

Received 31 January 2024, accepted 8 February 2024, date of publication 13 February 2024, date of current version 20 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3365777

APPLIED RESEARCH

MSFA-YOLO: A Multi-Scale SAR Ship Detection Algorithm Based on Fused Attention

ZHAO LIANGJUN¹, NING FENG², XI YUBIN¹, LIANG GANG¹,
HE ZHONGLIANG¹, AND ZHANG YUANYANG¹

¹School of Computer Science and Engineering, Sichuan University of Science and Engineering, Yibin 644000, China

²School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, China

Corresponding author: Zhao Liangjun (zhaoliangjun@suse.edu.cn)

This work was supported by the Science and Technology Plan Project of Sichuan Province, China, under Grant 2023YFS0371.

ABSTRACT Leveraging the excellent feature representation capabilities of neural networks, deep learning methods have been widely adopted for object detection in synthetic aperture radar (SAR) images. However, persistent challenges are encountered in SAR ship detection due to factors such as small ship sizes, high noise levels, multiple targets, and scale variations. To address these complexities, in this paper, the MSFA-YOLO algorithm, a novel multiscale SAR ship detection approach empowered by a fused attention mechanism, is presented. The proposed algorithm incorporates several key enhancements. The fused attention c2fSE module is integrated into the YOLOv8n baseline network to optimize feature extraction for SAR ships. In addition, the DenseASPP module is incorporated to enhance the model's adaptability to ships of varying scales, improving its capability to accommodate larger ships within lower model scales. Furthermore, the Wise-IoU loss function is adopted, and a dynamic non-monotonic focusing mechanism is employed for bounding box loss, significantly enhancing the model's ability to handle low-quality images. Extensive experiments conducted on benchmark datasets, namely SAR-Ship-Dataset, SSDD, and HRSID, validate the robustness and reliability of the proposed model. Experimental results demonstrate significant performance improvements over YOLOv8n: a 3.1% enhancement in mAP75 and a 2.1% boost in mAP50–95 on the SAR-Ship-Dataset, a 0.7% increase in mAP75 and a 0.5% increase in mAP50–95 on the SSDD dataset, and a 1.8% increase in mAP75 and a 0.7% increase in mAP50–95 on the HRSID dataset. Exhibiting superior performance to existing SAR ship detection models in terms of accuracy, the MSFA-YOLO algorithm represents a significant advancement, establishing itself as the current state-of-the-art algorithm in SAR ship detection.


INDEX TERMS Ship detection, SAR image, YOLO.

I. INTRODUCTION

With recent developments in synthetic aperture radar (SAR) technology, SAR images have become an important component in the field of ship detection. Although SAR has been employed in ocean research for over 50 years, efforts are still underway to comprehend and apply SAR in maritime radar scenes [1]. Several challenges persist in interpreting SAR images; however, its unique technical characteristics endow

SAR with an indispensable role in the field of maritime monitoring and security [2], [3], [4].

First, SAR images possess all-weather observation capability. Unlike optical remote sensing, SAR is not affected by weather conditions such as clouds, rain, or fog; thus, enabling the acquisition of high-quality image data under adverse weather conditions. This capability allows SAR to perform continuous maritime ship monitoring regardless of day or night, sunny or rainy conditions. Second, SAR images offer superior spatial and height resolutions, enabling clear capture of detailed information about a ship's shape, size, and structure; thus, aiding in more accurate identification

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Di Martino .

and classification of different types of vessels. This is of practical significance in maritime traffic management, border patrol, and maritime security [5]. Moreover, SAR technology enables multimode observation, offering different resolutions and coverage ranges through various operational modes, thereby making it suitable for different application scenarios. This multimode characteristic makes SAR technology more flexible and efficient for ship monitoring across large maritime areas [6].

However, compared to optical images, SAR images obtained from satellite and airborne platforms have lower resolutions and are more susceptible to background clutter and noise interference [7]. In addition, vessels of different sizes appear as objects represented by different pixels in SAR images, making accurate detection of vessels by using multiscale features a significant challenge [8], [9]. As a result, existing SAR image target detection methods rely on deep learning (DL) approaches. For example, Zheng et al. [10] proposed a hybrid representation learning enhanced SAR target detection algorithm HRLE-SARDet based on the unique features of SAR images to better extract the scattering information of small targets in SAR images and improve the detection accuracy.

DL methods provide advantages such as self-learning, self-improvement, and weight sharing. First-stage algorithms such as SSD [11], YOLO series [12], [13], [14], [15], and DL object detection models such as Faster R-CNN [16], and Cascade R-CNN [17] have been widely employed in object detection due to their high precision, efficiency, and robustness. These advantages offered by DL methods have positioned them as a new and preferred approach for the SAR vessel detection problem. For instance, to address the problem of the same loss value obtained for different predicted box sizes by using traditional loss functions based on centroid distance and aspect ratio, Zhou et al. [18] proposed a dual Euclidean distance loss function based on the corner coordinates between the predicted and real boxes, thereby enhancing the precision of ship detection. Miao et al. [19] proposed a deep hierarchical network architecture termed Contextual Region-based Convolutional Neural Network (CRCNN), which consists of a high-resolution Region Proposal Network (RPN) and a target detection network incorporating contextual features. Diverging from conventional RPN methodologies, this approach employs an intermediate layer in conjunction with downsampled shallow layers and upsampled deep layers for the generation of region proposals. Within the target detection network, proposed regions are projected onto multiple hierarchical levels through Region of Interest (ROI) pooling, facilitating the extraction of corresponding ROI features and surrounding contextual features.

In the traditional object detection process, the network's depth can lead to a decrease in training set accuracy due to information loss during convolutional and fully connected operations. This can lead to gradient explosion or vanishing,

resulting in decreased convergence speed or failure to converge. To address these issues, Zhu et al. [20] proposed the transformer prediction head detection head to improve the model's ability to detect objects of different scales. Shi et al. [21] proposed using the geometric transformation module and global context feature fusion module to improve recognition and positioning accuracy, thereby improving the feature extraction capability of the model.

Dealing with objects of different scales is a major challenge in object detection using DL models. The model's feature extraction tends to converge toward a specific scale due to the significant disparity in data volume across different target scales. Noh et al. [22] proposed the use of super-resolution techniques at feature hierarchies and employed generative adversarial learning to enhance the features of smaller objects, thus avoiding the problem of generating erroneous super-resolution features due to mismatched receptive fields. The Mosaic data augmentation method proposed by Alexey Bochkovskiy et al. [12] enhances data by using methods such as Mixup, Cutout, and CutMix, and is widely employed in mainstream object detection models. The above problems also exist in the field of ship detection. There are a few large ship models in the dataset, which leads to underfitting of the model to large ship targets. Therefore, the model is more sensitive to small targets and poorly detects large ships, and the model is not accurate enough for multi-scale targets. To address this problem, Li et al. [23] established the attention-guided balanced feature pyramid network, which better utilizes semantic and multilevel complementary features; thus, strengthening ship detection capabilities at different scales. Despite its merits, this approach exhibits diminished performance in detecting small targets and encounters challenges in achieving scale balance. In addition, Shao et al. [24] proposed the rotationally balanced feature alignment network to enhance multiscale ship detection capabilities. Nevertheless, it still performs poorly in coping with the problem of too few large ships in the dataset. In a related vein, Guo et al. [25] developed the deep self-adaptive spatial feature fusion neck module and SCYLLA-IoU (SIoU) loss function to improve the detection accuracy of multiscale targets. However, it is noteworthy that the aforementioned design primarily focuses on addressing the multi-scale predicament in the context of rotating target detection, exhibiting limited generalization for horizontal anchor frame detection tasks.

Furthermore, SAR images are highly susceptible to influences from the ocean, islands, coastal ports, and lighting, resulting in lower image quality in the dataset. To address this issue, Tang et al. [26] proposed a noise level classifier and STPAE module to extract complete regions of potential targets, categorizing high and low noise data to improve the accuracy of SAR ship image detection. However, this method still fails to address the model's ability to extract ship features in high-noise scenarios. Guo et al. [27] developed a single-stage detector called CenterNet++ to

balance foreground and background effectively. Nevertheless, a notable limitation is the absence of specific measures or enhancements to address the potential challenges posed by high-noise environments or scenarios. Sun et al. [28] improved the detection efficiency by performing pixel-by-pixel prediction on SAR images and avoided dense anchor points to enhance the model's adaptability in complex scenes. However, it is crucial to recognize that, despite the significant enhancement brought about by the pixel-by-pixel module, its efficacy in scenarios marked by elevated noise levels remains unexplored.

Lastly, the dataset contains a large number of small vessels, resulting in blurred ship targets due to low resolution. This makes it difficult to extract effective features, thereby lowering model accuracy. To address this issue, Zhang et al. [29] proposed the Quad-FPN module, which comprises four FPN structures, and conducted extensive experiments to enhance the model's capability in acquiring features from ship SAR images. Nevertheless, complete mitigation of the impact of low-resolution targets on feature extraction remains an ongoing challenge.

To overcome the aforementioned limitations, we proposed the MSFA-YOLO model in this paper, which is an efficient and fast network architecture. The key contributions of this paper are as follows:

- 1) In order to solve the problem of having a large number of small ships with fuzzy ship targets due to low resolution reasons, which cannot acquire effective features, we proposed the C2fSE module, utilizing the attention mechanism to replace the C2f module of the backbone network. The C2fSE module can acquire more image information without increasing the number of parameters and model size. And enhances the model's ability to acquire ship features.
- 2) To mitigate model bias toward small ship targets and improve the detection accuracy for large ship targets, we introduced the DenseASPP module, which improves the model's ability to adapt to multiscale features.
- 3) To address the problem that SAR images are affected by the ocean, islands, and coastal ports, resulting in low image quality, we introduced the Wise-IoUv3 module, which employs a dynamic non-monotonic focusing mechanism and utilizes the "outlier degree" to characterize the quality of the anchor frames. Consequently, Wise-IoU can target detection results more accurately, avoiding the problem of bias encountered in traditional IoU.
- 4) Extensive experiments were conducted on SSDD, SAR-Ship-Dataset, and HRSID open-source datasets, demonstrating excellent results and thus proving the effectiveness of the proposed model.

The rest of the paper is organized as follows. In Section II, the proposed methodology is described. In Section III, experimental results and analysis are presented. Finally, Section IV concludes the paper.

II. METHODS

A. NETWORK ARCHITECTURE

The overall design architecture of the proposed MSFA-YOLO model is illustrated in Figure 1. MSFA-YOLO comprises three main parts: the SAR-SHIP-NET backbone network, the neck network, and the detection part.

- 1) The SAR-SHIP-NET backbone network is based on YOLOv8 and yields enhanced feature extraction and multiscale semantic sensing capabilities. The backbone part includes the Conv module, C2fSE module, and DenseASPP module. Conv is the convolution module and encapsulates three functions: convolution (Conv2d), BN layer, and SiLU activation function. The C2fSE module combines the advantages of the C2f and SE attention mechanisms, and The DenseASPP module has a larger receptive field to enhance feature extraction for large vessels, thereby improving the model's semantic sensing ability and enabling MSFA-YOLO to obtain richer gradient stream information while ensuring a lightweight design.
- 2) In the neck part, the MSFA-YOLO model uses the same model structure as YOLOv8n. YOLOv8 utilizes the PAN-FPN structure and replaces the C3 and RepBlock modules in the YOLOv5 with the C2f module, which further improves the feature extraction capability of the model.
- 3) The detection part utilizes anchorless frames to separate the classification header from the detection header and determines positive and negative samples based on the weighted scores of classification and regression which effectively improves the model performance.

The structure of the MSFA-YOLO model is shown in Figure 1.

B. C2FSE MODULE

The C2f module in YOLOv8 yields richer information about the gradient flow while keeping the model lightweight. It uses 1×1 convolution and bottleneck blocks with residual concatenation to efficiently capture and process features, improving the feature extraction capability of the YOLOv8 architecture. The C2f module is illustrated in Figure 2.

However, in the traditional C2f module, the limited exchange of information between feature maps leads to poor target detection accuracy. Therefore, in this paper, we proposed the C2fSE module, which incorporates an attention mechanism to enhance the feature extraction capability of the model.

The attention mechanism allows the model to prioritize important features; thus, improving the target detection performance. We introduced the SE attention mechanism to improve the performance of the model by compressing and motivating input features [30]. The SE attention mechanism is depicted in Figure 3.

The SE attention mechanism consists of two steps: squeeze and excitation. In the squeeze step, the input feature map

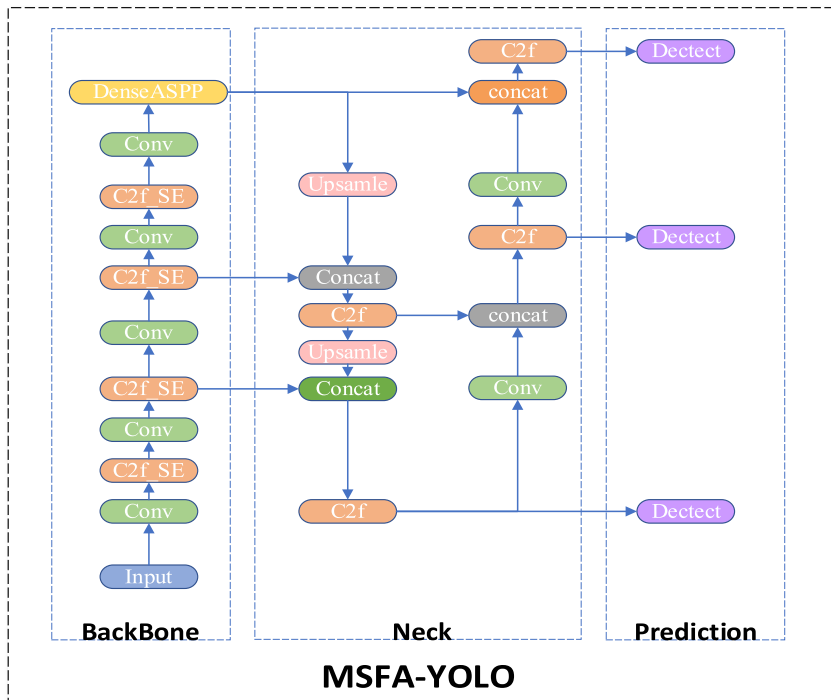


FIGURE 1. Structure of the MSFA-YOLO model.

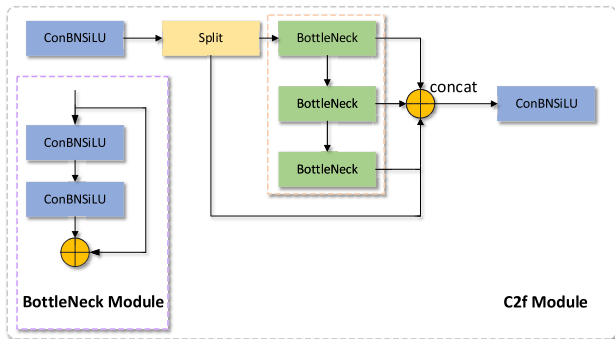


FIGURE 2. Structure of the C2f module.

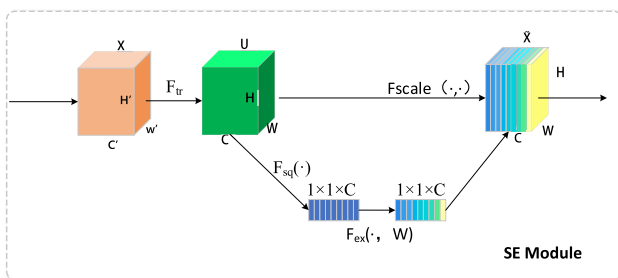


FIGURE 3. Structure of the SE module.

is compressed into a vector through global average pooling and then mapped to a smaller vector by a fully connected layer. In the excitation step, each element in the vector is compressed to a value between 0 and 1 by using a sigmoid

function. This value is then multiplied by the original input feature map to generate a weighted feature map. Through the SE attention mechanism, the model can adaptively learn the importance of each channel; thus, improving the model's performance. The derivation process of the SE module is as follows:

- 1) Given an input feature map X , let it undergo the F_{tr} operation to generate a feature map U .
- 2) The feature map undergoes global average pooling, resulting in a $1 \times 1 \times C$ vector such that each channel is represented by a single value. This global low-dimensional embedding implemented for U serves as a sensory field for each channel.

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (1)$$

- 3) Through two fully connected layers, the weight information is generated using the weights W , where W is obtained through learning and is used to model the feature correlation we need for the display.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \quad (2)$$

The vector z obtained in step 2 is processed through two fully-connected layers W_1 and W_2 to obtain the channel weight value s . After passing s through two fully-connected layers, different values in s denote the weight information of different channels.

- 4) The weight vector s generated in the third step is used to assign weights to the feature map U to obtain the

feature map X , whose size is exactly the same as the feature map as the SE module does not change the size of the feature map.

$$X_c = F_{scale}(u_c, s_c) = s_c u_c \quad (3)$$

To multiply the generated feature vector $s(1 \times 1 \times C)$ with the feature map $U(H \times W \times C)$, the corresponding channel, the $H \times W$ values of each channel in the feature map U are multiplied by the weights of the corresponding channel in s .

Combining the squeeze and excitation (SE) attention mechanism with YOLOv8's C2f module yields substantial advantages for target detection. This integrated module enhances model performance by dynamically learning feature channel weights, improving focus on critical information essential for target detection. This adaptability in feature modeling enhances the model's capability to discern targets of varying scales, shapes, and backgrounds. The C2fSE module is illustrated in Figure 4.

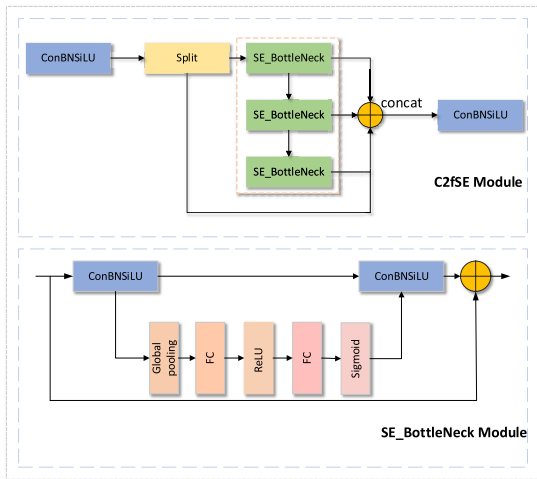


FIGURE 4. Structure of the C2fSE module.

This fusion not only enhances feature modeling capability but also reduces redundant information, thereby enabling the network to prioritize crucial data pertinent to target detection. Consequently, this helps mitigate the risk of false detections and improves the overall model accuracy. In addition, during training, the SE attention mechanism expedites model convergence, thereby reducing training time. The effectiveness of introducing the SE attention mechanism into the feature extraction stage for ship detection in SAR images can be attributed to several aspects:

Feature enhancement and emphasis on importance: SAR images exhibit characteristics such as noise interference and complex scattering. By focusing on and reinforcing the most representative ship features, the SE attention mechanism reduces interference and highlights crucial ship structural characteristics, thereby enhancing detection performance.

Adaptive feature modulation: The SE mechanism demonstrates adaptability by learning the interrelations between

feature channels and adjusting the weights of channels, thereby enabling the network to prioritize crucial features necessary for ship detection, suppressing irrelevant features and noise.

Enhancing model learning capabilities: By learning the inter-channel correlations among features, the SE attention mechanism enables the network to better adapt to various lighting conditions and changes in ship sizes and orientations, thereby enhancing the robustness and adaptability of the algorithm.

Improved generalization ability: The SE mechanism aids in better generalizing new, unseen data samples, thereby reducing the risk of overfitting and enhancing the practicality and reliability of the ship detection algorithm.

Therefore, integrating the SE attention mechanism into ship detection optimizes feature representation, emphasizes critical information, and enhances the model's robustness and generalization capabilities. This, in turn, improves the accuracy and reliability of ship detection in SAR images; thus, providing an effective approach to enhance the performance and generalization ability of object detection models with attention mechanisms. In practical applications, this translates into more accurate, robust, and efficient object detection systems that can be adopted for various scenarios and datasets.

C. DENSEASPP MODULE

Ship detection in SAR images is challenging due to the distinctive characteristics of SAR images. In SAR imagery, target vessels typically appear as low-contrast speckles, blending into the surrounding background. To accurately detect these targets, the model must capture multiscale features while densely covering the speckle-like ship targets.

Introducing DenseASPP can address this issue. Similar to its application in autonomous driving scenarios, DenseASPP generates denser multiscale feature representations by connecting dilated convolutional features with varying dilation rates. These features provide broader coverage and denser representation, thereby enabling better capture of ship targets. DenseASPP resolves the contradiction between feature map resolution and receptive field by employing dilated convolutions to enhance segmentation outcomes further. A base network is followed by multiple levels of dilated convolutional layers, integrating each dilated convolutional output in a dense manner [31]. The use of rational dilation rates in DenseASPP enables neurons to acquire progressively larger receptive fields while avoiding convolutional degradation caused by excessively large dilation rates.

Furthermore, in the case of a higher proportion of smaller vessels in the dataset, models focus on acquiring semantic information about smaller ships, resulting in lower detection capabilities for larger vessels. The DenseASPP module addresses this by employing multiple dilated convolutions with varying sampling rates to capture semantic information at different scales. These pieces of information are fused

together through dense connections, alleviating gradient vanishing issues and enabling better propagation and utilization of feature information. This design enables the network to better capture features at different scales during the object detection task, thereby enhancing the detection capability for larger vessels without compromising the ability to detect smaller vessels.

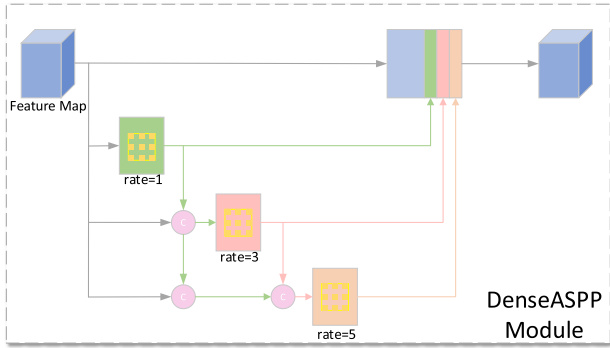


FIGURE 5. Structure of the DenseASPP module.

The effectiveness of the DenseASPP module in ship detection in SAR images can be attributed to the distinct characteristics of ship targets in SAR imagery. In SAR images, ships typically exhibit distinct shapes, sizes, clear edges, and textures, making them prominently visible against the background and providing the model with substantial and well-defined feature information.

SAR images possess remarkable high-resolution and surface feature representation capabilities. However, these images often contend with various types of noise interference, such as speckle noise and amplitude variations, posing challenges in ship target detection. The DenseASPP module effectively mitigates these interferences through its dilated convolutional structure, enhancing the clarity and quality of the images.

Concerning feature extraction and enhancement, DenseASPP efficiently extracts ship target features within SAR images through a sequence of dilated convolutional layers. These features encompass aspects such as shape, size, orientation, and texture, thereby aiding in distinguishing between various types and orientations of ships. Moreover, through dense connections and feature fusion, DenseASPP further enhances these feature representations, improving the model's performance in terms of classification accuracy and robustness. In this study, the null rate of DenseASPP is taken as 1, 3 and 5 for three sets of dilated convolutions.

D. LOSS FUNCTION

Due to the adoption of anchor-free concepts, the loss function in YOLOv8 has undergone significant changes compared to that used in the YOLOv5 series. Loss can be categorized as classification and regression losses. Binary cross-entropy loss is employed for classification loss, and distribution focal loss (DFL) and bounding box regression loss are employed

for regression loss. As SAR ship images involve only one loss category, the binary cross-entropy loss for classification loss is zero. The overall loss is the sum of these two losses weighted by certain proportions, represented as follows:

$$f_{loss} = \lambda_1 f_{DFL} + \lambda_2 f_{BBRL} \quad (4)$$

DFL optimizes the focal loss function by integrating discrete classification results into continuous outcomes. The expression is as follows:

$$f_{DFL}(S_i, S_{i+1}) = -((y_{i+1} - y) \log(s_i) + (y - y_i) \log(S_{i+1})) \quad (5)$$

where y_i and y_{i+1} respectively denote the values from the left and right side close to the successive labels y that satisfy $y_i < y < y_{i+1}$ and $y = \sum_{i=0}^n P(y_i)y_i$. P can be realized as $P(y_i)$ by means of a softmax layer, S_i of the above equation.

The bounding box loss function is a crucial component of object detection loss functions and thus greatly affects the performance of object detection models when well-defined. Recent studies have often assumed high-quality examples in training data and have focused on enhancing the fitting capability of bounding box losses.

However, avoiding the inclusion of low-quality images in training data is challenging; moreover, geometric metrics such as distance and aspect ratio can exacerbate errors in low-quality images, leading to a decrease in the model's generalization performance. An ideal loss function should reduce geometric errors when anchor boxes and target boxes have good overlap, minimizing excessive intervention in training results and thus enhancing the model's generalization ability.

The interference caused by environmental and geographical factors in SAR imagery results in a significant variation in the size and shape of vessels, affecting the image quality. Therefore, in this study, we adopted the Wise-IoU v3 loss function [32]. In contrast to its predecessor, Wise-IoU v3 does not involve the calculation of aspect ratios; instead, it incorporates a dynamic non-monotonic focusing mechanism based on attention-driven bounding box loss (Wise-IoU v1). It utilizes "outlierness" to describe the quality of anchor boxes, enabling Wise-IoU to more accurately assess object detection results and mitigate the bias issues associated with traditional IoU calculations.

The attention-driven bounding box loss function Wise-IoU v1 can be expressed as follows:

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \quad (6)$$

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \quad (7)$$

where L_{WIoUv1} is the boundary loss function Wise-IoU v1, L_{IoU} is the bounding box loss IoU, R_{WIoU} is the distance attention x and y respectively are the horizontal and vertical coordinates of the centroid of the prediction box, x_{gt} and y_{gt} respectively are the horizontal and vertical coordinates of the centroid of the real box, W_g and H_g respectively are the

width and height of the minimum outer connection matrix of the prediction box and the real box, respectively, and $*$ denotes the operation of separating the operation from the computational map to make it a constant with no gradient. Next, we use outlieriness to describe the quality of anchor boxes, defined as follows:

$$\beta = \frac{L_{IoU}^*}{L_{IoU}} \in [0, +\infty) \quad (8)$$

where β represents the moving average value.

Finally, by utilizing outlieriness to construct a non-monotonic focusing coefficient, the boundary box loss Wise-IoU v3 is derived. The calculation formula is as follows:

$$L_{WIoUv3} = kL_{WIoUv1} \quad (9)$$

where k is the non-monotonic focusing coefficient ($k = \frac{\beta}{\delta\alpha\beta - \delta}$) and is represented by hyperparameters α and δ (in this study, α and δ were set as 1.9 and 3, respectively).

III. RESULTS

A. EVALUATION METRICS

To evaluate the performance of the MSFA-YOLO algorithm in ship detection in SAR images by using the validation dataset, in this study, precision (P), recall (R), average precision (AP), and mean average recall (mAP) were employed as evaluation metrics. These metrics are calculated based on true positives (TP), false positives (FP), and false negatives (FN). TP indicates the count of predicted positive targets that are indeed positive; in other words, it signifies when MSFA-YOLO accurately detects and locates ship targets. FP indicates the count of predicted positive targets that are indeed negative. FN represents the count of predicted negative targets that are indeed positive. The calculation formulas for P , R , AP and mAP are as follows:

$$P = \frac{TP}{TP + FP} \quad (10)$$

$$R = \frac{TP}{TP + FN} \quad (11)$$

$$AP = \int_0^1 PRdR \quad (12)$$

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (13)$$

In the experiment, the YOLOv8 model served as the baseline model. The network was initially trained with a dataset in.txt format, followed by optimization using the SGD optimizer with a batch size of 16 for 300 epochs. The initial learning rate for the backbone network was set as 0.02, with a momentum of 0.9. All experiments were conducted on an NVIDIA GeForce RTX 3060Ti GPU. The software and hardware environment required for the experiment are listed in Table 1.

TABLE 1. Experimental hardware and software environment.

Item	Parameter
Operating System	Windows 11
Programming Language	Python3.8
CPU	Intel Core i5 12600KF
GPU	NVIDIA GeForce RTX 3060Ti
VRAM	8G
Algorithm Framework	Pytorch-2.0.1

B. DATASETS

In this study, the SAR-Ship-Dataset [33], SSDD [34], and HRSID [35] datasets were utilized for training and testing ship detection models. The SAR-Ship-Dataset combines data from the China Gaofen-3 SAR and Sentinel-1 SAR, comprising 102 Gaofen-3 and 108 Sentinel-1 SAR images, resulting in a collection of 43,819 images and 59,535 ship instances for use as high-resolution SAR ship target DL samples. The SSDD dataset includes 1160 SAR images, each with an average size of 500×500 pixels, and a total of 2358 ship instances. The HRSID dataset, released by the University of Electronic Science and Technology in January 2020 and designed for ship detection, semantic segmentation, and instance segmentation tasks in high-resolution SAR images, comprises 5604 high-resolution SAR images and 16,951 ship instances. Details of the number of images and ship instances in each dataset are provided in Table 2.

TABLE 2. The number of images in the dataset with the number of images.

datasets	Number of Images	Number of Ships
Sar-Ship-Dataset	43819	59535
SSDD	1160	2358
HRSID	5604	16951

The images of SAR ship datasets are typically grayscale, highlighting the target's structure and shape. The rich-texture characteristics enable clear outlines and details of ships in the images, providing additional information for identification. However, due to the absence of color information, further discrimination of targets is limited in certain scenarios. Additionally, SAR images may be influenced by factors such as terrain and surface scattering, leading to the appearance of artifacts or clutter, posing challenges to accurate ship detection. When analyzing SAR ship datasets, it is necessary to comprehensively consider both their advantages and limitations.

As illustrated in Figure 6, partial image data from three datasets are presented. Images 6(a)-6(c) are sourced from the SSDD dataset, 6(d)-6(f) from the SAR-Ship-Dataset, and 6(g)-6(i) from the HRSID dataset. Among these, 6(a), 6(c), 6(d), 6(g), and 6(i) represent images in complex scenes, while 6(b), 6(d), and 6(h) represent images in dense scenes.

Images 6(a), 6(f) depict large to medium-sized ship data, and 6(e) represents high-noise image data.

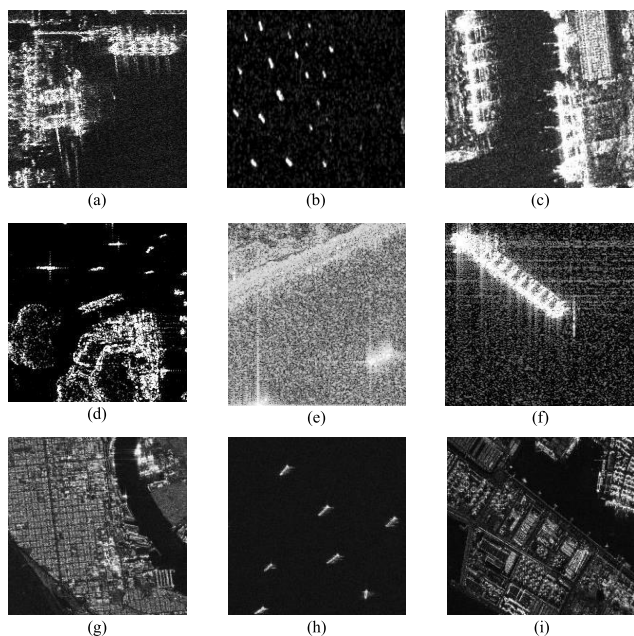


FIGURE 6. Instances of different dataset.

During the model training process, we divided SAR-Ship-Dataset, SSDD, and HRSID datasets into training, validation, and test sets in an 8:1:1 ratio. To ensure uniformity among the three datasets, image labels were uniformly formatted as .txt data.

Through an analysis of images from SAR-Ship-Dataset, SSDD, and HRSID datasets, the following common and different points were identified:

- 1) As can be seen in Figure 6-a, images with a single vessel are the most prevalent, while other images typically contain an average of 2–3 vessels, with some instances containing numerous vessel elements. This highlights the uneven spatial distribution of vessels, which may result in missed detections in images containing numerous vessel targets.
- 2) As can be observed in Figure 6-b, the majority of vessels are small-sized targets, with a few larger-sized vessel targets. This size imbalance may lead to underfitting issues when detecting larger vessel targets.
- 3) As can be seen in Figure 6-c, there is a higher proportion of small-sized vessels, and the majority of larger vessels exhibit an elongated and flat shape.
- 4) As can be observed in Figure 6-d, the bounding boxes outlining vessel targets are predominantly small squares or elongated shapes, with the majority of width-to-height ratios in the range of 1–2 for the true boxes.
- 5) The proportions of large vessels vary among the three datasets: SAR-Ship-Dataset has only 180 images containing large vessels, accounting for merely 0.302% of the total; SSDD has 2.63%, and HRSID has 1.894%.

This poses challenges for models detecting large vessels. (The classification of the size of ships is not based on their actual dimensions but is determined by the pixels occupied by the anchor box in the image. When the pixel area of the anchor box is less than 1024 ($32*32$), it is considered a small target. When the pixel area of the anchor box is greater than or equal to 1024 ($32*32$) and less than 9216 ($96*96$), it is classified as a medium-sized target. When the pixel area of the anchor box is greater than or equal to 9216 ($96*96$), it is considered a large target.)

C. ABLATION EXPERIMENTS

For ablation experiments, we used the open-source datasets SSDD, SAR-Ship-Dataset, and HRSID as testing benchmarks to assess the effectiveness of various model improvements across different datasets. We used YOLOv8 as the baseline algorithm to observe the practical impact of the modifications made to different modules. The performance of each module was measured using P, R, mAP50, mAP75, and mAP50–95. Details regarding the ablation experiment design and results are presented in Table 3.

In Table 3, YOLOv8n is the reference model. Taking the SAR-Ship-Dataset dataset as an example, its mAP50 is 96.5%, mAP75 is 77.3%, and mAP50–95 is 65.6%. Given the high level achieved by P, R and mAP50, our primary focus in the subsequent analysis is the comparison between mAP75 and mAP50–95.

The integration of the SE attention mechanism into the C2f module increased mAP50–95 by 0.6% and mAP75 by 1.7%. This integration allowed better feature learning for target regions, improving accuracy in detection and localization, especially for smaller targets. Furthermore, in scenarios with complex backgrounds, the SE attention mechanism aided in precise feature selection, reducing FPs and subsequently increasing the mAP.

The addition of the DenseASPP module resulted in a decrease in mAP on the SAR-Ship-Dataset, but no noticeable disparity was observed on the other two datasets. This discrepancy is due to the scarcity and low proportion of large vessels in the SAR-Ship-Dataset compared to the SSDD and HRSID.

Incorporating Wise-IoU resulted in a noticeable improvement in mAP. This improvement is attributed to the interference encountered in SAR images due to environmental and geographical factors, causing significant variations in vessel sizes and shapes in SAR images. Training data unavoidably contains low-quality images. The Wise-IoU loss function reduces geometric errors when anchor boxes align well with target boxes, thereby reducing intervention in low-quality images and enhancing the model's generalization capability.

D. COMPARATIVE EXPERIMENTS

To further assess the impact and effectiveness of the C2fSE and denseASPP modules and Wise-IoU loss function on

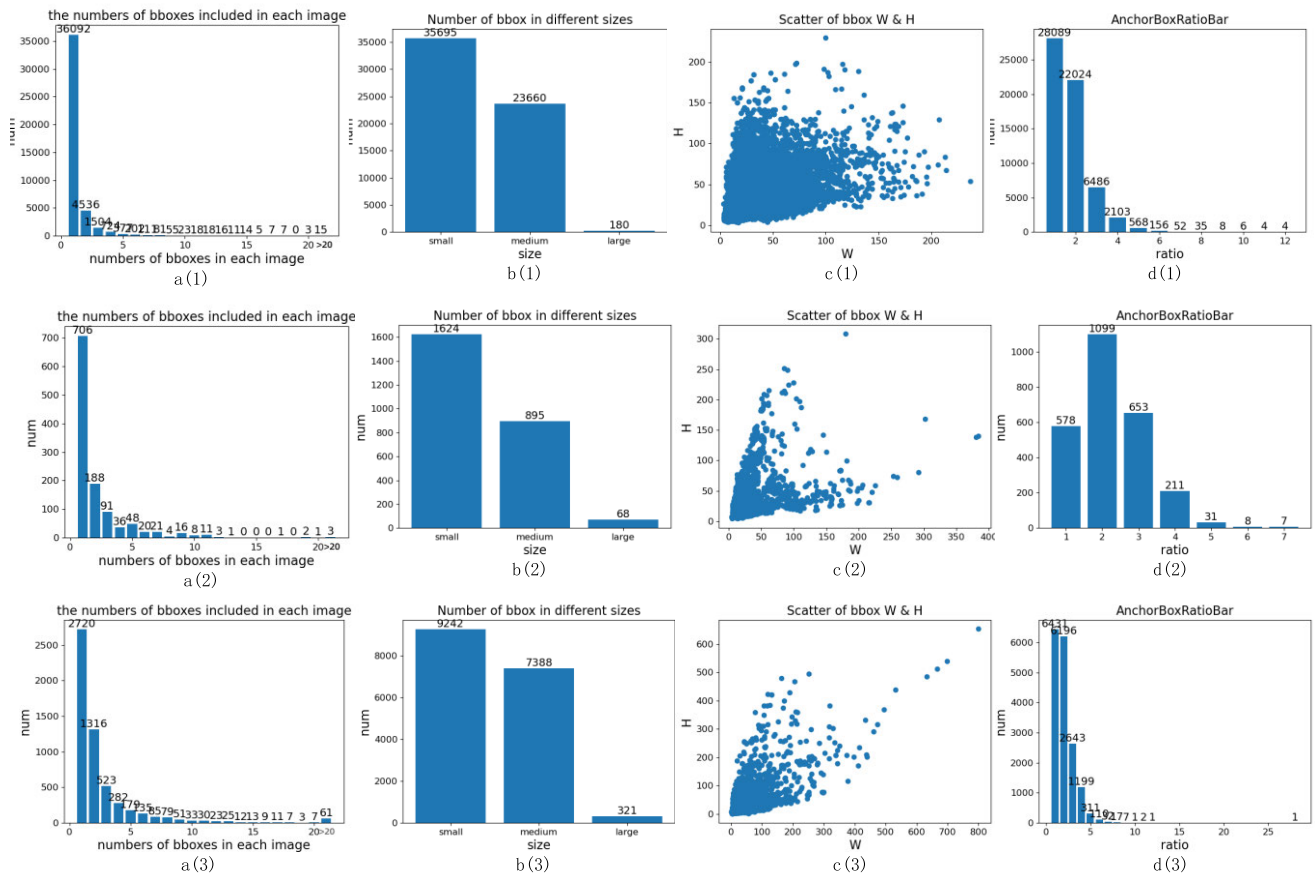


FIGURE 7. Data set analysis charts.

TABLE 3. Ablation experiment.

Methods	C2fSE	DenseASPP	Wise-IoU	Dataset	P	R	mAP50	mAP75	mAP50-95
YOLOv8n	-	-	-	Sar-Ship-Dataset	93.6	91.7	96.5	77.3	65.6
A1	√	-	-		94.5	93.8	96.8	79.0	66.2
B1	√	√	-		93.9	93.4	96.4	78.1	65.8
C1	√	√	√		94.5	93.3	96.9	80.4	67.7
YOLOv8n	-	-	-	SSDD	96.8	96.9	98.2	76.2	65.7
A2	√	-	-		97.4	97.3	98.1	76.5	65.9
B2	√	√	-		96.9	97.5	98.2	76.4	65.6
C2	√	√	√		97.7	98.0	98.7	76.9	66.2
YOLOv8n	-	-	-	HRSID	93.1	83.6	92.6	74.9	66.4
A3	√	-	-		93.0	85.1	92.6	75.4	66.4
B3	√	√	-		93.1	85.7	92.7	75.1	66.5
C3	√	√	√		92.4	86.0	92.7	76.7	67.1

the model, we conducted comparative experiments for each module.

1) C2FSE MODULE

In the backbone network, we integrated the SE attention into the C2f module of the baseline model, enhancing the model’s ability to extract image features. Consequently, we compared the initial module with the addition of CBAM [36] and ECA [37]. A comparison of the effects of incorporating different attention mechanisms into the baseline model is presented in Table 4.

The model with the integrated SE attention mechanism exhibited superior results. Specifically, when compared to the baseline model, on the SAR-Ship-Dataset, mAP75 and mAP50–95 increased by 1.7% and 0.6%, respectively. On the SSDD, mAP75 and mAP50–95 increased by 0.3% and 0.2%, respectively. On the HRSID, mAP75 increased by 0.5%, while mAP50–95 did not improve.

The advantage of the SE attention mechanism over the baseline model lies in increasing the importance of specific features, enabling the model to focus more on crucial regions, resulting in enhanced accuracy and precision of ship

TABLE 4. C2f module fusion different attention comparison table.

Methods	Sar-Ship-Dataset			SSDD			HRSID		
	mAP50	mAP75	mAP50-95	mAP50	mAP75	mAP 50-95	mAP50	mAP75	mAP 50-95
C2f	96.5	77.3	65.6	98.2	76.2	65.7	92.6	74.9	66.4
C2fCBAM	96.8	79.3	66.7	98.1	76.4	65.4	91.7	74.6	66.0
C2fECA	96.3	78.1	66.0	98.5	76.3	66.1	91.2	74.1	65.3
C2fSE	96.8	79.0	66.2	98.1	76.5	65.9	92.6	75.4	66.4

TABLE 5. Comparison of different pyramid structures.

Methods	Sar-Ship-Dataset				SSDD			HRSID		
	mAP50	mAP75	mAP50-95	mAP50	mAP75	mAP50-95	mAP50	mAP75	mAP50-95	
SPP[38]	96.7	77.1	67.0	97.9	76.7	65.6	92.4	74.3	65.7	
SPPF	96.5	77.3	65.6	98.2	76.2	65.7	92.6	74.9	66.4	
simSPPF[14]	96.8	77.6	66.9	97.3	74.2	62.8	92.1	74.1	64.8	
ASPP[39]	96.4	78.3	65.3	97.7	75.9	63.2	90.7	73.5	64.7	
SPPCSPC[15]	96.3	76.8	67.3	98.0	76.0	65.3	91.8	75.0	65.8	
Ours	96.8	80.1	67.4	98.3	76.1	65.9	92.7	75.1	66.8	

detection. In comparison to other attention mechanisms, such as efficient channel attention (ECA) and convolutional block attention module (CBAM), SE exhibits greater proficiency in capturing specific frequency or channel information within SAR images. In addition, SE demonstrates more significant advantages in model design and data adaptability. Therefore, the superiority of the SE attention mechanism stems from its effectiveness in feature extraction and its more adaptable nature in ship detection tasks.

2) FEATURE PYRAMID STRUCTURE

To validate the contributions of different attention modules in SAR ship image detection, we conducted comparative experiments between the CASPP module and existing mainstream architectures, including SPP [38], SPPF, simSPPF [14], ASPP [39], and SPPFCSPC [15]. Table 5 presents the performance of different feature pyramids across the SAR-Ship-Dataset, SSDD, and HRSID.

The improved DenseASPP module exhibited outstanding performance across all three datasets. Specifically, on SAR-Ship-Dataset, mAP75 and mAP50-95 increased by 2.8% and 1.8%, respectively. On the SSDD dataset, mAP75 did not improve, but mAP50-95 increased by 0.2%. On the HRSID dataset, mAP75 and mAP50-95 increased by 0.2% and 0.4%, respectively.

This improvement is attributed to the capability of the DenseASPP module to capture receptive fields and global information. Dilation convolutions within the DenseASPP module aid in enlarging the neurons' receptive fields, enabling the model to comprehensively acquire global information from images. This ability is crucial in ship detection as the shapes and positions of ships are often closely linked to their surrounding environment. By acquiring more global information, DenseASPP can more accurately locate and identify ship targets compared to the baseline model.

3) LOSS FUNCTION

To assess the contribution of the Wise-IoU module to SAR ship image detection, we conducted comparative experiments

against prevalent models, namely GIoU [40], DIOU [41], CIOU [41], SIOU [42], and EIOU [43]. The performance of different IoU metrics across the SAR-Ship-Dataset, SSDD, and HRSID datasets is presented in Table 6.

From Table 6, it is evident that Wise-IoU outperformed other IoU metrics across the three datasets. On the SAR-Ship-Dataset dataset, there was a 3.4% increase in mAP75 and a 2.4% increase in mAP50-95. On the SSDD dataset, there was a 0.4% increase in mAP75, while mAP50-95 remained unchanged. On the HRSID dataset, mAP75 and mAP50-95 increased by 1.4% and 0.4%, respectively.

Due to disturbances caused by environmental and geographical factors, ship sizes and shapes vary significantly in SAR images, leading to the inclusion of low-quality images in training data. Traditional bounding box loss functions assume high-quality training data. However, this assumption can cause a decrease in the model's generalization capability on low-quality images. The Wise-IoU loss function reduces geometric errors when anchor boxes align well with target boxes, minimizing intervention on low-quality images and thereby improving the model's generalization capability. This demonstrates the effectiveness of introducing the Wise-IoU loss function into the model.

E. COMPARISON OF DIFFERENT ALGORITHMS

A quantitative comparison of MSFA-YOLO against various SAR image ship detection algorithms is presented in Table 7. This comparison involves two-stage and one-stage object detection algorithms alongside SAR image ship detection algorithms against this proposed algorithm on the SAR-Ship-Dataset, SSDD, and HRSID datasets.

Among two-stage detection methods, most studies have utilized Faster R-CNN [16] and Cascade R-CNN [17] as benchmark models for improvements. CRTransSar [44] incorporates a backbone network based on Swin Transformer for context-aggregated representation learning but requires a large model size and parameter count. In the realm of single-stage detection methods, inspired by the YOLO series

TABLE 6. Effect of different IoUs on experimental results.

IoU	Sar-Ship-Dataset			SSDD			HRSID		
	mAP50	mAP75	mAP50-95	mAP50	mAP75	mAP50-95	mAP50	mAP75	mAP50-95
GIoU[40]	96.4	77.8	66.4	98.1	76.2	65.4	91.6	74.1	64.9
DIoU[41]	96.9	77.3	66.1	98.4	75.7	65.2	92.4	73.6	65.3
CIoU[41]	96.5	77.3	65.6	98.2	76.2	65.7	92.6	74.9	66.4
SIoU[42]	96.9	80.3	66.7	98.6	77.1	65.4	92.5	74.3	65.7
EIoU[43]	96.8	80.1	67.3	98.1	75.3	65.9	92.0	74.3	65.7
ours	96.8	80.7	68.0	98.0	76.6	65.7	92.7	76.3	66.8

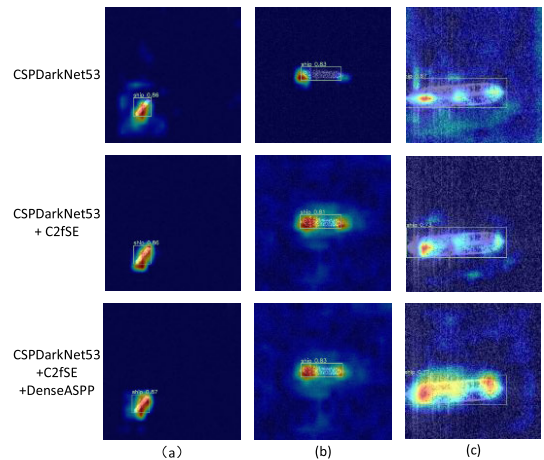
TABLE 7. Comparison of different algorithms.

Methods	Sar-Ship-Dataset					SSDD					HRSID				
	P	R	mAP 50	mAP P75	mAP50- 95	P	R	mAP5 0	mAP5 5	mAP5 0-95	P	R	mA P50	mAP7 5	mAP50 -95
Faster rnn[16]	80.7	82.4	96.3	72.6	64.2	81.0	94.2	97.1	71.3	61.0	87.2	89.1	89.1	64.7	56.1
Cascade rnn[17]	81.2	82.7	96.7	72.7	65.4	81.9	93.0	97.1	72.1	62.5	88.6	87.7	90.3	65.8	57.7
CRTransSar [44]	-	-	-	-	-	81.9	93.0	97.0	76.2	-	-	-	-	-	-
YOLOv5n	94.4	94.6	96.9	76.8	65.9	92.5	98.3	98.0	75.8	63.2	94.7	89.4	91.3	72.8	65.7
YOLOv8n	93.6	91.7	96.5	77.3	65.6	96.8	96.9	98.2	76.2	65.7	93.1	83.6	92.6	74.9	66.4
CRAS- YOLO[45]	-	-	-	-	-	97.3	95.5	98.7	-	61.1	-	-	-	-	-
CSD- YOLO[46]	-	-	-	-	-	95.9	95.9	98.6	-	-	93.2	80.4	86.1	-	-
FEPS- NET[47]	-	-	-	-	-	-	-	96.0	67.5	59.9	-	-	90.7	74.3	65.7
MSFA- YOLO(ous)	94.5	93.3	96.9	80.4	67.7	97.7	98.0	98.7	76.9	66.2	92.4	86.0	92.7	76.7	67.1

algorithms, several researchers have independently implemented related detection models. For instance, CRAS-YOLO [45], which incorporates one-stage algorithms, exhibits outstanding performance in detecting small objects; however, the overall precision, particularly mAP50-95, does not show significant improvement. CSD-YOLO [46] proposes a module that enhances the model's ability to process complex information and improves the model's adaptability to complex scenarios, but it is only comparable to the model proposed in this paper at mAP50. FEPS-NET [47] is mainly optimized for the detection of small ships, and the overall accuracy has not been effectively improved.

Compared to existing object detection methods, the proposed MSFA-YOLO model, which is a one-stage SAR image ship detection model, demonstrates superior performance, exhibiting higher efficiency in detecting small objects while enhancing the detection of larger objects.

In addition, we validate the model performance in inshore and offshore scenarios in the SSDD dataset. In the pelagic scenario, the background is simpler, and in the nearshore scenario, the background is complex and difficult to recognize. This is shown in Table 8 below. It can be seen that compared with the initial model, MSFA-YOLO has a relatively significant improvement in P, R, and mAP in complex scenarios. Specifically, in the inshore case, the mAP50 of the proposed method increased by 2.5%, the mAP75 by 8.2%, and the mAP50-95 by 4.5%, suggesting that MSFA-YOLO has an effective effect in coping with complex

**FIGURE 8. Contrasting heat maps of different backbone feature extraction capabilities.**

scenarios. In the offshore case, MSFA-YOLO improves mAP50 by 0.4%, mAP75 by 4.3%, and mAP50-95 by 1.7%. This indicates that the model has good results in the offshore case as well.

F. MULTI-SCALE TARGET ANALYSIS

To illustrate the enhancement of the model at multiple scales, we compared the model on SSDD, Sar-Ship-Dataset and HRSID datasets for small, medium and large targets

TABLE 8. Performance of the model on the SSDD dataset for the inshore versus offshore scenarios.

Method	inshore					offshore				
	P	R	mAP50	mAP75	mAP50-95	P	R	mAP50	mAP75	mAP50-95
YOLOv8n	92.5	89.5	93.4	52.4	52.6	96.7	96.5	98.0	62.5	58.2
MSFA-YOLO (proposed method)	94.0	90.3	95.9	60.6	57.1	97.8	95.5	98.4	66.8	59.9

TABLE 9. Performance of ships at multi scales on SSDD, Sar-Ship-Dataset and HRSID.

Method	SSDD			Sar-Ship-Dataset			HRSID		
	AP _S	AP _M	AP _L	AP _S	AP _M	AP _L	AP _S	AP _M	AP _L
YOLOv8n	56.2	70.4	64.6	58.3	69.7	57.7	52.4	80.8	62.1
MSFA-YOLO	56.6	71.0	66.7	59.9	71.3	66.3	53.7	80.6	64.0

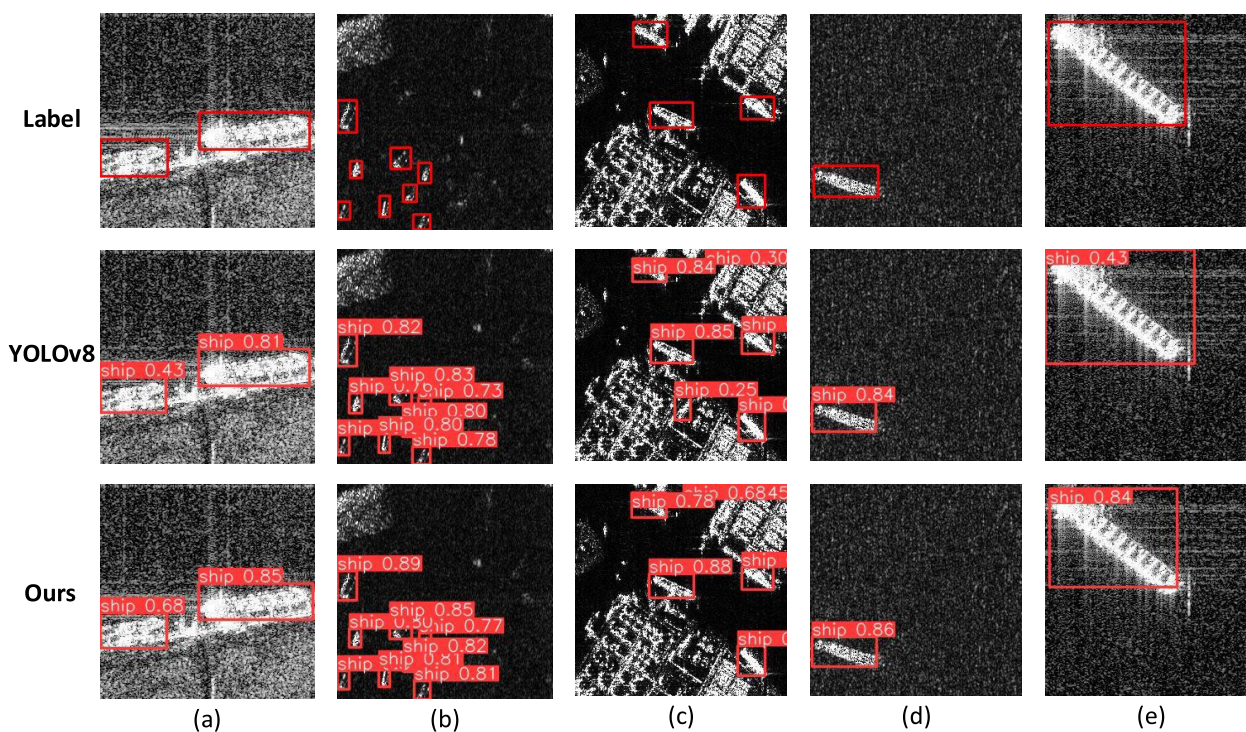


FIGURE 9. Comparative performance graph.

respectively. The resultant data are shown in the table 9 below. Specifically, the small target ships were improved by 0.4% in SSDD, 1.6% in Sar-Ship-Dataset, and 1.3% in HRSID in the SSDD dataset; the medium target was improved by 0.6%, 1.6%, and -0.2% in the three datasets, and the large target was improved by 2.1%, 8.6%, and 1.9%, respectively. This indicates that the model has good results for large, medium and small ships, with extraordinarily good results for large ships.

Furthermore, to demonstrate the backbone network’s superior semantic perception capability, we compared the results of image heatmaps at different scales. Heatmaps illustrate the model’s activity levels across different regions or features of the input data, reflecting the model’s perception of various semantic categories. The heatmaps revealed that the improved model extracts object boundaries and texture

features more meticulously than the baseline model [48]. The heatmaps of different backbone networks are depicted in Figure 8.

As can be seen in Figure 8(a), both the baseline and improved models performed well in detecting small-sized vessels. However, the enhanced model exhibited a more pronounced delineation of ship boundaries.

As can be seen in Figure 8(b), the heatmaps of both improved models (CSPDarkNet53 + C2fSE and CSPDarkNet53 + C2fSE + DenseASPP) were clearer for medium-sized vessels. This is due to the enhanced feature extraction capability of the models, particularly in capturing ship outlines and finer details.

As can be observed in Figure 8(c), CSPDarkNet53 + C2fSE + DenseASPP exhibited significant progress compared to the baseline and CSPDarkNet53 + C2fSE models.

This improvement can be attributed to the DenseASPP module, which utilizes dilated convolutions to capture image information across multiple scales. The dense feature extraction comprehensively captures various-sized and shaped ship features, enhancing the model's understanding of the image. Furthermore, the DenseASPP module incorporates more branches and pooling rates, enabling better fusion of features across different levels and scales. This robust feature fusion improves the model's ability to identify larger vessels by effectively combining multiple features.

G. EXPERIMENTAL EFFECT ANALYSIS

In practical scenarios, variations in image-capturing height, environmental noise, day–night transitions, and ship sizes lead to different semantic representations of ships in images. To evaluate the proposed model's performance across different scenarios, we selected SAR ship images of different sizes and scenes. The comparative performance graph is shown in Figure 9.

As can be seen in Figures 9(a), 9(b), and 9(c), the MSFA-YOLO model outperformed the baseline YOLOv8n in nearshore, dense, and complex scenes, demonstrating its adaptability across diverse scenarios. In particular, as can be seen in Figure 9(a) for coastal ship situations with lower image quality and higher noise, MSFA-YOLO outperformed the baseline YOLOv8n model, thus demonstrating its adaptability to low-quality images. As can be seen in Figure 9(b), for densely populated scenes with small objects, both the baseline model and MSFA-YOLO model performed well; however, MSFA-YOLO exhibited higher confidence. In the case of ship detection in complex scenes (Figure 9(c)), the baseline model yielded false positives, whereas MSFA-YOLO performed well; thus, demonstrating its adaptability to complex scenarios.

Furthermore, as can be seen in Figures 9(b), 9(d), and 9(e), both the baseline model and the proposed MSFA-YOLO model performed well in scenarios with small- and medium-sized objects. However, in scenarios with large objects, MSFA-YOLO significantly outperformed the baseline model. This improvement can be attributed to the use of the DenseASPP module, which utilizes various dilation rates in dilated convolutions to acquire contextual information with different receptive fields, thereby enhancing the model's ability to extract features at different scales and allowing the model to capture semantic information for objects of varying sizes, thereby improving the model's generalization ability.

IV. CONCLUSION

In this paper, we proposed a novel SAR ship detection model, MSFA-YOLO, to enhance the detection and localization of ships in various SAR imaging scenarios. In the experiments, the proposed model demonstrated superior performance across different ship scales and for strong noise interference, thus demonstrating its effectiveness.

The proposed method can be employed for the efficient and accurate detection and localization of maritime vessels. With ship detection at its core, this study is directly relevant to maritime safety and monitoring vessels in restricted areas.

The MSFA-YOLO model can be effectively utilized for SAR-based maritime ship detection tasks. For significant variations in ship size and noise due to differences in SAR imaging distance and environments in real-world applications, the MSFA-YOLO model not only enables enhanced ship feature extraction but also refines detection granularity; thus, ensuring a higher detection rate for small targets while improving the accuracy of large target detection.

In the current research framework, to further advance the research and application in the field of SAR image ship detection, future investigations can be focused on the following aspects. Firstly, the fusion of multi-source data is regarded as a crucial factor for enhancing the accuracy and robustness of ship detection. By integrating SAR images with data obtained from other sensors, such as optical images and radar images, it is anticipated to capture target information more comprehensively, particularly in adverse weather conditions. This research direction is expected to contribute to the development of a more comprehensive and reliable ship detection system.

Secondly, the application of transfer learning is considered an essential measure to improve the efficiency of SAR image ship detection models. Effectively utilizing models pretrained in other domains or tasks for transfer learning can accelerate and enhance the training process of the model. This approach holds promise for rapid deployment of SAR image ship detection models in scenarios where large-scale annotated data is lacking.

Simultaneously, real-time performance and efficiency optimization represent important directions tailored for practical applications. Researching how to further optimize models to achieve real-time performance and ensure efficient ship detection in large-scale SAR images holds significant practical implications, particularly in areas related to monitoring and emergency response.

These future research directions are poised to propel the development of SAR image ship detection, making it more adaptable to practical requirements. Through in-depth exploration of these directions, continuous advancements in the performance and practicality of ship detection technology can be achieved, providing robust support for research and practical applications in relevant domains.

REFERENCES

- [1] J. Zhang, S. Li, Y. Dong, B. Pan, and Z. Shi, "Hierarchical similarity alignment for domain adaptive ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3227626.
- [2] C. Zhang, P. Liu, H. Wang, and Y. Jin, "A review of recent advance of ship detection in single-channel SAR images," *Waves Random Complex Media*, vol. 33, nos. 5–6, pp. 1442–1473, Nov. 2023.
- [3] J. Zhang, Z. Liu, W. Jiang, Y. Liu, X. Zhou, and X. Li, "Application of deep generative networks for SAR/ISAR: A review," *Artif. Intell. Rev.*, vol. 56, no. 10, pp. 11905–11983, Oct. 2023.

- [4] K. Ouchi and T. Yoshida, "On the interpretation of synthetic aperture radar images of oceanic phenomena: Past and present," *Remote Sens.*, vol. 15, no. 5, p. 1329, Feb. 2023.
- [5] Z. Sun, X. Leng, Y. Lei, B. Xiong, K. Ji, and G. Kuang, "BiFA-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images," *Remote Sens.*, vol. 13, no. 21, p. 4209, Oct. 2021.
- [6] M. J. Er, Y. Zhang, J. Chen, and W. Gao, "Ship detection with deep learning: A survey," *Artif. Intell. Rev.*, vol. 56, pp. 1–41, Mar. 2023.
- [7] I. G. Rizaev, O. Karakuş, S. J. Hogan, and A. Achim, "Modeling and SAR imaging of the sea surface: A review of the state-of-the-art with simulations," *ISPRS J. Photogramm. Remote Sens.*, vol. 187, pp. 120–140, May 2022.
- [8] S. Zhao, Z. Zhang, W. Guo, and Y. Luo, "An automatic ship detection method adapting to different satellites SAR images with feature alignment and compensation loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3160727.
- [9] Z. Lin, K. Ji, X. Leng, and G. Kuang, "Squeeze and excitation rank faster R-CNN for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 751–755, May 2018.
- [10] Z. Zhou, J. Chen, Z. Huang, J. Lv, J. Song, H. Luo, B. Wu, Y. Li, and P. S. R. Diniz, "HRLE-SARDet: A lightweight SAR target detection algorithm based on hybrid representation learning enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3251694.
- [11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 21–37.
- [12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOV4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
- [13] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.
- [14] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOV6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.
- [15] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOV7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
- [16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [17] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
- [18] Y. Zhou, H. Liu, F. Ma, Z. Pan, and F. Zhang, "A sidelobe-aware small ship detection network for synthetic aperture radar imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3264231.
- [19] M. Kang, K. Ji, X. Leng, and Z. Lin, "Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection," *Remote Sens.*, vol. 9, no. 8, p. 860, Aug. 2017.
- [20] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOV5: Improved YOLOV5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.
- [21] G. Shi, J. Zhang, J. Liu, C. Zhang, C. Zhou, and S. Yang, "Global context-augmented object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10604–10617, Dec. 2021, doi: [10.1109/TGRS.2020.3043252](https://doi.org/10.1109/TGRS.2020.3043252).
- [22] J. Noh, W. Bae, W. Lee, J. Seo, and G. Kim, "Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9724–9733.
- [23] X. Li, D. Li, H. Liu, J. Wan, Z. Chen, and Q. Liu, "A-BFPN: An attention-guided balanced feature pyramid network for SAR ship detection," *Remote Sens.*, vol. 14, no. 15, p. 3829, Aug. 2022.
- [24] Z. Shao, X. Zhang, T. Zhang, X. Xu, and T. Zeng, "RBFA-Net: A rotated balanced feature-aligned network for rotated SAR ship detection and classification," *Remote Sens.*, vol. 14, no. 14, p. 3345, Jul. 2022.
- [25] Y. Guo, S. Chen, R. Zhan, W. Wang, and J. Zhang, "LMSD-YOLO: A lightweight YOLO algorithm for multi-scale SAR ship detection," *Remote Sens.*, vol. 14, no. 19, p. 4801, Sep. 2022, doi: [10.3390/rs14194801](https://doi.org/10.3390/rs14194801).
- [26] G. Tang, Y. Zhuge, C. Claramunt, and S. Men, "N-YOLO: A SAR ship detection using noise-classifying and complete-target extraction," *Remote Sens.*, vol. 13, no. 5, p. 871, Feb. 2021. [Online]. Available: <https://www.mdpi.com/2072-4292/13/5/871>
- [27] H. Guo, X. Yang, N. Wang, and X. Gao, "A CenterNet++ model for ship detection in SAR images," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107787.
- [28] Z. Sun et al., "An anchor-free detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7799–7816, Jul. 2021.
- [29] T. Zhang, X. Zhang, and X. Ke, "Quad-FPN: A novel quad feature pyramid network for SAR ship detection," *Remote Sens.*, vol. 13, no. 14, p. 2771, Jul. 2021.
- [30] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
- [31] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.
- [32] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-IoU: Bounding box regression loss with dynamic focusing mechanism," 2023, *arXiv:2301.10051*.
- [33] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sens.*, vol. 11, no. 7, p. 765, Mar. 2019, doi: [10.3390/rs11070765](https://doi.org/10.3390/rs11070765).
- [34] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *Proc. SAR Big Data Era, Models, Methods Appl. (BIGSAR DATA)*. IEEE, 2017, pp. 1–6.
- [35] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020.
- [36] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
- [37] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
- [38] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015, doi: [10.1109/TPAMI.2015.2389824](https://doi.org/10.1109/TPAMI.2015.2389824).
- [39] L. C. Chen, G. Papandreou, and I. Kokkinos, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Jun. 2017.
- [40] H. Rezatofghi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.
- [41] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12993–13000.
- [42] Z. Gevorgyan, "SIoU loss: More powerful learning for bounding box regression," 2022, *arXiv:2205.12740*.
- [43] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, Sep. 2022.
- [44] R. Xia, J. Chen, Z. Huang, H. Wan, B. Wu, L. Sun, B. Yao, H. Xiang, and M. Xing, "CRTransSar: A visual transformer based on contextual joint representation learning for SAR ship detection," *Remote Sens.*, vol. 14, no. 6, p. 1488, Mar. 2022.
- [45] W. Zhao, M. Syafrudin, and N. L. Fitriyani, "CRAS-YOLO: A novel multi-category vessel detection and classification model based on YOLOV5s algorithm," *IEEE Access*, vol. 11, pp. 11463–11478, 2023.
- [46] Z. Chen, C. Liu, V. Filaretov, and D. Yukhimets, "Multi-scale ship detection algorithm based on YOLOV7 for complex scene SAR images," *Remote Sens.*, vol. 15, no. 8, p. 2071, Apr. 2023.
- [47] L. Bai, C. Yao, Z. Ye, D. Xue, X. Lin, and M. Hui, "Feature enhancement pyramid and shallow feature reconstruction network for SAR ship detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1042–1056, Jan. 2023.
- [48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.



ZHAO LIANGJUN received the Ph.D. degree from the College of Resource and Environmental Sciences, Xinjiang University, China. He was an Associate Professor, a Senior Engineer, and a Master's Supervisor with the College of Computer Science and Engineering, Sichuan University of Science and Engineering. His research interests include satellite remote sensing, deep learning, and image processing.



LIANG GANG is currently pursuing the master's degree with the School of Computer Science and Engineering, Sichuan University of Science and Engineering. His main research interests include remote sensing, image processing, and target detection.



NING FENG is currently pursuing the master's degree with the College of Automation and Information Engineering, Sichuan University of Science and Engineering, China. His research interests include object detection and deep learning.



HE ZHONGLIANG is currently pursuing the master's degree with the School of Computer Science and Engineering, Sichuan University of Science and Engineering. His main research interests include remote sensing, image processing, and target detection.



XI YUBIN is currently pursuing the master's degree with the School of Computer Science and Engineering, Sichuan University of Science and Engineering. His main research interests include remote sensing, image processing, and building segmentation.



ZHANG YUANYANG is currently pursuing the master's degree with the School of Computer Science and Engineering, Sichuan University of Science and Engineering. His main research interests include remote sensing, image processing, and deep learning.

...