Heliyon 10 (2024) e26229

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A lightweight CNN for multi-source infrared ship detection from unmanned marine vehicles

Liqian Wang [a,b], Yakui Dong [a,b], Cheng Fei [a,b], Junliang Liu [a,b], Shuzhen Fan [a,b], Yunxia Liu [a,b], Yongfu Li [a,b,*], Zhaojun Liu [a,c], Xian Zhao [a,b]

[a] Key Laboratory of Laser & Infrared System, Shandong University, Ministry of Education, 72 Binhai Road, Qingdao, 266237, Shandong, China
[b] Center for Optics Research and Engineering, 72 Binhai Road, Qingdao, 266237, Shandong, China
[c] School of Information Science and Engineering, 72 Binhai Road, Qingdao, 266237, Shandong, China

ARTICLE INFO

Keywords:
Maritime surveillance
Ship detection
Infrared ship target
Autonomous ships
Convolutional Neural Network (CNN)

ABSTRACT

Infrared ship detection is of great significance due to its broad applicability in maritime surveillance, traffic safety and security. Multiple infrared sensors with different spectral sensitivity provide enhanced sensing capabilities, facilitating ship detection in complex environments. Nevertheless, current researches lack discussion and exploration of infrared imagers in different spectral ranges for marine objects detection. Furthermore, for unmanned marine vehicles (UMVs), e.g., unmanned surface vehicles (USVs) and unmanned ship (USs), detection and perception are usually performed in embedded devices with limited memory and computation resource, which makes traditional convolutional neural network (CNN)-based detection methods struggle to leverage their advantages. Aimed at the task of sea surface object detection on USVs, this paper provides lightweight CNNs with high inference speed that can be deployed on embedded devices. It also discusses the advantages and disadvantages of using different sensors in marine object detection, providing a reference for the perception and decision-making modules of USVs. The proposed method can detect ships in short-wave infrared (SWIR), long-wave infrared (LWIR) and fused images with high-performance and high-inference speed on an embedded device. Specifically, the backbone is built from bottleneck depth-separable convolution with residuals. Generating redundant feature maps by using cheap linear operation in neck and head networks. The learning and representation capacities of the network are promoted by introducing the channel and spatial attention, redesigning the sizes of anchor boxes. Comparative experiments are conducted on the infrared ship dataset that we have released which contains SWIR, LWIR and the fused images. The results indicate that the proposed method can achieve high accuracy but with fewer parameters, and the inference speed is nearly 60 frames per second (FPS) on an embedded device.

1. Introduction

Global shipping currently exceeds 80 percent of world merchandise trade [1]. Monitoring maritime ships timely and effectively is important to guarantee the safety of maritime transportation, trade, fishery and scientific investigation. LWIR imaging has been

* Corresponding author at: Center for Optics Research and Engineering, 72 Binhai Road, Qingdao, 266237, Shandong, China.
E-mail address: yfli@sdu.edu.cn (Y. Li).

https://doi.org/10.1016/j.heliyon.2024.e26229
Received 17 May 2023; Received in revised form 12 September 2023; Accepted 8 February 2024

Available online 13 February 2024

Fig. 1. Infrared images of different spectral bands captured in a same marine scenario. (a) SWIR image, (b) LWIR image, (c) fused image of SWIR and LWIR.

widely used in ship detection, it has significant advantages in night vision. SWIR imagers are complementary to LWIR imagers when it comes to vision enhancement and low visibility in poor weather conditions [2]. Imagery in the SWIR is similar to visible imagery, in that it senses reflected light, thus interpretation and scene analysis is improved over LWIR systems which have good detection abilities [2]. The combination of LWIR and SWIR imaging is better adjusted to the complex marine environment and facilitates ship detection. Fig. 1 presents ship infrared images captured by different infrared sensors in the same marine scenario. Fig. 1(a), Fig. 1(b), Fig. 1(c) are SWIR, LWIR and their fused images, respectively.

Unmanned vehicles have become more widely used in marine science, ocean engineering recently due to their ease of deployment, mobility and low cost [3]. For object detection tasks with unmanned marine platforms, e.g., USVs and USs, the platforms carried embedded devices with insufficient computing power and traditional convolutional neural network (CNN)-based methods hardly deployed.

Recently, a multitude of ship detection methods for infrared and visible images have been put forwarded by researchers. For infrared ship detection, the common methods are based on segmentation, such as Yang [4] proposing a probability induced intuitionistic FCM clustering algorithm for infrared ship segmentation. Bai [5] developed an improved fuzzy C-means (FCM) method based on the spatial information for IR ship target segmentation. Liu [6] designed a global background subtraction filter (GBSF) and an adaptive row mean subtraction filter (ARMSF) to suppress the background and enhance the target, and then segmented ship target using threshold and ship's geometric prior information. Mumtaz [7] presented a saliency-based ship detection method, at first it computes the saliency map of the input image using the Graph-Based Visual Saliency (GBVS) algorithm, then uses fuzzy C-means (FCM) to obtain fine ship regions. In addition, detection methods based on hand-crafted features have been extensively studied, Li [8] proposed a method for infrared ship detection using time fluctuation feature and space structure feature, the experiments were conducted on a computer. Li [9] incorporated morphological reconstruction and multi-feature analysis into infrared ship detection to improve the performance. Zhang [10] presented a ship detection method with visible images including horizon detection, background modeling and background subtraction, all of which are on Discrete Cosine Transform (DCT).

CNN-based methods have been successfully exploited to automatically and intelligently detect ships due to its powerful feature extraction and generalization ability of data. Liu [11] developed an enhanced CNN to improve ship detection under different weather conditions. Kim [12] proposed a probabilistic ship detection and classification system based on deep learning. Chen [13] proposed a deep learning based ship type recognition framework. Nie [14] adopted the synthetically-degraded images to enlarge the training datasets, and proposed an advanced YOLOv3 model to detect ships. Liu [15] proposed a global guided lightweight non-local depth feature (DG-Light-NLDF) model for detect infrared maritime salient objects. Deep learning combined with hand-crafted features further improves the robustness of detection, Shao [16] proposed a saliency-aware CNN framework for ship detection, comprising comprehensive ship discriminative features, such as deep feature, saliency map, and coastline prior. Song [17] presented an improved dim and small infrared ship detection network based on Haar wavelet. Chen [18] proposed a novel approach for achieving a pixel-wise ship segmentation and identification task through a novel design of U-Net deep learning architecture (denoted as EU-Net). The method has been validated on visible images and experimental results show that the ship segmentation accuracies were larger than 99 percent.

To summarize, current researches focused on ship detection in LWIR and visible images. Whereas the degradation of visible images caused by low-light and harsh conditions, and the lack of texture and structure data in LWIR images, are not conducive to ship detection. SWIR imagers penetrate fog, haze much better than detectors sensitive in visible spectral range and can provide wide dynamic imaging [2]. SWIR can enhance the perception and scene interpretation capabilities of LWIR imaging systems.

In addition, due to the limited hardware resource, unmanned marine vehicles have weak computational processing ability, which makes current methods difficult to deploy on embedded devices or difficult to play their capabilities.

Considering these issues, this paper proposed a lightweight CNN for ship detection with multiple infrared images including SWIR, LWIR and fused images. This method has high detection accuracy and fast inference speed on an embedded device, but with few parameters and low computation cost, making it suitable to be deployed on unmanned marine vehicles with limited hardware resources. In conclusion, given the current achievements, our method significantly differs from previous studies in the following aspects.

(1) We propose a lightweight CNN for multi-source infrared ship detection that is easy to deploy in embedded devices. It performs ship detection with LWIR, SWIR and fused images, has fast inference speed and high detection accuracy for applications in open-sea visual maritime surveillance, autonomous ships and navigation.

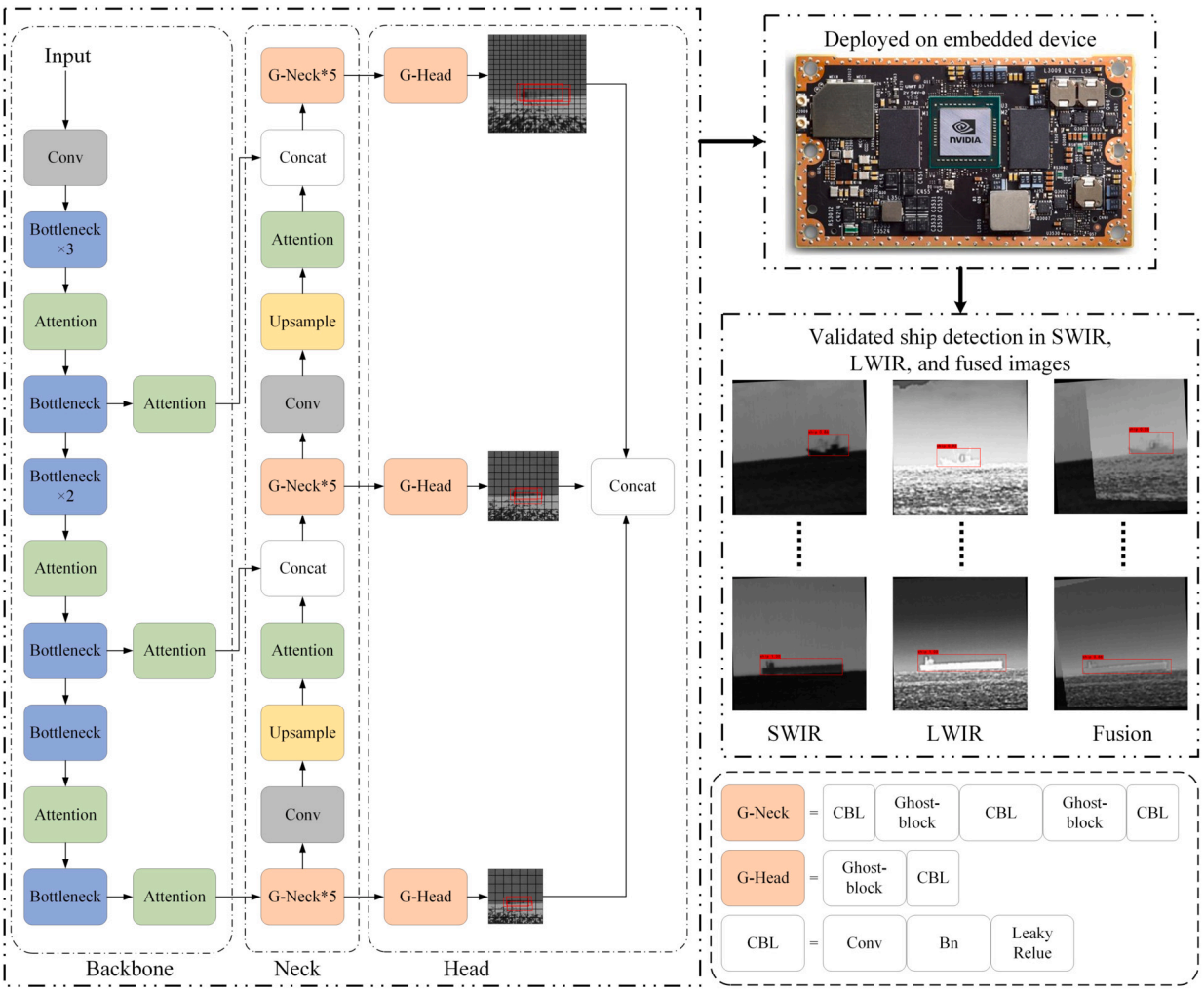

Fig. 2. Flowchart of the proposed method in this study. The proposed ship detection method is deployed on an embedded device and verified on SWIR, LWIR and fused image datasets.

(2) The proposed method employs bottleneck depth-separable convolution with residuals to build backbone and generates redundant feature maps at the head and neck networks with cheap linear operations, thus making it efficient and lightweight. Spatial and channel attention is introduced and anchor box sizes are redesigned to promote the performance and feature extraction capabilities of the network for infrared images.

(3) To our best knowledge, this paper is the first one to explore the application of multi-source infrared images in ship detection, including SWIR, LWIR, and their fused images, which provides a reference for maritime situational awareness in complicated environments. We conducted comparative experiments on the infrared ship dataset that we released, and the results demonstrate the superior performance of our proposed method.

The remainder of this article is organized as follows. Section 2 introduces the architecture of proposed method in detail. Section 3 presents the experimental results. The conclusion is presented in Section 4.

2. Methodology

2.1. Overview of the framework

The flowchart of the proposed IR ship detection method is shown in Fig. 2. The network architecture is lightweight in design and has enhanced feature extraction capabilities. Unmanned marine vehicles have limited hardware resources and typically carry embedded devices with limited computational capability. Hence, the proposed model is deployed on an embedded device to verify its computational efficiency. Comparative experiments were carried out on our released infrared ship dataset which contains SWIR, LWIR, and their fused images to verify its accuracy in ship detection under harsh marine environments. The network architecture is based on YOLOv3 [19], which considers the balance between accuracy and speed. To reduce the parameters of network and ease

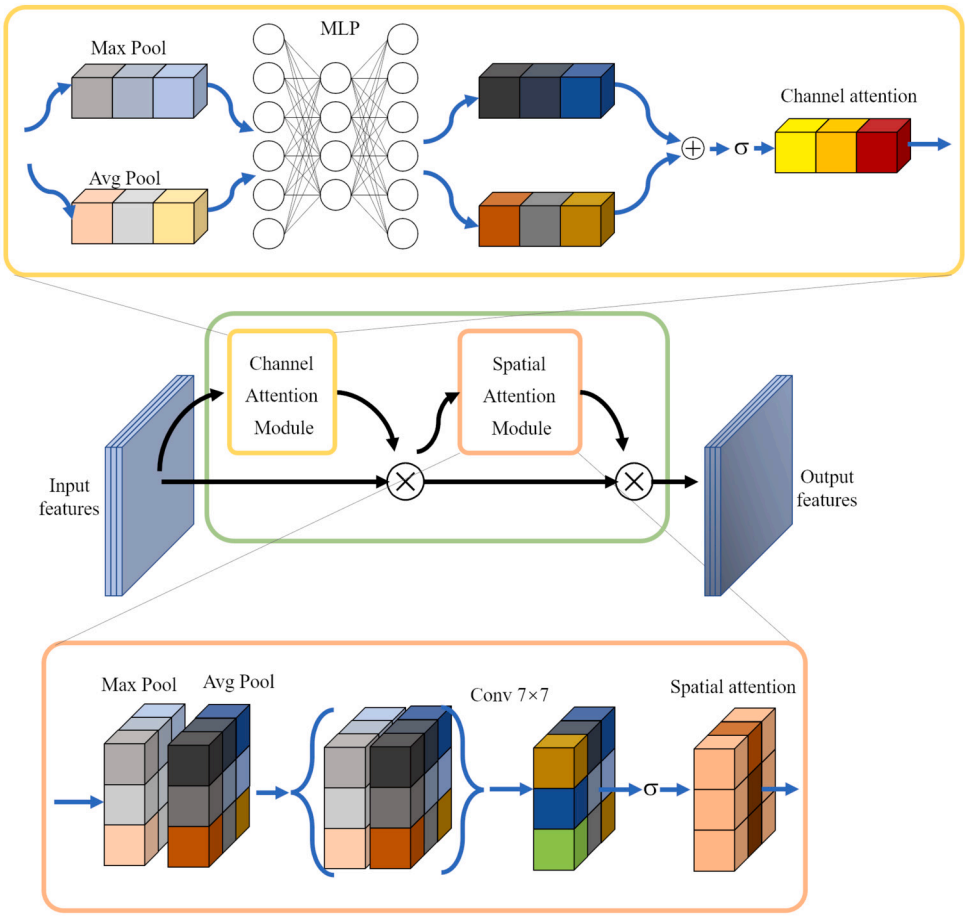

Fig. 3. Attention module enhances the extraction and refinement of features, so that the network focuses on meaningful information and important locations in infrared images.

deployment in embedded devices, the backbone is built using bottleneck depth-separable convolution with residuals. The head and neck network use cheap linear operations to generate redundant feature maps. We introduced spatial and channel attention and redesigned the sizes of anchor boxes to promote feature extraction capabilities of the network for infrared images.

2.2. Attention module

The visually-degraded infrared images under in severe marine environments make it difficult to distinguish ship targets from background. Additionally, CNNs can effectively extract high-level feature maps $f_d(d = 1, ..., D)$ from raw infrared images, where D is the dimension of feature maps, d is the d-th feature map. However, not all high-level features contribute to the discrimination of dissimilarity [20]. To enhance the representation power and discrimination ability of CNNs in infrared images, we introduce attention modules [21] to make the model more effectively utilize features in different dimensions. The attention module is shown in Fig. 3.

In infrared images, attention helps to focus on important features and suppress unnecessary ones. The convolution block attention module (CBAM) integrates channel and spatial attention [21], where channel attention focuses across channels of features to tell 'what' is important in images [22], spatial attention focuses 'where' is an informative region in images [23].

Channel attention module is detailed in Eq (1), it focuses on 'what' is meaningful in input images. The importance of each channel is encoded in channel attention maps. Spatial information of feature maps is squeezed by using average-pooling ($AvgPool$) and max-pooling ($MaxPool$) operations along spatial axis. The average-pooled features and max-pooled features are forwarded to a shared multi-layer perception (MLP). The output feature vectors are merged using element-wise summation, and then a channel attention map is generated by sigmoid function σ.

$$\mathrm{M}_c(f) = \sigma(MLP(AvgPool(f)) + MLP(MaxPool(f))) \tag{1}$$

Spatial attention module is computed as Eq (2). The importance of informative regions in feature maps is encoded in spatial attention maps. It applies average-pooling and max-pooling operations along to the channel axis. Average-pooled features and max-

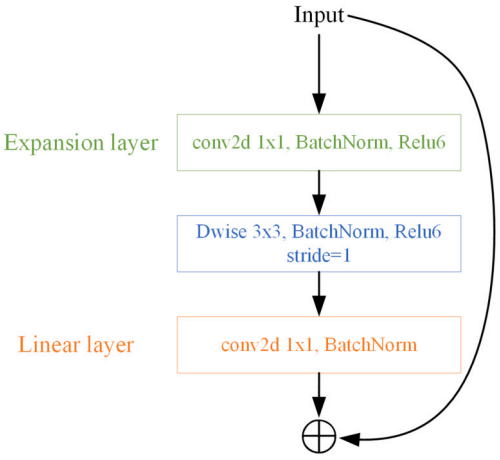

Fig. 4. The structure of Bottleneck block.

Table 1
Architecture of lightweight backbone. t is expansion factor; c is the number of output channels; n is repeated times of block; s is stride.

Operator	t	c	n	s
conv2d_3×3	-	32	1	2
bottleneck	1	16	1	1
bottleneck	6	24	2	2
bottleneck	6	32	2	2
attention	-	32	1	-
bottleneck	6	32	1	2
bottleneck	6	64	4	2
bottleneck	6	96	2	1
attention	-	96	1	-
bottleneck	6	96	1	1
bottleneck	6	160	3	2
attention	-	160	1	-
bottleneck	6	320	1	1

pooled features are concatenated and convoluted by a convolution operation ($k^{7\times7}$) with the filter size of 7×7.σ is the sigmoid function.

$$\mathrm{M}_s(f) = \sigma(k^{7\times7}([AvgPool(f); MaxPool(f)])) \tag{2}$$

2.3. Lightweight backbone with enhanced feature extraction

Backbone network is acting as the basic feature extractor for ship detection task, which generates feature maps from input images [24]. Deeper and densely connected backbones generally perform better. Nevertheless, considering the limited hardware resources and the degradation of infrared images in complicated marine scenarios, we developed a lightweight backbone with enhanced feature extraction, with architecture that was modified from MobileNetv2 [25] and that mainly consists of residual bottleneck [25] and attention modules in Section 2.2 as shown in Table 1. Bottleneck is lightweight and efficient for feature extraction. To enhance the representation power of the backbone, we introduce a lightweight attention module, with an overhead of parameters and computation that are negligible compared to other blocks.

The detailed structure of the bottleneck is shown as Fig. 4. Its structure is similar to residual connections [26], which helps improve the ability of a gradient to propagate across multiplier layers [25]. In the bottleneck, an efficient depthwise separable convolution [27] is used to reduce the amount of computation.

The standard convolution operation can be formulated as Eq (3), where * is the convolution operation, b is the bias term, $f \in \mathbb{R}^{h\times w\times c}$ is the input data, c is the number of input channels, h and w is the height and width of the input data. $f' \in {}^{h'\times w'\times c'}$ is the output feature map with c' channels, h' and w' are the height and width of the output feature map. Standard convolutions have the computational cost of Eq (4).

$$f' = f * k + b \tag{3}$$

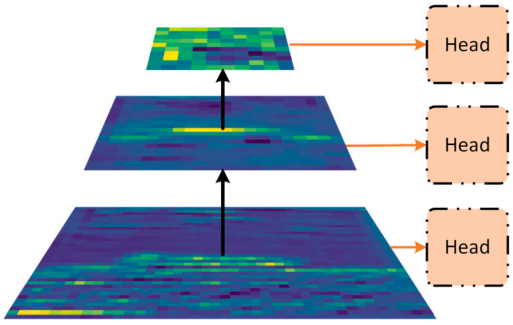

Fig. 5. Multi-scale feature extraction with feature pyramid network (FPN) and lightweight detection head.

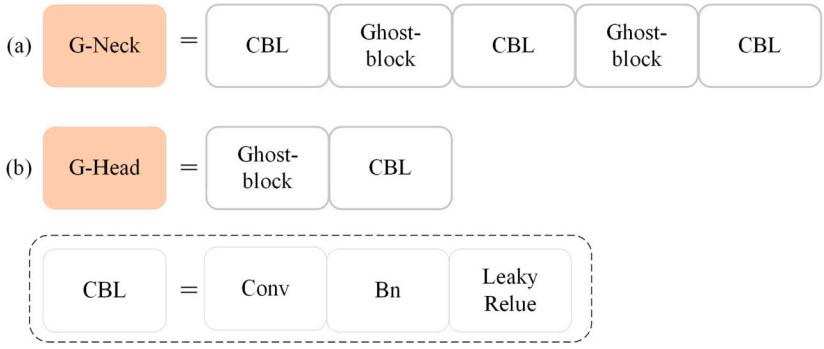

Fig. 6. (a) G-Neck and (b) G-Head with cost efficient feature extraction operation, can generate redundant features utilizing linear operations.

$$h' \cdot w' \cdot c \cdot c' \cdot k \cdot k \tag{4}$$

In the bottleneck, depthwise separable convolution is an efficient convolution operation consisting of two steps. 1) Depthwise convolution uses a single convolution kernel for each channel of the input data, the computation cost is Eq (5). 2) The pointwise convolution applies a 1×1 convolution to number of the channels of the output feature map, its computation cost is Eq (6). So the total computation cost of depthwise separable convolution is Eq (7). We get a reduction in computation of Eq (8).

$$h' \cdot w' \cdot c \cdot k \cdot k \tag{5}$$

$$h' \cdot w' \cdot c \cdot c' \tag{6}$$

$$h' \cdot w' \cdot c \cdot \left(k^2 + c'\right) \tag{7}$$

$$\frac{h' \cdot w' \cdot c \cdot \left(k^2 + c'\right)}{h' \cdot w' \cdot c \cdot c' \cdot k \cdot k} = \frac{1}{c'} + \frac{1}{k^2} \tag{8}$$

For the bottleneck, the size of the input feature map is $h \times w$, expansion factor is t, kernel size is k, c input channels and c' output channels. The total computation cost of bottleneck is Eq (9).

$$\begin{aligned} &h \cdot w \cdot c \cdot 1 \cdot 1 \cdot tc + \\ &h \cdot w \cdot tc \cdot k \cdot k \cdot 1 + \\ &h \cdot w \cdot c \cdot 1 \cdot 1 \cdot c' \\ &= h \cdot w \cdot c \cdot t(c + k^2 + c') \end{aligned} \tag{9}$$

2.4. Cost-efficient neck network and detection head

Ships have large variance in scale and aspect ratios. It is challenging to detect ships across large ranges of sizes at a single scale feature map. For deep convolution networks, features in shallow layers have high resolution with rich spatial information that is suitable to detect small targets [28], while features in deep layers have rich semantic information and large receptive fields, making them suitable for detecting large targets [28]. To accurately locate ships of different sizes in images, the neck network adopts feature pyramid network (FPN), as shown in Fig. 5, to capture features from different layers in backbone.

The detection head and neck network also adopt a lightweight and efficient architecture. The feature maps generated by deep neural networks exist many similar pairs, like a ghost of each other [29]. Redundancy in these feature maps often guarantees a comprehensive understanding of the input data, which is beneficial for accurate ships detection. Nevertheless, generating redundant feature maps consumes large computation resources. In this study, Ghost modules [29] are introduced in the neck network and

Table 2
Width and height of different anchor sizes.

Anchor Size	Width	Height
Anchor size 1	197.3	41.6
	211.2	48.0
	260.3	59.7
Anchor size 2	105.6	48.0
	137.6	38.4
	164.3	40.5
Anchor size 3	48.0	26.7
	80.0	28.8
	119.5	35.2

detection head, which can generate ghost feature maps with a cost-efficient operation. Specifically, the G-neck and G-head modules are designed to construct the neck network and detection head. The architecture of the G-neck is shown in Fig. 6 (a) and that of the G-head is shown in Fig. 6 (b). The Ghost block is used in G-neck and G-head instead of traditional convolutional operation to generate redundant feature maps with low computation.

Ghost block uses cheap linear operations to generate ghost features, as shown in Eq (10). f_i is the intrinsic feature map, $\Phi_{i,j}$ is the linear operation for generating the j-th ghost feature map f'_{ij} in an efficient way. The last $\Phi_{i,s}$ is the identity mapping, so the number of effective linear operations is $s-1$. With linear operations, one intrinsic feature can generate s($s \geq 1$) features.

$$f'_{ij} = \Phi_{i,j}(f_i), \quad \forall i = 1, ..., m, j = 1, ..., s, \tag{10}$$

The intrinsic feature map is $f_i \in \mathbb{R}^{h' \times w' \times m}$, the output feature map is $f'_{ij} \in \mathbb{R}^{h' \times w' \times n}$, with s times of linear operation, m intrinsic feature maps can generate n feature maps, so $n = m * s$. Due to the effective linear operation is $s-1$, so we can get that $m \cdot (s-1) = \frac{n}{s} \cdot (s-1)$. In linear operations, kernel size is d, and $d \times d$ is of similar size to $k \times k$. We get a reduction in computation of Eq (11).

$$\frac{h' \cdot w' \cdot c \cdot k \cdot k \cdot n}{\frac{n}{s} \cdot h' \cdot w' \cdot c \cdot k \cdot k + (s-1) \cdot \frac{n}{s} \cdot h' \cdot w' \cdot d \cdot d} = \frac{c \cdot k \cdot k}{\frac{1}{s} c \cdot k \cdot k + \frac{s-1}{s} \cdot d \cdot d} \approx \frac{s \cdot c}{s+c-1} \approx s \tag{11}$$

2.5. Redesigning the sizes of anchor boxes

The size of the input image is 320×320. FPN generates three scaled feature maps with sizes of 10×10, 20×20, 40×40. Due to feature maps from different levels in a network have different receptive field, so they are respectively responsible for detecting ships with various sizes ships. We used the anchor box-based detection method, where the numbers and sizes of anchor boxes are predefined before network training, and each scaled feature map is matched with three anchor boxes. We use k-means clustering to generate 9 priors, mentioned in YOLO [19], which correspond to three different scaled ships. Fig. 7 illustrates prior values of anchor boxes obtained from clustering in SWIR, LWIR and fusion datasets. Table 2 shows the width and height of different anchor sizes.

3. Experimental results and analysis

Models are trained on a computer with an Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz and Nvidia Tesla V100 GPU. The optimizer uses stochastic gradient descent (SGD) and the cosine learning rate decay strategy to train the network. The initial learning rate is 0.01, momentum is 0.9, weight decay is 0.0005. The network was trained for 250 epochs. Comparative experiments are conducted on the embedded device Jetson TX2 with 8Gb memory and 256 CUDA cores.

3.1. Description of infrared ship datasets

In our previous study, we released the infrared ship dataset including SWIR, LWIR and their fusion images [30]. There are 1044 images of each type, for a total of 1044×3 images. For image fusion, the premise is to register source images using feature matching methods. Image registration is the process of aligning two or more images of the same scene obtained from different viewpoints, at different times, or from different sensors. It can geometrically warp the sensed image into the common spatial coordinate system of the reference image and align their common area pixel-to-pixel. More details about the method of image fusion can be obtained from our published article [31].

SWIR images were captured by indium gallium arsenide (InGaAs) uncooled infrared focal plane array (FPA) detector with the resolution of 320×256 and using a zoom lens with focal lengths of 16 to 160 mm. The spectral ranges from 0.9 to 1.7 μm and the pixel spacing is 30 μm. LWIR images were captured by vanadium oxide (VOx) uncooled infrared FPA detector with the resolution of 640×512 and using a prime lens with a focal length of 50 mm. The spectral ranges from 8 to 14 μm. Fig. 8 (a) shows the structural

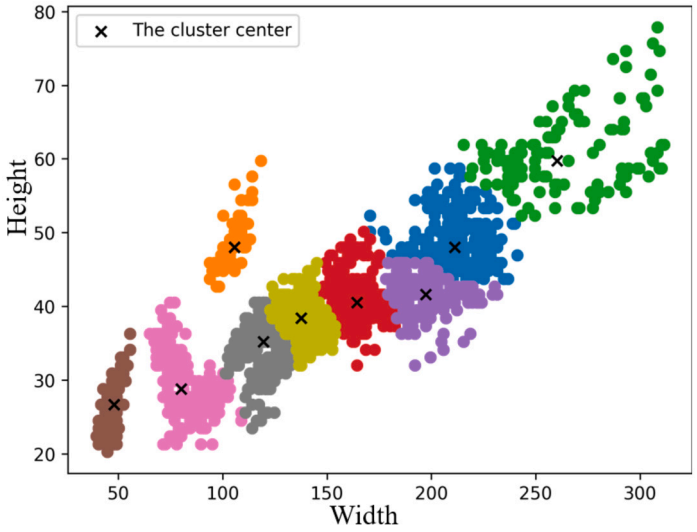

Fig. 7. The height and width distributions of the ground truth obtained by clustering from the original SWIR, LWIR and fusion ship datasets. Cluster centers are used as a prior for anchor boxes.

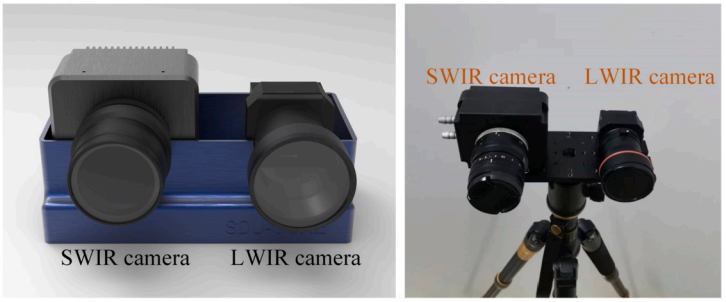

Fig. 8. The images acquisition device. (a) The structural diagram of the device, including the mechanical housing of SWIR and LWIR cameras and the holder of the device. The structural model is rendered to make it more intuitive. (b) The real hardware of the images acquisition device consists of SWIR and LWIR cameras, mounted on a tripod.

Fig. 9. Partial infrared images of ships in multiple bands. (a) SWIR, (b) LWIR and (c) Fusion of SWIR and LWIR.

Table 3
True positive (TP), True negative (TN), False positive (FP), False negative (FN).

Ground Truth	Prediction	
	Ship	Non-Ship
Ship	TP	FN
Non-Ship	FP	TN

Table 4
Complexity comparison of different methods, including parameters and computation cost and inference speed.

Model	Faster-RCNN	YOLOv3	YOLOv3-Efficientnet	YOLOv3-SPP	YOLOv4	YOLOv5-x	YOLOv5-l	YOLOv5-m	Ours
Input Size	300 × 300	320 × 320	320 × 320	320 × 320	320 × 320	320 × 320	320 × 320	320 × 320	320 × 320
Params (M)	28.275M	61.523	10.552	62.573	63.938	87.244	46.631	21.056	4.848
FLOPs (G)	448.29	19.38	1.10	19.49	17.68	54.449	28.640	12.650	1.29
FPS	-	21.66	29.28	21.34	20.67	13.22	24.35	44.34	59.89

diagram of the device, Fig. 8 (b) shows the real hardware of the images acquisition consists of SWIR and LWIR cameras. Fig. 9 (a)(b)(c) shows the partially acquired SWIR, LWIR, and fused images respectively.

The SWIR, LWIR, and Fusion datasets were divided into three parts in this experiment with a ratio of 0.8:0.1:0.1, respectively. That is 835×3 images for training, 104×3 images for validation, 104×3 images for testing.

3.2. Evaluation criteria

In this study, the metrics including Precision, Recall, F1 and Average Precision (AP), as shown in Eq (12), Eq (13), Eq (14) and Eq (15) respectively, frames per second (FPS), parameters and floating-point operations (FLOPs) were used to evaluate the performance of the proposed method. The meanings of true positive (TP), true negative (TN), false positive (FP), false negative (FN) are shown in Table 3. Precision is the probability of correctly predicting positive samples among positive predictions, as shown in Equation. Recall represents the probability of correctly classifying a true positive sample, as shown in Equation. F1 is a comprehensive indicator that combines precision and recall to evaluate the performance of the model. Average precision is the area enclosed under the precision-recall curve calculated by integration, as shown in Equation. Frames per second (FPS) is the number of frames processed by the model in a second, which represents the inference speed and real-time performance of the model. FLOPs represent the computation cost of the model, it with parameters evaluate the complexity of the model.

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{14}$$

$$AP = \int_0^1 P(R) dR \tag{15}$$

3.3. Computational complexity and inference speed

Compared with other methods, the proposed method has fewer parameters, low FLOPs and faster inference speed, as shown in the Table 4. We deployed all models on the embedded device Jetson TX2, and tested the inference speed with 10,000 images. The results illustrated that our method's inference speed close to 60 FPS with limited computing resources.

3.4. Comparative experiments on SWIR, LWIR and fusion datasets

To evaluate the detection performance, we quantitatively compare our method with other representative object detection methods, such as YOLOv3 [19], YOLOv3-SPP (YOLOv3 with SPP module [32]), YOLOv4 [33], on our previously published datasets.

3.4.1. Experiments on SWIR images

SWIR imaging system relies on receiving reflected light for imaging, so disfavored illumination and dense sun glint, may cause poor contrast and targets are easily drowned in background noise. Seawater absorbs SWIR light, so the boundary between ocean and sky is usually obvious in SWIR images. Fig. 10 illustrates the results of different methods for ship detection in SWIR images. Under poor illumination conditions or specific detection angles, structural and texture information about the target is lost in SWIR images. For some small size ships, accurate detection is difficult and can easily cause missed detection. Our method is effective in detecting

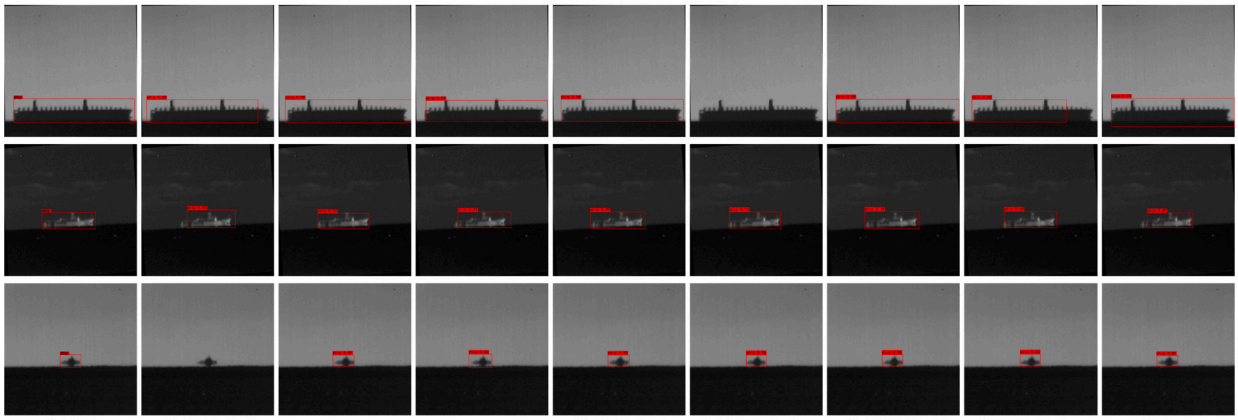

Fig. 10. The visual comparisons of different methods for ship detection on SWIR images. There is less sea clutter and other noise in SWIR images, and there is a more prominent contrast difference between ship and background. (a) is Ground true (GT) image. The variable size of ships brings challenges to the detection. YOLOv3 failed to detect the ship of small size, while YOLOv5-x failed in detecting large-size ships. Our method accurately detected ships of different sizes on SWIR images.

Table 5
Performance comparison of the proposed method with other representative networks on SWIR dataset.

Model	Input Size	SWIR			
		AP @0.7	F1	Recall	Precision
YOLOv3	320 × 320	96.48%	0.93	90.48%	95.96%
YOLOv3-Efficientnet	320 × 320	97.99%	0.98	98.10%	98.10%
YOLOv3-SPP	320 × 320	95.36%	0.94	91.43%	96.00%
YOLOv4	320 × 320	96.31%	0.97	97.14%	96.23%
YOLOv5-x	320 × 320	97.06%	0.94	89.52%	97.92%
YOLOv5-l	320 × 320	98.71%	0.95	93.33%	97.03%
YOLOv5-m	320 × 320	93.61%	0.94	93.33%	94.23%
Ours	320 × 320	95.05%	0.96	95.24%	96.15%

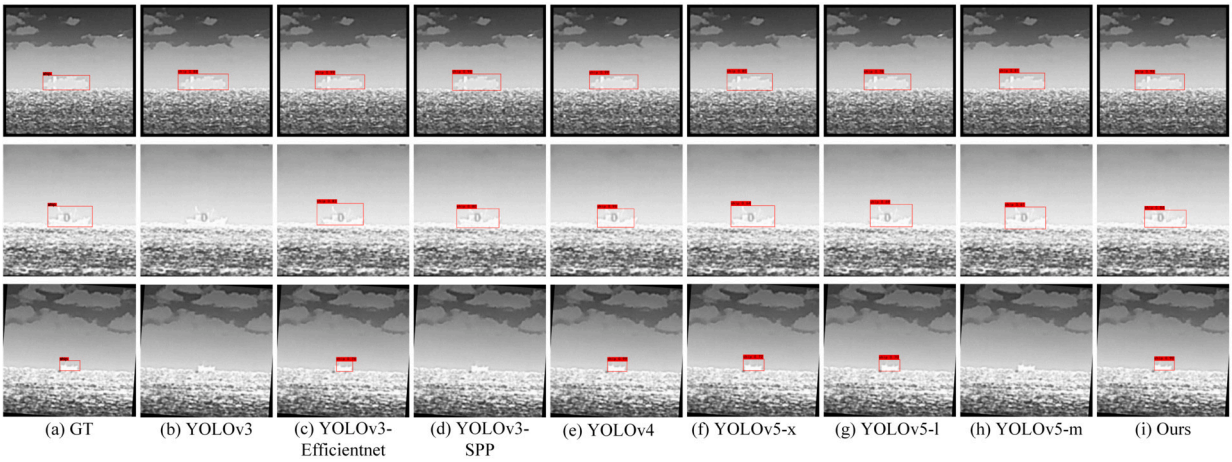

Fig. 11. The visual comparisons of different methods for ship detection on LWIR images. (a) is Ground true (GT) image. The LWIR image has low contrast, it is difficult to distinguish between the ship and the background (sky, sea clutter), due to they have similar gray level distribution. YOLOv3, YOLOv3-SPP and YOLOv5-m failed to detect ships in some LWIR images. Our method can precisely detect ships in LWIR images with low contrast.

small size ships or ships with lacking structural and texture information. Table 5 indicates a quantitative comparison of different methods for ship detection in SWIR images.

Table 5 shows the performance comparison of the proposed method with other representative methods on the SWIR dataset. Experimental results demonstrate that the performance of the proposed method is close to the state-of-the-art method with fewer parameters and low computation cost.

Table 6
Performance comparison of the proposed method with other representative networks on LWIR dataset.

Model	Input Size	LWIR			
		AP @0.7	F1	Recall	Precision
YOLOv3	320 × 320	91.32%	0.89	82.86%	95.60%
YOLOv3-Efficientnet	320 × 320	95.87%	0.97	97.14%	97.14%
YOLOv3-SPP	320 × 320	96.00%	0.93	87.62%	98.92%
YOLOv4	320 × 320	94.84%	0.95	95.24%	95.24%
YOLOv5-x	320 × 320	93.57%	0.94	95.24%	92.59%
YOLOv5-l	320 × 320	97.75%	0.93	95.24%	90.09%
YOLOv5-m	320 × 320	94.67%	0.93	95.24%	91.74%
Ours	320 × 320	92.69%	0.93	92.38%	94.17%

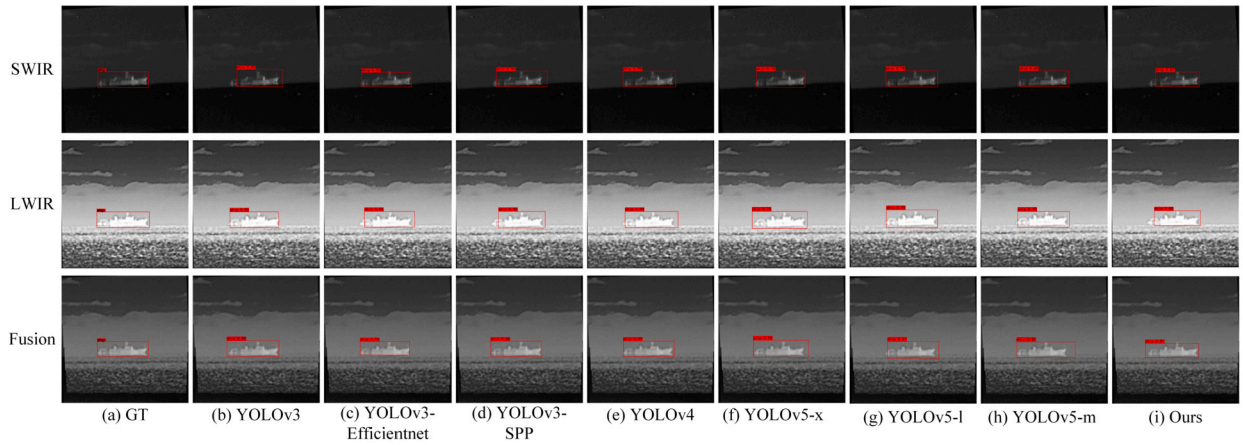

Fig. 12. The fused images have more information than SWIR and LWIR, which is beneficial to locate the ship's position more precisely, as shown in the figure. However, fused images also bring drawbacks from SWIR and LWIR images, such as increased noise and decreased contrast. There is more noise from LWIR images in fused images. (a) is Ground true (GT) image.

3.4.2. Experiments on LWIR images

LWIR images exhibit intensity inhomogeneity due to the uneven heat dissipation. In addition, ship target and background (e.g. sea clutter, sky and clouds) have similar intensity distributions in LWIR images. These issues bring challenges to ship detection with LWIR images. Fig. 11 illustrates the visual comparisons of different methods for ship detection in LWIR images. The proposed method can correctly detect ships in LWIR images with low contrast and interference from the background. Table 6 indicates a quantitative comparison of different methods for ship detection in LWIR images. The proposed method is close to being the state-of-the-art method in various metrics. From experimental results, we can see that the accuracy of LWIR images is lower compared with the ship detection accuracy of SWIR images. Although LWIR system has better imaging quality for targets at night or under poor illumination conditions. However, in daytime ship detection, LWIR imaging has no obvious advantages [34]. All of these issues contribute to making ship detection in LWIR difficult.

3.4.3. Experiments on fused images

Fusion images can provide comprehensive information about marine scenarios, which can help make more robust decisions. Several ship detection results of fusion, SWIR and LWIR images are visually illustrated in Fig. 12. However, fused images also bring drawbacks from SWIR and LWIR images. Disturbed by sea clutter and limited by the fusion methods, ships and clutters may have inapparent difference, which is not conducive to ship detection. Table 7 shows the quantitative comparison of different methods on the fusion dataset. Compared with SWIR and LWIR datasets, the detection accuracy of all methods decreases. While fused images can provide more information in a variety of harsh illumination conditions and environments, fused images bring weaknesses in SWIR and LWIR images. In order to exploit the advantages of fused images, better image fusion strategies are needed.

3.4.4. Discussion of SWIR, LWIR, fused images in ship detection

From the experimental results, SWIR images have the highest accuracy in ship detection. This is due to the characteristics of SWIR imaging. Seawater absorbs SWIR light, which suppresses interference from sea clutter, resulting in strong contrast between ships and the background. SWIR imagers rely on reflected light for imaging; thus, SWIR images obtain more information than thermal infrared such as texture information. In addition, SWIR imaging has strong penetration ability for haze and fog. For low light level condition, it can utilize the night glow phenomenon and it can provide wide dynamic imaging (from daylight to overcast night conditions) [2].

Table 7
Performance comparison of the proposed method with other representative networks on Fusion dataset.

Model	Input Size	Fusion			
		AP @0.7	F1	Recall	Precision
YOLOv3	320 × 320	77.93%	0.81	79.05%	83.00%
YOLOv3-Efficientnet	320 × 320	91.79%	0.94	92.38%	95.10%
YOLOv3-SPP	320 × 320	92.47%	0.89	83.81%	95.65%
YOLOv4	320 × 320	91.84%	0.88	88.57%	87.74%
YOLOv5-x	320 × 320	84.68%	0.87	82.86%	90.62%
YOLOv5-l	320 × 320	93.20%	0.90	85.71%	93.75%
YOLOv5-m	320 × 320	89.61%	0.88	83.81%	93.62%
Ours	320 × 320	86.34%	0.85	80.00%	91.30%

Table 8
Discussion on the advantages and disadvantages of SWIR, LWIR and fused images in ship detection.

Type of image	Advantages	Disadvantages
SWIR	(1) SWIR imaging exhibits strong penetration capabilities for haze and fog. (2) SWIR can utilize the night glow phenomenon and it can provide wide dynamic imaging. (3) SWIR images display high contrast and abundant detail information.	SWIR imaging is susceptible to interference from strong sunlight reflection.
LWIR	LWIR can imaging without any natural or artificial illumination required.	(1) LWIR imaging may not effectively capture the target during day and night alternation. (2) LWIR imaging is susceptible to excessive sea clutter under direct sunlight.
Fusion	Fused images can provide more information in a variety of harsh illumination conditions and environments.	(1) Fused images exhibit relatively low contrast. (2) Fused images contain noise from the LWIR image.

Table 9
Ablation experiments on the effect of Bottleneck, Ghost, CBAM modules on model parameters, FLOPs, and inference speed. Model1 is the original YOLOv3 network.

Model	Bottleneck	Ghost module	CBAM	Parameters (M)	FLOPs (G)	FPS
Model1				61.523	19.38	21.66
Model2	✓			22.273	5.32	43.54
Model3	✓	✓		4.750	1.29	62.50
Model4	✓	✓	✓	4.848	1.29	59.89

LWIR can image without any natural or artificial light sources, which is advantageous for all-weather monitoring. However, LWIR imaging may not effectively capture targets during day-night transitions. Under direct sunlight, LWIR images may contain too much sea clutter, which is not conducive to ship detection.

Fused images provide more information in a variety of harsh illumination conditions and environments. However, fused images also bring drawbacks from SWIR and LWIR images. Disturbed by sea clutter and limited by the fusion methods, ships and clutters may have inapparent difference, which is not conducive to ship detection.

On the basis of the above analysis, Table 8 summarizes the advantages and disadvantages of SWIR, LWIR and fused images for ship detection in the marine environment.

3.5. Ablation experiments

To validate the effectiveness of the method and submodules, ablation experiments were conducted in the paper. Table 9 shows the effect of different submodules on network parameters, computational complexity, and inference speed. Model1 is the original YOLOv3 network. The application of bottleneck in the backbone network greatly reduces the parameters and computational complexity of the model, and improves the inference speed. The Ghost module is used in the Neck and Head networks to further reduce the network parameters and computational complexity, and to further improve the model's inference speed. The addition of the CBAM module did not significantly increase the model's parameters and computational complexity.

Tables 10, 11, and 12 show the effect of different submodules on the accuracy of ship detection in SWIR, LWIR, and fused images, respectively. The use of Bottleneck and Ghost modules reduces the model's parameters and computational complexity and improves the model's inference speed. However, as the parameters decrease, the ship detection accuracy of the model in SWIR, LWIR, and fused images also decreases. The use of CBAM improves the ship detection accuracy while maintaining the lightweight and high inference speed of the model.

Table 10
Ablation experiments on the effect of Bottleneck, Ghost, and CBAM modules on ship detection in SWIR images. Model1 is the original YOLOv3 network.

Model	Bottleneck	Ghost module	CBAM	AP@0.7	F1	Recall	Precision
Model1				96.48%	0.93	90.48%	95.96%
Model2	✓			92.52%	0.91	90.48%	91.35%
Model3	✓	✓		90.72%	0.90	89.52%	89.52%
Model4	✓	✓	✓	95.05%	0.96	95.24%	96.15%

Table 11
Ablation experiments on the effect of Bottleneck, Ghost, and CBAM modules on ship detection in LWIR images. Model1 is the original YOLOv3 network.

Model	Bottleneck	Ghost module	CBAM	AP@0.7	F1	Recall	Precision
Model1				91.32%	0.89	82.86%	95.60%
Model2	✓			87.66%	0.88	85.71%	90.91%
Model3	✓	✓		85.28%	0.86	83.81%	88.00%
Model4	✓	✓	✓	92.69%	0.93	92.38%	94.17%

Table 12
Ablation experiments on the effect of Bottleneck, Ghost, and CBAM modules on ship detection in fused images. Model1 is the original YOLOv3 network.

Model	Bottleneck	Ghost module	CBAM	AP@0.7	F1	Recall	Precision
Model1				77.93%	0.81	79.05%	83.00%
Model2	✓			73.83%	0.77	73.33%	81.91%
Model3	✓	✓		71.71%	0.69	60.00%	80.77%
Model4	✓	✓	✓	86.34%	0.85	80.00%	91.30%

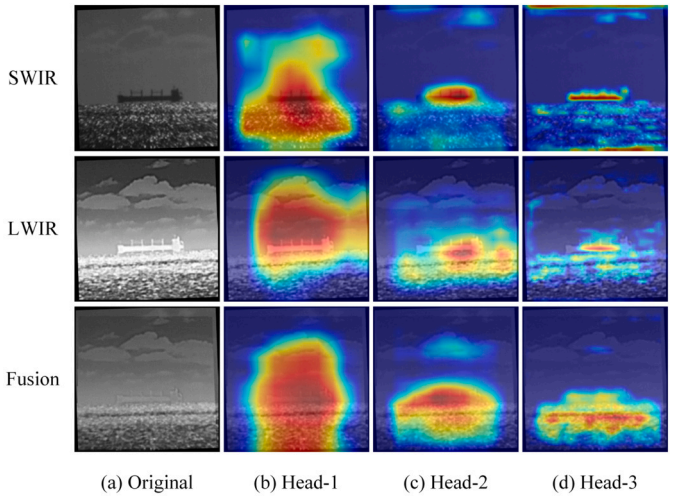

Fig. 13. Visualization results of the proposed network for three detection heads in SWIR, LWIR and fusion images. For heads at different scales, the network has different sizes of regions of interest (ROI). From head 1 to 3, the size of the ROI decreases sequentially.

3.6. Network visualization

For the qualitative analysis, we apply the Eigen-CAM [35] to generate visual explanation, as shown in Fig. 13. Class activation map (CAM) focuses on making sense of what the model learns from the data, calculating the importance of the spatial locations in the convolutional networks. Eigen-CAM can produce a location map highlighting the important regions in the input image, providing interpretability and transparency to the network.

Fig. 13 shows the visualization results of the proposed network for three detection heads in the SWIR, LWIR and fusion images. Fig. 13 (a) is the original image and Fig. 13 (b)(c)(d) shows the ROI of the detection heads Head-1, Head-2, Head-3, respectively. Head 1 has the largest receptive field, from head 1 to 3, the size of the ROI decreases sequentially. This is advantageous for accurate detection of ships in different scales. For the SWIR image, there is a significant grayscale difference between the ship and background and heads in different scales can precisely locate the ship's location in the image. In LWIR images, there is more interference from sea

clutter and clouds, and the contrast between the ship and background is low. The network cannot locate the ship and its surroundings in LWIR image. For fusion image, the grayscale distribution between the ship and the image is analogous, which makes it difficult for the network to locate the ship's location. The sea clutter in the fused image is even more pronounced than the ship. We can see that in head 3 of the fused image, the network focus on the sea cluster, which also explains why the detection accuracy of the fusion image is lower than that of LWIR and SWIR images. In subsequent studies, better image fusion strategies are needed to suppress noise in SWIR and LWIR images and improve the contrast of ships in images.

4. Conclusion

Ship detection is the core component for realizing the application of autonomous ships and improve maritime traffic safety. Infrared imaging systems with different spectral ranges can provide robust decision for ship detection in marine environments. In the open sea, unmanned marine vehicles (e.g., unmanned surface vehicles and unmanned ships) have limited hardware resources, and complicated methods are difficult to deploy on a computationally limited platform.

In this paper, we have presented a lightweight CNN for ship detection in SWIR, LWIR and fused images. Specially, the backbone is built from bottleneck depth-separable convolution with residuals. Redundant feature maps are generated using cheap linear operations in the neck and head networks. The network's learning and representational capacities are promoted by introducing channel and spatial attention, and redesigning the sizes of anchor boxes. Experiments have verified the accuracy and reliability of the algorithm adopted in this paper. The proposed method has high inference speed in an embedded device with limited computing resources, and the detection accuracy is close to that of the state-of-the-art method.

In future research, we will explore improved strategies for fusing infrared images in various spectral bands to improve the ship detection accuracy in fusion images. Moreover, combining multispectral, radar and infrared data will provide richer target information and improve the performance of ship detection. In coming research, the proposed method will be deployed to other platforms, such as digital signal processor (DSP), field-programmable gate array (FPGA), to evaluate algorithm performance and prepare for practical applications in marine environments. The proposed method will contribute to automation and autonomy for marine vehicles and systems and enhance the maritime traffic safety in practical applications.

CRediT authorship contribution statement

Liqian Wang: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper. Yakui Dong: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data. Cheng Fei: Conceived and designed the experiments; Performed the experiments; Contributed reagents, materials, analysis tools or data. Junliang Liu: Conceived and designed the experiments; Performed the experiments. Shuzhen Fan: Conceived and designed the experiments; Performed the experiments. Yunxia Liu: Analyzed and interpreted the data. Yongfu Li: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data. Zhaojun Liu: Contributed reagents, materials, analysis tools or data. Xian Zhao: Contributed reagents, materials, analysis tools or data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data included in article/supp. material/referenced in article.

Funding statement

This work was partially supported by "Natural Science Foundation of Shandong Province (ZR2022MF323, ZR2022LLZ002)", "Key R&D Plan of Shandong Province (2020JMRH0101)", "the Fundamental Research Funds for the Central Universities (2021JCG018)", "Opening Foundation of Key Laboratory of Infrared Imaging Materials and Devices, Chinese Academy of Sciences (IIMDFJJ-21-11)", "Shandong University Equipment Development Cultivation Project (zy202004)".

References

[1] M. Akdağ, P. Solnør, T.A. Johansen, Collaborative collision avoidance for maritime autonomous surface ships: a review, Ocean Eng. 250 (2022) 110920.
[2] D. Perić, B. Livada, Analysis of swir imagers application in electro-optical systems, in: Proceedings of the Proceedings of 4th International Conference on Electrical, Electronics and Computing Engineering, IcETRAN, Kladovo, Serbia, 2017, pp. 5–8.
[3] S. Yuan, Y. Li, F. Bao, H. Xu, Y. Yang, Q. Yan, S. Zhong, H. Yin, J. Xu, Z. Huang, et al., Marine environmental monitoring with unmanned vehicle platforms: present applications and future prospects, Sci. Total Environ. (2022) 159741.
[4] F. Yang, Z. Liu, X. Bai, Y. Zhang, An improved intuitionistic fuzzy c-means for ship segmentation in infrared images, IEEE Trans. Fuzzy Syst. (2020).
[5] X. Bai, Z. Chen, Y. Zhang, Z. Liu, Y. Lu, Infrared ship target segmentation based on spatial information improved fcm, IEEE Trans. Cybern. 46 (12) (2015) 3259–3271.

[6] Z. Liu, F. Zhou, X. Chen, X. Bai, C. Sun, Iterative infrared ship target segmentation based on multiple features, Pattern Recognit. 47 (9) (2014) 2839–2852.
[7] A. Mumtaz, A. Jabbar, Z. Mahmood, R. Nawaz, Q. Ahsan, Saliency based algorithm for ship detection in infrared images, in: 2016 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST), IEEE, 2016, pp. 167–172.
[8] L. Li, G. Liu, Z. Li, Z. Ding, T. Qin, Infrared ship detection based on time fluctuation feature and space structure feature in sun-glint scene, Infrared Phys. Technol. 115 (2021) 103693.
[9] Y. Li, Z. Li, Y. Zhu, B. Li, W. Xiong, Y. Huang, Thermal infrared small ship detection in sea clutter based on morphological reconstruction and multi-feature analysis, Appl. Sci. 9 (18) (2019) 3786.
[10] Y. Zhang, Q.-Z. Li, F.-N. Zang, Ship detection for visual maritime surveillance from non-stationary platforms, Ocean Eng. 141 (2017) 53–63.
[11] R.W. Liu, W. Yuan, X. Chen, Y. Lu, An enhanced cnn-enabled learning method for promoting ship detection in maritime surveillance system, Ocean Eng. 235 (2021) 109435.
[12] K. Kim, S. Hong, B. Choi, E. Kim, Probabilistic ship detection and classification using deep learning, Appl. Sci. 8 (6) (2018) 936.
[13] X. Chen, Y. Yang, S. Wang, H. Wu, J. Tang, J. Zhao, Z. Wang, Ship type recognition via a coarse-to-fine cascaded convolution neural network, J. Navig. 73 (4) (2020) 813–832.
[14] X. Nie, M. Yang, R.W. Liu, Deep neural network-based robust ship detection under different weather conditions, in: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), IEEE, 2019, pp. 47–52.
[15] Z. Liu, X. Zhang, T. Jiang, T. Zhang, B. Liu, M. Waqas, Y. Li, Infrared salient object detection based on global guided lightweight non-local deep features, Infrared Phys. Technol. 115 (2021) 103672.
[16] Z. Shao, L. Wang, Z. Wang, W. Du, W. Wu, Saliency-aware convolution neural network for ship detection in surveillance video, IEEE Trans. Circuits Syst. Video Technol. 30 (3) (2019) 781–794.
[17] Z. Song, J. Yang, D. Zhang, S. Wang, Z. Li, Semi-supervised dim and small infrared ship detection network based on haar wavelet, IEEE Access 9 (2021) 29686–29695.
[18] X. Chen, X. Wu, D.K. Prasad, B. Wu, O. Postolache, Y. Yang, Pixel-wise ship identification from maritime images via a semantic segmentation model, IEEE Sens. J. 22 (18) (2022) 18180–18191, https://doi.org/10.1109/JSEN.2022.3195959.
[19] J. Redmon, A. Farhadi, Yolov3: an incremental improvement, preprint, arXiv:1804.02767, 2018.
[20] S. Saha, F. Bovolo, L. Bruzzone, Unsupervised deep change vector analysis for multiple-change detection in vhr images, IEEE Trans. Geosci. Remote Sens. 57 (6) (2019) 3677–3693.
[21] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.
[22] J. Hu, L. Shen, S. Albanie, G. Sun, A. Vedaldi, Gather-excite: exploiting feature context in convolutional neural networks, Adv. Neural Inf. Process. Syst. 31 (2018).
[23] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.-S. Chua Sca-cnn, Spatial and channel-wise attention in convolutional networks for image captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5659–5667.
[24] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, R. Qu, A survey of deep learning-based object detection, IEEE Access 7 (2019) 128837–128868.
[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520.
[26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
[27] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: efficient convolutional neural networks for mobile vision applications, preprint, arXiv:1704.04861, 2017.
[28] X. Wu, D. Sahoo, S.C. Hoi, Recent advances in deep learning for object detection, Neurocomputing 396 (2020) 39–64.
[29] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, Ghostnet: More features from cheap operations, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1580–1589.
[30] C. F. O. RESEARCH, ENGINEERING, Infrared Ship Dataset, Website, http://www.gxzx.sdu.edu.cn/info/1133/2174.htm, 2020.
[31] Y. Dong, C. Fei, G. Zhao, L. Wang, Y. Liu, J. Liu, S. Fan, Y. Li, X. Zhao, Registration method for infrared and visible image of sea surface vessels based on contour feature, Heliyon 9 (3) (2023).
[32] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.
[33] A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, Yolov4: optimal speed and accuracy of object detection, preprint, arXiv:2004.10934, 2020.
[34] J.D. Stets, F.E. Schöller, M.K. Plenge-Feidenhans, R.H. Andersen, S. Hansen, M. Blanke, Comparing spectral bands for object detection at sea using convolutional neural networks, J. Phys. Conf. Ser. 1357 (2019) 012036, IOP Publishing.
[35] M.B. Muhammad, M. Yeasin, Eigen-cam: class activation map using principal components, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7.