electronics

MDPI

Article

Ship Classification Based on Attention Mechanism and Multi-Scale Convolutional Neural Network for Visible and Infrared Images

Yongmei Ren [1,2], Jie Yang [1,*], Zhiqiang Guo [1], Qingnian Zhang [3] and Hui Cao [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.); caohui@whut.edu.cn (H.C.)

[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: 7 November 2020; Accepted: 26 November 2020; Published: 30 November 2020

check for updates

Abstract: Visible image quality is very susceptible to changes in illumination, and there are limitations in ship classification using images acquired by a single sensor. This study proposes a ship classification method based on an attention mechanism and multi-scale convolutional neural network (MSCNN) for visible and infrared images. First, the features of visible and infrared images are extracted by a two-stream symmetric multi-scale convolutional neural network module, and then concatenated to make full use of the complementary features present in multi-modal images. After that, the attention mechanism is applied to the concatenated fusion features to emphasize local details areas in the feature map, aiming to further improve feature representation capability of the model. Lastly, attention weights and the original concatenated fusion features are added element by element and fed into fully connected layers and Softmax output layer for final classification output. Effectiveness of the proposed method is verified on a visible and infrared spectra (VAIS) dataset, which shows 93.81% accuracy in classification results. Compared with other state-of-the-art methods, the proposed method could extract features more effectively and has better overall classification performance.

Keywords: ship classification; feature fusion; attention mechanism; convolutional neural network; infrared image; visible image

1. Introduction

Ship classification plays an important role in military and civilian fields, such as maritime traffic, fishing vessel monitoring, maritime search and rescue, etc. [1,2]. However, in real life, ship classification results are very susceptible to background settings and recognition of intra-class differences among various types of ship has proven difficult. Therefore, ship classification has become one of the research hotspots in pattern recognition.

The main types of ship image are synthetic aperture radar (SAR) images, visible images and infrared images. After the launch of SEASAT in the 1970s, SAR began to be used in marine environmental research. SAR images are immune to light and weather conditions, but they have low resolution and are susceptible to electromagnetic interference as they are radar signals. Visible images, on the other hand, have high resolution and possess detailed texture, but they are easily affected by light conditions. When illumination is insufficient, the acquired image details drop significantly. Infrared images are not affected by light conditions either. Although the resolution is not very high, it has a clear target contour. Moreover, it has practical advantage as an infrared sensor can produce stable imaging. Therefore, combining visible and infrared images can improve the practicability of ship classification system.

Ship classification methods can be generalized into two categories, one is traditional handcrafted feature-based, and the other is convolutional neural network (CNN)-based. Traditional handcrafted features mainly include a histogram of oriented gradients (HOG) [3], local binary pattern (LBP) [4], Hu invariant moments [5] and scale-invariant feature transform (SIFT) [6], etc. Handcrafted features are only suitable for use in specific applications and rely on expert knowledge.

Nowadays, with the rapid development of deep learning technology, CNN has become a research hotspot in computer vision, with successful adoption in the fields of image classification [7], object detection [8], and traffic sign recognition [9]. Ding et al. [10] proposes a deep CNN method combining three types of data enhancement for ship object recognition. In literature [11] a CNN using extreme learning machine [12,13] is proposed to recognize infrared ship images. Not only does it need an extreme learning machine method to learn CNN features, it also requires additional integrated extreme learning machine for classification, which doubles complexity. Li et al. [14] proposes a CNN-based ship classification method which designed two networks built on AlexNet and GoogleNet, and used a pre-trained model on the ImageNet dataset for transfer learning. This method achieves good classification performance. Zhang et al. [15] proposes a multi-feature structure fusion method for maritime vessels classification based on spectral regression discriminant analysis (SF-SRDA), which combines structure fusion and linear discriminant analysis. However, it can only perform separate training and testing for visible or infrared images, without integration between results from visible and infrared images. Liu et al. [16] proposes a fusion recognition method based on CNN. The method designs a sensible network model to extract features of three band images and effectively fuse them. It then uses a feature selection method based on mutual information to sort the features according to their importance, which can eliminate redundant information and improve computational efficiency. Chen et al. [17] propose a from-coarse-to-fine CNN ship-type recognition method. The training method of the 'coarse' step is similar to traditional CNN. The 'fine' step introduces a regularization mechanism to extract more inherent ship features, and improved recognition performance can be obtained by fine-tuning parameter settings. Shi et al. [18] propose a deep learning framework that integrates low-level features for ship classification. Aziz et al. [19] propose a robust recognition method for maritime images based on multi-modal deep learning. Huang et al. [20] propose a ship classification model that combines multiple deep CNN features and use a fusion strategy to explore the relationship between multi-scale features. Jia et al. [21] propose a maritime ship recognition method based on two cascaded CNNs. A shallow network is used for speedy removal of the background area to reduce the computational cost, and a deep network is used to classify the ship types in the remaining areas.

Most of the existing ship classification methods use a single band of infrared or visible images to classify ships, without taking into account the complementary information within images obtained by different sensors. There is relatively little research on ship classification methods based on visible and infrared images fusion. The accuracy of ship classification method needs to be further improved. Although the CNN can automatically learn high-level features from ship images, a single-scale convolution kernel may lose some detailed information when extracting the ship image features. In addition, attention mechanism [22] (refer Section 2.4) can focus on object area and suppress other useless information. Applying attention mechanism to CNN can improve the quality of convolutional feature mapping [23]. Considering also that a single feature may not be comprehensive to represent ship images, the study proposes ship classification based on attention mechanism and multi-scale convolutional neural network (MSCNN). Firstly, the MSCNN has been proposed to extract the visible image features and infrared image features. Then the visible image features and infrared image features are fused to make full use of the complementarity of different features and obtain more comprehensive ship information. Lastly, we use the attention mechanism to enhance fusion feature representation capability, so as to achieve more accurate ship classification results.

Major contributions of this study can be summarized as follows: (1) a two-stream symmetric MSCNN feature extraction module is proposed to extract the features of visible and infrared images. The module can selectively extract those deep features of visible and infrared images with more detailed information. (2) The visible image features and infrared image features are concatenated to allow further use of the complementary information within different modal images, such that a more detailed ship object description can be obtained. (3) The attention mechanism is applied to the concatenated fusion layer to enhance important local details in the feature map, thereby improving overall classification capability of the model.

The remainder of this paper is organized as follows. Section 2 describes the proposed classification method in details. Section 3 introduces the visible and infrared spectra (VAIS) dataset [24] and parameter settings, and analyzes experimental results. Section 4 summarizes conclusions and the prospects of future work.

2. Proposed Ship Classification Method

2.1. Framework of Proposed Approach

Visible images are quite vulnerable to light conditions, which is the reason that ship classification relying only on single-senor images is subject to many limitations and deficiencies. On the other hand, a deep learning algorithm can automatically acquire higher-level and more abstract features in the images and an attention mechanism can enhance the feature representation with more effective information in the feature map. However, ship image features extraction using a single-scale convolution kernel is prone to detail omission. Therefore, this study proposes a ship classification method based on attention mechanism and MSCNN for the visible and infrared images, so as to combine the respective advantages of different types of sensor image and improve accuracy of ship classification results. The specific flow chart is shown in Figure 1. The proposed method consists of a feature extraction module, attention mechanism and feature fusion module, and classification module. The feature extraction module uses a two-stream symmetric MSCNN to extract the features of the preprocessed visible and infrared image, respectively. The attention mechanism and feature fusion module first concatenates extracted visible image features and infrared image features, and then obtains attention weights by applying attention mechanism to the concatenated fusion feature layer to enhance key local features, suppress unimportant features, and improve feature expression results of the model. The classification module is composed of three fully connected layers and a Softmax output layer. The ship classification results are obtained through Softmax output layer function. Image preprocessing in the figure refers to image size adjustment (refer Section 2.2).

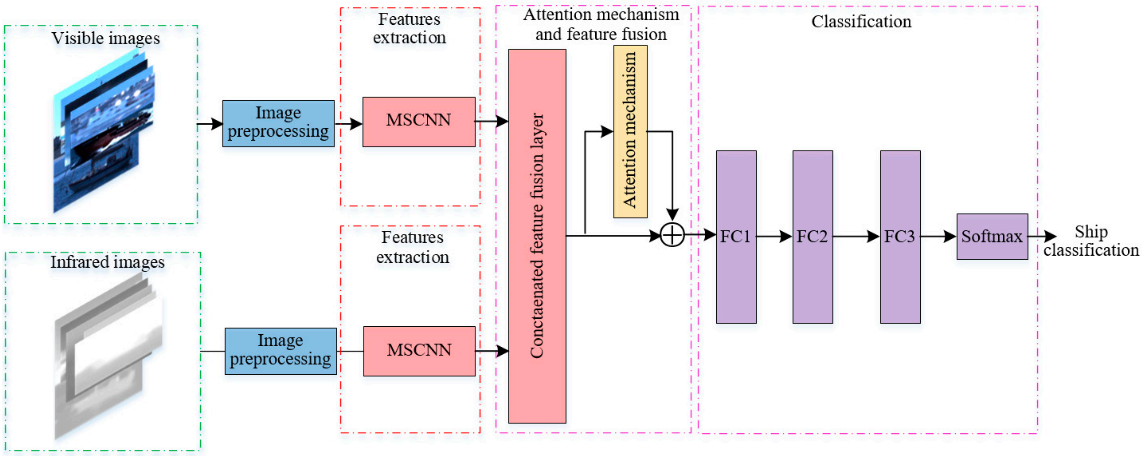


Figure 1. The framework of the proposed ship classification method. ⊕ denotes element-wise addition.

2.2. Two-Stream Symmetric Multi-Scale Convolutional Neural Network (MSCNN) Feature Extraction Module

CNN is a feed-forward neural network architecture, which adopts local perception and weight sharing methods. The traditional CNN contains convolutional layers, pooling layers and fully connected layers, etc. Among them, network architecture is one of the core factors that determines classification performance. Choosing an appropriate CNN framework to extract ship image features effectively is a prerequisite for classification performance improvement. However, image features extraction using single-scale convolution kernel is prone to details omission. Inspired by InceptionNet [25], this study uses convolution kernels of different sizes in the convolution layer to extract features of different scales, thereby enriching ship image features.

The proposed MSCNN feature extraction module consists of two identical parallel multi-scale CNNs. MSCNN is mainly composed of 4 convolutional layers (Conv1–Conv4), 3 pooling layers (Max Pooling1–Max Pooling3), 3 fully connected layers (FC1–FC3) and a Softmax output layer. After Conv3, two sets of convolution kernels of different sizes (i.e., 3 × 3 and 5 × 5) are used for parallel convolution to obtain Conv4_1 and Conv4_2, and then Conv4_1 and Conv4_2 are concatenated, and the deep features of the ship image are extracted by using two sets of convolution kernels of different sizes. The obtained feature map contains more detailed information to reduce information loss in image processing. The rectified linear unit (ReLU) function is used in both convolutional layer and fully connected layers as it can prevent gradient disappearance, make the network sparse, and it is more efficient than the sigmoid function. In order to prevent over-fitting, Dropout technology is used in 3 fully connected layers. In each iteration, the dropout technology can randomly hide certain neurons, the selected hidden neurons do not contribute to the parameter updates, which can effectively avoid overfitting and enhance the generalization ability of the network. Their respective parameters are shown in Table 1. It can be seen that the input image size is 227 × 227 × 3. Since the original infrared image is a grayscale image, we create a pseudo-red, green and blue (RGB) image using the same method mentioned in literature [24], where a single infrared channel is duplicated three times. The pooling layer adopts a maximum pooling of size 3 × 3, which reduces the dimensionality of upper layer convolution result, simplifies calculation process, and at the same time retains more image texture. "Padding: 1" refers to edge expansion with one circle of 0. Conv4 refers to the result that Conv4_1 and Conv4_2 are concatenated, such that output size is 13 × 13 × 512. In this study we compare the classification results of visible ship images and results of infrared ship images with a MSCNN and the network (referred to as CNN) that only uses 256 convolution kernels of size 3 × 3 for the convolution operation in Conv4. Both networks are used as baseline methods (refer Section 3.4) to compare with and validate the proposed method.

Table 1. Specific parameters of the multi-scale convolutional neural network (MSCNN).

Layer	Input Size	Kernel Size	Filters No.	Stride	Padding	Output Size
Conv1	227 × 227 × 3	11 × 11	64	4	2	56 × 56 × 64
Max Pooling1	56 × 56 × 64	3 × 3	-	2	-	27 × 27 × 64
Conv2	27 × 27 × 64	5 × 5	192	1	2	27 × 27 × 192
Max Pooling2	27 × 27 ×192	3 × 3	-	2	-	13 × 13 × 192
Conv3	13 × 13 × 192	3 × 3	384	1	1	13 × 13 × 384
Conv4_1	13 × 13 × 384	3 × 3	256	1	1	13 × 13 × 256
Conv4_2	13 × 13 × 384	5 × 5	256	1	2	13 × 13 × 256
Conv4	Conv4_1, Conv4_2	-	-	-	-	13 × 13 × 512
Max Pooling3	13 × 13 × 512	3 × 3	-	2	-	6 × 6 × 512
FC1	6 × 6 × 512	-	-	-	-	4096
FC2	4096	-	-	-	-	4096
FC3	4096	-	-	-	-	2048
Softmax	2048	-	-	-	-	6

Feature visualization of different convolutional layers in MSCNN is shown in Figure 2, with the visible image only as an example. It can be seen that the number of feature maps in each layer equals the number of filters in Table 1. Different convolution kernels respond differently to various ship image positions. The shallow layer extraction focuses mainly on texture features, with obtained feature maps closer to the original image. Deeper layer extraction focuses more on features such as contours and shapes, which are in general more abstract and representative. The deeper the layer goes, the lower the resolution of the feature map.

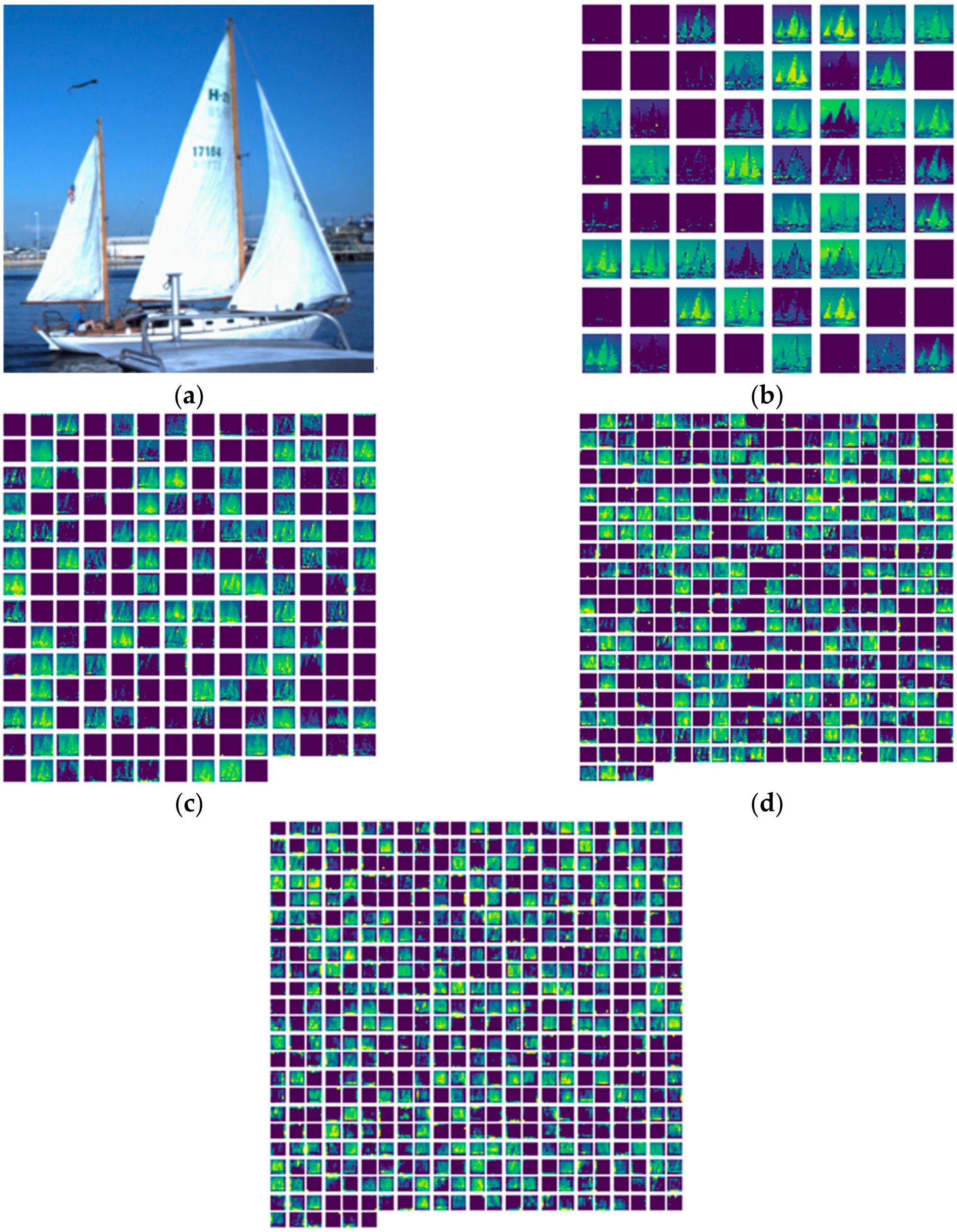
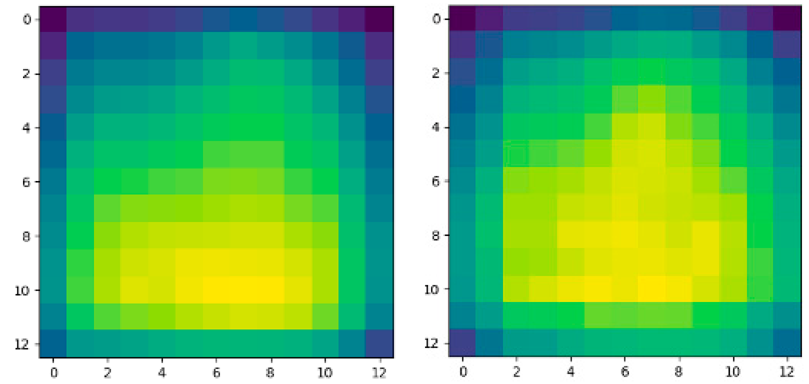


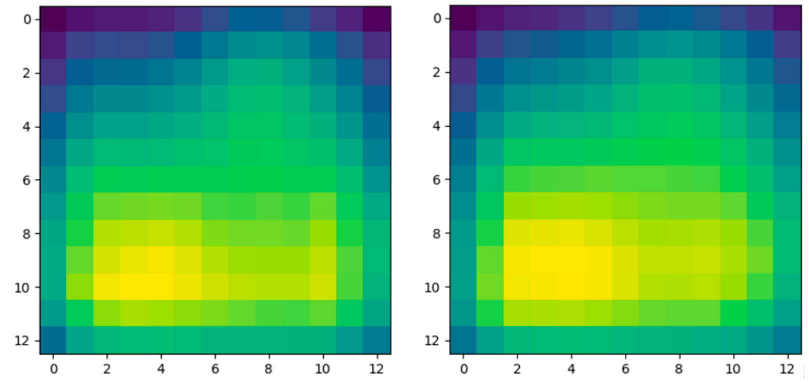
Figure 2. Feature visualization of different convolutional layers in MSCNN: (**a**) input image; (**b**) conv1; (**c**) conv2; (**d**) conv3; and (**e**) conv4.

Since the feature maps obtained by different convolution kernels of the same layer are complementary to each other when describing ship images, one can obtain an overall feature map by fusing these individual ones from the layer in a ratio of 1:1. Figures 3 and 4 show comparison of the overall feature map of visible image and corresponding infrared image at CNN's conv4 and MSCNN's conv4. The feature map size is consistent with the output size of conv4 and route layer. From below, it can be seen that MSCNN-based feature extraction responds better to ship area, i.e., the yellow area highlighted in the feature map in Figures 3b and 4b is darker and wider than that of map in Figures 3a and 4a. This would play a positive role in our subsequent fusion classification of visible image features and infrared image features. In addition, it can be seen that the position of strongest response to ship area is different between the visible image feature and infrared image feature. Therefore, the fusion of these two features can effectively use the complementary information within different modal images, enrich the fused information, and improve the ship classification performance.



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

Figure 3. Comparison of the overall feature map: (**a**) CNN's conv4; and (**b**) MSCNN's conv4 (The input image is visible image).



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

Figure 4. Comparison of the overall feature map: (**a**) CNN's conv4; and (**b**) MSCNN's conv4 (The input image is infrared image).

It is known that features extracted by the deep convolutional layer contain abundant useful information. Furthermore, medium-term feature fusion method has achieved better classification results according to literature [26]. In view of the above, this study focuses on fusion using features obtained by the Max Pooling3 layer of MSCNN. Max Pooling3 layer reduces the dimensionality of the last convolution layer (i.e., Conv4). The two-stream symmetric MSCNN feature extraction module is shown in Figure 5. The visible and infrared image of the same ship object are respective inputs of the two-stream network for feature extraction, and the two-stream features are processed via the concatenated feature fusion layer after the Max Pooling 3 layer to obtain the concatenated fusion feature. In summary, three fully connected layers and one Softmax output layer in the multi-scale CNN are used to classify the fusion features, with Softmax output node equal to 6 (the number of ship types in the VAIS dataset).

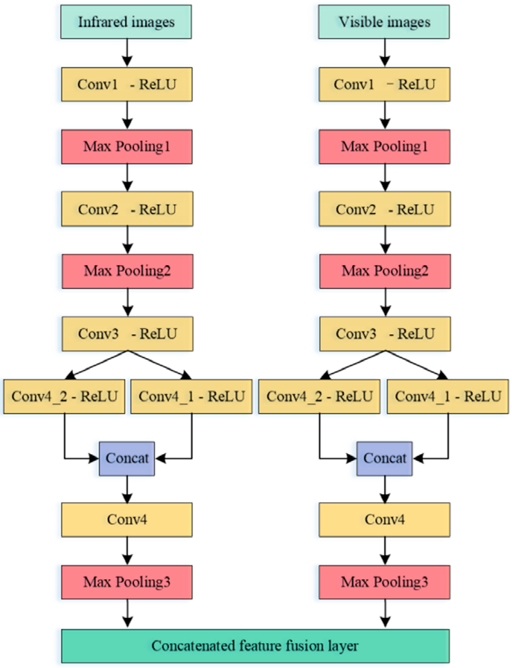


Figure 5. Two-stream symmetric MSCNN feature extraction module.

The proposed ship classification model in this study also includes a training process and a testing process. During training, the visible image and infrared image of the same object (with the same label) are preprocessed and input into the two-stream network to extract features and conduct simultaneous training. Then the error is calculated between the true class labels and the predicted class labels obtained by the Softmax function. After that, the weight and bias are adjusted by back propagation process to minimize the error. Lastly, the optimal model is saved. In the testing phase, the visible image and infrared image of the same object (with the same label) are also preprocessed and input into the two-stream network to extract features, and call upon the optimal model to test features, and the predicted labels of the ship images are output. The purpose of preprocessing is to randomly crop the training images into 227 pixels × 227 pixels. Such random cropping can not only increase the training data, but also improve the generalization ability of the model. In this study, stochastic gradient descent algorithm (SGD) [27] is used to minimize the cross entropy loss function. The test images are cropped into 227 pixels × 227 pixels by center cropping. We also use data enhancement techniques such as random horizontal flipping and z-score standardization [28] to ensure sample randomness and avoid model overfitting.

2.3. Feature Fusion

In this study, the visible image features and infrared image features extracted by the MSCNN are fused, and the effective information of the two features can be combined through feature fusion to obtain a more comprehensive feature representation of the ship object. Common feature fusion [29] methods include additive fusion, maximum fusion and concatenated fusion, etc. The expression of feature fusion can be defined as:

$$\mathbf{F} = f(\mathbf{X}, \mathbf{Y}), \tag{1}$$

where $\mathbf{X}$ and $\mathbf{Y}$ represent the visible image features and infrared image features extracted by the two-stream symmetric MSCNN respectively. $\mathbf{F}$ indicates the fusion feature, $\mathbf{X}, \mathbf{Y}, \mathbf{F} \in \mathbb{R}^{HWC}$. H, W and C indicate the height, width and number of channels of the feature map respectively.

Additive fusion is achieved by adding element values at the corresponding positions of the two feature maps, with total number of channels in the fusion feature map unchanged. If a visible image feature is denoted as $\mathbf{X} = [x_1, x_2, \ldots, x_n]$, and infrared image feature is denoted as $\mathbf{Y} = [y_1, y_2, \ldots, y_n]$, then additive fusion can be denoted as:

$$\mathbf{F} = f(\mathbf{X}, \mathbf{Y}) = \mathbf{X} + \mathbf{Y} = [x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n], \tag{2}$$

Maximum fusion is to take the element of higher value at the corresponding position of the two feature maps as the fusion result, which can be expressed as:

$$\mathbf{F} = f(\mathbf{X}, \mathbf{Y}) = \max\{\mathbf{X}, \mathbf{Y}\} = [max(x_1, y_1), max(x_2, y_2), \ldots, max(x_n, y_n)], \tag{3}$$

It should be noted that both additive fusion and maximum fusion are only applicable to feature maps fusion of same dimension.

On the other hand, concatenated fusion connects two feature maps directly, which can be applied to feature maps of any dimension. The total number of fusion feature channels is the sum of all visible image feature channels and infrared feature channels. Concatenated fusion can be expressed as:

$$\mathbf{F} = f(\mathbf{X}, \mathbf{Y}) = \mathbf{X}, \mathbf{Y} = [x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n], \tag{4}$$

Through experimental comparison, it can be seen that concatenated fusion of the visible image features and infrared image features can achieve better classification effect (refer Section 3.4.1), and concatenated fusion retains all the elements of the feature maps, this study adopts the concatenated fusion method. From Table 1, it can be seen that the output size of each feature map after Max Pooling3 is 6 × 6 × 512, where 512 represents the number of channels. Hence the size of the feature map after concatenated fusion is 6 × 6 × 1024. The detailed concatenated fusion process is shown in Figure 6.

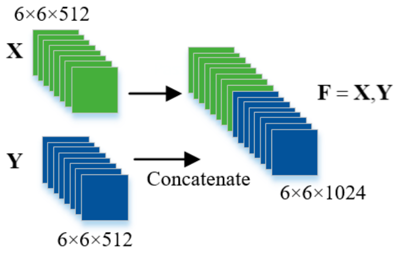


Figure 6. The concatenated fusion process of visible image features and infrared image features.

2.4. Feature Fusion Layer Based on Attention Mechanism

Attention mechanism (AM) [30] is a cognitive mechanism that mimics the human brain. In visual perception, it pays attention mainly to the features of interest, and suppresses redundant information. The attention mechanism can be integrated into the CNN framework with negligible overhead, and trained together with CNN [31]. Inspired by the convolutional block attention module (CBAM) [32] which allows auto-learning of pixel correlation among different feature maps, we add the attention mechanism after feature fusion layer. By combining the learned attention weights and the original concatenated fusion features, the model can greatly enhance local details of ship images, thereby improving the representation ability of the existing fusion feature maps. Subsequent experiments also prove that the addition of such module can improve the ship classification performance.

The structure diagram of the attention mechanism and feature fusion module is shown in Figure 7. It can be seen from Figure 7 that the attention mechanism includes a channel attention module and a spatial attention module, where the two modules are connected in series. Concatenated fusion features are used as the input of channel attention, and the channel attention weight is calculated and multiplied with the concatenated fusion feature to each channel to obtain the feature map of the channel attention module, which is then used as the input of spatial attention. After spatial attention weights being obtained, it is also multiplied with the feature map of the channel attention module to arrive at the final attention mechanism feature map. To avoid features loss and performance degradation, the attention mechanism feature map and the concatenated fusion feature are added element by element to achieve the final refined feature.

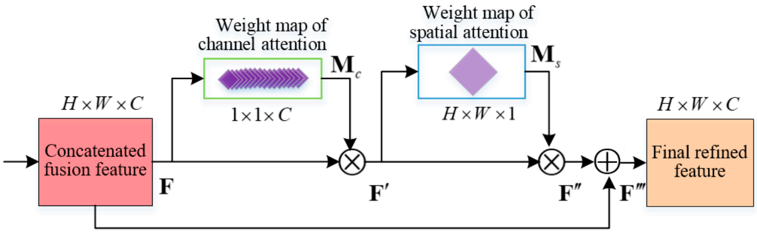


Figure 7. Structure diagram of attention mechanism and feature fusion module. ⊗ denotes element-wise multiplication, ⊕ denotes element-wise addition.

2.4.1. Channel Attention Module

The channel attention module establishes a weight map to evaluate the importance of each channel. The channel that contains more important information has higher weight, and vice versa. It focuses on "what" it views as meaningful. The channel attention module is shown in Figure 8.

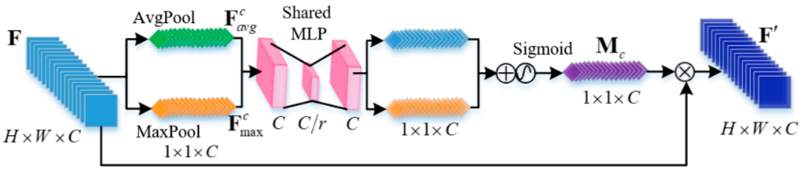


Figure 8. Diagram of channel attention module. σ denotes sigmoid function.

Firstly, concatenated fusion feature **F** serves as input feature with dimension $H \times W \times C$. Both average pooling and maximum pooling are then applied to input features to aggregate spatial information of the feature map, and the channel attention descriptors $\mathbf{F}^c_{avg}$ and $\mathbf{F}^c_{\max}$ of dimension $1 \times 1 \times C$ are obtained. These two channel attention descriptors are fed into a multi-layer perceptron (MLP) with a hidden layer. To reduce number of parameters used, the activation size of the hidden layer is $\mathbb{R}^{C/r\times1\times1}$, in which r is the compression ratio. In this study, r is set to 8 (refer Section 3.4.2). The features of the two parallel branches are then added and processed by sigmoid activation function to obtain the final channel attention weight map $\mathbf{M}_c$. Finally, multiplication between the channel attention weight map $\mathbf{M}_c$ and the original concatenated fusion feature **F** is done to obtain the channel-refined feature $\mathbf{F}'$. The calculation process of channel attention weight map $\mathbf{M}_c$ and channel-refined feature $\mathbf{F}'$ can be expressed as:

$$\mathbf{M}_c(\mathbf{F}) = \sigma(MLP(AvgPool(\mathbf{F})) + MLP(MaxPool(\mathbf{F}))), \tag{5}$$

$$\mathbf{F}' = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F}, \tag{6}$$

where σ denotes the Sigmoid activation function, ⊗ denotes element-wise multiplication.

2.4.2. Spatial Attention Module

Spatial attention module obtains the weight map of features in spatial dimension, which focuses on "where" useful information can be found, supplementing channel attention. The details of the spatial attention module is shown in Figure 9.

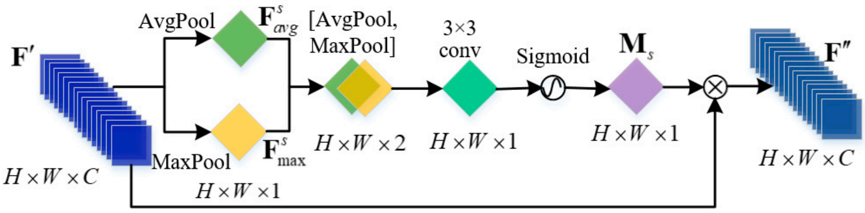


Figure 9. Diagram of spatial attention module.

Firstly, the input feature is $\mathbf{F}^{'}$ of dimension $H \times W \times C$. Maximum pooling and average pooling are performed in parallel on the input feature in the channel dimension to obtain two descriptors $\mathbf{F}^s_{avg}$ and $\mathbf{F}^s_{\max}$ of dimension $H \times W \times 1$, which are then concatenated. After that, the concatenated fusion descriptor is convolved by the convolution kernels of size 3×3 and processed by the sigmoid activation function to obtain the spatial attention weight map $\mathbf{M}_s$. Finally, the spatial attention weight map $\mathbf{M}_s$ and $\mathbf{F}^{'}$ are multiplied to obtain the features $\mathbf{F}^{''}$ after spatial attention. The calculation process of spatial attention weight map $\mathbf{M}_s$ and the feature $\mathbf{F}^{''}$ after spatial attention can be expressed as:

$$\begin{aligned} \mathbf{M}_s(\mathbf{F}) &= \sigma(f^{3\times3}([AvgPool(\mathbf{F}), MaxPool(\mathbf{F})])) \\ &= \sigma(f^{3\times3}([\mathbf{F}^s_{avg}, \mathbf{F}^s_{\max}])) \end{aligned}, \tag{7}$$

$$\mathbf{F}^{''} = \mathbf{M}_s(\mathbf{F}) \otimes \mathbf{F}^{'}, \tag{8}$$

where $f^{3\times3}$ denotes convolution kernels of size 3×3.

From the above analysis, we can see that both average pooling and maximum pooling are used in the attention mechanism. Average pooling takes the average information on each channel, whereas maximum pooling only considers the most significant information of each channel of the feature map. Through the combination of the two pooling operations, the attention mechanism is able to focus on the important channel and spatial feature information of the ship image, and filter out the unimportant feature information. As a result, more discriminative features can be obtained to improve ship classification performance.

3. Experimental Results and Analysis

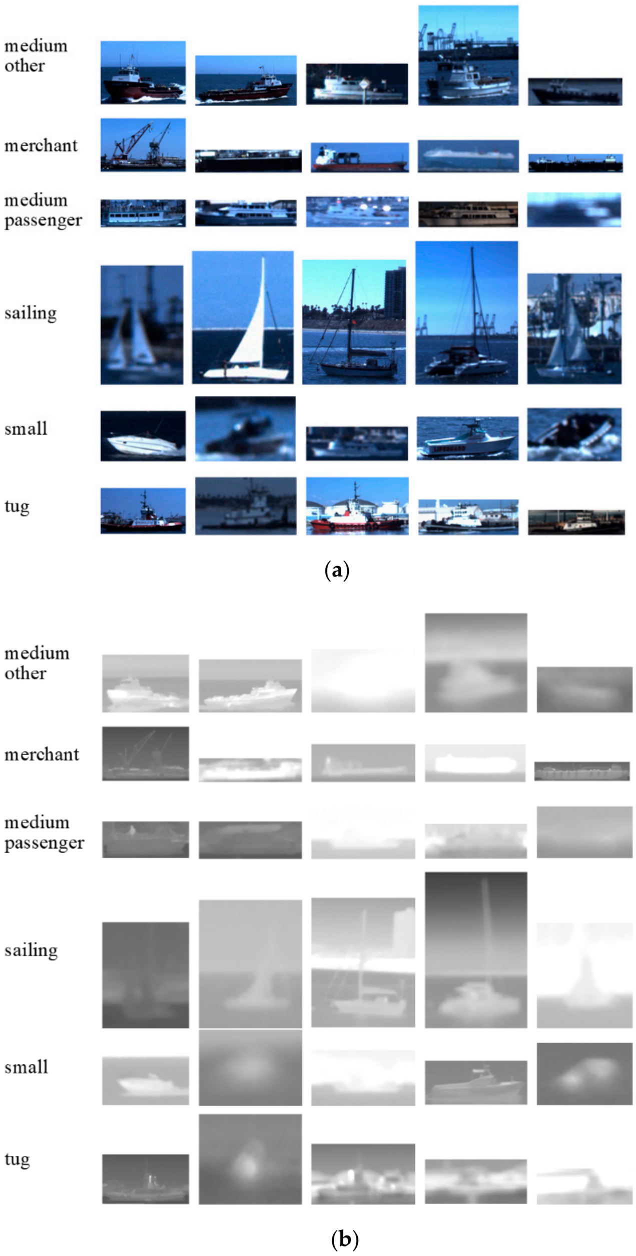
3.1. Experimental Environment and Parameter Setting

The experimental environment used in this study is computer with Intel(R) Core(TM) i9-7980XE@2.6GHz processor, GPU of NVIDIA TITAN Xp Pascal and 32GB memory. Python3.5 and deep learning open source framework Pytorch programming are used to perform all experiments.

Experimental parameter settings are as follows. The batch size is set to 32, and learning rate at 0.001. The optimization method is the stochastic gradient descent (SGD) algorithm. The momentum parameter is set to 0.9, with weight coefficient 0.0001. The dropout is at 0.5, and learning epochs of 500.

3.2. Experimental Dataset

The dataset used in this study is the VAIS dataset [24], which is the only available public dataset of paired visible and long wave infrared ship images. These images were captured using a multimodal stereo camera rig on harbours. The RGB global shutter camera was a ISVI IC-C25. The long wave infrared camera was Sofradir-EC Atom 1024, which has a spectral range of 8–12 microns. The cameras are tightly mounted next to each other and checked to ensure no interference. The dataset consists of 2865 images (1623 visible images and 1242 infrared images) including 1088 paired visible and infrared images. There are a total of 154 nighttime infrared images. The number of unique ships is 264. For most ships, only one orientation was captured; for a few, up to 5 to 7 orientations. This way, we avoid duplicates in the dataset. The image format is png. The dataset can be divided into 6 coarse-grained categories, namely Medium "other" ships, Merchant ships, Medium passenger ships, Sailing ships, Small boats and Tugboats, as shown in Figure 10. It can be seen that the background of the ship is complex, with uneven illumination and various size. The area of visible bounding boxes ranges from 644 to 4,478,952 pixels, with mean of 181,319 pixels and median 9983 pixels. The area of infrared bounding boxes ranges from 594 to 137,240 pixels, with mean of 8544 pixels and median 1610 pixels. In this study, only 1088 pairs of visual and infrared images are chosen for experiment purposes. According to the number of "official" training and test sample. A total of 539 pairs are randomly selected as training images, and the remaining 549 pairs are test images. The number of samples in the training set and test set are listed in Table 2. There are 138 pairs of images in Medium-other, 146 pairs of images in Merchant, 117 pairs of images in Medium-passenger, 284 pairs of images in Sailing, 353 pairs of images in Small boats, and 50 pairs of images in Tugboats. The bicubic interpolation method presented in a previous study is used here to uniformly adjust the size of the ship image to 256 pixels × 256 pixels.



(**a**)

(**b**)

Figure 10. Five visible samples from each of the main classes of the visible and infrared spectra (VAIS) dataset: (**a**) Visible images; and (**b**) infrared images.

Table 2. Number of training and test samples for the VAIS dataset.

No.	Class	Train	Test
1	Medium-other	62	76
2	Merchant	83	63
3	Medium-passenger	58	59
4	Sailing	148	136
5	Small	158	195
6	Tug	30	20
	Total	539	549

3.3. Evaluation Metrics

The evaluation metrics of ship image classification results adopted by this study include classification accuracy, F1-score and average feature extraction time consumption per image.

Classification accuracy is defined as the ratio of correctly classified samples to the total number of samples. The higher the ratio is, the better the classification performance. The classification accuracy can be expressed as:

$$Acc = \frac{\mathrm{TP} + \mathrm{FP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}, \tag{9}$$

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

F1-score is a comprehensive measure of classification performance, which is the weighted average of precision ratio and recall ratio. The maximum F1-score is 1 and the minimum is 0. It can be defined as:

$$F1 = \frac{2 \times P \times R}{P + R}, \tag{10}$$

Precision ratio is calculated using the number of true positives divided by predicted positive samples. Recall ratio is computed by taking the number of true positives divided by all positive samples. Precision and recall ratios can be expressed as:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{11}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{12}$$

To further analyze the classification results, the confusion matrix was used to visualize them. The confusion matrix represents the mistakes caused by the classifier when dealing with multi-class problems. The horizontal axis represents the predicted category and the vertical axis gives the true category. Hence diagonal elements are those correctly classified ship images of each type. The diagonal element of the normalized confusion matrix represents classification accuracy achieved by each ship type.

3.4. Experimental Results and Analysis

3.4.1. Classification Performance Comparison

To validate its classification performance, the proposed method was compared with the baseline method and other state-of-the-art methods under the same experimental conditions.

Comparison with Baseline Method and Feature Fusion Method

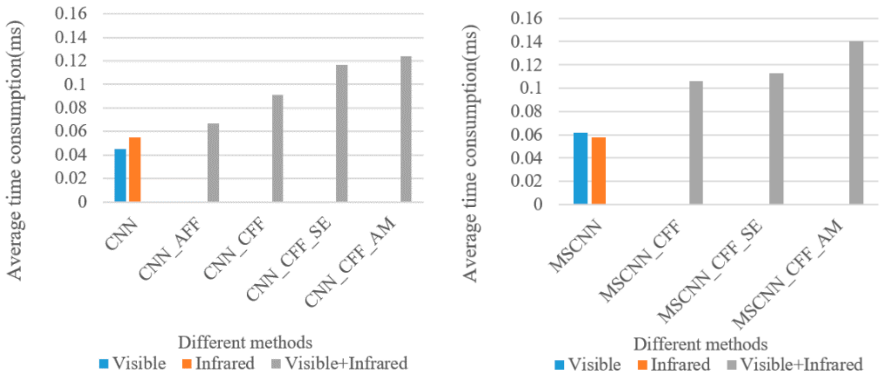
Table 3 lists out evaluation metrics, namely classification accuracy and average feature extraction time consumption per image, and their respective results when the baseline method, feature fusion method and proposed method are applied to VAIS dataset. Herein, the baseline methods are CNN and

MSCNN. Herein, CNN represents Conv4 with only 256 convolution kernels of size 3 × 3. CNN_AFF refers to additive feature fusion of infrared image features and visible image features extracted by CNN. CNN_CFF and MSCNN_CFF represent concatenated feature fusion of infrared image features and visible image features extracted by CNN and MSCNN, respectively. CNN_CFF_SE and MSCNN_CFF_SE denote applying the channel attention mechanism in literature [22] to concatenated fusion features. CNN_CFF_AM and MSCNN_CFF_AM indicate the proposed attention mechanism being applied to process concatenated fusion features. It can be observed that classification accuracy of visible images is higher than that of infrared images, mainly due to lower resolution and less texture information of infrared images. MSCNN has higher classification accuracy than CNN for both visible and infrared images, which indicates that the proposed MSCNN can extract more detailed information with enriching ship image features. Classification accuracy using feature fusion methods is higher than the baseline method, which means that fusing visible image features and infrared image features can complement information from multiple sources and improve classification performance. Furthermore, CNN_CFF attains higher classification accuracy than CNN_AFF, as concatenated feature fusion overcomes information offset caused by addition of elements in spatial dimension. Therefore, we did not conduct additive fusion experiments for MSCNN method. It can be seen that the classification accuracy of the CNN_CFF_SE, MSCNN_CFF_SE, CNN_CFF_AM and MSCNN_CFF_AM methods is higher than that of the feature fusion method, and the MSCNN_CFF_AM method (the proposed method) achieved the highest classification accuracy. This can be explained as fused features after attention mechanism modification can highlight key local features, suppress useless features, and significantly enhance feature expression. The SE (squeeze and excitation) module in literature [22] focuses on the channel information of feature map, and ignores the importance of spatial location. The attention mechanism in this study combines channel attention and spatial attention modules, so that each of the modules can learn "what" and "where" to see in the channel and spatial dimensions. Hence, the proposed method can achieve better classification performance.

Table 3. Classification accuracy (%) and average time consumption for feature extraction per image (ms) of the proposed method, baseline methods and feature fusion methods on VAIS dataset.

Method		Accuracy	Feature Extraction Time
CNN(baseline)	Visible	90.53	0.045
	Infrared	85.43	0.055
CNN_AFF	Visible + Infrared	91.26	0.067
CNN_CFF	Visible + Infrared	91.99	0.091
CNN_CFF_SE	Visible + Infrared	92.71	0.117
CNN_CFF_AM	Visible + Infrared	93.26	0.124
MSCNN(baseline)	Visible	91.44	0.062
	Infrared	86.52	0.058
MSCNN_CFF	Visible + Infrared	92.53	0.106
MSCNN_CFF_SE	Visible + Infrared	93.08	0.113
MSCNN_CFF_AM(Proposed method)	Visible + Infrared	93.81	0.140

Figure 11 depicts the average feature extraction time consumption per image of the proposed method, baseline methods and feature fusion methods on the VAIS dataset. It can be seen that in the proposed method, the additions of feature fusion and attention mechanism bring the average time consumption per image slightly higher than that of the baseline method; however, the former has obvious advantages in classification accuracy and the time consumption of 0.140 ms was also relatively short.



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

Figure 11. Comparison of the average time consumption for feature extraction per image of the proposed method, baseline methods and feature fusion methods on the VAIS dataset: (**a**) CNN (baseline); and (**b**) MSCNN (baseline).

Classification accuracy obtained per ship class using baseline methods, feature fusion methods and proposed method on VAIS dataset are listed in Table 4. As observed, the proposed method demonstrated best classification results compared to the other methods. The classification accuracy of CNN_CFF_AM is better than that of the CNN method for each type. The classification accuracy of the proposed method is higher than that of MSCNN for all ship types except Merchant. Overall, the proposed method has the highest classification accuracy for medium-passenger, small and tug. The classification accuracy of the proposed method for medium-other and merchant is slightly lower than that of MSCNN_CFF_SE, but the classification accuracy of the proposed method for medium-passenger is 6.78% higher than that of MSCNN_CFF_SE. The classification accuracy for tugs is 100%, because of the larger difference in appearance between the tug and other ship types, and the better image quality of the tugs in VAIS dataset. In conclusion, the proposed method achieved the highest overall classification accuracy on the VAIS dataset.

Table 4. Class-specific accuracy (%) of the proposed method, baseline methods and feature fusion methods on the VAIS dataset.

Method		**Class**					
		Medium-Other	**Merchant**	**Medium-Passenger**	**Sailing**	**Small**	**Tug**
CNN(baseline)	Visible	76.32	93.65	79.66	92.65	96.41	95.00
	Infrared	64.47	93.65	76.27	94.12	86.15	100.00
CNN_CFF	Visible + Infrared	77.63	90.48	88.14	96.32	95.38	100.00
CNN_CFF_SE	Visible + Infrared	73.68	96.83	86.44	97.06	96.92	100.00
CNN_CFF_AM	Visible + Infrared	77.63	96.83	86.44	97.06	96.92	100.00
MSCNN(baseline)	Visible	77.63	96.82	79.66	92.65	96.92	100.00
	Infrared	65.79	95.24	74.58	93.38	89.23	100.00
MSCNN_CFF	Visible + Infrared	76.31	93.65	88.14	95.59	96.92	100.00
MSCNN_CFF_SE	Visible + Infrared	78.95	98.41	86.44	96.32	95.90	100.00
MSCNN_CFF_AM (Proposed method)	Visible + Infrared	77.63	95.24	93.22	96.32	97.44	100.00

In order to verify the classification capability of the proposed method, we compare the F1-scores based on baseline, feature fusion and proposed method, as shown in Table 5. It can be seen that the proposed method achieves the highest average F1-score among 6 ship types. CNN_CFF_SE gives best F1-score for merchant, and CNN_CFF_AM beats the rest for sailing. However the proposed one gives the highest F1-score for all other four ship types, and is doing almost equally well as CNN_CFF_AM for sailing. This can be largely attributed to the addition of feature fusion and attention mechanisms in the proposed method, in which complementary information from multi-source images are effectively utilized and more characteristic features are properly extracted. It greatly enhances overall model classification capability.

Table 5. F1-score of the proposed method, baseline methods and feature fusion methods on the VAIS dataset.

Method		Class						
		Medium-Other	**Merchant**	**Medium-Passenger**	**Sailing**	**Small**	**Tug**	**Avg. Total**
CNN(baseline)	Visible	0.8169	0.9440	0.8785	0.9545	0.8974	0.9268	0.9030
	Infrared	0.7481	0.8939	0.7563	0.9176	0.8638	0.8333	0.8355
CNN_CFF	Visible + Infrared	0.8429	0.9421	0.8889	0.9458	0.9231	1.0000	0.9238
CNN_CFF_SE	Visible + Infrared	0.8358	0.9683	0.9027	0.9462	0.9333	0.9756	0.9269
CNN_CFF_AM	Visible + Infrared	0.8613	0.9313	0.8947	0.9706	0.9356	1.0000	0.9322
MSCNN(baseline)	Visible	0.8489	0.9173	0.8868	0.9474	0.9220	0.9091	0.9052
	Infrared	0.7752	0.8889	0.7857	0.9203	0.8766	0.8163	0.8438
MSCNN_CFF	Visible + Infrared	0.8345	0.9516	0.9123	0.9559	0.9265	0.9756	0.9261
MSCNN_CFF_SE	Visible + Infrared	0.8571	0.9612	0.9027	0.9632	0.9280	0.9756	0.9313
MSCNN_CFF_AM (Proposed method)	Visible + Infrared	0.8676	0.9375	0.9244	0.9704	0.9383	1.0000	0.9397

Comparison with Other State-of-the-Art Methods

To further verify the effectiveness of the proposed method, we compare the proposed method with other state-of-the-art methods developed in recent years. We reimplement the state-of-the-art methods on the VAIS dataset, and the partition method of the dataset is consistent with that of the proposed method. The results are shown in Tables 6–8. Table 6 gives out classification accuracy of different methods on the VAIS dataset. Table 7 illustrates the classification accuracy of each class with different methods on VAIS dataset. Table 8 lists the F1-score using different methods on VAIS dataset. Among them, traditional methods (HOG + SVM, LBP + SVM), AlexNet, Method [33], Method [10] and Method [14] only process either visible images or infrared images of a single band, whereas Method [19] uses two parallel CNNs to extract the features of visible images and infrared images, respectively, and classifies them after feature fusion in the last fully connected layer. It can be seen from Table 6 that, compared with other methods, the proposed one achieves the best classification accuracy on both single-band and multi-source images. It can be seen from Table 7 that the proposed method had the highest classification accuracy for medium-passenger, sailing, small and tug. Although for Medium-other category, the proposed method falls slightly behind Method [14] in classification accuracy, it still beats all other methods. It can be seen from Table 8 that the average F1-score of the proposed method is the highest. Method [19] had the highest F1-score for Merchant, and the proposed method had the highest F1-score for all five other types. It can be concluded that the proposed method achieves overall best classification performance, which is due to effective extraction and fusion of visible image and infrared images and inclusion of the attention mechanism.

Confusion Matrix and Confusion Matrix Normalization of the Classification Results

Although compared with other methods, the proposed method improves the classification performance greatly, there are still some cases of misclassification. The confusion matrix and its normalization process using the VAIS dataset are depicted in Figure 12. It can be seen that initially 12 Medium-other ships were misjudged as Small, with an inter-class error of 15.9%, and 3 Medium-passenger ships were misjudged as Small, with an inter-class error of 5.1%, indicating that both

Medium-other and Medium-passenger are the classifications that can be easily confused with Small. This is not surprising as their shapes resemble each other closely, especially when the image resolution is low, as shown in Figure 10. Figure 13 illustrates misclassified examples of Medium-passenger ships and Small ships.

Table 6. Classification accuracy (%) of different methods on VAIS dataset.

Method	**Accuracy**		
	Visible	**Infrared**	**Visible + Infrared**
HOG + SVM	88.34	82.15	-
LBP + SVM	82.33	69.76	-
AlexNet	90.35	85.06	-
Method [33]	86.52	76.14	-
Method [10]	88.89	71.58	-
Method [14]	89.25	85.43	-
Method [19]	90.16	85.25	91.62
MSCNN_CFF_AM (Proposed method)	91.44	86.52	93.81

Table 7. Class-specific accuracy (%) of different methods on VAIS dataset.

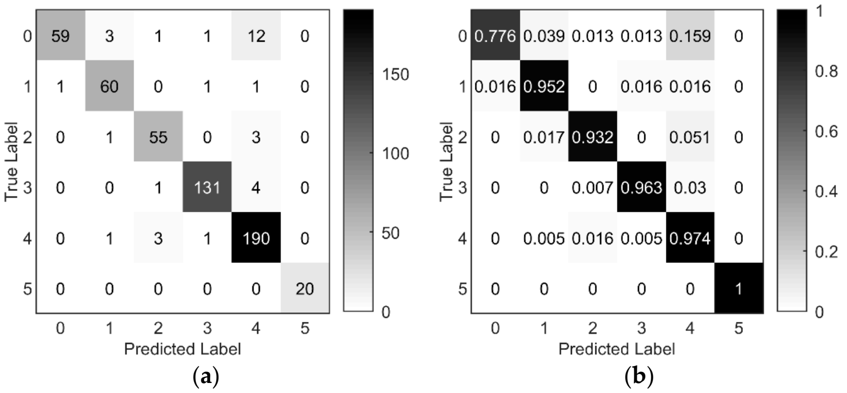
Method		**Class**					
		Medium-Other	**Merchant**	**Medium-Passenger**	**Sailing**	**Small**	**Tug**
AlexNet	Visible	73.68	98.41	77.97	94.85	93.85	100.00
	Infrared	61.84	90.48	69.49	96.32	87.69	100.00
Method [33]	Visible	71.05	96.83	77.97	88.97	89.23	95.00
	Infrared	43.42	93.65	72.88	80.88	80.00	85.00
Method [10]	Visible	73.68	98.41	71.19	91.18	95.38	90.00
	Infrared	69.74	98.41	72.88	93.38	94.36	95.00
Method [14]	Visible	82.89	95.24	83.05	96.32	86.67	90.00
	Infrared	53.95	90.48	74.58	95.59	90.77	100.00
Method [19]	Visible + Infrared	75.00	96.83	76.27	96.32	97.44	95.00
MSCNN_CFF_AM (Proposed method)	Visible + Infrared	77.63	95.24	93.22	96.32	97.44	100.00

Table 8. F1-score of different methods on VAIS dataset.

Method		**Class**						
		Medium-Other	**Merchant**	**Medium-Passenger**	**Sailing**	**Small**	**Tug**	**Avg. Total**
AlexNet	Visible	0.8058	0.9538	0.8288	0.9451	0.9219	0.8333	0.8815
	Infrared	0.7068	0.8906	0.7257	0.9424	0.8571	0.8511	0.8289
Method [33]	Visible	0.7397	0.8531	0.8519	0.9167	0.8878	0.8444	0.8489
	Infrared	0.5197	0.8027	0.6056	0.8429	0.8298	0.7556	0.7260
Method [10]	Visible	0.8058	0.9051	0.8155	0.9151	0.9231	0.8000	0.8608
	Infrared	0.7994	0.9254	0.8269	0.9137	0.9246	0.7917	0.8603
Method [14]	Visible	0.8182	0.9231	0.8909	0.9161	0.9086	0.7826	0.8732
	Infrared	0.6891	0.8906	0.7333	0.9386	0.8762	0.8000	0.8213
Method [19]	Visible + Infrared	0.8507	0.9457	0.8333	0.9493	0.9246	0.9500	0.9089
MSCNN_CFF_AM (Proposed method)	Visible + Infrared	0.8676	0.9375	0.9244	0.9704	0.9383	1.0000	0.9397

3.4.2. Influence of Compression Rate r on the Classification Accuracy

We carry out experiments with different r in order to find the optimal r suitable for the attention mechanism. Figure 14 shows classification accuracy and average feature extraction time consumption per image using CNN_CFF_AM (refer Section 3.4.2) method under different r. The purpose of r is to reduce the number of parameters used. Since we design the attention mechanism in the feature fusion layer to be lightweight, the difference in the number of parameters caused by different compression rates is ignored. In Figure 14, it can be seen that when *r* is 8, a good balance is achieved between classification accuracy and average time consumption for feature extraction per image. As such, *r* is set to 8 in this study.



	0	1	2	3	4	5
0	59	3	1	1	12	0
1	1	60	0	1	1	0
2	0	1	55	0	3	0
3	0	0	1	131	4	0
4	0	1	3	1	190	0
5	0	0	0	0	0	20

(**a**)

	0	1	2	3	4	5
0	0.776	0.039	0.013	0.013	0.159	0
1	0.016	0.952	0	0.016	0.016	0
2	0	0.017	0.932	0	0.051	0
3	0	0	0.007	0.963	0.03	0
4	0	0.005	0.016	0.005	0.974	0
5	0	0	0	0	0	1

(**b**)

Figure 12. 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.

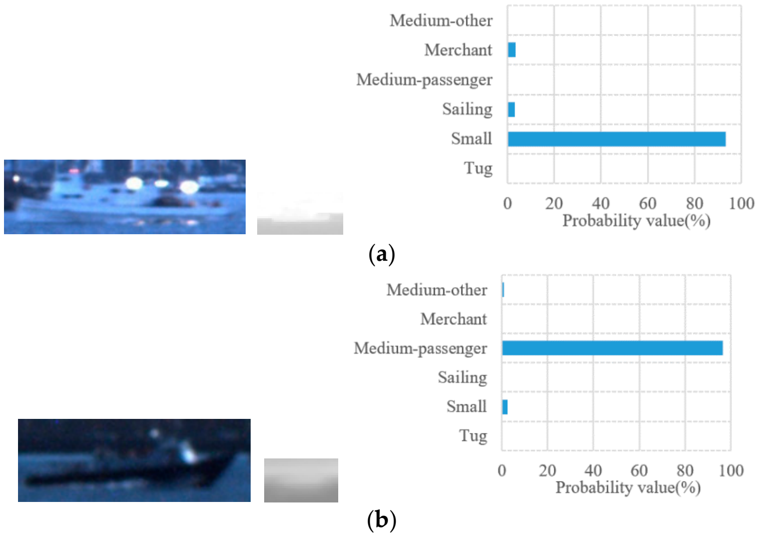


Figure 13. Example of misclassified ships. (**a**) medium-passenger is classified as small; and (**b**) small is classified as medium-passenger. Their corresponding classification probabilities into six categories are shown as indicated by blue bar chart.

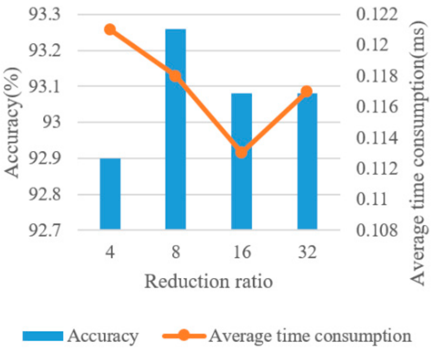


Figure 14. Classification accuracy and the average feature extraction time consumption per image of the CNN_CFF_AM (refer Section 3.4.2) method under different *r*.

4. Conclusions

In this study, the authors proposed the use of an attention mechanism and MSCNN method for accurate ship classification. Firstly, a two-stream symmetric MSCNN is adopted to extract the features of visible images and infrared images, and the two features are concatenated such that complementary features can be effectively utilized. After that, the attention mechanism is applied to the concatenated fusion layer to obtain more effective feature representation. Lastly, fused features after attention mechanism modification are sent to fully connected layers and the Softmax output layer to obtain the final classification result. In order to verify the effectiveness of the proposed method, we conduct experiment on the VAIS dataset. The results show that, compared with existing methods, the proposed method can achieve better classification performance, with a classification accuracy of 93.81%. Results from F1-score and confusion matrix further validate the effectiveness of the proposed method. However, in the presence of high intra-class similarity, the proposed method still results in some degree of misclassification, and increases the average feature extraction time consumption per image slightly. In future research, we will consider exploring and researching how to select the fused features while maintaining high classification accuracy to improve the efficiency of the method.

Author Contributions: Y.R. conceived the manuscript and conducted the whole experiments. J.Y. supervised the experiments and helped discuss the proposed the method. Z.G. and Q.Z. contributed to give helpful suggestions on this research. H.C. supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Nature Science Foundation of China (Grant No. 51879211), the National Key Research and Development Project of China (2020YFB1710800), 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. Sharifzadeh, F.; Akbarizadeh, G.; Seifi Kavian, Y. Ship classification in SAR images using a new hybrid CNN-MLP classifier. *J. Indian Soc. Remote Sens.* **2019**, *47*, 551–562. [CrossRef]
2. Kanjir, U.; Greidanus, H.; Oštir, K. Vessel detection and classification from spaceborne optical images: A literature survey. *Remote Sens. Environ.* **2018**, *207*, 1–26. [CrossRef] [PubMed]
3. Xu, G.J.; Wang, J.Y.; Qi, S.X. Ship detection based on rotation-invariant HOG descriptors for airborne infrared images. In Proceedings of the SPIE 10609, MIPPR: Pattern Recognition and Computer Vision, Xiangyang, China, 28–29 October 2017; p. 1060912. [CrossRef]
4. Li, W.; Chen, C.; Su, H.; Du, Q. Local binary patterns and extreme learning machine for hyperspectral imagery classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3681–3693. [CrossRef]
5. Xu, F.; Han, S.K. Ship targets recognition method based on moments and SVM. *Transducer Microsyst. Technol.* **2018**, *3*, 43–45.
6. Parameswaran, S.; Rankey, K. Vessel classification in overhead satellite imagery using weighted "bag of visual words". In Proceedings of the SPIE 9476, Automatic Target Recognition XXV, Baltimore, MD, USA, 20–22 April 2015; p. 947609. [CrossRef]
7. Zhang, S.; Wu, G.S.; Gu, J.H.; Han, J.O. Pruning convolutional neural networks with an attention mechanism for remote sensing image classification. *Electronics* **2020**, *9*, 1209. [CrossRef]
8. Qi, L.; Li, B.Y.; Chen, L.K.; Wang, W.; Dong, L.; Jia, X.; Huang, J.; Ge, C.W.; Xue, G.M.; Wang, D. Ship target detection algorithm based on improved Faster R-CNN. *Electronics* **2019**, *8*, 959. [CrossRef]
9. Zhang, J.; Wang, W.; Lu, C.; Wang, J.; Sangaiah, A.K. Lightweight deep network for traffic sign classification. *Ann. Telecommun.* **2020**, *75*, 369–379. [CrossRef]
10. 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]
11. 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]

12. Kasun, L.L.C.; Zhou, H.M.; Huang, G.B.; Vong, C.M. Representational learning with ELMs for big data. *IEEE Intell. Syst.* **2013**, *28*, 31–34.
13. Yoo, Y.W.; Oh, S.Y. Fast training of convolutional neural network classifiers through extreme learning machines. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 1702–1708. [CrossRef]
14. Li, Z.Z.; Zhao, B.J.; Tang, L.B.; Li, Z.; Feng, F. Ship classification based on convolutional neural networks. *J. Eng.* **2019**, *21*, 7343–7346. [CrossRef]
15. Zhang, E.H.; Wang, K.L.; Lin, G.F. Classification of Marine Vessels with Multi-Feature Structure Fusion. *Appl. Sci.* **2019**, *9*, 2153. [CrossRef]
16. Liu, F.; Shen, T.S.; Ma, X.X. Convolutional Neural Network Based Multi-Band Ship Target Recognition with Feature Fusion. *Acta Opt. Sin.* **2017**, *37*, 1015002. [CrossRef]
17. Chen, X.Q.; Yang, Y.S.; Wang, S.Z.; Wu, H.; Tang, J.; Zhao, J.; Wang, Z. Ship type recognition via a coarse-to-fine cascaded convolution neural network. *J. Navig.* **2020**, *73*, 813–832. [CrossRef]
18. 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]
19. Aziz, K.; Bouchara, F. Multimodal deep learning for robust recognizing maritime imagery in the visible and infrared spectrums. In Proceedings of the International Conference Image Analysis and Recognition 2018 (ICIAR 2018), LNCS 10882, Póvoa de Varzim, Portugal, 27–29 June 2018; pp. 235–244. [CrossRef]
20. Huang, S.Z.; Xu, H.S.; Xia, X.Z.; Yang, F.; Zou, F.H. Multi-feature fusion of convolutional neural networks for fine-grained ship classification. *J. Intell. Fuzzy Syst.* **2019**, *37*, 125–135. [CrossRef]
21. Jia, H.R.; Ni, L. Marine ship recognition based on cascade CNNs. In Proceedings of the SPIE 11427, Second Target Recognition and Artificial Intelligence Summit Forum, Changchun, China, 31 January 2020; p. 114270A. [CrossRef]
22. Hu, J.; Shen, L.; Gang, S. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. [CrossRef]
23. Zhao, X.L.; Zhang, J.; Tian, J.M.; Zhuo, L.; Zhang, J. Residual dense network based on channel-spatial attention for the scene classification of a high-resolution remote sensing image. *Remote Sens.* **2020**, *12*, 1887. [CrossRef]
24. 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 (CVPRW), Boston, MA, USA, 7–12 June 2015; pp. 10–16. [CrossRef]
25. Christian, S.; Liu, W.; Jia, Y.Q.; Pierre, S.; Scott, R.; Dragomir, A.; Dumitru, E.; Vincent, V.; Andrew, R. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
26. Liu, F.; Shen, T.S.; Ma, X.X.; Zhang, J. Ship recognition based on multi-band deep neural network. *Opt. Precis. Eng.* **2017**, *25*, 2939–2946. [CrossRef]
27. Bottou, L. Large-scale machine learning with stochastic gradient descent. In Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT 2010), Paris, France, 22–27 August 2010; pp. 177–186. [CrossRef]
28. Shalabi, L.A.; Shaaban, Z.; Kasasbeh, B. Data Mining: A Preprocessing Engine. *J. Comput. Sci.* **2006**, *2*, 735–739. [CrossRef]
29. Liu, W.B.; Zou, Z.Y.; Xing, W.W. Feature Fusion Methods in Pattern Classification. *J. Beijing Univ. Posts Telecommun.* **2017**, *40*, 1–8.
30. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent Models of visual attention. In Proceedings of the Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 13 December 2014; pp. 2204–2212.
31. Wu, P.; Cui, Z.; Gan, Z.; Liu, F. Residual group channel and space attention network for hyperspectral image classification. *Remote Sens.* **2020**, *12*, 2035. [CrossRef]
32. Woo, S.; Park, J.C.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]

33. 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, Baltimore, MD, USA, 17–21 April 2016; p. 984409. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.