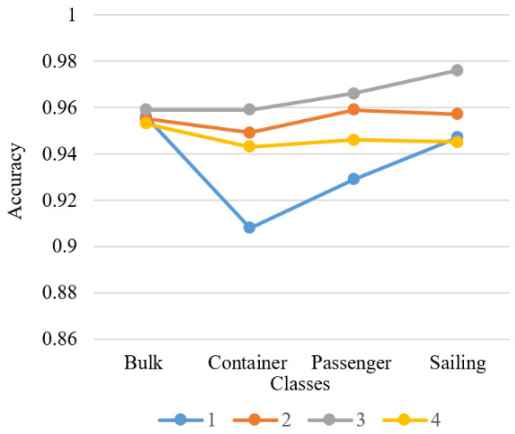


Figure 3. Classification accuracy (%) of each ship class corresponding to four cases listed in Table 1.

As depicted in Figure 3, the classification accuracy of the container ship was observed to be the lowest in the first case, whereas that of a sailing ship was the lowest in the fourth case. Classification

accuracies of four ship classes were observed to be the highest in the third case, and these were closely followed by those corresponding to the second case. In view of these results, the third case was used to design the improved CNN-based method proposed in this study. Additionally, convolution kernel sizes of 5×5 and 3×3 were considered to make the extraction of ship features subtle and comprehensive.

Figure 4 depicts the visualization of feature maps within different convolutional layers. There exists 32 feature maps in Figure 4a,c, whereas Figure 4b,d comprises of 64 feature maps. It is easy to conclude that convolution kernels could describe ship-image characteristics based on different aspects. The first and second convolutional layers mainly extracted features pertaining to the texture and detail of ship images. Moreover, these features were very close to the original image (i.e., container ship). All features extracted from these convolutional kernels could be combined to characterize the ships more comprehensively. The shallow networks contained more features and possessed the ability to extract key features (for example, the container feature extracted from the fifth feature map in Figure 4a). The third and fourth convolutional layers mainly extracted the contour, shape, and other strong features (e.g., container area). The greater the depth of these layers, the more abstract these features were observed to be. The said features could be regarded as a combination of features extracted within previous layers. Features obtained using different convolution kernels can be complementary with regard to the description of ship images. By combining the features extracted by the different convolution kernels, information concerning ship images can be accurately represented using the proposed method.

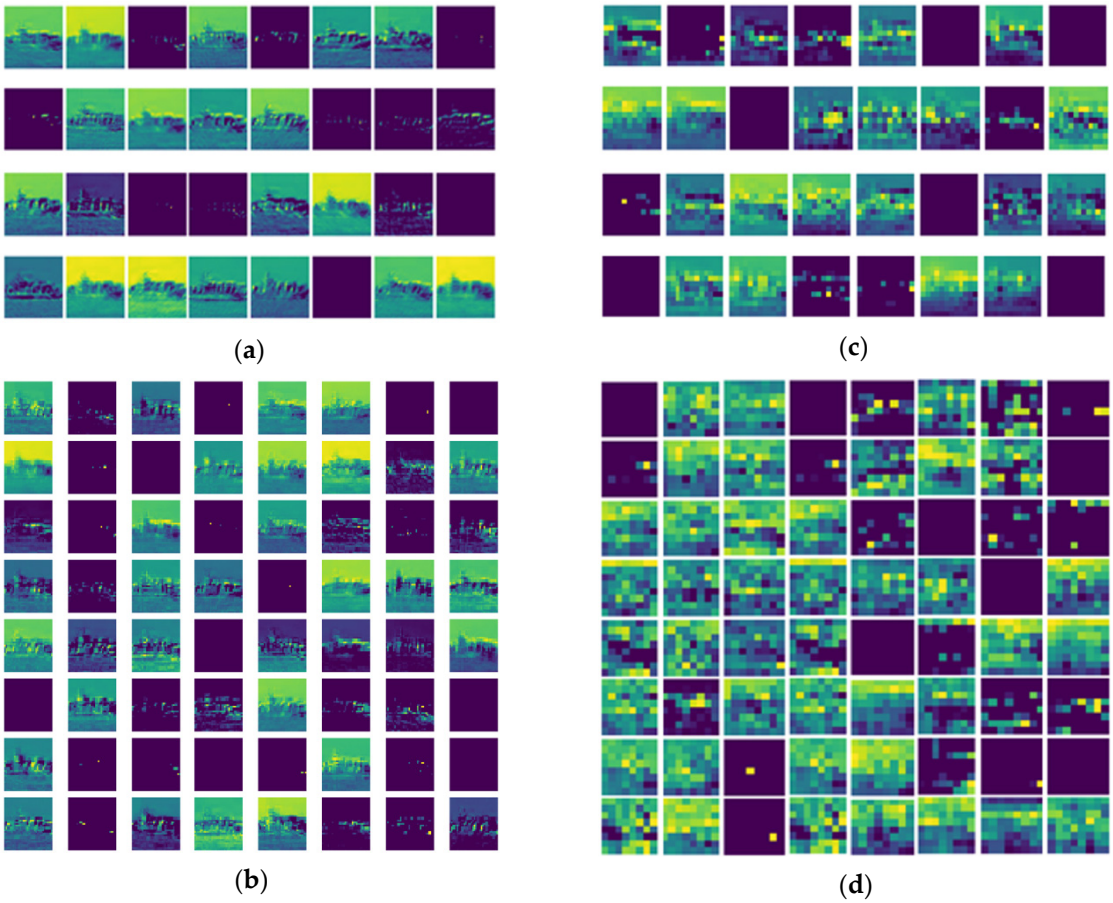


Figure 4. Visualization of feature maps within different convolutional layers: (**a**) First; (**b**) second; (**c**) third; and (**d**) fourth.

3.1.2. Improved CNN

Based on the above analysis, the detailed architecture of the proposed improved CNN described in this paper has been depicted in Figure 5. Table 2 describes specific parameters.

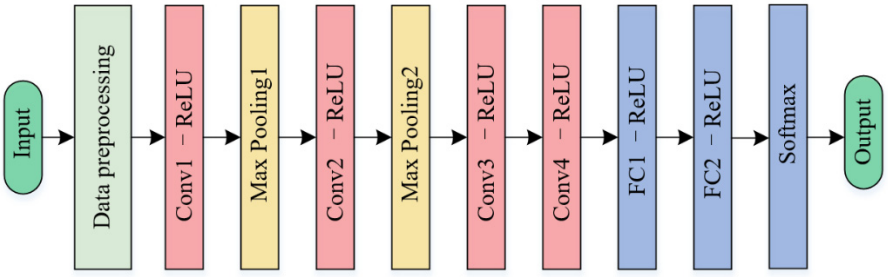


Figure 5. Structure of the proposed improved CNN.

Table 2. Details concerning proposed CNN architecture.

Layer	Kernel Size	Filter Number	Stride	#Parameters
Conv1	5 × 5	32	1	2432
Max Pooling1	2 × 2	-	2	0
Conv2	3 × 3	64	1	18,496
Max Pooling2	2 × 2	-	2	0
Conv3	3 × 3	32	1	18,464
Conv4	5 × 5	64	1	51,264
FC1	384	-	-	884,736
FC2	192	-	-	73,729
Softmax	4	-	-	769

As described in Figure 5 and Table 2, the proposed network comprises of four convolution layers, two max pooling layers, two fully connected layers, and the Softmax layer. After preprocessing (refer Section 3.1.3), the size of ship images equaled 56 pixels × 56 pixels. The first convolution layer (i.e., Conv1) filters the processed image with 32 convolution kernels of size 5 × 5 with a stride equal to 1, thereby generating 32 feature maps of size 52 × 52. The kernel size of the first pooling layer (i.e., Max Pooling1) equaled 2 × 2 with stride equal to 2. Upon completion of the pooling operation, 32 feature maps of size 26 × 26 were generated. The second convolution layer (i.e., Conv2) took the output of the first pooling layer as input and filtered it with 64 convolution kernels of size 3 × 3 with a stride equal to 1, thereby generating 64 feature maps of size 24 × 24. The kernel size of the second pooling layer (i.e., Max Pooling2) again equaled 2 × 2 with a stride equal to 2. Upon completion of the second pooling operation, 64 feature maps of size 12 × 12 were generated. Likewise, the third and fourth convolution layers (i.e., Conv3 and Conv4) used 32 and 64 convolution kernels of size 3 × 3 and 5 × 5, respectively, to perform convolution operation with a stride equal to 1, thereby generating 32 and 64 feature maps each of size 10 × 10 and 6 × 6, respectively. Since the output size after convolution was relatively small, the number of max pooling layers need not be increased. The number of convolution kernels in the fourth convolution layer was considered twice in the previous layer to facilitate improvement in the overall feature-extraction result. Wang et al. [34] has proved that the number of neural units in the fully connected layer has little impact on the classification results. Therefore, the first fully connected layer (i.e., FC1) connects all feature maps generated by the fourth convolutional layer to obtain a 384-dimensional feature vector; likewise, the second fully connected layer (i.e., FC2) comprises of a 192-dimensional feature vector. Few weights need to be learnt to avoid overfitting. The last layer (i.e., Softmax layer) employed the Softmax function to obtain output decision classes. The self-built dataset contained four output classes, whereas the VAIS dataset contained six classes. The number of output classes identified was found to be consistent with the number of ship classes contained within the ship dataset.

The activation function used in the convolution and fully connected layers was the ReLU function, which offers one-sided suppression and sparse characteristics. Additionally, it is more efficient compared to the sigmoid function and is capable of accelerating convergence speeds up to a certain extent. A dropout function has been added in FC1 to avoid overfitting, and the corresponding drop parameter was set to 0.5. Local response normalization is to be performed after completion of operation of the Conv1 and Conv2 layers of the network to achieve local suppression and enhance generalization ability. The response-normalized activity $b^i_{x,y}$ can be expressed as:

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} \left(a^i_{x,y}\right)^2 \right)^{\beta}, \quad (1)$$

where $a^i_{x,y}$ represents neuron activity computed by applying kernel i at position (x, y); N denotes the total number of kernels; and n denotes the number of adjacent nuclear maps located at the same position. Constants k, n, α, and β are considered hyper-parameters, values of which were set identical to those reported in a previous study [22]; i.e., $k = 2$, $n = 5$, $\alpha = 0.0001$, and $\beta = 0.75$.

The convolution-kernel size of the proposed improved CNN was considered reasonable, and the maximum number of convolution kernels equaled 64. The proposed CNN with parameters contains only seven layers—Conv1, Conv2, Conv3, Conv4, FC1, FC2, and Softmax. The spatial complexity (parameter amount) of the convolutional layer can be expressed as:

$$Space \sim O\left(\sum_{l=1}^{D} K_l^2 \cdot C_{l-1} \cdot C_l + \sum_{l=1}^{D} M^2 \cdot C_l \right), \quad (2)$$

where K denotes the convolution kernel size; C_l denotes the number of convolution kernels comprising layer l; C_{l-1} denotes the number of output channels in the $l-1$ layer; M denotes the output feature map side length, and D denotes the number of convolutional layers. In accordance with Equation (2), the total number of parameters equals approximately 1 million. However, the fine-AlexNet network (refer Section 3.4) contains eight layers with parameters, and the total number of parameters equals nearly 84 million. Likewise, the fine-VGG-16 and fine-ResNet-18 networks (refer Section 3.4) contain 16 and 18 layers with corresponding total number of parameters equal to roughly 16 million and 10 million, respectively. Thus, compared to the three above-mentioned fine-tuned networks, the proposed CNN comprises of shallow layers and fewer total parameters, thereby simultaneously ensuring attainment of high classification accuracy and reduced computational complexity. Results of experiments reported in Section 3 demonstrated that the proposed CNN achieved higher classification accuracy when operating on ship images.

The training and testing phases of the specific model have been described in Algorithms 1 and 2, respectively, as under.

Algorithm 1 Training phase

Requirement: A well-prepared training dataset
Step 1: Set model-parameter values.
Step 2: Execute preprocessing for training-set ship images.
Step 3: Perform feature extraction via forward propagation of CNN and use of the Softmax function to obtain predicted image classes.
Step 4: Calculate the error between the predicted and true classes, followed by weight and bias adjustment via back propagation to minimize the error.

Algorithm 2 Testing phase

Requirement: A prepared testing set.
Step 1: Set model parameters values.
Step 2: Preprocess ship images within the testing dataset.
Step 3: Perform feature extraction via CNN forward propagation, and call upon the optimum training model to test it.
Step 4: Obtain classification output for predicted classes (i.e., classification results), and evaluate classification results based on evaluation metrics.

3.1.3. Preprocessing

Ship images captured in actual scenarios often contain complex backgrounds. Additionally, differences in illumination and viewing angles influence image-feature extraction. If the ship images are directly fed to CNNs, the classification result may get adversely impacted. In this study, therefore, ship images were subjected to a series of preprocessing procedures that tend to weaken the influence of background noise as well as ensure sample randomness along with an increase in generalization ability and stability of the model.

The said preprocessing operations involve the use of the bicubic interpolation method to adjust the size of ship images to 64 pixels × 64 pixels. Subsequently, all training images were randomly cropped. This helped the elimination of the effect of illumination and viewing angles, thereby enhancing classification performance. However, center cropping is only performed on the set of test images. The image size after random and center cropping equaled 56 pixels × 56 pixels.

3.1.4. CNN Parameter Adjustment and Optimization

Despite the general ability of CNNs to self-learn and share weights whilst adopting sparse connections, there exist certain shortcomings, such as long training time, low accuracy, low generalization ability, and high overfitting tendency, which must be addressed. To this end, further adjustment and optimization of CNN parameters were performed in this study. The proposed optimized network was trained using the ship dataset to ensure better classification performance of the proposed CNN network. The following points concerning the proposed CNN must be noted.

Weight decay: Addition of a regularization term after the cross entropy loss function used in this paper served to reduce overfitting of the CNN model to some extent. The said regularization term contains a weight-decay coefficient, the value of which was set as 0.0005.

The learning rate affects the convergence speed and network-training performance. The learning rate was set as 0.001 in accordance with the stochastic gradient descent algorithm (SGD) [35], and the momentum parameter was set as 0.9.

The batch size was set to 32 with the maximum number of iterations within the self-built and VAIS datasets equal to 15,500 and 3700, respectively.

3.2. Feature Extraction Based on HOG and LBP

CNN automatically learns features through layer-by-layer propagation, but it may lose some important low-level information. Therefore, low-level features extracted by HOG and LBP are used to supplement contour, edge features, and spatial texture feature. The combination of three features can more accurately represent the features of ship images.

3.2.1. HOG

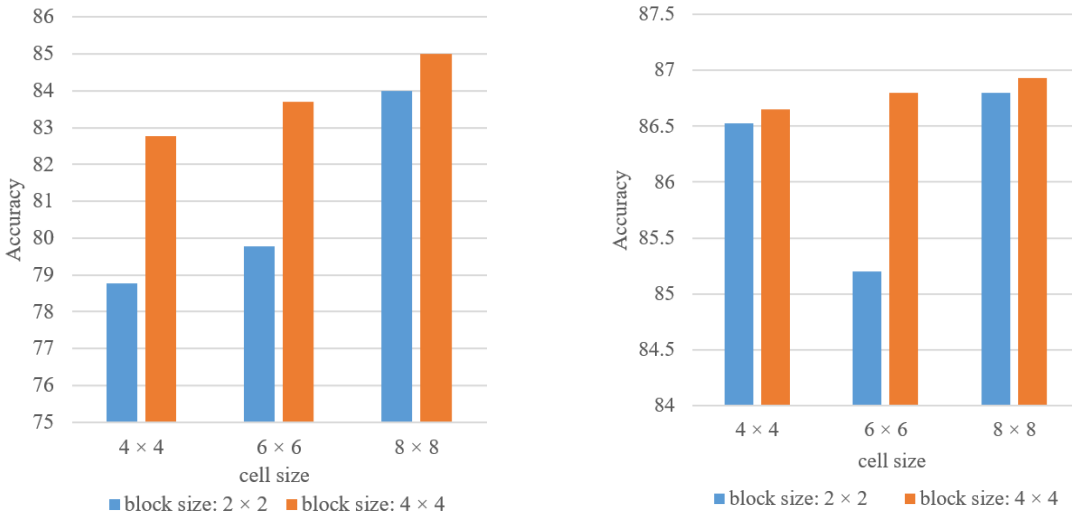
The HOG feature has been widely used in computer vision, such as pedestrian detection and vehicle classification. HOG features are formed by calculations using the statistics of the histogram of the gradient direction in the local area of the image, and it can maintain good invariance to both geometric and optical deformations of the image. The gradient mainly exists at the edge, so the

histogram of gradient direction can be used to extract the edge and contour features of ship images. Therefore, this paper supplemented the CNN feature with the HOG feature to extract more accurate edge features and global information of the ship images. The basic composition unit of the HOG detection window is a cell of $n \times n$ pixels, and then a block is composed of $m \times m$ cells, and finally a window is composed of block. The extraction process of HOG feature is shown in Algorithm 3.

Algorithm 3 The extraction process of HOG feature

Input: A well-prepared training dataset.
Step 1: Convert the training dataset images to grayscale images.
Step 2: Normalized images by the Gamma correction method.
Step 3: Calculate the gradient of each pixel of the image.
Step 4: Divide the image into cells, and calculate the gradient histogram of each cell.
Step 5: Every few cells form a block, and the normalized gradient histogram is contained within the block.
Step 6: The HOG feature descriptors of all blocks are concatenated to obtain the HOG feature of the image.
Output: HOG feature.

The parameters directly affect the final classification accuracy. In this study, we tuned parameters based on the available training data, and reported the experiment results in Figure 6. Here, optimum parameters were set as follows. The block size was first set to 4 × 4, followed by setting of the cell size 8 pixels × 8 pixels. Next, the direction of each gradient was divided into nine intervals, and lastly, the block histogram normalization method was used to perform L2-Hys-norm normalization. The HOG features obtained by using two datasets (see Section 4.1) are shown in Figures 7 and 8. As can be seen, the edge features of the ship images were well extracted.



(**a**) (**b**)

Figure 6. Classification accuracy (%) with varying parameters of histogram of oriented gradients (HOG) for two experimental data: (**a**) The self-built dataset and (**b**) the visible and infrared spectrums (VAIS) dataset.

(**a**) (**b**) (**c**)

Figure 7. Illustration of HOG feature using the self-built dataset: (**a**) Original image; (**b**) grayscale image; and (**c**) HOG feature.

(**a**) (**b**) (**c**)

Figure 8. Illustration of HOG feature using the VAIS dataset: (**a**) Original image; (**b**) grayscale image; and (**c**) HOG feature.

3.2.2. LBP

The LBP feature is used as a local texture feature descriptor to extract spatial texture features of ship images [36], which has the advantages of rotation invariance and gray invariance, and has been widely used in texture classification [37] and ship classification. Given a pixel, its grayscale value is g_c. Its m neighborhood pixels are on a circle of radius equal to r equidistant from a given pixel. The LBP value of g_c is defined as:

$$LBP_{m,r}(g_c) = \sum_{j=0}^{m-1} s(g_j - g_c)2^j, \tag{3}$$

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}, \tag{4}$$

where g_j is the gray value of the neighbors, and $g_j - g_c$ represents the difference between the center pixel and each neighbor. m is the total number of involved neighbors. LBP records the difference between the pixel at the center and the pixel in the neighborhood. When the illumination transformation causes the same increase or decrease in pixel gray value, LBP changes slightly, so it is insensitive to the illumination change. In this study, in order to describe ship image features more accurately, LBP features were combined with CNN features to supplement spatial texture features. The extraction process of LBP feature is shown in Algorithm 4.

Algorithm 4 The extraction process of LBP feature

Input: A well-prepared training dataset.
Step 1: Convert the training set images to grayscale images.
Step 2: Tune parameters (m, r) and select the optimal parameters.
Step 3: Calculate the LBP feature.
Output: LBP feature.

In this study, the classification performance was best when $(m,r) = (8, 1)$ for the two datasets. The LBP features obtained by using two datasets are shown in Figures 9 and 10. As can be seen, the spatial texture features of the ship images are well extracted.

(**a**) (**b**) (**c**)

Figure 9. Illustration of the local binary pattern (LBP) feature using the self-built dataset: (**a**) Original image; (**b**) grayscale image; and (**c**) LBP feature.

(**a**) (**b**) (**c**)

Figure 10. Illustration of the LBP feature using the VAIS dataset: (**a**) Original image; (**b**) grayscale image; and (**c**) LBP feature.

3.3. Multi-Feature Fusion

In order to obtain more comprehensive ship image feature representation, the high-level features extracted by the improved CNN and handcrafted features of HOG and LBP were considered to be fused. The improved CNN can extract the structure and semantic information of ship images. Based on the low-level features obtained from the first convolution layer, CNN learns features through layer-by-layer propagation, which may lose some important low-level information. Therefore, handcrafted features such as HOG and LBP were fused to supplement edge features and spatial texture features to obtain more comprehensive feature representation.

After feature extraction according to Algorithms 1, 3 and 4, in the last fully connected layer these three types of features were concatenated into a composite vector with the weight of 1:1:1 and fed into the Softmax layer for final classification. The procedure of multi-feature fusion strategy is shown in Figure 11.

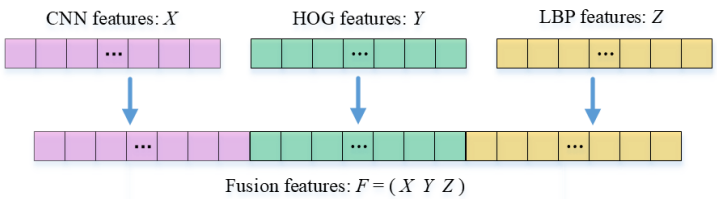


Figure 11. The procedure of the multi-feature fusion strategy.

3.4. Fine-Tuning Deep CNN

Image sizes within the self-built dataset equaled 64 pixels × 64 pixels with the size of the image provided as an input to the first convolution layer of deep CNNs, such as AlexNet, VGG-16, and ResNet-18, exceeding 64 pixels × 64 pixels. To compare the proposed method against other CNNs and validate its effectiveness, the three CNNs were fine-tuned to fit the image size of 64 pixels × 64 pixels to facilitate ship classification.

3.4.1. Fine-AlexNet

To facilitate effective extraction of image features and considering the ship-image size, the size of the convolution kernel was fine-tuned to 3 pixels × 3 pixels, and the number of filters contained within convolution layers Conv1–Conv5 equaled 32, 64, 64, 128, and 256, respectively. The size of the pooling kernel equaled 2 × 2 and the number output layers equaled 4 or 6 corresponding to the number of ship classes contained within the two self-built and VAIS datasets. The ReLU nonlinear activation function was employed after each convolution and fully connected layer. The specific network architecture has been described in Table 3.

Table 3. Structure of fine-AlexNet network.

Layer	Kernel Size	Filter Number	Stride
Conv1	3 × 3	32	1
Max Pooling1	2 × 2	-	2
Conv2	3 × 3	64	1
Max Pooling2	2 × 2	-	2
Conv3	3 × 3	64	1
Conv4	3 × 3	128	1
Conv5	3 × 3	256	1
Max Pooling3	2 × 2	-	2
FC1	4096	-	-
FC2	4096	-	-
FC3-Softmax	4(6)	-	-

3.4.2. Fine-ResNet-18

The convolution kernel size of the first convolutional layer (i.e., Conv1) equaled 3 × 3. Since the output size of Conv5_x equaled 8 × 8 × 512, the kernel size of the global average pooling layer was fine-tuned to 8 × 8, and the number of output layers within the fully connected layer (i.e., FC1) equaled 4 or 6.

3.4.3. Fine-VGG-16

For an input image size of 64 × 64 layers, the output size of Max Pooling5 equaled 2 × 2 × 512. The output of the first and second fully connected layers (i.e., FC1 and FC2) was fine-tuned to obtain a 512-dimensional vector, and the output of FC3 layer equaled 4 or 6 depending on the dataset considered.

4. Experimental Results and Analysis

4.1. Experimental Datasets

The first dataset used in this study corresponds to the self-built dataset, which contained 3000 RGB original images of different sizes. The said images were partly collected from ship-image databases available on websites of the China Shipping Service and Baidu. The remaining images were collected from the Yangtze River channel between the Zhonghua road and Wuhan wharfs, which define an inland waterway with the largest cargo volume in the world. The said ship images were collected during daytime between 9 am and 5 pm to ensure uniform illumination conditions during image collection. The acquisition area is depicted in Figure 12. The circulation of ships within the said region was large, and there not only existed bulk carriers but also passenger ships, and container ships. This facilitated collection of different types of ship images under different environments. Each image within the dataset was manually labeled as belonging to one of four classes—bulk carriers, container ships, passenger ships, and sailing boats. A few image samples are depicted in Figure 13. In order to improve the recognition ability, each image was rotated counterclockwise 15 degrees and mirrored image respectively to expand the dataset. Therefore, the self-built dataset contained 9000 ship images after augmentation. There were 7201 training images, accounting for roughly 80% of the

expanded image dataset, that were randomly selected from the said four classes, and the remaining 1799 images were considered test images. To ensure data balance, the ratio of images comprising the training dataset to those comprising the test dataset for each class approximately equaled 4:1. The number of training and testing samples are listed in Table 4. To further verify the correctness of the proposed method, we also used the original images in test images as original images test dataset. The original images test dataset is shown in Table 5.

Figure 12. Scene diagram of the image-acquisition area.

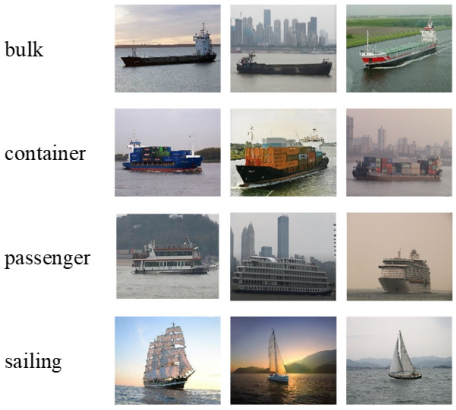


Figure 13. Sample images from each class of self-built dataset.

Table 4. Number of training and test samples comprising the self-built dataset.

No.	Class	Train	Test
1	Bulk	1385	346
2	Container	2381	595
3	Passenger	1632	408
4	Sailing	1803	450
	Total	7201	1799

Table 5. Original images test dataset.

No.	Class	Test
1	Bulk	120
2	Container	98
3	Passenger	93
4	Sailing	96
	Total	407

The second dataset (i.e., VAIS dataset), which is the world's first publicly available dataset comprised of 2865 images (1623 visible and 1242 infrared), including 1088 corresponding pairs. These images were captured using a multimodal stereo camera rig. The dataset included six coarse-grained classes (or 15 fine-grained classes)—merchant ships (26 cargo ships and nine barges), medium passenger ships (11 ferries and four tour boats), sailing ships (41 sails up and 24 sails down), small boats (28 speedboats, six jet-skis, 25 small pleasure boats, and 13 large pleasure boats), 19 tugboats, and medium "other" ships (eight fishing and 14 medium other), as depicted in Figure 14. For each image within the dataset, bounding boxes were manually labeled. The area of visible bounding boxes occupied 644–4,478,952 pixels with corresponding mean and median values of 181,319 pixels and 9983 pixels, respectively. The dataset was divided into "official" training and testing parts. Since the authors were interested in generalization, all images were greedily assigned from each named ship to either partition. This resulted in the creation of 539 image pairs and 334 singletons for training and 549 image pairs and 358 singletons for testing. In this study, only the visible ship imagery was chosen. Table 6 lists the number of training and test samples. Each image was resized to 64 pixels × 64 pixels via bicubic interpolation, which was performed in a manner similar to that described in a previous study [27].

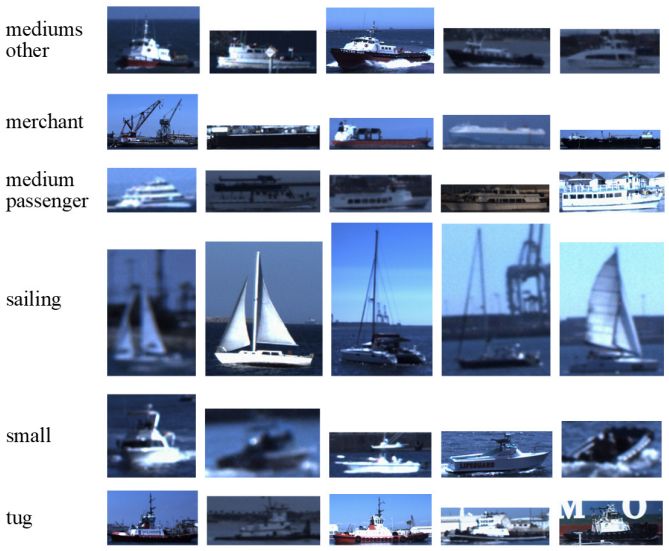


Figure 14. Five visible samples from each of the main classes of the VAIS dataset.

4.2. Simulation Environment

To validate the performance of the proposed method, several experiments were performed on the two ship datasets described in the previous section to facilitate training and evaluation of the proposed improved CNN. The experimental environment comprised of an Inter(R) Core(TM) i9-7980XE@2.6GHz processor along with an NVIDIA TITAN Xp Pascal graphics card. All experiments were performed using Python 3.5, MATLAB, and the TensorFlow framework.

Table 6. Number of training and test samples using the VAIS dataset.

No.	Class	Train	Test
1	Medium-other	99	86
2	Merchant	103	71
3	Medium-passenger	78	62
4	Sailing	214	198
5	Small	342	313
6	Tug	37	20
	Total	873	750

4.3. Experimental Methods and Evaluation Metrics

The experiment was divided into the training and testing phases.

Training phase: Ship images within training sets of both datasets were trained to obtain an optimum model.

Testing phase: Testing was performed in two cases. In the first case, all samples in the testing set were tested to obtain the average classification accuracy, whereas the second case involved testing of different classes of ship images to obtain classification accuracy per class of ship images.

In this study, the classification accuracy, F1-score, confusion matrix, and the average time consumption used for feature extraction per image were considered evaluation metrics concerning image-classification results.

Classification accuracy corresponds to the ratio of the number of correctly classified samples to the total number of samples. The F1-score was considered a comprehensive measure of classification performance of the proposed method. By definition, the F1-score corresponds to the weighted average of the precision and recall, and its value lies in the range of 0 and 1. Mathematically, the F1-score can be expressed as:

$$F1 = \frac{2 \times P \times R}{P + R}. \tag{5}$$

Precision concerns the ratio of the number of true positives to the number of predicted positive samples. Recall can be defined as the ratio of the number of true positives to all positive samples. For the two-class problem, precision and recall can be evaluated as:

$$P = \frac{TP}{TP + FP'}, \tag{6}$$

and

$$R = \frac{TP}{TP + FN'}, \tag{7}$$

respectively, where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively.

The confusion matrix, as the name suggests, represents the confusion caused by the classifier when dealing with multiclass problems. Herein, the rows and columns represent the prediction and true classes, respectively, whereas diagonal elements denote the correct quantity of each ship class.

4.4. Classification Results and Analysis

To validate its classification performance, the proposed method was compared against other state-of-the-art techniques under identical experimental conditions. All of the experiments have been done several times, and we listed the average results of the experiments. The comparison experiments involved many parameters. They were set as follows. Support vector machine (SVM) toolbox selected sklearn. svm. For HOG + SVM, the kernel was set to 'linear'. For LBP + SVM, the kernel was set to 'rbf'. Gamma parameters and penalty coefficients were obtained by grid optimization. The value of the penalty coefficient C was set to 100, and γ was set to 1.

4.4.1. Comparison of Classification Accuracy

Tables 7 and 8 list the classification accuracy and number of misclassification samples for the different methods considered in this study for the self-built dataset and VAIS dataset. As can be observed, the proposed method achieved superior classification performance compared to all other existing methods. When applied to the self-built dataset, the average accuracy of the proposed method exceeded those of the fine-AlexNet and a previously proposed CNN approach [38] by nearly 4.28% and 6.56%, respectively. Additionally, the number of misclassification samples corresponding to the proposed method was the least compared to other existing methods. When applied to the VAIS dataset, the accuracy exceeded that of the multiple feature learning (MFL) (feature-level) + SVM [36], and

ME-CNN [32] methods by 8.27% and 6.27%, respectively. Therefore, the improved method had obvious advantages, as the improved method fully considers the specific features of ship images and combines the high-level and low-level features that are conducive to ship classification. The improved CNN had good image description ability. In addition, HOG and LBP features could describe the edge and spatial texture features of ship images, respectively. Thus, the fusion of these three types of features could yield more distinguishing features and improve the ship classification accuracy. That said, no particular method demonstrated attainment of exceptionally high classification accuracy when applied to the VAIS dataset. This is because ship images contained within the VAIS dataset were of very low resolution and exhibited glare. Figure 14 illustrates that the quality of images was not good, and this played an important role in regards to the performance of the classification method. This further strengthens the argument that VAIS was a very challenging dataset to work with.

Table 7. Classification accuracy and the number of misclassification samples associated with different methods when applied to the self-built dataset.

Method	Accuracy (%)	Number of Misclassification
HOG + SVM	84.99	270
LBP + SVM	77.10	412
CNN [29]	87.77	220
Fine-AlexNet	93.22	122
Fine-ResNet-18	91.50	153
Fine-VGG-16	92.61	133
CNN [38]	90.94	163
HOG + CNN	92.88	128
LBP + CNN	92.94	127
Improved CNN	96.50	63
Proposed Method	97.50	45

Table 8. Classification accuracy and number of misclassifications performed by different methods when applied to VAIS dataset.

Methods	Accuracy (%)	Number of Misclassification
HOG + SVM	86.93	98
LBP + SVM	78.80	159
MFL (feature-level) + SVM [36]	85.33	110
CNN [19]	85.75	107
CNN [27]	81.90	136
CNN [38]	86.00	105
ME-CNN [32]	87.33	95
Fine-VGG-16	86.40	102
Fine-ResNet-18	90.40	72
HOG + CNN	90.40	72
LBP + CNN	91.07	67
Improved CNN	92.13	59
Proposed Method	93.60	48

In addition, to further verify the discrimination ability of the improved method, we compared the classification accuracy obtained per ship class of different methods. The experiment results are listed in Tables 9 and 10. As observed, compared to existing approaches, the proposed method demonstrated better classification performance compared to all other methods. For the self-built dataset, the proposed method had the highest classification accuracy for sailing, LBP + CNN and improved CNN methods also obtained high accuracy for sailing. Due to sailing's distinct shape and size characteristics, it was easy to classify. For the VAIS dataset, the proposed method had the highest classification accuracy for each type, although the number of tug-ship samples in VAIS was quite small. The classification

accuracies of the HOG + CNN and LBP + CNN methods for Tug were 95%, which was 60% higher than the classification accuracy of the traditional LBP + SVM method for Tug.

Table 9. Class-specific accuracy (%) of different methods for the self-built dataset.

Method	Class			
	Bulk	Container	Passenger	Sailing
HOG + SVM	83.53	82.52	81.62	93.33
LBP + SVM	69.08	79.50	70.34	86.22
HOG + CNN	93.64	91.43	92.16	94.89
LBP + CNN	89.88	93.28	91.91	95.78
Improved CNN	95.95	95.97	96.56	97.56
Proposed Method	96.82	96.30	97.79	99.33

Table 10. Class-specific accuracy (%) of different methods for the VAIS dataset.

Method	Class					
	Medium-Other	Merchant	Medium-Passenger	Sailing	Small	Tug
HOG + SVM	70.93	85.92	75.81	93.94	92.33	75.00
LBP + SVM	55.81	76.06	67.74	81.82	88.82	35.00
HOG + CNN	76.74	87.32	83.87	95.96	92.33	95.00
LBP + CNN	77.91	87.32	87.10	94.95	93.61	95.00
Improved CNN	80.23	87.32	88.71	94.95	94.89	100.00
Proposed Method	82.56	88.73	90.32	97.47	95.53	100.00

4.4.2. Comparison of the F1-Score and Confusion Matrix

To further validate the performance of the improved method, values concerning the F1-score obtained using these different CNN methods are listed in Tables 11 and 12. As observed, compared to other techniques, the proposed method demonstrated attainment of the highest average F1-score along with high classification accuracy for each class. For the self-build dataset, the CNN method presented in extant studies [29] demonstrates poor classification accuracy for bulk ships owing to the similarity between some bulk ships and container ships, which makes it difficult to distinguish them accurately. The fine-AlexNet, CNN [38], and fine-VGG-16 methods demonstrated the better classification performance for sailing ships. For the VAIS dataset, the average values of F1-scores corresponding to the proposed method were the highest. While the other methods demonstrated better classification performance with regard to sailing ships, their performance with regard to the classification of the medium-other and medium-passenger ship types was rather poor. Since the appearance of the medium-other and medium-passenger ship types was more complex and that there existed considerable difference between sailing and other ship classes, it was much easier to distinguish sailing ships. Compared with other methods, the proposed method had an improved F1-score for each category. On the one hand, this indicates that fine-VGG-16 and fine-ResNet-18 caused overfitting owing to the small ship dataset, and they were better suited to deal with complex classification task. On the other hand, it demonstrated that the proposed method could better extract ship features and that it offered superior ship classification performance after combining CNN high-level features with handcrafted features of HOG and LBP.

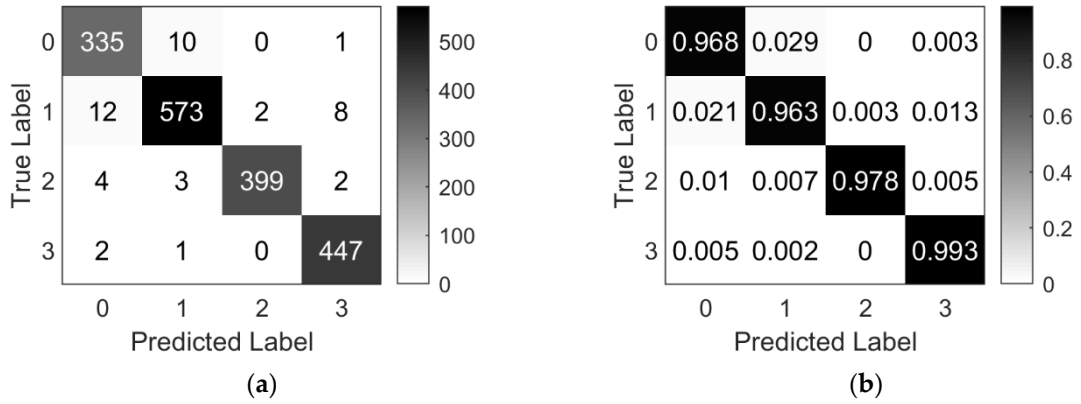
Figure 15 depicts the confusion matrix and confusion matrix normalization corresponding to the proposed method for the self-built dataset. As can be seen, diagonal elements of the confusion matrix and its normalized form denoted the correct quantity of each ship class and classification accuracy achieved per class, respectively. It was easily found that major confusion occurred between class 0 (i.e., bulk ships) and class 1 (i.e., container), or between class 1 and class 3 (i.e., sailing). It was observed that the length of bulk ships was similar to that of container carriers, some images of bulk were similar to these of the container.

Table 11. F1-score obtained using different CNN methods when applied to the self-built dataset.

Class	Method						
	CNN [29]	Fine-AlexNet	Fine-ResNet-18	Fine-VGG-16	CNN [38]	Improved CNN	Proposed Method
Bulk	0.8493	0.9124	0.9032	0.9284	0.8693	0.9513	0.9585
Container	0.8637	0.9260	0.9217	0.9097	0.8938	0.9654	0.9695
Passenger	0.8630	0.9377	0.9167	0.9042	0.9160	0.9621	0.9864
Sailing	0.9338	0.9503	0.9132	0.9667	0.9537	0.9777	0.9846
Avg. total	0.8775	0.9316	0.9137	0.9272	0.9082	0.9641	0.9748

Table 12. F1-score obtained using different CNN methods when applied to the VAIS dataset.

Class	Method				
	ME-CNN [32]	Fine-VGG-16	Fine-ResNet-18	Improved CNN	Proposed Method
Medium-other	0.6364	0.7211	0.7821	0.8264	0.8402
Merchant	0.8695	0.7862	0.8429	0.9117	0.9197
Medium-passenger	0.7155	0.7123	0.8160	0.8943	0.9180
Sailing	0.9824	0.9440	0.9494	0.9495	0.9650
Small	0.8830	0.8966	0.9372	0.9369	0.9492
Tug	0.8333	0.9048	0.8837	0.9091	0.9524
Avg. total	0.8200	0.8275	0.8685	0.9047	0.9241



(**a**) (**b**)

Figure 15. Confusion matrix and its normalization of the proposed method using the self-built dataset. Notes: Numbers 0, 1, 2, and 3 correspond to the bulk, container, passenger, and sailing ship types; (**a**) confusion matrix; and (**b**) confusion matrix normalization.

Figure 16 depicts the confusion matrix and confusion matrix normalization corresponding to the proposed method for the VAIS dataset. As observed, confusion primarily occurred within class 0 (i.e., medium-other) and class 4 (i.e., small), or class 1 (i.e., merchant) and class 3 (i.e., sailing), or between classes 2 (i.e., medium-passenger) and 4. According to the composition of the VAIS dataset introduced in Section 4.1 and Figure 14, some small ships and medium-other ships had relatively high similarity. Similarly, there were similarities between the mast of sails down and merchant hoisting equipment. The other was that some small and medium-passengers existed a similarity. Hence, the aforementioned classes were prone to confusion.

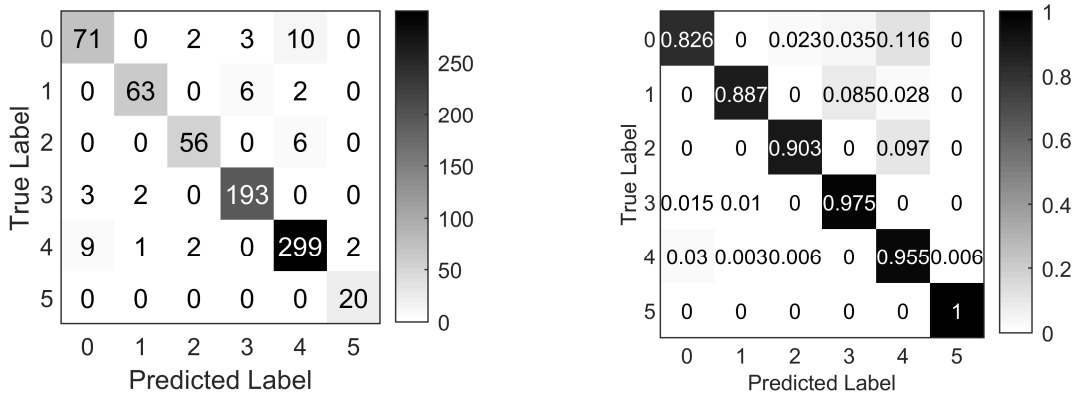


Figure 16. Confusion matrix and its normalization of the proposed method using the VAIS dataset. Note that numbers 0–5 denote medium-other, merchant, medium-passenger, sailing, small, and tug ship types, respectively; (**a**) confusion matrix; and (**b**) confusion matrix normalization.

4.4.3. Comparison of the Average Time Consumption of Feature Extraction per Image

Tables 13 and 14 list the average time consumption used for feature extraction per image of the different methods considered in this study for the self-built dataset and VAIS dataset. As can be observed, the feature extraction speed of improved CNN was faster than the other methods due to its shallow layers and few parameters. Due to the combination of HOG and LBP features, the average time consumption of feature extraction per image increased a little. However, the HOG and LBP features were fused with the high-level features extracted by the improved CNN network to further supplement the contour and spatial texture feature of ship images. The combination of three features could more accurately represent the features of ship images, and obtained a more robust ship classification model. In addition, experimental results also show that the classification accuracy of the proposed method was the best, and the time consumption of 14.074 ms and 15.329 ms was also relatively fast. Meanwhile, the CNN method used mini-batch technology, so the feature extraction speed was faster than traditional methods.

Table 13. The average time consumption of feature extraction per image for the self-built dataset.

Method	Average Time Consumption of Feature Extraction (ms)
HOG + SVM	4.061
LBP + SVM	1.460
CNN [29]	0.056
Fine-AlexNet	0.682
Fine-ResNet-18	0.829
Fine-VGG-16	0.772
HOG + CNN	7.504
LBP + CNN	7.410
Improved CNN	0.047
Proposed Method	14.074

Table 14. The average time consumption of the feature extraction per image for the VAIS dataset.

Methods	Average Time Consumption of Feature Extraction (ms)
HOG + SVM	4.265
LBP + SVM	1.409
CNN [38]	0.241
Fine-VGG-16	0.756
Fine-ResNet-18	0.913
HOG + CNN	8.784
LBP + CNN	8.548
Improved CNN	0.104
Proposed Method	15.329

4.4.4. Performance in the Original Images Test Dataset

We also did an experiment with the original images test dataset, and the results are shown in Table 15.

Table 15. Experiment with the original images test dataset.

Method	Accuracy (%)	Classified/Real	Average Time Consumption of the Feature Extraction Per Image (ms)
Proposed Method	98.33	589/599	16.341

The results show that the proposed method could also achieve good classification performance with the original images test dataset.

5. Conclusions

In this paper, the authors proposed the use of a multi-feature fusion with a CNN method for ship classification. To facilitate the training and performance evaluation of the proposed multi-feature fusion CNN framework, the authors used the VAIS dataset to test the proposed method. Simultaneously, they established their own ship dataset comprising a combination of ship images captured along the Yangtze River channel and those obtained from ship-image databases of the China Shipping Service and Baidu websites. Compared to fine-VGG-16, fine-ResNet-18, and other deep CNNs, the proposed improved CNN was characterized by shallow layers and relatively few parameters, thereby reducing its computational complexity. HOG features are used to extract edge features and LBP features are used to extract texture features. These two handcrafted features were adopted to compensate for the shortcomings of CNN (that is, partial local features are lost) to more accurately describe ship images. In addition, the advantages of these three types of features were considered, and they were fused to obtain a robust ship classification model. Results of experiments performed in this study demonstrated that the average classification accuracy of the proposed method was equal to 97.50% and 93.60%, respectively, when applied to the limited number of self-built and VAIS datasets, respectively. Additionally, a consideration of evaluation metrics, such as the F1-score, classification accuracy of each class, confusion matrix, and average time consumption of the feature extraction per image, revealed that classification performance of the proposed method was superior to other state-of-the-art methods. This also implies that the proposed method performed better at extracting features from ship images. Compared with other deep networks, the improved CNN also had better classification ability. However, results obtained concerning the VAIS dataset demonstrated that there still existed room for improvement in classification performance of the proposed method. As a future endeavor, the authors intend to enhance further the classification ability of the proposed method from the viewpoints of the extended dataset and transfer learning.

Author Contributions: Y.R. conceived the manuscript and conducted the whole experiments. J.Y. supervised the experiments and helped discuss the proposed the method. Q.Z. and Z.G. contributed to the organization of the paper and gave helpful suggestion on this research. All authors read and approved the final manuscript.

Funding: This work was supported by the National Science Foundation of China (Grant No. 51879211), and the Scientific Research Project of the Hunan Provincial Education Department (Grant No. 18C0900).

Acknowledgments: The authors would like to thank the anonymous reviewers for their very competent comments and helpful suggestions.

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

References

1. Huang, S.Z.; Xu, H.S.; Xia, X.Z. Active deep belief networks for ship recognition based on BvSB. *Optik* **2016**, *127*, 11688–11697. [CrossRef]
2. Sun, X.; Wang, G.; Fan, Y.; Mu, D.; Qiu, B. An Automatic Navigation System for Unmanned Surface Vehicles in Realistic Sea Environments. *Appl. Sci.* **2018**, *8*, 193. [CrossRef]
3. Xu, F.; Wang, H.P.; Song, Q.; Ao, W.; Shi, Y.Q.; Qian, Y.T. Intelligent ship recognition from synthetic aperture radar images. In Proceedings of the 38th IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018; pp. 4387–4390.
4. Lu, C.Y.; Zou, H.X.; Sun, H.; Zhou, S.L. Combing rough set and RBF neural network for large-scale ship recognition in optical satellite images. In Proceedings of the 35th International Symposium on Remote Sensing of Environment (ISRSE35), Beijing, China, 22–26 April 2013; pp. 1–7.
5. Guo, W.Y.; Xia, X.Z.; Wang, X.F. Variational approximate inferential probability generative model for ship recognition using remote sensing data. *Optik* **2015**, *126*, 4004–4013. [CrossRef]
6. Yue, Q.; Ma, C.W. Hyperspectral data classification based on flexible momentum deep convolution neural network. *Multimed. Tools Appl.* **2018**, *77*, 4417–4429. [CrossRef]
7. Park, S.; Cho, C.J.; Ku, B.; Lee, S.; Ko, H. Simulation and ship detection using surface radial current observing compact HF radar. *IEEE J. Ocean. Eng.* **2017**, *42*, 544–555. [CrossRef]
8. Harguess, J.; Rainey, K. Are face recognition methods useful for classifying ships? In Proceedings of the 2011 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 11–13 October 2011; pp. 11–13. [CrossRef]
9. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 886–893. [CrossRef]
10. Lowe, D.G. Distinctive image features from scale-invariant key points. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
11. Ahonen, T.; Hadid, A.; Pietikäinen, M. Face recognition with local binary patterns. In Proceedings of the European Conference on Computer Vision, Prague, Czech Republic, 11–14 May 2004; pp. 469–481. [CrossRef]
12. Rainey, K.; Parameswaran, S.; Harguess, J.; Stastny, J. Vessel classification in overhead satellite imagery using learned dictionaries. In Proceedings of the SPIE 8499, Applications of Digital Image Processing XXXV, 84992F, San Diego, CA, USA, 15 October 2012; pp. 1–12. [CrossRef]
13. Arguedas, V.F. Texture-based vessel classifier for electro-optical satellite imagery. In Proceedings of the IEEE International Conference on Image Processing, Quebec City, QC, Canada, 27–30 September 2015; pp. 3866–3870. [CrossRef]
14. Parameswaran, S.; Rainey, K. Vessel classification in overhead satellite imagery using weighted "bag of visual words. In Proceedings of the SPIE 9476, Automatic Target Recognition XXV, 947609, Baltimore, MD, USA, 22 May 2015. [CrossRef]
15. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
16. Chen, Y.F.; Xie, H.; Shin, H. Multi-layer fusion techniques using a CNN for multispectral pedestrian detection. *IET Comput. Vis.* **2018**, *12*, 1179–1187. [CrossRef]
17. Liu, Y.; Yin, B.C.; Yu, J.; Wang, Z.F. Image classification based on convolutional neural networks with cross-level strategy. *Multimed. Tools Appl.* **2017**, *76*, 11065–11079. [CrossRef]

18. Natarajan, S.; Annamraju, A.K.; Baradkar, C.S. Traffic sign recognition using weighted multi-convolutional neural network. *IET Intell. Transp. Syst.* **2018**, *12*, 1396–1405. [CrossRef]
19. Rainey, K.; Reeder, J.D.; Corelli, A.G. Convolution neural networks for ship type recognition. In Proceedings of the SPIE 9844, Automatic Target Recognition XXVI, 984409, Baltimore, MD, USA, 12 May 2016; pp. 1–11. [CrossRef]
20. Bentes, C.; Velotto, D.; Tings, B. Ship classification in TerraSAR-X images with convolutional neural networks. *IEEE J. Ocean. Eng.* **2018**, *43*, 258–266. [CrossRef]
21. Khellal, A.; Ma, H.B.; Fei, Q. Convolutional neural network based on extreme learning machine for maritime ships recognition in infrared images. *Sensors* **2018**, *18*, 1490. [CrossRef] [PubMed]
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the Neural Information Processing Systems Conference, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
23. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
24. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. [CrossRef]
25. Shi, Q.Q.; Li, W.; Zhang, F.; Hu, W.; Sun, X.; Gao, L.R. Deep CNN with multi-scale rotation invariance features for ship classification. *IEEE Access* **2018**, *6*, 38656–38668. [CrossRef]
26. Rainey, K.; Stastny, J. Object recognition in ocean imagery using feature selection and compressive sensing. In Proceedings of the 2011 IEEE Applied Imagery Pattern Recognition Workshops, Washington, DC, USA, 11–13 October 2011; pp. 1–6. [CrossRef]
27. Zhang, M.M.; Choi, J.; Daniilidis, K.; Wolf, M.T.; Kanan, C. VAIS: A dataset for recognizing maritime imagery in the visible and infrared spectrums. In Proceedings of the 2015 IEEE Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 10–16. [CrossRef]
28. Liu, Y.; Cui, H.Y.; Li, G.Q. A novel method for ship detection and classification on remote sensing images. In Proceedings of the Artificial Neural Networks and Machine Learning, Alghero, Italy, 11–14 September 2017; pp. 556–564. [CrossRef]
29. Zhao, L.; Wang, X.F.; Yuan, Y.T. Research on ship recognition method based on deep convolutional neural network. *Ship Sci. Technol.* **2016**, *38*, 119–123. (In Chinese)
30. Cao, X.F.; Gao, S.; Chen, L.C.; Wang, Y. Ship recognition method combined with image segmentation and deep learning feature extraction in video surveillance. *Multimed Tools Appl.* **2018**, 1–16. [CrossRef]
31. Zhang, E.H.; Wang, K.L.; Lin, G.F. Classification of Marine Vessels with Multi-Feature Structure Fusion. *Appl. Sci.* **2019**, *9*, 2153. [CrossRef]
32. Shi, Q.Q.; Li, W.; Tao, R.; Sun, X.; Gao, L.R. Ship classification based on multifeature ensemble with convolutional neural network. *Remote Sens.* **2019**, *11*, 419. [CrossRef]
33. Zhuo, L.; Zhu, Z.Q.; Li, J.F.; Jiang, L.Y.; Zhang, H.; Zhang, J. Feature extraction using lightweight convolutional network for vehicle classification. *J. Electron. Imaging* **2018**, *27*, 051222. [CrossRef]
34. Wang, Y.Y.; Wang, C.; Zhang, H. Ship Classification in High-Resolution SAR Images Using Deep Learning of Small Datasets. *Sensors* **2018**, *18*, 2929. [CrossRef]
35. Bottou, L. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*; Grégoire, M., Geneviève, B.O., Müller, K.R., Eds.; Springer: New York, NY, USA, 2012; Volume 7770, pp. 421–436. [CrossRef]
36. Huang, L.H.; Li, W.; Chen, C.; Zhang, F.; Lang, H.T. Multiple features learning for ship classification in optical imagery. *Multimed. Tools Appl.* **2018**, *77*, 13363–13389. [CrossRef]
37. Porebski, A.; Vandenbroucke, N.; Hamad, D. LBP histogram selection for supervised color texture classification. In Proceedings of the 2013 IEEE International Conference on Image Processing, Melbourne, Australia, 15–18 September 2013; pp. 3239–3243. [CrossRef]
38. Ding, J.; Chen, B.; Liu, H.W.; Huang, M.Y. Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 364–368. [CrossRef]