Scientific Research Publishing

Journal of Computer and Communications, 2020, 8, 23-43
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Deep Convolutional Feature Fusion Model for Multispectral Maritime Imagery Ship Recognition

Xiaohua Qiu[1,2], Min Li[1], Liqiong Zhang[1], Rui Zhao[2]

[1]Xi'an Research Institute of Hi-Tech, Xi'an, China
[2]School of Information Engineering, Engineering University of PAP, Xi'an, China
Email: qxh_1025@163.com, proflimin@163.com

How to cite this paper: Qiu, X.H., Li, M., Zhang, L.Q. and Zhao, R. (2020) Deep Convolutional Feature Fusion Model for Multispectral Maritime Imagery Ship Recognition. *Journal of Computer and Communications*, **8**, 23-43.
https://doi.org/10.4236/jcc.2020.811003

Received: October 9, 2020
Accepted: November 9, 2020
Published: November 12, 2020

Abstract

Combining both visible and infrared object information, multispectral data is a promising source data for automatic maritime ship recognition. In this paper, in order to take advantage of deep convolutional neural network and multispectral data, we model multispectral ship recognition task into a convolutional feature fusion problem, and propose a feature fusion architecture called Hybrid Fusion. We fine-tune the VGG-16 model pre-trained on ImageNet through three channels single spectral image and four channels multispectral images, and use existing regularization techniques to avoid over-fitting problem. Hybrid Fusion as well as the other three feature fusion architectures is investigated. Each fusion architecture consists of visible image and infrared image feature extraction branches, in which the pre-trained and fine-tuned VGG-16 models are taken as feature extractor. In each fusion architecture, image features of two branches are firstly extracted from the same layer or different layers of VGG-16 model. Subsequently, the features extracted from the two branches are flattened and concatenated to produce a multispectral feature vector, which is finally fed into a classifier to achieve ship recognition task. Furthermore, based on these fusion architectures, we also evaluate recognition performance of a feature vector normalization method and three combinations of feature extractors. Experimental results on the visible and infrared ship (VAIS) dataset show that the best Hybrid Fusion achieves 89.6% mean per-class recognition accuracy on daytime paired images and 64.9% on nighttime infrared images, and outperforms the state-of-the-art method by 1.4% and 3.9%, respectively.

Keywords

Deep Convolutional Neural Network, Feature Fusion, Multispectral Data, Object

Recognition

1. Introduction

By integrating complementary information from visible (VIS) and infrared (IR) images, multispectral data has recently received much attention in machine learning and computer vision [1] [2] [3] [4] [5]. VIS images are sensitive to variation illumination and unfavourable weather conditions, which degrade the performance of computer vision systems built on these images. Thermal camera can ameliorate the problem, but it cannot provide image with the same high-resolution as visible camera, and often exhibit a decrease in image quality during daytime due to a high background temperature. Therefore, multispectral images have been successfully used to face recognition [6] [7] [8] [9], and are also widely applied to object recognition [10], person re-identification [11], pedestrian detection [12], and object tracking [13] by exploiting deep learning in recent years.

As known to all, after the breakthrough research by Krizhevsky *et al.* [14], deep convolutional neural networks (CNN) have achieved remarkable success for a large variety of tasks, and quickly became the dominant tool in computer vision. Meanwhile, some well-known deep CNN models have been reported, such as Oxford VGG Model [15], Google Inception Model [16] and Microsoft ResNet Model [17]. One factor for the dramatic improvement in performance of deep CNN is that many challenging datasets for training with millions of labeled examples are harvested from the web, such as ImageNet [18]. However, a large-scale training set is expensive or difficult to collect in the real world, and training a large neural network on a small dataset would lead to poor performance due to the problem of overfitting. The lack of a large-scale training set forces the computer vision community to find practical workarounds. Much recent effort [19] [20] [21] has been dedicated to developing methods that fine-tune the well-known pre-trained deep CNN models or directly take these models as feature extractors. Research in vision tasks based on multispectral data follows the same trend, e.g., action recognition [22], pedestrian detection [23], object recognition [10]. In the previous works on multispectral data, whether fine-tuning after feature fusion or directly extracting feature without fine-tuning, features are produced at the same layer of the pre-trained deep CNN model for VIS and IR images. However, due to the aforementioned difference between VIS and IR images, features extracted from the same layer may not both be the best, so feature fusion cannot fully take advantage of multispectral data. Therefore, how features of VIS and IR images can be properly fused in pre-trained or fine-tuned deep CNN model to achieve the best performance in vision task remains to be solved.

In this paper, we focus on using the pre-trained or fine-tuned deep CNN

model to extract features of VIS and IR image, and propose a novel feature fusion architecture called Hybrid Fusion for multispectral maritime ship recognition. We firstly model the multispectral maritime ship recognition task to a convolutional feature fusion problem, and then evaluate the feature representation ability of the pre-trained or fine-tuned deep CNN model for multispectral data. Thirdly, Hybrid Fusion and the other three feature fusion architectures are investigated. Finally, we compare Hybrid Fusion with the other reported methods. Our idea is that combining high-level feature of VIS image and middle-level feature of IR image can provide rich multispectral information to the classifier for the final prediction. Due to the large gap of feature values at different layers, a features normalization method is exploited. Meanwhile, based on different feature extractors used by VIS and IR images, three combinations are also investigated.

Our major contribution is fourfold: First, we propose a feature fusion architecture named Hybrid Fusion, which combines high-level feature of VIS image and middle-level feature of IR image. Second, we investigate four distinct feature fusion architectures, namely Early Fusion, Halfway Fusion, Late Fusion and Hybrid Fusion, and evaluate these fusion architectures on the public multispectral maritime ship images, the VAIS dataset [10]. Third, we fine-tune the pre-trained VGG-16 model on both single spectral image and multispectral images, and also exploit three existing regularization techniques to avoid over-fitting problem. Fourth, the best Hybrid Fusion performs 89.6% mean per-class recognition accuracy on the daytime paired images of VAIS dataset, outperforms the state-of-the-art method by 1.4%, and also achieves 64.9% on nighttime and 68.6% on all time IR images.

2. Related Work

Object recognition with deep convolutional feature fusion. Initializing with transferred features whether features are transferred from the low-level, middle-level or high-level of the pre-trained deep CNN, can improve generalization performance even after substantial fine-tuning on a new task [24]. Schwarz *et al.* [25] presented feature fusion model for multi-modal object recognition, a pre-trained AlexNet model [14] is exploited to extract features from the last two fully connected layers. An extension of the fusion model further improves object recognition accuracy by fine-tuning the pre-trained AlexNet with multi-modal training data [19]. Furthermore, Zia *et al.* [26] proposed a hybrid 2D/3D convolutional neural network initialized by the pre-trained VGG-16 model [15], and fused the features separately extracted from the fully connected layers of three network architectures. Another interesting work [27] presented an unsupervised feature learning framework. In this framework, the pre-trained VGG-f model [28] is taken as a feature extractor, and then recursive neural network [29] is used to reduce dimension of the extracted features and learn high-level features. The aforementioned methods focus not only on convolutional feature fusion but

also on the processing of modal data. The goal of our work is how to leverage convolutional feature fusion and limited multispectral data to maximize ship recognition accuracy.

Ship recognition on single spectral image. Kanjir *et al.* [30] provided an overview of existing literature on ship detection and classification from optical satellite imagery. However, most of the reviewed methods are performed on optical remote sensing images, and our work focuses on ship recognition from VIS and IR images. Therefore, we mainly review the works of vessel/ship recognition on VIS image due to little of ones on IR image. Khellal *et al.* [31] used extreme learning machine (ELM) to learn discriminative CNN feature for IR image maritime ship recognition. Fouad *et al.* [32] presented an experimental study to investigate the ability of deep CNN features to catch details of VIS image maritime ships for fine-grained classification. Cuong *et al.* [33] trained and tested AlexNet with a dataset of 130,000 VIS images of maritime ships, which are collected from website *ShipSpotting*[1]. Gundogdu *et al.* [34] [35] introduced a large-scale VIS image maritime vessels dataset, namely MARVEL, for the fine-grained visual categorization, recognition, retrieval and verification tasks. To achieve the baseline results, both extracting feature from pre-trained VGG-f model and training AlexNet model have been used to perform the aforementioned tasks. To improve the performance of the tasks on MARVEL dataset, Solmaz *et al.* [36] exploited a multi-task learning framework based on deep CNN models to accompany deep metric learning with a proposed loss function. Milicevic *et al.* [37] used the training dataset of MARVEL to fine-tune VGG-19 model [15] pre-trained on ImageNet, then boosted the recognition accuracy by 3%. Huang *et al.* [38] exploited low-level and high-level features to classify ship categories on VIS images. The proposed method learns the high-level features via fine-tuning pre-trained deep CNN model, and incorporates the multi-scales rotation invariant features obtained by Gabor filter and multi-scale completed local binary patterns (MS-CLBP), then these features are fed into support vector machine (SVM) classifier. This method was extended to improve recognition performance by replacing SVM with ELM classifier in [39]. Shi *et al.* [40] proposed a classification framework consists of a multi-feature ensemble based on convolutional neural network (ME-CNN).

Ship recognition on multispectral images. Currently, there are few literature about multispectral maritime ship recognition due to the lack of corresponding multispectral data. VAIS dataset including VIS and IR images is the only public multispectral maritime ship dataset for image classification or object recognition. Zhang *et al.* [10] reported the VAIS dataset in detail, and combined the results of gnostic fields and deep CNN to provide the baseline recognition accuracy on this dataset, 87.4% mean per-class recognition accuracy during the daytime and 61% at nighttime. They also tried to fine-tune the pre-trained VGG-16 model, but failed in improving recognition performance. Aziz *et al.* [41] used a

[1]www.shipspotting.com.

large-scale visible ship dataset to train a deep CNN, and then fine-tuned their pre-trained CNN model with the training images of VAIS dataset. Santos *et al.* [42] proposed a decision level fusion of convolutional neural networks using a probabilistic model, in which features are extracted from the last convolutional activate map of the pre-trained VGG-19 model. Zhang *et al.* [43] presented a multi-feature structure fusion based on spectral regression discriminant analysis (SF-SRDA) by combining structural fusion with linear discriminant analysis, and used the pre-trained models VGG-19 and ResNet-152 [17] to achieve a promising result. The above work has achieved good ship recognition performance. However, they did not consider the difference between each convolutional layer of the pre-training or fine-tuning models for different spectrum image ship recognition, and our work considers this difference and proposes a Hybrid fusion model based on this difference.

3. Proposed Feature Fusion Method

Intuitively, VIS and IR images provide auxiliary visual information to each other in depicting ship objects. Encouraged by the recent tremendous advances in deep learning techniques, as well as inspired by the work of multispectral pedestrian detection [44], we explore the effectiveness of using the VGG-16 model pre-trained on ImageNet dataset and fine-tuned on VAIS dataset to perform multispectral ship recognition. The structure of our method is shown in **Figure 1**.

The proposed fusion framework mainly includes four stages:

1) **Image preprocessing:** as the pre-trained VGG-16 model expects 224×224 pixels and three channels images as input, we simply clone the single IR channel three times. Meanwhile, both VIS and IR images are resized to 224×224 using nearest interpolation.

2) **Feature extraction:** there are two feature extraction branches, visible image branch (shorted as VGG-16-VIS) and infrared image branch (shorted as VGG-16-IR), as shown in **Figure 1**. Each branch takes the pre-trained or fine-tuned VGG-16 model as feature extractor. Besides, image features of both branches are extracted from the same layer or different layers of VGG-16 model according to feature fusion architectures.

3) **Feature fusion:** the features extracted from the two branches are flattened to feature vectors, and then are concatenated to produce a multispectral feature vector representing the maritime ship.

4) **Classification:** before the fused feature vector is fed into a linear SVM classifier for the final prediction, feature vector is normalized by l2-norm (shorted as L2) normalization method. According to Hybrid Fusion, feature vector should be normalized before feature fusion because of the large gap of feature values at different layers.

Additionally, the training samples of VIS and IR images in the VAIS dataset are used to fine-tune the pre-trained VGG-16 model in an end-to-end way, respectively. Then, the two fine-tuned VGG-16 models are also taken as feature extractors.

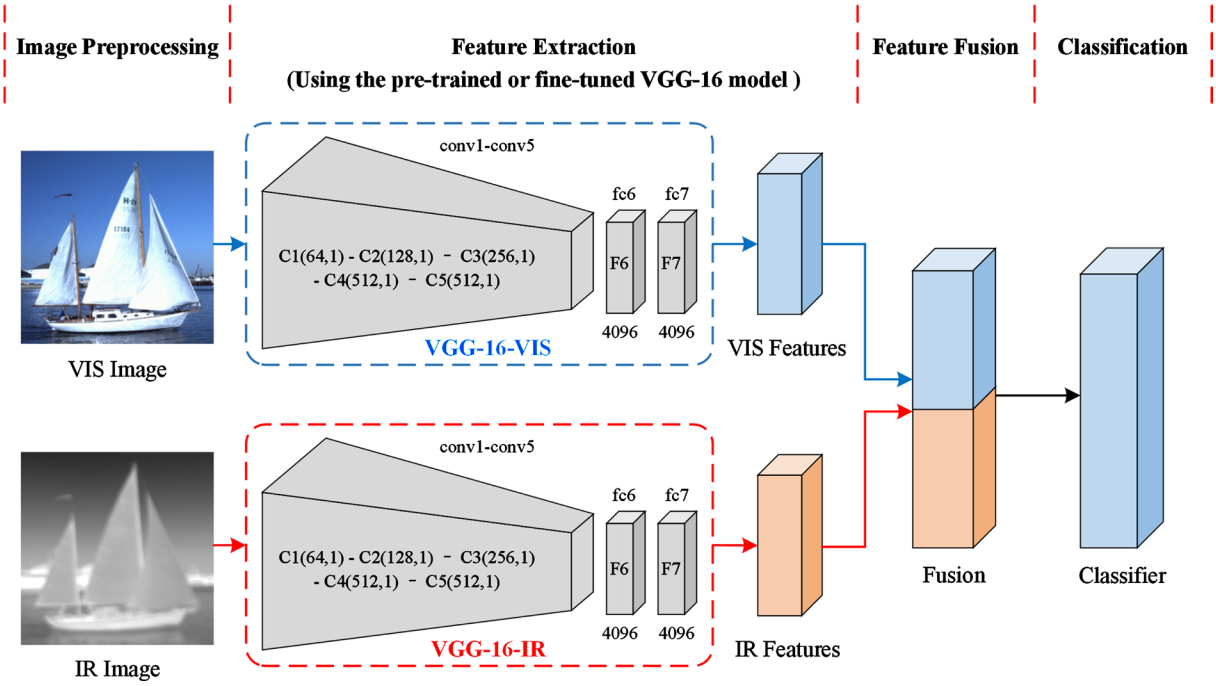

Figure 1. The illustration of overview pipeline for multispectral maritime ship recognition. The proposed fusion framework takes VIS-IR image pair as the input and outputs ship image category. VGG-16-VIS and VGG-16-IR are the VGG-16 model pre-trained on ImageNet dataset or fine-tuned on VAIS dataset. Features of VIS and IR images are extracted from VGG-16-VIS and VGG-16-IR, respectively. C1 denotes the first convolutional layer, the same as to C2, C3, C4 and C5. F6 and F7 represent the first and second fully connected layers, respectively.

3.1. Feature Fusion Architecture

Due to features at different levels of VGG-16 correspond to various levels of semantic information and fine visual details [45], feature fusion at different layers would lead to different recognition results. Therefore, the multispectral ship recognition task is modelled into a convolutional feature fusion problem, *i.e.*, which feature fusion architecture could get best recognition performance. Then, we propose a feature fusion architecture called Hybrid Fusion, which combines high-level feature of VIS image and middle-level feature of IR image. We investigate Hybrid Fusion as well as Early Fusion, Halfway Fusion and Late Fusion. These fusion architectures integrate two-branch convolutional features at different layers of VGG-16 model, as shown in **Figure 2**. Each branch represents a single spectral image.

Early Fusion combines the feature maps from VIS and IR images immediately after the first and second convolutional layers (C1 and C2 layers) followed by a Max Pool layer (this fusion architecture is ignored in **Figure 2**). Since C1 and C2 layers capture low-level visual features, such as color, corners and line segments. This fusion architecture fuses features at low-level.

Halfway Fusion also implements feature fusion at convolutional layers. Different from Early Fusion, it fuses the features after the third, fourth and fifth convolutional layers (C3 - C5 layers) followed by a Max Pool layer, as shown in **Figure 2(a)**. Features from C3 - C5 layers contain more semantic information

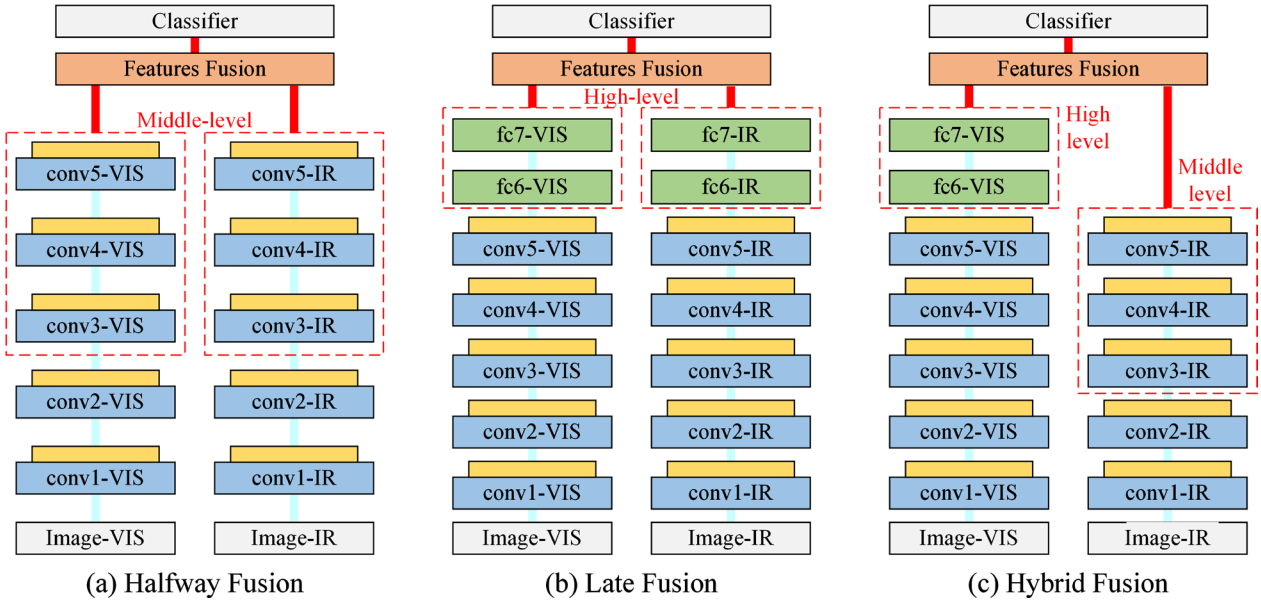

Figure 2. Feature fusion architectures for multispectral maritime ship recognition. Blue and green boxes represent convolutional and fully connected layers including ReLU layers, respectively. Yellow boxes represent Max Pool layer. Orange boxes denote feature fusion at this location. (Best viewed in color.)

than C1 and C2 layers features, while retain some fine visual details. The fusion architecture fuses features at middle-level.

Late Fusion combines features extracted from the first and second fully connected layers (F6 and F7 layers) followed by an activation layer named ReLU, which performs feature fusion at fully connected stage, as shown in **Figure 2(b)**. Conventionally, F6 and F7 layers features are used as new representations of ship objects. This fusion architecture executes high-level feature fusion.

Hybrid Fusion combines high-level feature of VIS image and middle-level feature of IR images, that is F6 and F7 layers features of VIS images and C3-C5 layers features of IR images, as shown in **Figure 2(c)**, due to different feature representation of the VGG-16 model at different levels for each spectral image. Hybrid Fusion leverages the feature representation of different levels for multispectral images.

3.2. Feature Fusion Method

After extracting two-branch convolutional features from different levels of the pre-trained or fine-tuned VGG-16 model, each branch features are flattened to a feature vector. Following the work based on multispectral data in [46], concatenation fusion method is used to fuse two feature vectors. The fusion goal is to integrate two feature vectors F^{VIS} and F^{IR} to a fused feature vector F^{F}, where F^{VIS}, F^{IR} denote feature vector of VIS and IR images, respectively. The concatenation fusion method is to directly concatenate two feature vectors, which can be defined as:

$$F^{F} = f^{concat}\left(F^{VIS}, F^{IR}\right), \tag{1}$$

$$F_{d_1}^{F} = F_{d_1}^{VIS} \tag{2}$$

$$F_{D_1+d_2}^{F} = F_{d_2}^{IR} \tag{3}$$

where $F_{d_1}^{VIS}$ denotes the d_1^{th} value of F^{VIS}, $F_{d_2}^{IR}$ denotes the d_2^{th} value of F^{IR}, $F_{d_1}^{F}$ is the d_1^{th} value of F^{F}, $F_{D_1+d_2}^{F}$ is the $\left(D_1+d_2\right)^{th}$ value of F^{F}. $1 \le d_1 \le D_1$, $1 \le d_2 \le D_2$ and $F^{VIS} \in \mathbb{R}^{D_1}$, $F^{IR} \in \mathbb{R}^{D_2}$, $F^{F} \in \mathbb{R}^{D_1+D_2}$. This fusion method concatenates the dimensions of the two input feature vectors.

3.3. Normalization and Classification

In order to evaluate the multispectral ship recognition performance of four feature fusion architectures, we exploit a linear SVM as classifier. It is crucial to normalize the feature vector before putting it into the linear SVM classifier. The reason is threefold: First, it avoids the feature characteristic of small value range to be over-branched by the feature characteristic of large value range, so as to improve the performance of linear SVM classifier. Second, it adjusts values measured on different scales to a same scale, and then facilitates data comparison and common processing. Third, it reduces numerical value complexity in calculation. To normalize the features of train data and test data, we use L2 normalization method. L2 normalization is a normalization method commonly used in machine learning. The main idea is to divide each element in a vector by the L2 norm of the vector, that is defined as formulas Equation (4).

$$x_d^{L2} = \frac{x_d}{\sqrt{\sum_{d=1}^{D} \left|x_d\right|^2}} \tag{4}$$

where x_d^{L2} represents the d^{th} value of the D dimension feature vector after L2 normalization, x_d is the d^{th} value of the D dimension feature vector, $\left|\ \right|$ denotes an absolute operator and $1 \le d \le D$.

4. Experiments

4.1. Dataset

To investigate our four feature fusion architectures for ship category recognition, we use the publicly available VAIS dataset [10]. For now as we know, it is the only existing public database of paired VIS and IR ship imagery. The dataset contains 2865 images (1623 VIS images and 1242 IR images), in which 1088 "VIS-IR" unregistered images pairs and 154 nighttime IR images, and includes 6 categories: cargo ships, medium-other ships, passenger ships, sailing ships, small boats and tug boats. However, the images are captured at different distance and various times of one day, including dusk and dawn. Therefore, some images are high-resolution while a part of images may appear dim and hard to recognize even with manual inspection. In the dataset, the paired VIS-IR image set is partitioned into 539 image pairs for training and 549 image pairs for testing. A sample pairs from VAIS is illustrated in **Figure 1** and **Table 1** shows the number of train and test samples for each class. As followed the baseline method [10], the

Table 1. The number of train and test images for each class in the paired images, nighttime and all time IR images of VAIS dataset.

VAIS	Time	Cargo	Medium-other	Passenger	Sailing	Small	Tug	Total
Train set	Daytime	83	62	58	148	158	30	539
Test set	Daytime	63	76	59	136	195	20	549
	Nighttime	34	14	12	15	30	49	154
	All time	97	90	71	151	225	69	703

same train data and test data are used and the mean per-class recognition accuracy is taken as the evaluation measurement in the experiments.

4.2. Implementation Platform and Details

Our processing platform is a personal computer with Ubuntu 16.04, with a single CPU (4.20 GHz) of an Intel Core i7-7770K with 16 GB random access memory (RAM). An NVIDIA GTX1080Ti Graphics PU is used for deep CNN computations. The computation environment is a Keras environment with TensorFlow backend, which is a high-level neural network application programming interface written in Python. Our experiment is divided into two stages: features are first extracted and stored subsequently, then are fed into linear SVM classifier. We use LibSVM toolbox [47], which has been packaged as a module of *scikit-learn*[2], as classifier to implement ship classification, the relaxation coefficient C is set to 10, kernel function is set to *linear*. Due to limited RAM, we did not perform experiments on feature fusion at the first convolutional layer, but experiments at the second convolutional layer can reflect the performance of Early Fusion.

It is not easy to fine-tune the pre-trained VGG-16 model end to end on small-scale dataset like VAIS, especially on IR images. The main problem is how to avoid over-fitting and take into account model convergence during fine-tuning model. Some existing regularization techniques [48], such as data argumentation, dropout and L^2 parameter regularization known as weight decay, are used to fine-tune the pre-trained model on VIS and IR images. Additionally, in order to investigate whether the VGG-16 model learn fusing inputs implicitly, 4 channels multispectral image consisting of VIS and IR images (shorted as 4C VIS-IR) are also taken as inputs to fine-tune the model. In fine-tuning experiment, the initial learning rate is set as 0.001 for VIS images and 0.0001 for IR images and 4C VIS-IR images. Stochastic gradient descent optimizer is utilized for optimization, the momentum is set to 0.9, and the decay is set as 0.00001. The train step is set to 50 epochs, the batch size is set to 32. Random horizontal flip, random vertical flip are used for online data argumentation. Dropout is applied after the second fully connected layer and its rate is set to 0.5. L^2 weight decay is applied on the last fully connected layer and its value is set to 0.1.

[2]https://scikit-learn.org.

4.3. Experimental Results

4.3.1. Evaluation of the Pre-Trained and Fine-Tuned Models

Firstly, we evaluate the effects of existing regularization techniques during fine-tuning VGG-16 model. Table 2 shows the comparison of recognition performance with and without regularization techniques. Accuracy evaluation uses the average value together with standard deviation in 10 groups of fine-tuning experiments. Figure 3 and Figure 4 give the accuracy and loss curves of fine-tuning VGG-16 model on VIS and IR images in one group of experiments, respectively. As shown in Table 2, using data argumentation greatly improves the recognition accuracy of 4C VIS-IR images, and combining three regularization techniques achieves the best results. However, compared to using data argumentation for VIS and IR images, fine-tuning with dropout or L^2 weight decay has slightly higher average value and smaller standard deviation. A combination of two or more regularization techniques cannot significantly improve the performance of fine-tuned model. Combining dropout and data argumentation even leads to model degradation when fine-tuning on VIS images. Furthermore, it can be observed from Figure 3 and Figure 4 that over-fitting problem is worse on IR images than VIS images. Over-fitting on VIS images is easy to overcome by using any of the three regularization techniques, as shown in Figure 3. However, data argumentation and dropout make accuracy and loss fluctuate too much for IR images, and the being fine-tuned model is difficult to converge, as shown in Figure 4. L^2 weight decay can restrain the loss ascension of IR images after about 20 epochs, and model starts to converge. In summary, considering over-fitting problem and model convergence, we exploit dropout regularization technique for fine-tuning model on VIS images, and L^2 weight decay for fine-tuning model on IR images. Meanwhile, the fine-tuned models on VIS and IR images, in which the accuracy is close to the corresponding average value of 10 groups experiments, are chosen as feature extractor in our fusion method.

Secondly, we analyze the feature representation ability of different layers on the pre-trained VGG-16 model for VIS and IR images. As the horizontal axis shown in Figure 5(a), C2 is low-level layer, C3 - C5 are middle-level layers, and

Table 2. The comparison of recognition performance (%) with and without regularization techniques during fine-tuning VGG-16 model.

Type	DR	L^2	Without data argumentation			With data argumentation		
			4C VIS-IR	VIS	IR	4C VIS-IR	VIS	IR
Type 1	0.0	0.0	81.2 ± 2.2	85.6 ± 2.7	67.0 ± 2.0	82.5 ± 1.5	85.7 ± 2.0	66.2 ± 2.3
Type 2	0.5	0.0	81.7 ± 1.0	86.2 ± 1.3	67.5 ± 1.2	83.4 ± 1.5	83.9 ± 3.5	65.9 ± 2.0
Type 3	0.0	0.1	81.6 ± 1.4	85.7 ± 1.7	67.5 ± 1.6	83.1 ± 1.4	85.5 ± 2.3	67.5 ± 2.4
Type 4	0.5	0.1	82.9 ± 1.7	85.4 ± 1.7	68.0 ± 2.2	83.6 ± 1.2	83.6 ± 3.4	66.5 ± 1.3

Notes: Accuracy evaluation uses the average value together with standard deviation in 10 times. Setting dropout rate (DR) or L^2 weight decay (L^2) to 0.0 means that it does not use dropout or L^2 weight decay regularization technique.

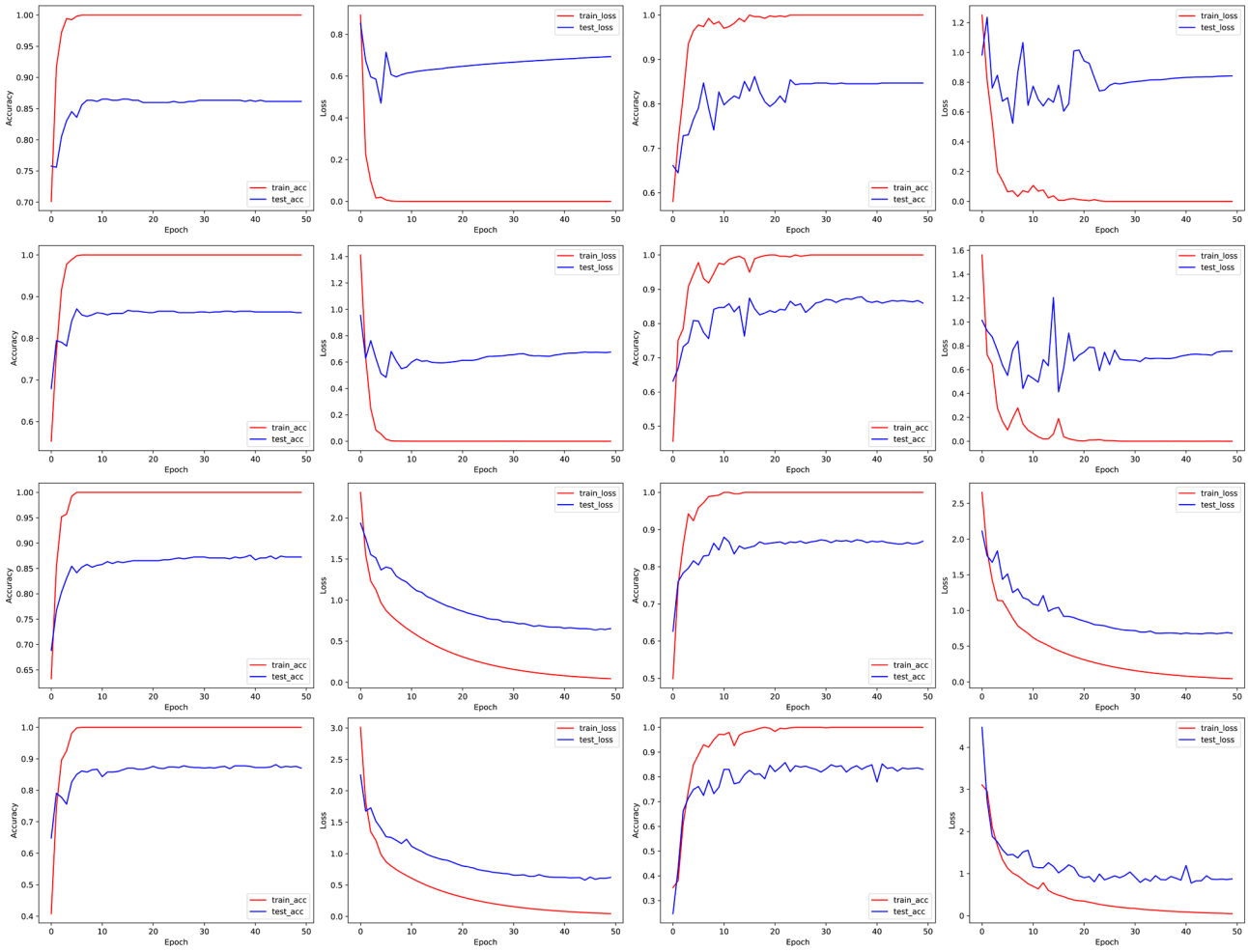

Figure 3. The accuracy and loss curves of the pre-trained VGG-16 model with fine-tuning on VIS images. Rows 1 - 4 correspond Type 1 - 4 in **Table 2**, respectively. Columns 1 - 2 are accuracy and loss curves of model fine-tuned on VIS images without data argumentation, and columns 3 - 4 are accuracy and loss curves of model fine-tuned on VIS images with data argumentation.

F6 - F7 are high-level layers. VIS image obtains the more feature representation at high-level layers (see block line with squares in **Figure 5(a)**) due to the pre-trained VGG-16 model is trained by a later-scale dataset of VIS image. However, IR image obtains more rich features at middle-level layers (see block dotted line with squares in **Figure 5(a)**) than high-level layers for ship recognition. The main reason is that IR images are different from VIS images, such as high contrast, low resolution and insufficient details. Meanwhile, we evaluate the effect on recognition accuracy of the two feature vector normalization methods. **Figure 5(a)** shows that L2 normalization improves the recognition performance of IR image at almost all layers (see blue dotted line with diamonds in **Figure 5(a)**). The main reason may be that IR images have more noise and are more blurry than VIS images, and L2 normalization eliminates the influence of these small values.

Thirdly, we evaluate the recognition performance of the fine-tuned models. For convenience, the pre-trained VGG-16 model without fine-tuning is shorted

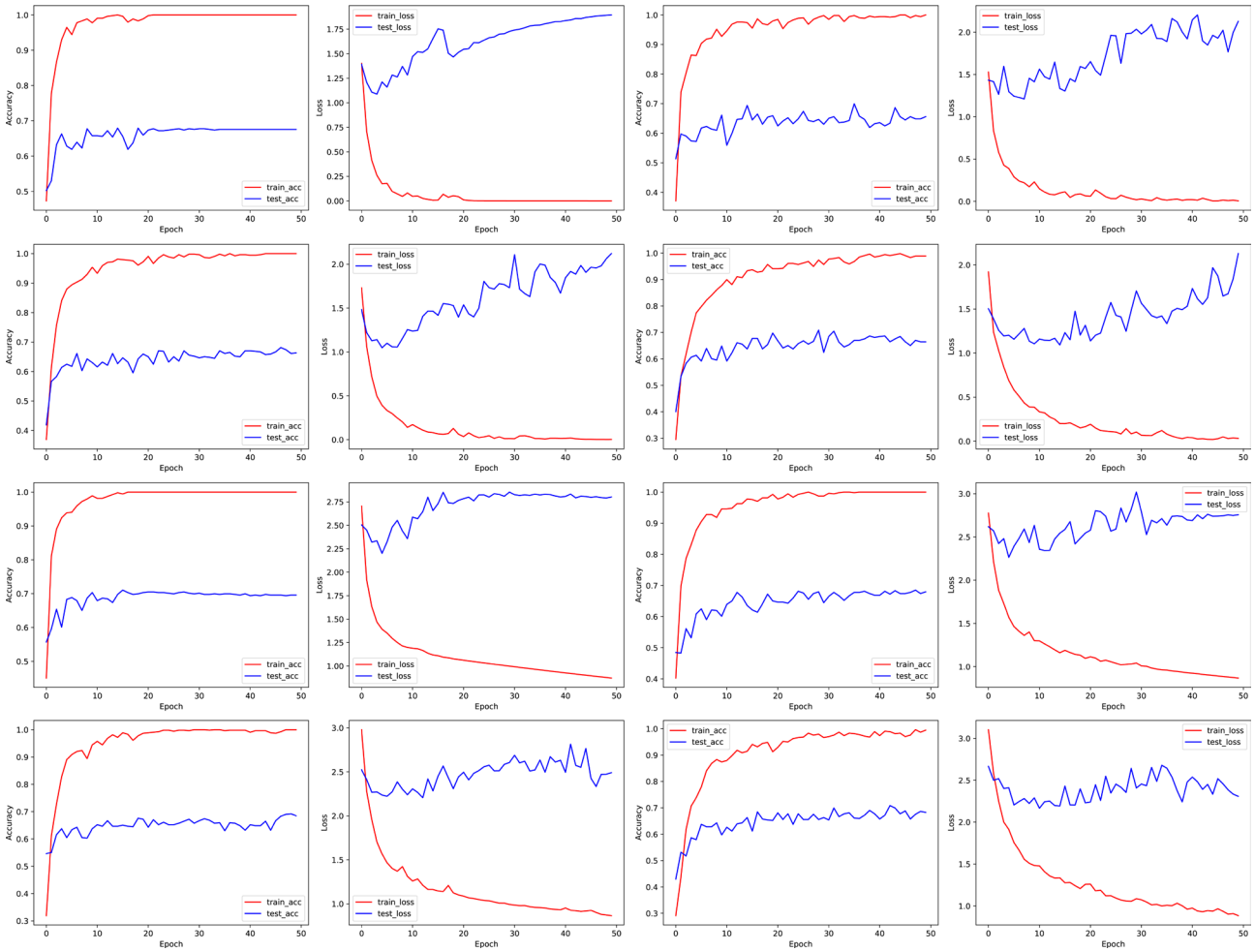

Figure 4. The accuracy and loss curves of the pre-trained VGG-16 model with fine-tuning on IR images. Rows 1 - 4 correspond Type 1 - 4 in **Table 2**, respectively. Columns 1 - 2 are accuracy and loss curves of model fine-tuned on IR images without data argumentation, and columns 3 - 4 are accuracy and loss curves of model fine-tuned on IR images with data argumentation.

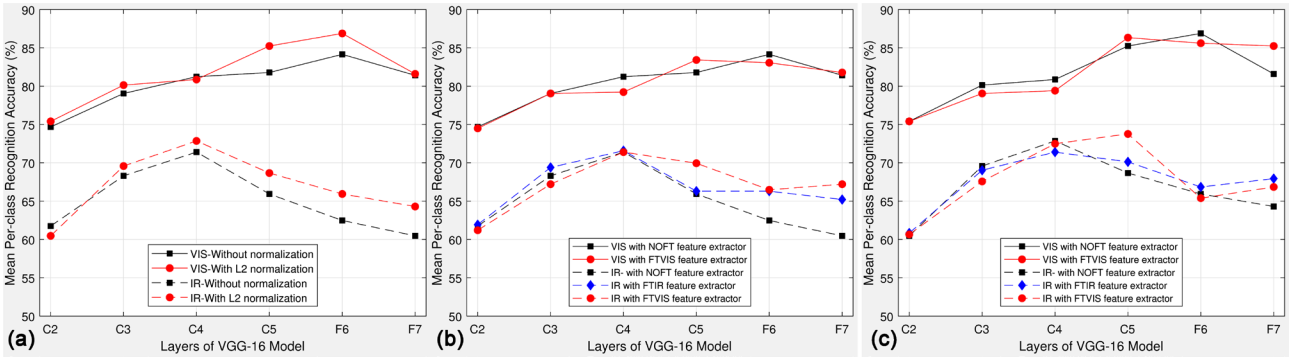

Figure 5. Evaluation feature representation ability of different layers on the pre-trained and fine-tuned VGG-16 model for VIS and IR images. (a) shows the evaluation of L2 normalization based on the pre-trained VGG-16 model without fine-tuning, (b) and (c) show the performance evaluation of fine-tuned models. (a) Without fine-tuning; (b) Without normalization; (c) L2 normalization.

as NOFT, the pre-trained VGG-16 model with fine-tuning on VIS images is shorted as FTVIS, and the pre-trained VGG-16 model with fine-tuning on IR images is shorted as FTIR. **Figure 5(b)** and **Figure 5(c)** show the comparison of

fine-tuned and pre-trained models without or with normalizations. Fine-tuning model on VIS images doesn't obviously improve the performance of layers on VGG-16 model (see red line with rounds in Figure 5(b) & Figure 5(c)). Fine-tuning model on IR images also doesn't obviously improve the performance of C2 - C4 layers on VGG-16 model, therefore it indicates that the low-level and middle-level layers of pre-trained VGG-16 model has strong generalization performance. However, it can be found that the recognition accuracy of C5, F6 and F7 layers on FTVIS and FTIR fine-tuned models are better than those of the NOFT model (see blue dotted line with diamonds and red dotted line with rounds in Figure 5(b) and Figure 5(c)). Thus, NOFT is taken as the feature extractor of VIS images, but the feature extractors of IR images are NOFT, FTIR and FTVIS. Therefore, the three combinations of feature extractors for VIS and IR images are investigated, as shown in Table 3.

4.3.2. Evaluation of Four Fusion Architectures

Firstly, we investigate the recognition performance of Early Fusion, Halfway Fusion and Late Fusion by using L2 normalization method along with three combinations. Due to feature extraction and feature fusion at the same layer for these three fusion architectures, feature is normalized for SVM classifier after features are fused. Figure 6 shows the recognition accuracy of three fusion architectures by using L2 normalization method. For an intuitive comparison, that of feature

Table 3. The three combinations of the different VGG-16 models taken as feature extractors for VIS and IR image.

Combination	Feature extractors of VIS and IR images	
	VIS	IR
Combination 1	NOFT	NOFT
Combination 2	NOFT	FTIR
Combination 3	NOFT	FTVIS

Notes: NOFT denotes the pre-trained VGG-16 model without fine-tuning, FTVIS denotes the pre-trained VGG-16 model with fine-tuning on VIS images, and FTIR denotes the pre-trained VGG-16 model with fine-tuning on IR images.

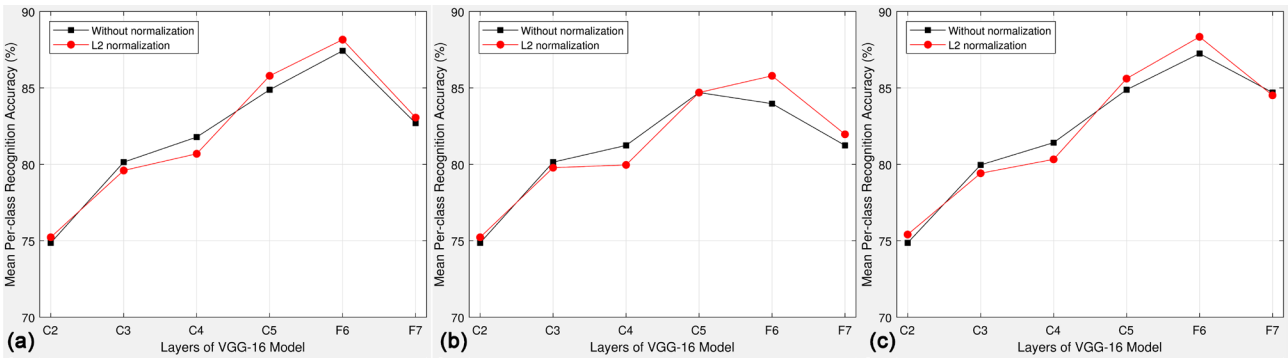

Figure 6. Evaluation recognition accuracy of Early Fusion, Halfway Fusion and Late Fusion in three combinations. (a) Combination 1 (VIS-NOFT, IR-NOFT), (b) Combination 2 (VIS-NOFT, IR-FTIR), (c) Combination 3 (VIS-NOFT, IR-FTVIS). (a) Combination 1; (b) Combination 2; (c) Combination 3.

fusion without normalization are also shown in Figure 6. As shown in the figure, using L2 normalization method greatly degenerate the recognition accuracy of feature fusion at C4 layer, but significantly improve at C5, F6 and F7 layers. It indicates that L2 normalization method facilitates feature representation of semantic information. Moreover, Late Fusion at F6 layer using L2 normalization almost achieves the best recognition accuracy among the three fusion architectures.

Secondly, Hybrid Fusion is compared to Late Fusion, which is the best of the above three fusion architecture. Hybrid Fusion integrates high-level feature of VIS image and middle-level feature of IR image, and there is the large gap of values at different layers, thus the extracted features are normalized before being fused. Table 4 shows the recognition accuracy of Late Fusion and Hybrid Fusion with L2 normalization method in three combinations. For each combination, Hybrid Fusion (F6C3) and Hybrid Fusion (F6C4) are better than Late Fusion (F6F6), but Hybrid Fusion (F6C5) is worse than Late Fusion (F6F6). Besides, Hybrid Fusion at F7 layer are also better than Late Fusion (F7F7) in all combinations.

4.3.3. Comparison with Other Reported Methods

We compare the proposed Hybrid Fusion with four methods for paired images: 1) the baseline method (CNN + Gnostic Fields) [10], 2) Multimodal CNN [41], 3) DyFusion [42], 4) SF-SRDA [43], and with three methods for VIS images in the paired images: 5) MFL (feature-level) + ELM [38], 6) CNN + Gabor + MS-CLBP [36], 7) ME-CNN [40], and with one method for all time IR images: 8) ELM-CNN [31]. Table 5 shows the comparison results using the mean pre-class recognition accuracy as evaluation measure. As shown in Table 5, Hybrid Fusion (F6C3) is 2.2% higher than the baseline method, outperforms the state-of-the-art (DyFusion) by 1.4% in daytime, and boosts the baseline method by 3.9% on

Table 4. The recognition accuracy (%) of Late Fusion and Hybrid Fusion with L2 normalization method in three combinations.

Fusion Architecture	Combination 1			Combination 2			Combination 3		
	VIS	IR	VIS + IR	VIS	IR	VIS + IR	VIS	IR	VIS + IR
LF(F6F6)	86.9	65.9	88.2	86.9	66.9	85.8	86.9	65.4	88.3
HF(F6C3)	86.9	69.6	88.7	86.9	69.0	**89.6**	86.9	67.6	88.5
HF(F6C4)	86.9	72.9	**89.1**	86.9	71.4	88.3	86.9	72.5	**88.9**
HF(F6C5)	86.9	68.7	85.8	86.9	70.1	85.3	86.9	73.8	87.4
LF(F7F7)	81.6	64.3	83.1	81.6	67.9	82.0	81.6	66.8	84.5
HF(F7C3)	81.6	69.6	85.6	81.6	69.0	86.7	81.6	67.6	85.4
HF(F7C4)	81.6	72.9	88.0	81.6	71.4	86.5	81.6	72.5	87.2
HF(F7C5)	81.6	68.7	84.3	81.6	70.1	84.3	81.6	73.8	86.3

Notes: Abbreviated symbol LF (F6F6) represents Late Fusion combining F6 layer features of VIS and IR images, the same as to LF (F7F7). Abbreviated symbol HF (F6C3) represents Hybrid Fusion combining F6 layer feature of VIS and C3 layer feature of IR, the same as to others. Bold denotes the recognition accuracy is the best one in the same combination.

Table 5. Comparison of recognition accuracy (%) with other results reported on the VAIS dataset.

Method	Daytime			Nighttime IR	All time IR
	VIS	IR	VIS + IR		
CNN + Gnostic Fields [10]	81.0	56.8	87.4	61.0	-
MFL (feature-level) + ELM [38]	87.6	-	-	-	-
CNN + Gabor + MS-CLBP [39]	88.0	-	-	-	-
ME-CNN [40]	87.3	-	-	-	-
ELM-CNN [41]	-	-	-	-	61.2
Multimodal CNN [41]	80.2	63.5	86.7	-	-
DyFusion [42]	-	-	88.2	-	-
SF-SRDA [43]	87.6	74.7	88.0	57.8	71.0
Combination 1 Late Fusion (F6F6)	86.9	65.9	88.2	46.8	51.7
Combination 1 Hybrid Fusion (F6C4)	86.9	72.9	89.1	57.1	68.4
Combination 2 Hybrid Fusion (F6C3)	86.9	69.0	89.6	64.9	68.6

Notes: For ship recognition on nighttime and all time IR images, Hybrid Fusion (F6C3) extracts the features of IR images from the F6 and C3 layers of the per-trained or fine-tuned VGG-16 model, the same as F6C4, but Late Fusion (F6) extracts the features of IR images only from the F6 layer. Bold denotes the best one.

nighttime IR images. Furthermore, Hybrid Fusion performs better than Late Fusion (F6F6) on daytime, nighttime and all time IR images. Although SF-SRDA method achieves higher accuracy than our proposed fusion models on daytime and all time IR images, Hybrid Fusion (F6C3) outperforms it by 1.6% on multispectral image and by 7.1% on nighttime IR images. Therefore, the proposed fusion models are more suitable for ship recognition on multispectral image than other methods. Note that combination 1 requires no training at feature extraction stage and is efficient, but combination 2 and combination 3 need a long time to fine-tune VGG-16 model, and some training tricks should be well used during fine-tuning on small-scale dataset.

In addition, normalized confusion matrices for Hybrid Fusion (F6C3) of combination 2 are shown in **Figure 7**. As shown in **Figure 7(a)**, all categories except for medium-other and tug are above 92% accuracy. Medium-other achieves only 64% because it is often confused with passenger and small ships. Besides, tug achieves only 60% in Hybrid Fusion (F6C3) due to it has less train samples than other classes (see **Table 1**) and being also confused with passenger and small ships. Nighttime IR images provide contour and few details of ship due to blur, low resolution and large pixels range, it is difficult to classify ship category on them. For normalized confusion matrix on nighttime IR images shown in **Figure 7(b)**, Hybrid Fusion (F6C3) performs worst on medium-other, which is misclassified as cargo by 50% and as passenger by 43% because they have similar contours. This also affects the recognition accuracy on all time IR images, as shown in **Figure 7(c)**. **Figure 8** gives some visual examples, which are

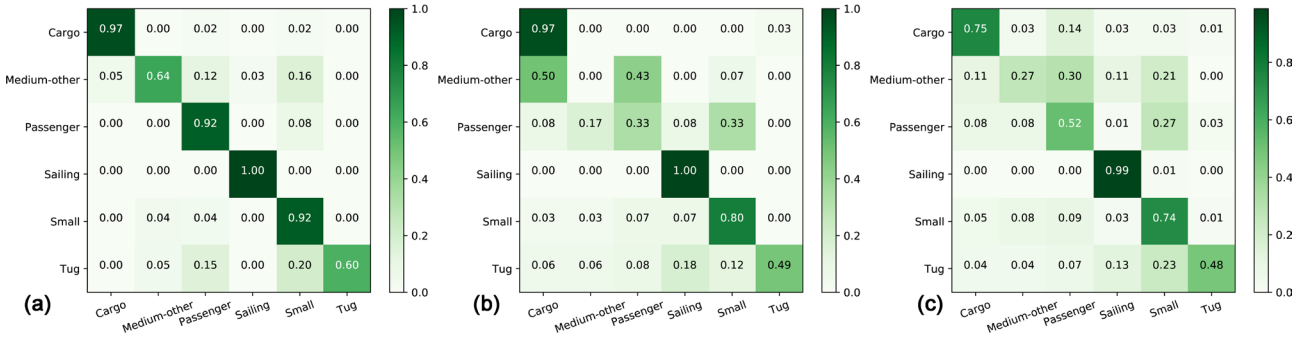

Figure 7. Normalized confusion matrices for the best proposed performing recognition models. (a) Normalized confusion matrix for Hybrid Fusion (F6C3) in combination 2, (b) and (c) are normalized confusion matrices for Hybrid Fusion (F6C3) on nighttime and all time IR images, respectively. (a) Hybrid Fusion (F6C3); (b) On nighttime IR images; (c) On all time IR images.

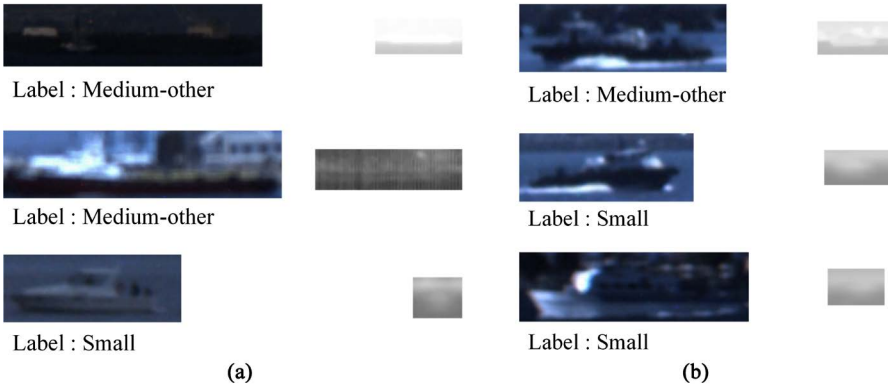

Figure 8. Some visual examples are misclassified by using either of VIS and IR images while correctly classified by using multispectral images. (a) Hybrid Fusion (F6C3) in combination 2; (b) Hybrid Fusion (F6C4) in combination 1.

misclassified by using either of VIS and IR images, but correctly classified by using multispectral images.

5. Discussion

It is not an easy work to use small-scale dataset to fine-tune the VGG-16 model trained on large-scale dataset like ImageNet, specifically for IR images because of the differences in imaging principle and technique between them. Some existing regularization techniques are used to avoid over-fitting problem in our experiment, however, data argumentation using image transformation causes strong correlation between samples, and don't improve the performance of model for small-scale dataset. Besides, early stopping is also taken as regularization technique to avoid over-fitting in our experiment, but it often stops before the model converges when we fine-tune the pre-trained model on IR images. L^2 weight decay is sensitive to the manual value, if the value is set too small, it cannot restrain the loss raise. In a word, if small-scale dataset like VAIS is used to fit the large-scale parameters of deep CNN like VGG-16, data argumentation is necessary. Generative adversarial networks [49] [50] may be an effective tool to produce the paired VIS-IR images and increase training samples.

We investigate the feature fusion performance of the pre-trained and fine-tuned VGG-16 models for multispectral images, and no techniques are used to project high-dimension feature into fewer dimension space. Our work is meaningful to future work for multispectral maritime ship recognition. For example, our work based on the pre-trained VGG-16 model, can be easily extended on the other pre-trained deep CNN models, such as VGG-19 model, Inception model [16] and ResNet model [17]. Besides, researchers can leverage unsupervised feature learning methods to reduce feature dimension, such as principal components analysis, and also embed Network-in-Network (NIN) [51] for fine-tuning the well-known pre-trained deep CNN models. Furthermore, the baseline method and the state-of-the-art method [42] adapt the decision level fusion for ship recognition, and extract features from the last fully connected layer of the pre-trained VGG-16 model and the last convolutional layer of the pre-trained VGG-19 model, respectively. Based on our experimental results, features extracted from the same layer of the pre-trained deep CNN model are not the best for both VIS and IR images. We believe that our work can be further investigated in the decision level fusion.

6. Conclusion

In this paper, we take advantage of the deep CNN model and multispectral data, and model multispectral ship recognition task into a convolutional feature fusion problem. We propose a feature fusion architecture, namely Hybrid Fusion, and investigate it as well as other three feature fusion architectures by exploiting L2 normalization method. Meanwhile, we use existing regularization techniques to fine-tune the pre-trained VGG-16 model on VIS and IR images in VAIS dataset, and investigate the ship recognition performance of three combinations. Experimental results demonstrate that feature representation ability is strong at high level of the pre-trained VGG-16 model for VIS image, and middle level for IR image. In the four feature fusion architectures, Hybrid Fusion performs better recognition accuracy than the other three feature fusion architectures. Besides, fine-tuning the pre-trained VGG-16 model can learn semantic information of ship, and slightly improve the recognition performance of Hybrid Fusion. The best Hybrid Fusion achieves 89.6% mean per-class recognition accuracy, and outperforms the state-of-the-art method. Our future work focuses on unsupervised feature learning and decision level fusion.

Funding Statement

This work is partly supported by the National Natural Science Foundation of China (Grant No. 61102170 and No. 62006240) and the National Social Science Foundation of China (Grant No. 15GJ003-243).

Data Availability Statement

The Excel data used to support the findings of this study are included within the

supplementary information file.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Guo, K., Wu, S. and Xu, Y. (2017) Face Recognition Using Both Visible Light Image and Near-Infrared Image and a Deep Network. *CAAI Transactions on Intelligence Technology*, **2**, 39-47. https://doi.org/10.1016/j.trit.2017.03.001

[2] Dai, P., Ji, R., Wang, H., Wu, Q. and Huang, Y. (2018) Cross-Modality Person Re-Identification with Generative Adversarial Training. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Stockholm, 13-19 July 2018, 677-683. https://doi.org/10.24963/ijcai.2018/94

[3] Li, C., Song, D., Tong, R. and Tang, M. (2019) Illumination-Aware Faster R-CNN for Robust Multispectral Pedestrian Detection. *Pattern Recognition*, **85**, 161-171. https://doi.org/10.1016/j.patcog.2018.08.005

[4] Cho, Y.R., Shin, S., Yim, S.H., Kong, K., Cho, H.W. and Song, W.J. (2019) Multistage Fusion with Dissimilarity Regularization for SAR/IR Target Recognition. *IEEE Access*, **7**, 728-740. https://doi.org/10.1109/ACCESS.2018.2885736

[5] Li, C., Wu, X., Zhao, N., Cao, X. and Tang, J. (2018) Fusing Two-Stream Convolutional Neural Networks for RGB-T Object Tracking. *Neurocomputing*, **281**, 78-85. https://doi.org/10.1016/j.neucom.2017.11.068

[6] Hong, C., Koschan, A., Abidi, M., Kong, S.G. and Won, C.-H. (2008) Multispectral Visible and Infrared Imaging for Face Recognition. 2008 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Anchorage, 24-26 June 2008, 1-6. https://doi.org/10.1109/CVPRW.2008.4563054

[7] Shoja Ghiass, R., Arandjelovic, O., Bendada, A. and Maldague, X. (2014) Infrared Face Recognition: A Comprehensive Review of Methodologies and Databases. *Pattern Recognition*, **47**, 2807-2824. https://doi.org/10.1016/j.patcog.2014.03.015

[8] Hermosilla, G., Rojas, M., Mendoza, J., Farias, G., Pizarro, F.T., San Martin, C. and Vera, E. (2018) Particle Swarm Optimization for the Fusion of Thermal and Visible Descriptors in Face Recognition Systems. *IEEE Access*, **6**, 42800-42811. https://doi.org/10.1109/ACCESS.2018.2850281

[9] Peng, C., Wang, N., Li, J. and Gao, X. (2019) DLFace: Deep Local Descriptor for Cross-Modality Face Recognition. *Pattern Recognition*, **90**, 161-171. https://doi.org/10.1016/j.patcog.2019.01.041

[10] Zhang, M.M., Choi, J., Daniilidis, K., Wolf, M.T. and Kanan, C. (2015) Vais: A Dataset for Recognizing Maritime Imagery in the Visible and Infrared Spectrums. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Boston, 7-12 June 2015, 10-16. https://doi.org/10.1109/CVPRW.2015.7301291

[11] Kniaz, V.V., Knyaz, V.A., Hladuvka, J., Kropatsch, W.G. and Mizgi-nov, V. (2019) ThermalGAN: Multimodal Color-to-Thermal Image Translation for Person Re-Identification in Multispectral Dataset. *Computer Vision ECCV* 2018 *Workshops*, Volume 11134, 606-624. https://doi.org/10.1007/978-3-030-11024-6_46

[12] Zhang, L., Liu, Z., Zhang, S., Yang, X., Qiao, H., Huang, K. and Hussain, A. (2019) Cross-Modality Interactive Attention Network for Multispectral Pedestrian Detec-

tion. *Information Fusion*, **50**, 20-29. https://doi.org/10.1016/j.inffus.2018.09.015

[13] Zhang, L., Gonzalez-Garcia, A., van de Weijer, J., Danelljan, M. and Khan, F.S. (2019) Synthetic Data Generation for End-to-End Thermal Infrared Tracking. *IEEE Transactions on Image Processing*, **28**, 1837-1850. https://doi.org/10.1109/TIP.2018.2879249

[14] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems*, Curran Associates Inc., Red Hook, 1097-1105.

[15] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. http://arxiv.org/abs/1409.1556

[16] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) Rethinking the Inception Architecture for Computer Vision. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[17] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[18] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L. (2009) ImageNet: A Large-Scale Hierarchical Image Database. 2009 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Miami, 20-25 June 2009, 8. https://doi.org/10.1109/CVPR.2009.5206848

[19] Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M. and Burgard, W. (2015) Multimodal Deep Learning for Robust RGB-D Object Recognition. 2015 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Hamburg, 28 September-3 October 2015, 681-687. https://doi.org/10.1109/IROS.2015.7353446

[20] Bui, H.M., Lech, M., Cheng, E., Neville, K. and Burnett, I.S. (2016) Object Recognition Using Deep Convolutional Features Transformed by a Recursive Network Structure. *IEEE Access*, **4**, 10059-10066. https://doi.org/10.1109/ACCESS.2016.2639543

[21] Ren, S., He, K., Girshick, R., Zhang, X. and Sun, J. (2017) Object Detection Networks on Convolutional Feature Maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1476-1481. https://doi.org/10.1109/TPAMI.2016.2601099

[22] Wang, T., Chen, Y., Zhang, M., Chen, J. and Snoussi, H. (2017) Internal Transfer Learning for Improving Performance in Human Action Recognition for Small Datasets. *IEEE Access*, **5**, 17627-17633. https://doi.org/10.1109/ACCESS.2017.2746095

[23] Park, K., Kim, S. and Sohn, K. (2018) Unified Multi-Spectral Pedestrian Detection Based on Probabilistic Fusion Networks. *Pattern Recognition*, **80**, 143-155. https://doi.org/10.1016/j.patcog.2018.03.007

[24] Yosinski, J., Clune, J., Bengio, Y. and Lipson, H. (2014) How Transferable Are Features in Deep Neural Networks? *Proceedings of the* 27*th International Conference on Neural Information Processing Systems*, Volume 2, 3320-3328.

[25] Schwarz, M., Schulz, H. and Behnke, S. (2015) RGB-D Object Recognition and Pose Estimation Based on Pre-Trained Convolutional Neural Network Features. 2015 *IEEE International Conference on Robotics and Automation* (*ICRA*), Seattle, 26-30 May 2015, 1329-1335. https://doi.org/10.1109/ICRA.2015.7139363

[26] Zia, S., Yuksel, B., Yuret, D. and Yemez, Y. (2017) RGB-D Object Recognition Using Deep Convolutional Neural Networks. 2017 *IEEE International Conference on Computer Vision Workshops* (*ICCVW*), Venice, 22-29 October 2017, 887-894.

https://doi.org/10.1109/ICCVW.2017.109

[27] Caglayan, A. and Can, A.B. (2019) Exploiting Multi-Layer Features Using a CNN-RNN Approach for RGB-D Object Recognition. In: *Computer Vision ECCV* 2018 *Workshops*, Springer International Publishing, Cham, 675-688. https://doi.org/10.1007/978-3-030-11015-4_51

[28] Chatfield, K., Simonyan, K., Vedaldi, A. and Zisserman, A. (2014) Return of the Devil in the Details: Delving Deep into Convolutional Nets. *Proceedings of the British Machine Vision Conference* 2014, Nottingham, 1-5 September 2014, 6.1-6.12. https://doi.org/10.5244/C.28.6

[29] Socher, R., Huval, B., Bhat, B., Manning, C.D. and Ng, A.Y. (2012) Convolutional-Recursive Deep Learning for 3D Object Classification. *Proceedings of the* 25*th International Conference on Neural Information Processing Systems*, Volume 1, 656-664. http://dl.acm.org/citation.cfm?id=2999134.2999208

[30] Kanjir, U., Greidanus, H. and Ostir, K. (2018) Vessel Detection and Classification from Spaceborne Optical Images: A Literature Survey. *Remote Sensing of Environment*, **207**, 1-26. https://doi.org/10.1016/j.rse.2017.12.033

[31] Khellal, A., Ma, H. and Fei, Q. (2018) Convolutional Neural Network Based on Extreme Learning Machine for Maritime Ships Recognition in Infrared Images. *Sensors*, **18**, 1490. https://doi.org/10.3390/s18051490

[32] Bousetouane, F. and Morris, B. (2015) Off-the-Shelf CNN Features for Fine-Grained Classification of Vessels in a Maritime Environment. In: *Advances in Visual Computing*, Volume 9475, Springer International Publishing, Cham, 379-388. https://doi.org/10.1007/978-3-319-27863-6_35

[33] Dao, C.D., Hua, X.H. and Morere, O. (2015) Maritime Vessel Images Classification Using Deep Convolutional Neural Networks. *Proceedings of the Sixth International Symposium on Information and Communication Technology*, Hue City, 3-4 December 2015, 1-6. https://doi.org/10.1145/2833258.2833266

[34] Solmaz, B., Gundogdu, E., Yucesoy, V. and Koc, A. (2017) Generic and Attribute-Specific Deep Representations for Maritime Vessels. *IPSJ Transactions on Computer Vision and Applications*, **9**, 22. https://doi.org/10.1186/s41074-017-0033-4

[35] Gundogdu, E., Solmaz, B., Yucesoy, V. and Koc, A. (2017) MARVEL: A Large-Scale Image Dataset for Maritime Vessels. *Computer Vision ACCV* 2016, Volume 10115, 165-180. https://doi.org/10.1007/978-3-319-54193-8_11

[36] Solmaz, B., Gundogdu, E., Yucesoy, V., Koc, A. and Alatan, A.A. (2018) Fine-Grained Recognition of Maritime Vessels and Land Vehicles by Deep Feature Embedding. *IET Computer Vision*, **12**, 1121-1132. https://doi.org/10.1049/iet-cvi.2018.5187

[37] Milicevic, M., Zubrinic, K., Obradovic, I. and Sjekavica, T. (2019) Application of Transfer Learning for Fine-Grained Vessel Classification Using a Limited Dataset. In: *Applied Physics, System Science and Computers* III, Volume 574, Springer International Publishing, Cham, 125-131. https://doi.org/10.1007/978-3-030-21507-1_19

[38] Huang, L., Li, W., Chen, C., Zhang, F. and Lang, H. (2018) Multiple Features Learning for Ship Classification in Optical Imagery. *Multimedia Tools and Applications*, **77**, 13363-13389. https://doi.org/10.1007/s11042-017-4952-y

[39] Shi, Q., Li, W., Zhang, F., Hu, W., Sun, X. and Gao, L. (2018) Deep CNN with Multi-Scale Rotation Invariance Features for Ship Classification. *IEEE Access*, **6**, 38656-38668. https://doi.org/10.1109/ACCESS.2018.2853620

[40] Shi, Q., Li, W., Tao, R., Sun, X. and Gao, L. (2019) Ship Classification Based on Multifeature Ensemble with Convolutional Neural Network. *Remote Sensing*, **11**, 419. https://doi.org/10.3390/rs11040419

[41] Aziz, K. and Bouchara, F. (2018) Multimodal Deep Learning for Robust Recognizing Maritime Imagery in the Visible and Infrared Spectrums. In: Campilho, A., Karray, F. and ter Haar Romeny, B., Eds., *Image Analysis and Recognition*, Springer International Publishing, Berlin, 235-244. https://doi.org/10.1007/978-3-319-93000-8_27

[42] Santos, C.E. and Bhanu, B. (2018) Dyfusion: Dynamic IR/RGB Fusion for Maritime Vessel Recognition. 2018 25*th IEEE International Conference on Image Processing* (*ICIP*), Athens, 7-10 October 2018, 1328-1332. https://doi.org/10.1109/ICIP.2018.8451745

[43] Zhang, E., Wang, K. and Lin, G. (2019) Classification of Marine Vessels with Multi-Feature Structure Fusion. *Applied Sciences*, **9**, 2153. https://doi.org/10.3390/app9102153

[44] Liu, J., Zhang, S., Wang, S. and Metaxas, D. (2016) Multispectral Deep Neural Networks for Pedestrian Detection. *Procedings of the British Machine Vision Conference* 2016, York, 19-22 September 2016, 73.1-73.13. https://doi.org/10.5244/C.30.73

[45] Zeiler, M.D. and Fergus, R. (2014) Visualizing and Understanding Convolutional Networks. *Computer Vision ECCV* 2014, Volume 8689, 818-833. https://doi.org/10.1007/978-3-319-10590-1_53

[46] Chen, Y., Xie, H. and Shin, H. (2018) Multi-Layer Fusion Techniques Using a CNN for Multispectral Pedestrian Detection. *IET Computer Vision*, **12**, 1179-1187. https://doi.org/10.1049/iet-cvi.2018.5315

[47] Chang, C.C. and Lin, C.J. (2011) LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 27:1-27:27. https://doi.org/10.1145/1961189.1961199

[48] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press, Cambridge. http://www.deeplearningbook.org

[49] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative Adversarial Networks.

[50] Isola, P., Zhu, J.Y., Zhou, T. and Efros, A.A. (2017) Image-to-Image Translation with Conditional Adversarial Networks. *CVPR* 2017, Hawaii, 22-25 July 2017, 1642-1650. https://doi.org/10.1109/CVPR.2017.632

[51] Lin, M., Chen, Q. and Yan, S. (2013) Network in Network.