MYU Tokyo

S & M 3126

Deep Learning Method for Ship Detection in Nighttime Sensing Images

Yunfeng Nie,[1*] Yejia Tao,[1] Wantao Liu,[1,2] Jiaguo Li,[2] and Bingyi Guo[1]

[1]School of Information Engineering, Nanchang Hangkong University, Nanchang 330063, China
[2]Aerospace Information Institute, Chinese Academy of Sciences, Beijing 100094, China

(Received July 25, 2022; accepted December 6, 2022)

Keywords: ship detection, nighttime remote sensing, size expansion, attention mechanism, feature pyramid network, modified CycleGAN

Nighttime ship detection is challenging due to the complicated interference of the nighttime background and the weak characteristics of ship targets, and research in this area is relatively scarce. In this study, we proposed a network called Size Expansion Attention Fusion Faster R-CNN (SEAFF), which is based on the Faster R-CNN deep convolutional network integrated with size expansion (SE), the attention mechanism (AM), and the feature pyramid network (FPN). Firstly, SE is adopted to enhance the spatial features of nighttime ship targets. Secondly, the AM is embedded to extract the features of nighttime ship targets from their channel and spatial dimensions. Lastly, the FPN is combined to compensate for the lack of feature extraction at different levels. In the data preprocessing, we first choose images generated by a Luojia 1-01 nighttime high-resolution sensor, then we adopt a modified cycle-consistent adversarial network (CycleGAN) to augment the dataset through a sample generation experiment. Our experiment on ship detection demonstrated that (1) the SE module improved the detection of weak and small ship targets; (2) the AM module plays an important role in reducing the impact of complex backgrounds; (3) the FPN module has a significant effect on suppressing the missed detection of nighttime ship targets. Moreover, compared with the mainstream object detection methods of a single-shot multibox detector, YOLOv5, and Faster R-CNN, the AP@0.50, AP@0.75, and AP@0.50:0.95 indicators of SEAFF were improved by 0.032, 0.048, and 0.029, respectively. The advantages of our network indicate its potential use in complex nighttime scenes.

1. Introduction

The variety of remote sensing images has grown with advances in remote sensing technology, with studies on nighttime remote sensing attracting increasing attention. Compared with regular daytime remote sensing images, nighttime remote sensing images can depict human activities more directly, because human-created nightlights are primary sources of nighttime remote sensing data. Thus, nighttime ship detection is of great value in combating illegal fishing,(1) detecting the invasion of military targets,(2) and evaluating port commerce activities.(3) Despite

[*]Corresponding author: e-mail: 70064@nchu.edu.cn

https://doi.org/10.18494/SAM4037

the proposal of various ship detection methods, nighttime ship detection remains difficult due to the complexity of nighttime scenes.

At present, images for ship detection mainly consist of synthetic aperture radar (SAR) images, IR images, and optical remote sensing images.(4–6) SAR images are used to detect ships by analyzing the amplitude and phase information of affected objects to obtain ship information in grayscale. SAR images have the major advantages of anti-interference and detection range, but the color and texture information of ship targets cannot be detected, and multifeature information is lost during the training, which reduces the detection performance.(7) IR images are formed by obtaining the IR band radiation of the targets. Although they are suitable for detection in all types of weather, IR images also have several shortcomings, such as low pixel resolution, low signal-to-noise ratio, and blurred image edges, making this approach prone to misdetection, false detection, and other problems.(8) Optical remote sensing is greatly affected by the weather and depends on the sun as an external light source; thus, it cannot be used for target detection in nighttime scenes. Nighttime remote sensing can obtain near-IR electromagnetic wave information emitted from the earth's surface under cloudless conditions. Much of this information is generated by human activities on the surface, with the most important being human nighttime lighting, fishing boats at sea, and forest fires. Compared with images from common remote sensing satellites, nighttime remote sensing images reflect human activities more directly.

Many sensors have the ability to detect the light reflected from the earth's surface at night, including the Operational Linescan System (OLS) sensor carried by the Defense Meteorological Satellite Program (DMSP) US military weather satellite, the Visible IR Imaging Radiometer Suite (VIIRS) sensor carried by the Suomi NPP(c) satellite, and China's Jilin-1 satellite. With the proliferation of multisource data, an increasing number of studies are focusing on ship detection in complex nighttime scenes. Ruiz *et al.* combined automatic identification system (AIS) and VIIRS data to monitor major global fisheries and discovered that the courses of fishing vessels were highly consistent in time and space, confirming the importance of remote sensing monitoring for fisheries.(9) Li *et al.* combined VIIRS nighttime remote sensing data to expand the AIS dataset and monitor fisheries in the north of the South China Sea.(10) These methods provide an important foundation for the oversight of fishery management systems.

Increasingly efficient detection algorithms, such as region proposal with convolution neural networks (R-CNNs),(11) spatial pyramid pooling networks (SSP-NETs),(12) and Faster R-CNN, have been proposed, which employ deep convolutional neural networks (CNNs)(13) in object detection. Faster R-CNN provides a region proposal network (RPN) and boosts detection efficiency while achieving end-to-end training. In contrast to the methods that depend on region proposals, You Only Look Once (YOLO)(14) and the single-shot multibox detector (SSD)(15) estimate the object region directly and enable true real-time detection. The above visual detection methods are also commonly utilized in ship detection via remote sensing. Zhang *et al.*(16) proposed a new ship detection method that combines a CNN with an enhanced saliency detection method. Liu *et al.*(17) proposed a framework for a sea-land segmentation-based convolutional neural network (SLS-CNN) for ship detection. The rotation dense feature pyramid networks (FPNs) proposed by Yang *et al.*(18) have achieved state-of-the-art performance. Liu *et* *al.*[19] applied the method of light spot detection and tracking in ship detection and tracking and improved maritime surveillance efficiency. The difficulties of detecting nighttime ship targets are mainly related to the complex background with interference, including similar nighttime targets such as nearshore, island, and offshore oil and gas platforms. In addition, nighttime ship targets are extremely small, which makes their detection challenging.

In this paper, we first propose a modified cycle-consistent adversarial network (CycleGAN) to solve the checkerboard artifact of generated samples. We adopt the nearest neighbor sampling method to replace the deconvolution structure to improve the network of the generator, and we introduce a new perceptual loss function to improve the diversity of generated samples. Then, we propose a network called Size Expansion Attention Fusion Faster R-CNN (SEAFF), which is based on Faster R-CNN and combines size expansion (SE), the attention mechanism (AM), and the FPN. The SE module is used to enhance the spatial distribution of nighttime ship targets, the AM module is adopted to filter the effective features from the channel and spatial dimensions, and the FPN module is introduced to fuse extracted features at different levels to compensate for the missing features. Compared with mainstream object detection methods such as SSD, YOLOv5, and Faster R-CNN, our method is more suitable for ship detection and achieves higher detection accuracy.

The paper is organized as follows. In Sect. 2, we introduce details of data acquisition and preprocessing. Section 3 describes the workflow and details of our proposed network. In Sect. 4, we analyze our network using experimental results. Section 5 concludes the paper.

2. Data Preprocessing

2.1 Region selection

As the region considered, we chose the navigable inshore area of the Beibu Gulf, China, which is an economically developed part of the country with trade links to more than 80 countries. It is located in China's well-known Beibu Gulf fishing ground, in which fishing often occurs at night, meeting the necessary conditions for the extraction of nighttime ship images.

2.2 Data acquisition

In 2018, Wuhan University in China launched Luojia 1-01, a high-resolution nighttime remote sensing satellite capable of both nighttime remote sensing and navigation enhancement. It is designed for high-resolution imaging, compression, and storage integration and has a high spatial resolution, temporal resolution, radiation resolution, and signal-to-noise ratio, effectively solving the problem of the supersaturation effect in DMSP satellites.[20,21] The spatial resolution of Luojia 1-01 nighttime light data is greatly improved with on-board calibration, making it ideal for our requirements.

Note that the data generated by Luojia 1-01 (available at http://59.175.109.173:8888) is released with geometric and radiometric correction and is stretched exponentially to facilitate data storage. Thus, before labeling the data, it must be stretched reversely to its original radiance using the formula

$$L = DN^{3/2} \cdot 10^{-10}, \tag{1}$$

where L is the absolute radiance corrected for radiation and DN is the gray value of the image.

2.3 Data augmentation

Deep learning methods require a large number of labeled samples for training, testing, and validation. Unfortunately, there are no public datasets of nighttime radiation sources of ship targets, making it necessary to manually establish a dataset for our requirements.

Techniques such as rotation, mirroring, and adding noise are widely used in computer vision for data augmentation. We adopted some of these techniques as shown in Fig. 1. These techniques also improve the generalization ability of our network.

Common data augmentation techniques are regular transformations based on the original images. To solve the problem of insufficient samples for training, Goodfellow *et al.* proposed a generative adversarial network (GAN),(22) inspired by a two-player zero-sum game. It aims to fit the distribution of samples then output highly qualified generated samples. Since then, many networks based on GAN have been proposed, such as DCGAN(23) and StackGAN.(24) Among them, CycleGAN(25) can be trained without paired examples. The network is trained in an

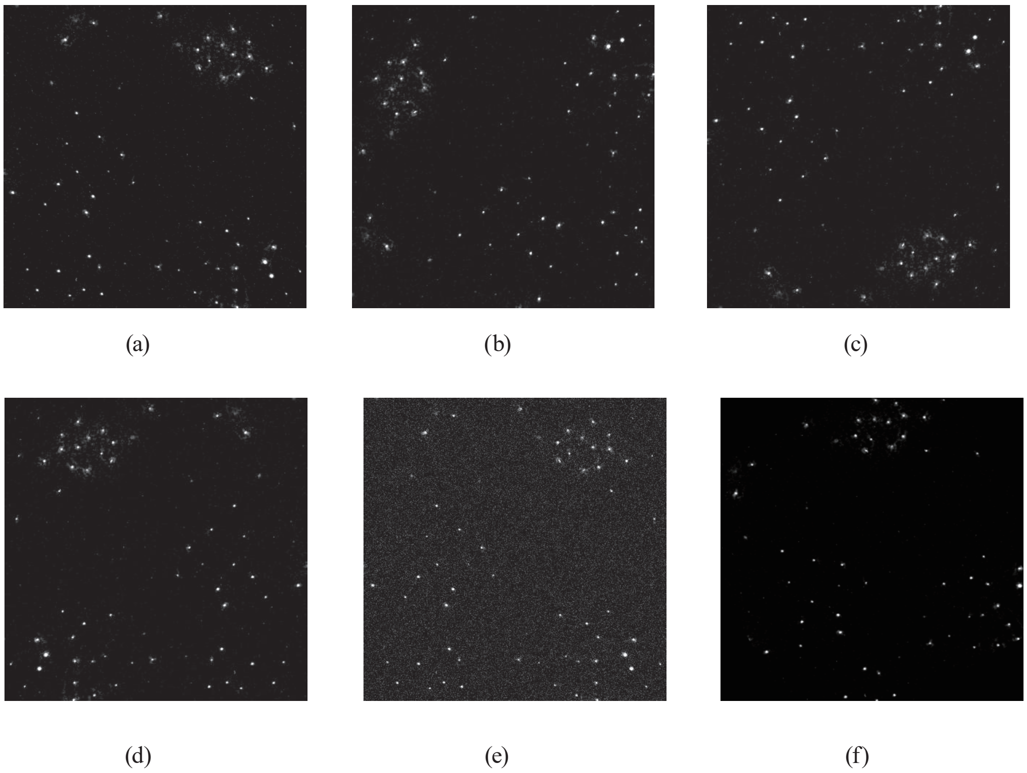

Fig. 1. Examples of common data augmentation techniques. (a) Original image, (b) rotation, (c) mirror vertically, (d) mirror horizontally, (e) addition of noise, and (f) affine transformation.

unsupervised manner using a collection of images from the source and target domains that need not be related in any way. In this paper, we adopted CycleGAN to further augment samples.

Deconvolution in the CycleGAN decoder is mainly used to amplify the feature map. However, because the kernel size and stride size are not divisible, there are pseudo pixels called checkerboard artifacts,(26) as shown in Fig. 2. In CycleGAN, deconvolution with a stride size of 2 and a 3 × 3 convolution kernel is continuously used, resulting in checkerboard artifacts.

To address this problem, we propose a modified structure of the generator as shown in Fig. 3. In this paper, the deconvolution layer is replaced by upsampling of nearest neighbor interpolation with the convolution layers. In this way, the structures of both the encoder and decoder are modified. The modified generator network structure is shown in Fig. 3.

The parameter setting for each layer of the modified generator network is shown in Table 1. The discriminator is used to judge whether samples generated by the generator are real or false. The discriminator adopts a fully convolutional network, as shown in Fig. 4.

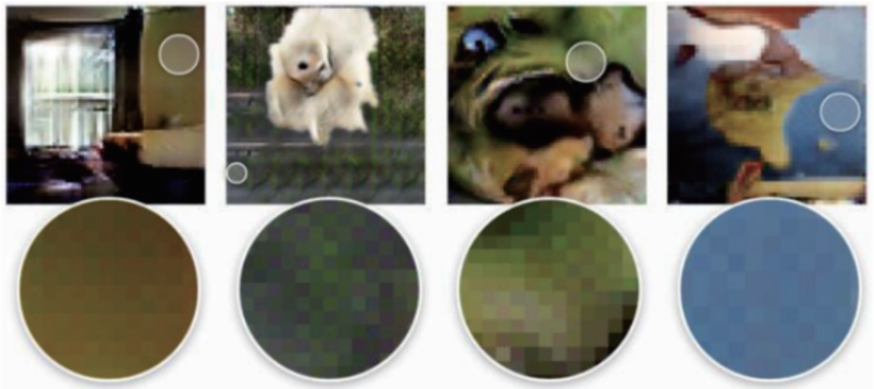

Fig. 2. (Color online) Checkerboard artifacts.

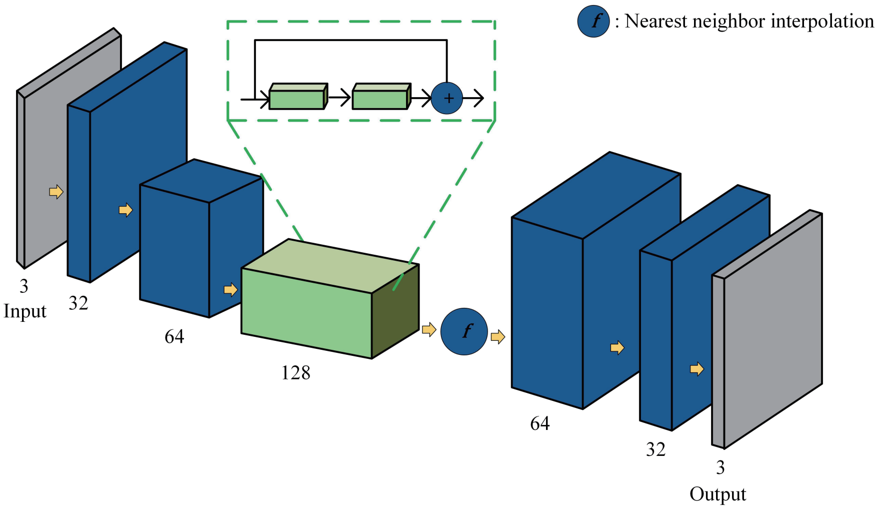

Fig. 3. (Color online) Structure of modified generator network.

Table 1
Specific parameters of modified generator network.

Instrument	#Layers	Operation	Kernel	#Channels	Stride	Padding	Activation function	Batch normal-ization
Encoder	1	Conv	7 × 7	32	1	2	ReLU	Yes
	2	Conv	3 × 3	64	2	1	ReLU	Yes
Transformer	3	ResNet	3 × 3	128	1	1	ReLU	Yes
Decoder	4	Conv	3 × 3	64	2	1	ReLU	Yes
	5	Conv	7 × 7	32	1	2	ReLU	—

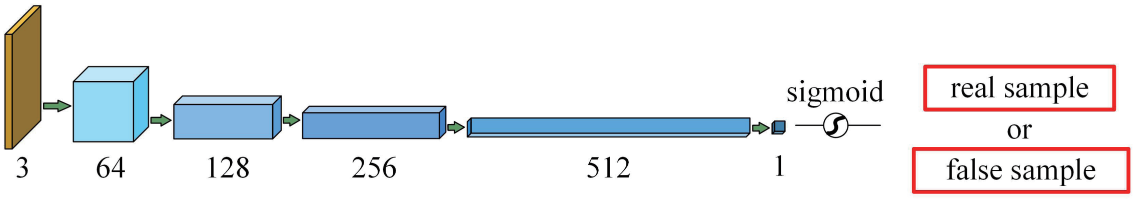

Fig. 4. (Color online) Structure of modified discriminator network.

The discriminator is composed of five convolutional layers. The parameter setting of each convolutional layer is shown in Table 2. The loss function of traditional CycleGAN includes the adversarial loss function and the cycle consistency loss function. In this paper, we propose a novel perceptual loss function for the feature-level consistency supervision of target networks, which increases the diversity of samples under the constraint of similar generative styles. First, *x*, the input image passes through two sets of adversarial networks to obtain the output *c*, then *x* and *c* are transferred to the encoder E_X, the penultimate encoder feature is output, and the L1 norm is used to constrain the two feature maps. As shown in Eq. (2), ϕ_i represents the feature parameter of the *i*th layer encoder. Generally, the perceptual loss function is activated after 30–50 training cycles. This is because the encoder does not perform multicycle training in the early stage of feature encoding and cannot discriminate target features. Thus, when calculating the loss, the parameters of the encoder E_X and E_Y remain unchanged.

$$\min_{G_X,G_Y} L_{per}\left(G_X,G_Y\right)=\left[\left\|\phi_i\left(G_X\left(E_Y\left(G_Y\left(E_X\left(x\right)\right)\right)\right)\right)-\phi_i\left(x\right)\right\|_1\right] + \left[\left\|\phi_i\left(G_Y\left(E_X\left(G_X\left(E_Y\left(y\right)\right)\right)\right)\right)-\phi_i\left(y\right)\right\|_1\right] \tag{2}$$

Adding up the above, the total loss function is finally defined as

$$L_{GAN}\left(G,D_Y,X,Y\right)+L_{GAN}\left(F,D_X,Y,X\right)+aL_{cyc}\left(G,F\right)+bL_{per}\left(G_X,G_Y\right). \tag{3}$$

We adopted the Frechet inception distance (FID), which is a commonly used assessment criterion in GANs, to measure the quality of images generated by the generator. The FID reflects

Table 2
Parameters of modified discriminator network.

#Layers	Kernel	#Channels	Stride	Padding	Activation function	Batch normalization
1	5 × 5	64	2	2	ELU	Yes
2	3 × 3	128	1	1	ELU	Yes
3	3 × 3	256	1	1	ELU	Yes
4	3 × 3	512	1	1	ELU	Yes
5	3 × 3	1	1	0	Sigmoid	No

the distance between two data distributions; the smaller the value, the better the generated images. The FID is expressed as

$$FID = \left\| \mu_r - \mu_g{}^2 \right\| + Tr\left(\sum r + \sum g - 2\left(\sum r \sum g \right)^{1/2} \right), \tag{4}$$

where *Tr* is the sum of the diagonal elements of the matrix, μ_r is the mean of the real image features, μ_g is the mean of the features of the generated image, Σr is the vector covariance of the real image features, and Σg is the vector covariance of the features of the generated image.

Note that CycleGAN requires two different styles of samples to perform the experiment. In addition to samples generated by Luojia 1-01, we also chose IR images generated by the GM IR satellite. We used both CycleGAN and modified CycleGAN to perform the style transformation from IR ship targets to nighttime ship targets. The FIDs of images generated by both GAN networks are shown in Table 3 and the experimental results of sample generation are shown in Fig. 5. It can be observed that owing to the improved network structure and loss function, the quality of generated images is improved.

3. SEAFF Network Construction

In the SEAFF network, the SE module is adopted to enlarge the ship targets since nighttime ship targets are only a point light source with a size of 5–15 pixels and a brightness of 0.0025–0.0175 W/(m^2 · sr · μm). The AM module focuses on the most relevant information to the current task among the numerous input information and filters out irrelevant information, thus improving the efficiency and accuracy of detection. The FPN module uses both high-resolution, low-level features and high-level features with high semantic information to enhance the detection performance by fusing features of different layers.

3.1 Size expansion

Bilinear interpolation is a method of 2D interpolation on a rectangle. Compared with nearest neighbor interpolation, bilinear interpolation requires more computation but the resulting image quality is higher, and the shortcoming of the nearest neighbor interpolation of discontinuous gray values is basically overcome. Bilinear interpolation is a weighted average of the values at the four corners of a rectangle. For a position (*x*, *y*) inside the rectangle, the weights are determined by the distance from the point to the four corners. Corners that are closer to the point are given a higher weighting.

Table 3
FIDs of images for the sample generation experiment.

Method	FID (%)
CycleGAN	66.5
Modified CycleGAN	64.7

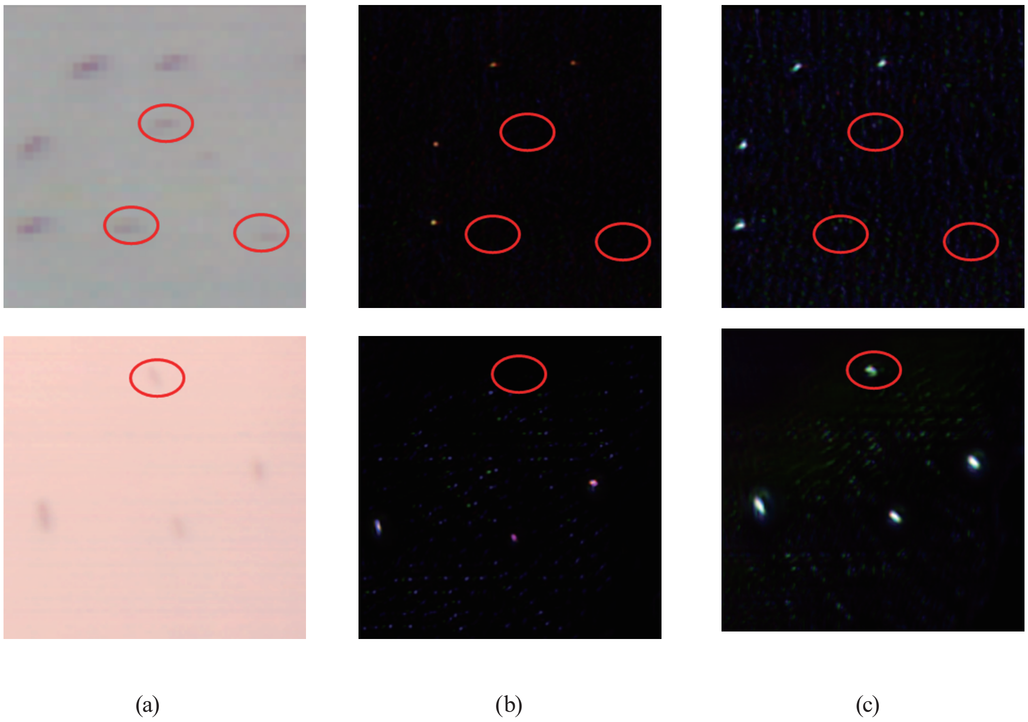

Fig. 5. (Color online) Results of sample generation experiment. (a) Original IR images, (b) images generated by CycleGAN, and (c) images generated by modified CycleGAN.

As shown in Fig. 6, to calculate the extended coordinate point P, four adjacent points Q_{11}, Q_{12}, Q_{13}, and Q_{14} with a pixel distance of one are selected to obtain the pixel values $f(Q_{11})$, $f(Q_{12})$, $f(Q_{13})$, and $f(Q_{14})$ of these coordinate points, respectively. The contribution of adjacent pixels to the pixel of coordinate point P is allocated according to the proportion of the distance between P and the adjacent coordinate points in the horizontal and vertical directions. In the horizontal direction, the computational equations are expressed as

$$f\left(x,y_1\right)\approx\frac{x_2-x}{x_2-x_1}f\left(Q_{11}\right)+\frac{x-x_1}{x_2-x_1}f\left(Q_{21}\right), \tag{5}$$

$$f\left(x,y_2\right)\approx\frac{x_2-x}{x_2-x_1}f\left(Q_{12}\right)+\frac{x-x_1}{x_2-x_1}f\left(Q_{22}\right), \tag{6}$$

where $f(x, y_1)$ and $f(x, y_2)$ represent the pixel values of R_1 and R_2, respectively, as shown in Fig. 6.

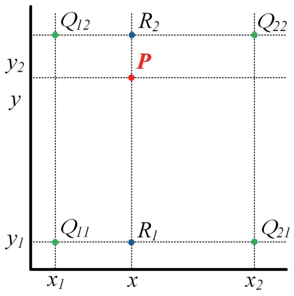

Fig. 6. (Color online) Bilinear interpolation.

In the vertical direction, the computational equation is expressed as

$$f\left(x,y\right) \approx \frac{y_2 - y}{y_2 - y_1} f\left(x,y_1\right) + \frac{y - y_1}{y_2 - y_1} f\left(x,y_2\right), \tag{7}$$

where *f*(*x*, *y*) is the pixel value of the desired point *P*.

3.2 Attention mechanism

For complex nighttime backgrounds, we must exclude factors such as coastal lights and seawater reflection lights, which could affect the accuracy of detection. The essence of the AM module is to locate the information of interest and suppress redundant information. The results are usually presented in the form of a probability graph or probability feature vector. In principle, attention modules can be divided into three types: a spatial attention module, a channel attention module, and a spatial and channel combined attention module.

The convolutional block attention module (CBAM)(27) is an attention module for CNNs that combines both channel and spatial modules. Its structure is shown in Fig. 7.

The channel attention module can filter the features suitable for object discrimination in the channel dimension. After feature extraction of the backbone network, different channels of the feature map have multidimensional feature information of the object. Thus, enhancing the important channels of ship features can improve the identification of ships. As shown in Fig. 8, the feature map F passes through a max-pooling layer f_m and an average-pooling layer f_a in the spatial dimension, which denote max-pooled features and average-pooled features, respectively. Then, after feature transformation by the multilayer perceptron, the channel weight vector of feature F is obtained by σ (the sigmoid activation function). The channel weight vector is multiplied by the feature map F along the channel dimension to complete the feature map F weighted by the channel, as expressed by

$$F_c = \sigma\left(W_1\left(W_0\left(f_a\left(F\right)\right)\right) + W_1\left(W_0\left(f_m\left(F\right)\right)\right)\right), \tag{8}$$

where $F \in \mathbb{R}^{C\times H\times W}$, $W_0 \in \mathbb{R}^{C/(r\times C)}$, and $W_1 \in \mathbb{R}^{C\times\frac{C}{r}}$. $\mathbb{R}^z$ represents a real number field and z

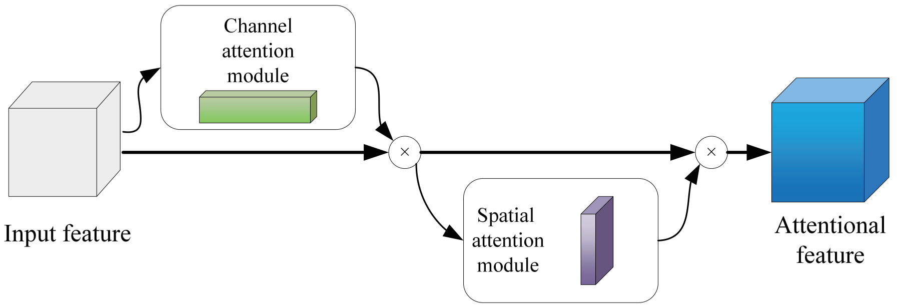

Fig. 7. (Color online) Structure of CBAM.

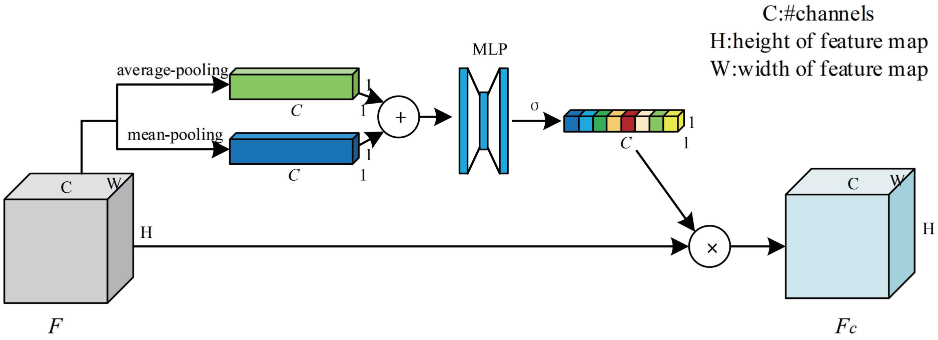

Fig. 8. (Color online) Channel AM module.

denotes the range of $\mathbb{R}$. W_0, W_1 are the weight vectors of the multilayer perceptron, C is the feature channel dimension, H and W represent the height and width of the feature, respectively, and r represents the decay rate.

The spatial AM utilizes spatial relations to generate a spatial attention feature map, which enhances the key features of objects in the spatial dimension and improves the semantic information of nighttime ship object mining. As shown in Fig. 9, the feature map F passes through the max-pooling layer f_m and average-pooling layer f_a in the channel dimension. By splicing the average-pooled features and max-pooled features, the pooling feature map is obtained, which represents the response degree of object features in different regions in the spatial dimension. Then, by using a 7 × 7-dimensional convolutional kernel, a convolution feature map is obtained. The spatial weight vector of feature F is obtained by σ. Finally, the weighted spatial feature map is obtained by weighting the spatial weight vector of feature F with F_c along the spatial dimension using

$$F_s = \sigma\left(f^{7\times7}\left(\left[f_a\left(F_c\right); f_m\left(F_c\right)\right]\right)\right), \tag{9}$$

where $F_s \in \mathbb{R}^{1\times H\times W}$ and $f^{7\times7}$ is the convolutional operation with a 7 × 7-dimensional convolutional kernel.

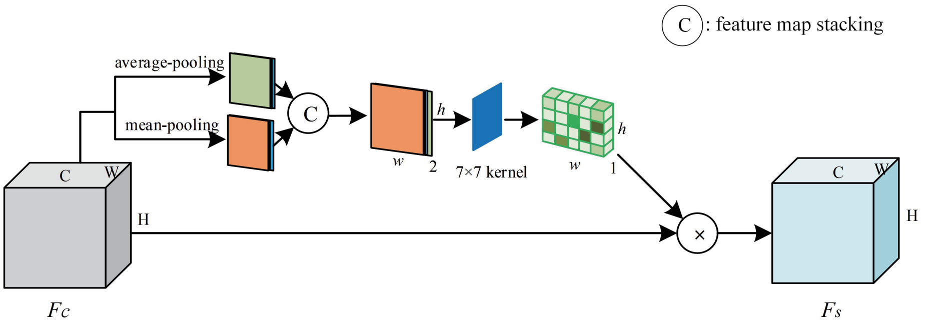

Fig. 9. (Color online) Spatial AM module.

3.3 Feature pyramid network

An FPN is a feature extractor that takes a single-scale image as the input and outputs proportionally sized feature maps at multiple levels in a fully convolutional fashion. This process is independent of the backbone convolutional architecture. It therefore acts as a generic solution for building feature pyramids inside deep convolutional networks to be used in tasks such as object detection. Here, we adopt an FPN to improve the utilization of features of nighttime ship targets by the neural network.

As shown in Fig. 10, since ship targets are small and Faster R-CNN significantly reduces or even omits information of targets in the compression of feature information at the spatial scale, the FPN is introduced to deepen the semantic feature association of different layers and retain the features of small targets.

3.4 SEAFF

As described above, the SEAFF network mainly consists of three parts. (1) An SE is used to increase the feature size of ship targets. (2) An AM module, which enhances the features of nighttime ship targets, is introduced to extract feature maps from different convolution blocks. (3) An FPN is adopted to fuse the extracted semantic feature maps at different levels to improve the utilization of nighttime ship features. The overall structure of the SEAFF network is shown in Fig. 11. The input image is expanded to 1.5 times its original image by the SE operation. Feature maps are extracted from convolution blocks Conv1–5 in the trunk network, and each convolution block is integrated with the AM module to enhance ship features. Each integrated convolution block is used to obtain the feature maps, then feature maps M4, M3, and M2 are fused by high-semantic feature maps and the adjacent low-semantic feature maps through the 1 × 1 convolutional adjustment channel of feature maps. Then, the fused feature maps M4, M3, and M2 are extracted by the 3 × 3 convolutional kernel to obtain multiscale feature maps P2, P3, P4, and P5. An RPN uses feature map M5 to mine positive and negative candidate anchor boxes and calculate anchor boxes corresponding to P2, P3, P4, and P5. ROI Align is used for feature sampling on feature maps M2, M3, M4, and M5 to obtain feature vectors of ship targets

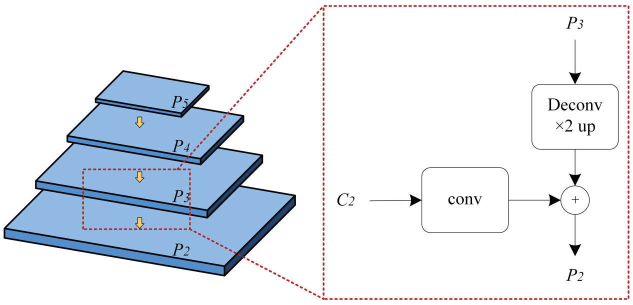

Fig. 10. (Color online) Structure of FPN.

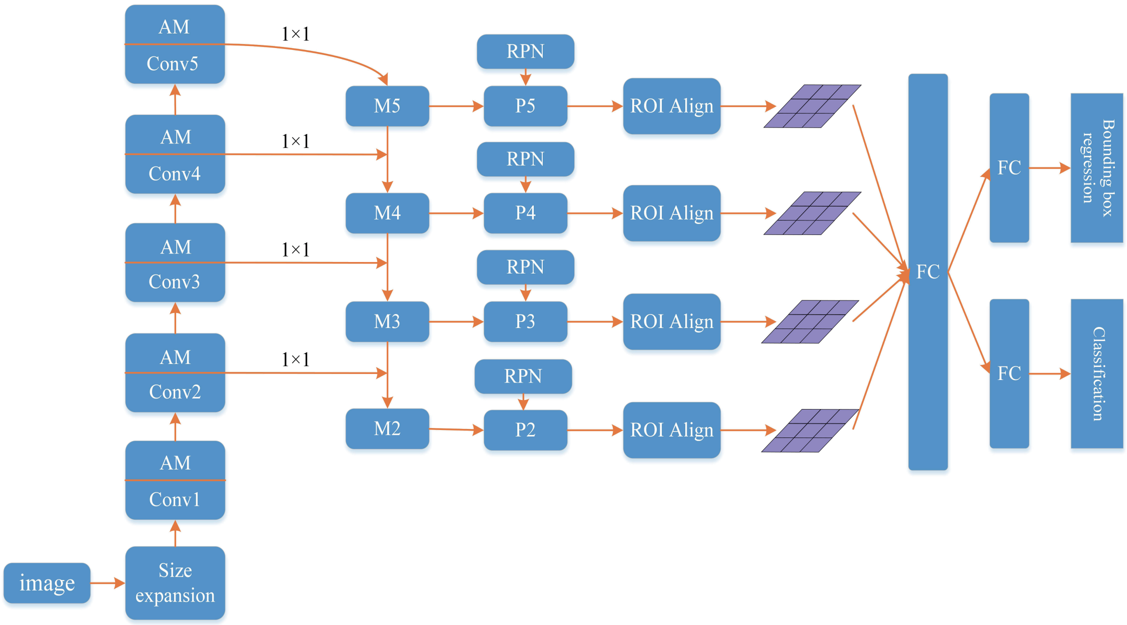

Fig. 11. (Color online) Structure of SEAFF network.

of different scales. Through the fully connected (FC) neural network, the object features of different scales are supervised for object classification and bounding box regression. The candidate boxes with high confidence values are selected and the object detection anchor boxes are obtained by non-maximum suppression (NMS).

4. Experiments and Analysis

4.1 Dataset construction

Figure 12 shows the flow of the dataset construction. We first downloaded 25 raw nighttime remote sensing images generated by Luojia 1-01 (http://59.175.109.173:8888/), then stretched them reversely to their original radiance as discussed earlier. The average scale of these images was about 2900 × 2300, making them unsuitable for training; thus, we clipped them to a size of 416 × 416. After eliminating small images that only contained land or had no visible ship targets,

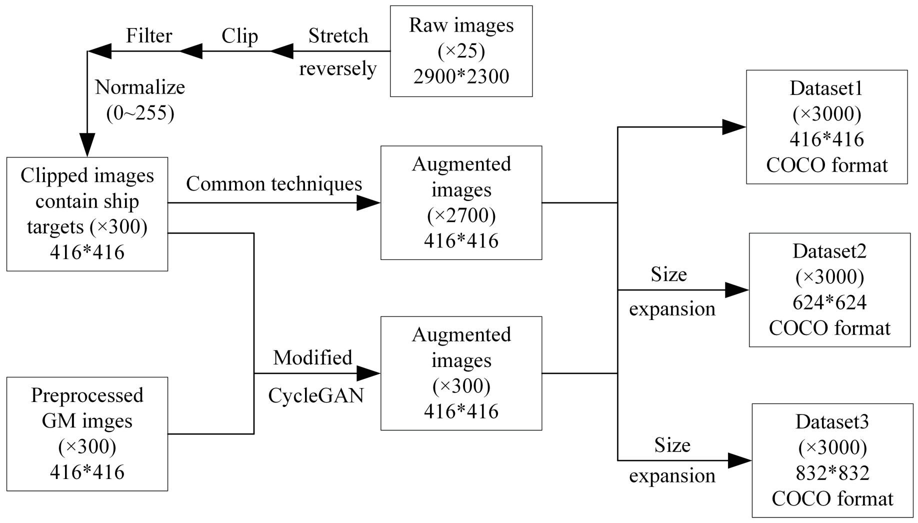

Fig. 12. Flow of dataset construction.

we obtained 300 clipped images. Since the pixel values of these clipped images can reach 10^4, we also normalized them to make them suitable for our experiments. Then, we adopted the common techniques introduced above for data augmentation and obtained 2700 images by combining the clipped images with images generated by CycleGAN to give a total of 3000 images. To study the influence of different image sizes on nighttime ship target detection, we expanded the 416 × 416 images to 624 × 624 and 832 × 832 and divided every dataset into training, test, and validation sets with the ratio of 8:1:1. Finally, we constructed the three datasets in COCO format.

4.2 Setup and training details

In the experiments, we used an Intel Corei9-9900k@3.60 GHz CPU with an NVIDIA GeForce RTX 2080Ti 11 GB CPU and 64 GB RAM. We used a PyTorch deep learning framework to carry out the experiments. Table 4 shows the details of the training.

4.3 Assessment metrics

Commonly used metrics for object detection are the precision–recall (PR) curve and average precision (AP). A drawback of the *P*–*R* curve is that it can be affected by the proportion of positive and negative samples; thus, we adopted AP to assess experimental results. By plotting the recall rate as the abscissa and precision rate as the ordinate, AP is calculated as the area under the curve. Precision *P* and recall *R* are calculated as

$$P = \frac{TP}{TP + FP} \times 100\%, \tag{10}$$

$$R = \frac{TP}{TP + FN} \times 100\%, \tag{11}$$

Table 4
Details of training in experiments.

Experiment		Learning rate	Momentum	Weight decay	Batch size
SE experiment		0.0025	0.9	0.0001	16
Ablation experiment		0.0025		0.0001	16
Network contrast experiment	SSD	0.00002		0.0005	64
	YOLOv3	0.0001		0.0005	64
	Faster R-CNN	0.0025		0.0001	16
	SEAFF	0.0025		0.0001	16

where *TP* is the number of positive samples correctly classified, *FP* is the number of samples incorrectly classified as positive, and *FN* is the number of samples incorrectly classified as negative.

In our experiments, we adopted Intersection over Union (IOU) as the threshold for defining positive and negative samples of nighttime ship targets. IOU is defined as

$$IOU = \frac{A_{pred} \cap A_{gt}}{A_{pred} \cup A_{gt}}, \tag{12}$$

where A_{pred} and A_{gt} are the areas of the predicted bounding box and ground-truth bounding box, respectively.

AP@0.50 and AP@0.75 denote the AP values when IOU is set to 0.5 and 0.75, respectively, and AP@0.50:0.95 indicates the average value of 10 AP values obtained with IOU increased from 0.5 to 0.95 in steps of 0.05.

4.4 SE experiment

To study the effect of expanding the image on nighttime ship target detection, we expanded the original images from 416 × 416 to 624 × 624 and 832 × 832, then used the SEAFF network to train the three datasets. Table 5 shows the experimental results for the three indicators for different image sizes. The best results were obtained when the images were expanded to 624 × 624. We believe that excessive image expansion is not conductive to separating targets from the background and causes the original features of ship targets to be lost, resulting in reduced detection performance.

4.5 Ablation experiment

To observe how each module affects the experimental results, we also conducted ablation experiments. Table 6 shows the values of the three indicators obtained in the ablation experiments. Table 6 shows that when the network lacks the SE module, AP@0.50, AP@0.75, and AP@0.50:0.95 decrease by 0.026, 0.024, and 0.034, respectively, and the accuracy of the network decreases significantly. When the network lacks the FPN module, the three indicators

Table 5
Values of three indicators in SE experiments.

Image size	AP@0.50	AP@0.75	AP@0.50:0.95
416 × 416	0.913	0.330	0.432
624 × 624	0.939	0.354	0.466
832 × 832	0.901	0.327	0.425

Table 6
Values of three indicators in ablation experiments.

	SE	FPN	AM	AP@0.50	AP@0.75	AP@0.50:0.95
A	×	√	√	0.913	0.330	0.432
B	√	×	√	0.928	0.341	0.447
C	√	√	×	0.911	0.325	0.420
SEAFF	√	√	√	0.939	0.354	0.466

decrease by 0.011, 0.013, and 0.019. Although the average decrease of the three indicators is about 0.014, the FPN module greatly improves the detection performance of nighttime ship targets at multiple scales. When the network lacks the AM module, the three indicators decrease by 0.028, 0.029, and 0.046. In complex nighttime backgrounds, the AM module helps filter out complex light sources such as coastal light and the light reflected by seawater, thus greatly increasing the accuracy.

We selected images that were affected by light reflected from seawater to study the performance of each component as shown in Fig. 13. Weak and small nighttime ship targets were often missed without the SE module. The FPN module improved the accuracy of our network at multiple scales. Meanwhile, the AM module made our network less susceptible to complex backgrounds. With the AM module, the nighttime ship targets surrounded by light reflected from the seawater could also be detected. Nevertheless, some targets were missed by our network, as can be seen by comparison with the ground truth of the result of the SEAFF experiment.

4.6 Comparison experiment

To further verify the performance of the SEAFF network, we conducted experiments with the mainstream SSD, YOLOv3, and Faster R-CNN networks. It can be seen from Table 7 that SEAFF had the highest values of AP@0.50, AP@0.75, and AP@0.50:0.95, which were 0.032, 0.048, and 0.029 higher than those for Faster R-CNN, respectively. Figure 14 shows results of 150 iterations for the different networks. AP@0.50 stabilized after 15 rounds of iteration for all networks. Both AP@0.75 and AP@0.50:0.95 converged most slowly for SEAFF. This is considered to be because of the characteristic of the SEAFF network: it first performs an approximate detection of ship targets, then focuses on fine detection with the aim of improving the accuracy. Overall, the SEAFF network strengthens the structure of the network to enhance the learning of ship features, improves the detection performance of nighttime ship targets, and improves the detection accuracy.

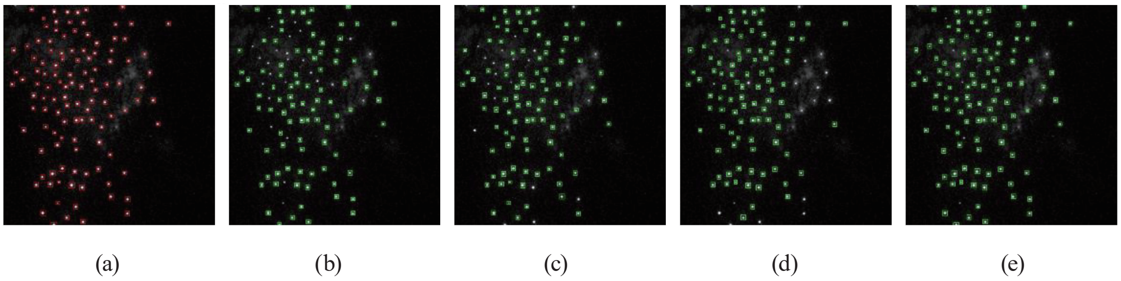

Fig. 13. (Color online) Comparison of ablation experiments. (a) Ground truth, (b) result of experiment A, (c) result of experiment B, (d) result of experiment C, and (e) result of SEAFF experiment.

Table 7
Values of three indicators in comparative experiments.

Network	AP@0.50	AP@0.75	AP@0.50:0.95
SSD	0.832	0.175	0.343
YOLOv3	0.904	0.272	0.436
Faster R-CNN	0.907	0.306	0.437
SEAFF	0.939	0.354	0.466

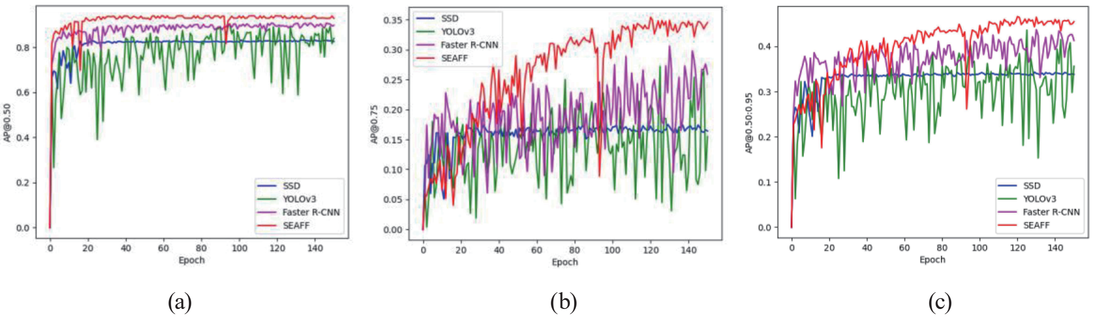

Fig. 14. (Color online) Results of 150. (a) Results for AP@0.5, (b) Results for AP@0.75, and (c) Results for AP@0.5:0.95.

Figure 15 shows a comparison of the results for different networks for nighttime ship targets. By comparison with the ground truth, it was found that all four networks can detect nighttime ship targets. Among them, SSD has many false detections, with most light sources identified as ship targets. YOLOv5, Faster R-CNN, and SEAFF have much fewer false detections but still fail to detect ship targets, especially when the density of nighttime ship targets is high. Overall, the SEAFF network strengthens the learning of features of nighttime ship targets and improves the detection accuracy of nighttime ship targets.

4.7 Discussion

Through the results of SE, ablation, and comparison experiments, we concluded the following: (1) When there are many weak and small nighttime ship targets, the SE module enlarges the pixel size of ship targets to make them easier to detect. (2) The FPN module

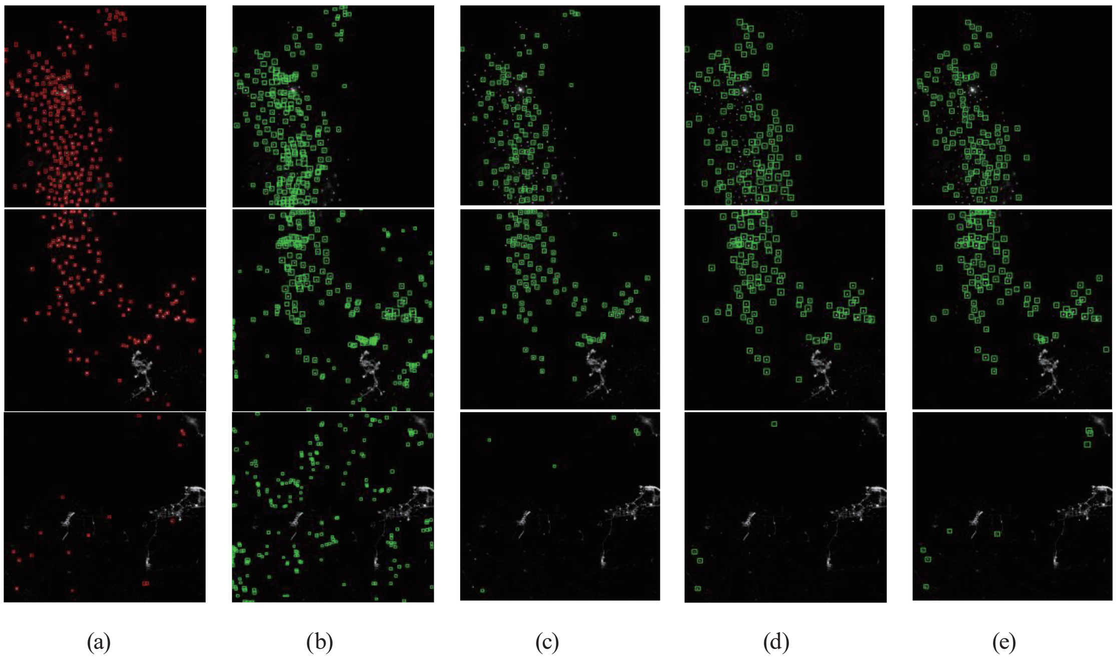

Fig. 15. (Color online) Comparison of results for different detection networks. (a) Ground truth, (b) SSD, (c) YOLOv3, (d) Faster R-CNN, and (e) SEAFF.

markedly improves the detection performance of nighttime ship targets at multiple scales. (3) The AM is beneficial for eliminating the influence of a complex background. The SEAFF network integrating the SE, FPN, and AM modules had the highest performance in comparative experiments. However, in the case of dense nighttime ship targets, some ship targets remain undetected. Moreover, for weak and small ship targets in complex nighttime scenes, the SEAFF network is still insufficient.

5. Conclusions

Owing to the high cost of object detection and its difficulty in a large region, we applied deep learning methods to verify the feasibility of detecting nighttime ship targets. Starting with a dataset of original nighttime images generated by the Luojia 1-01 remote sensing satellite, we augmented the dataset by implementing common data augmentation techniques and modified CycleGAN. Then, we conducted experiments on several mainstream networks. The experimental results verified the feasibility of deep learning methods for detecting nighttime ship targets. Then, we constructed the SEAFF network, with SE, AM, and FPN modules integrated into Faster R-CNN, and conducted experiments. The experimental results showed that SEAFF has superior detection performance to mainstream networks such as SSD, YOLOv3, and Faster R-CNN. In the future, we plan to combine multiple remote sensing images for joint detection to improve the overall discrimination accuracy of nighttime ship targets.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 41101426), the Natural Science Foundation of Jiangxi Province (No. 20202BABL202040), and the Graduate Innovation Fund of Nanchang Hangkong University (No. YC2022-052). The authors would like to thank the handling editor and anonymous reviewers for their reading and remarks.

References

1 D. J. Agnew, J. Pearce, G. Pramod, T. Peatman, R. Watson, J. R. Beddington, and T. J. Pitcher: PLoS One **4** (2009) e4570. https://doi.org/10.1371/journal.pone.0004570
2 P. Qin, Y. Cai, J. Liu, P. Fan, and M Sun: IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **14** (2921) 11058. https://doi.org/10.1109/jstars.2021.3123080
3 C. Lim, G. Cheng, J. Luo, S. Li, and J. Ye: Int. J. Remote Sens. **38** (2017) 6007. https://doi.org/10.1080/01431161.2017.1312034
4 M. Kang, K. Ji, X. Leng, and Z. Lin: Remote Sens. **9** (2017) 860. https://doi.org/10.3390/rs9080860
5 J. Wu, S. Mao, X. Wang, and T. Zhang: Opt. Eng. **50** (2011) 057207. https://doi.org/10.1117/1.3578402
6 Y. Tian, J. Liu, S. Zhu, F. Xu, and G. Bai: Remote Sens. **14** (2022) 3347. https://doi.org/10.3390/rs14143347
7 Y. Chang, A. Anagaw, L. Chang, Y. Wang, C.Hsiao, and W. Lee: Remote Sens. **11** (2019) 786. https://doi.org/10.3390/rs11070786
8 B.Jiang, X. Ma, Y. Lu, L. Feng, and Z. Shi: IR Phys. Technol. **97** (2019) 229. https://doi.org/10.1016/j.IR.2018.12.040
9 J, Ruiz, I, Caballero, and G. Navarro: Remote Sens. **12** (2020) 32. https://doi.org/10.3390/rs12010032
10 X. Li, Y, Xiao, F. Su, W. Wu, and L. Zhou: ISPRS Int. J. Geo-Inf. **10** (2021) 277. https://doi.org/10.3390/ijgi10050277
11 R. Girshick, J. Donahue, T. Darrell, and J. Malik: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2014) 580–587.
12 K. He, X. Zhang, S. Ren, and J. Sun: IEEE Trans. Pattern Anal. Mach. Intell. **37** (2015) 1904. https://doi.org/10.1109/TPAMI.2015.2389824
13 S. Ren, K. He, R. Girshick, and J. Sun: Proc. Advances in Neural Information Processing Systems (NIPS, 2015).
14 J. Redmon, S. Divvala, R. Girshick, and A. Farhadi: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2016) 779–788
15 W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg: ECCV (2016) 21. https://doi.org/10.1007/978-3-319-46448-0_2
16 Q. Zhang, J. Yao, K. Zhang, C. Feng, and J. Zhang: ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. **41** (2016) 423. https://doi.org/10.5194/isprsarchives-XLI-B7-423-2016
17 Y. Liu, M. Zhang, P. Xu, and Z. Guo: Int. Workshop RSIP (2017) 1. https://doi.org/10.1109/RSIP.2017.7958806
18 X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo: Remote Sens. **10** (2018) 132. https://doi.org/10.3390/rs10010132
19 L. Liu, G. Liu, X. Chu, Z. Jiang, M. Zhang, and J. Ye: J. Phys. Conf. Ser. **1187** (2019) 042074. https://doi.org/10.1088/1742-6596/1187/4/042074
20 X. Li, X. Li, D.Li, X. He, and M. Jendryke: Remote Sens. Lett. **10** (2019) 526. https://doi.org/10.1080/2150704X.2019.1577573
21 D. Li, G. Zhang, X. Shen, X. Zhong, Y. Jiang, T. Wang, J. Tu, and Y. Li: J. Remote Sens. **23** (2019) 1011. http://doi.org/10.11834/jrs.20199327
22 I. J. Goodfellow, J. Pouget-Adadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio: Commun. ACM **63** (2020) 139. https://doi.org/10.1145/3422622
23 A. Radford, L. Metz, and S. Chintala: arXiv **1511** (2015) 06434. https://doi.org/10.48550/arXiv.1511.06434
24 H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas: Proc. IEEE Int. Conf. Computer Vision (ICCV) (2017) 5907–5915.
25 J. Zhu, T. Park, P. Isola, and A. A. Efros: Proc. IEEE Int. Conf. Computer Vision (ICCV) (2017) 2223–2232.
26 O. Augustus, D. Vincent, and O. Chris: Distill (2016). http://doi.org/10.23915/distill.00003
27 S. Woo, J. Park, J. Lee, and I. Kweon: Proc. European Conf. Computer Vision (ECCV) (2018) 3–19.