YOLO-v8 with Multidimensional Attention and Upsampling Fusion for Small Air Target Detection in Radar Images

JIANG Zhenyu[1*], *LI Xiaodong*[1], *DU Chen*[2,3], *CHEN An*[2,3],
HAN Yanqiang[2,3], *LI Jinjin*[2,3]

1. College of Aerospace Science, National University of Defense Technology, Changsha 410073, P. R. China;
2. National Key Laboratory of Advanced Micro and Nano Manufacture Technology, Shanghai Jiao Tong University, Shanghai 200240, P. R. China;
3. Department of Micro/Nano Electronics, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, P. R. China

(Received 20 September 2024; revised 30 November 2024; accepted 5 December 2024)

Abstract: This study presents an innovative approach to improving the performance of YOLO-v8 model for small object detection in radar images. Initially, a local histogram equalization technique was applied to the original images, resulting in a notable enhancement in both contrast and detail representation. Subsequently, the YOLO-v8 backbone network was augmented by incorporating convolutional kernels based on a multidimensional attention mechanism and a parallel processing strategy, which facilitated more effective feature information fusion. At the model's head, an upsampling layer was added, along with the fusion of outputs from the shallow network, and a detection head specifically tailored for small object detection, thereby further improving accuracy. Additionally, the loss function was modified to incorporate focal-intersection over union (IoU) in conjunction with scaled-IoU, which enhanced the model's performance. A weighting strategy was also introduced, effectively improving detection accuracy for small targets. Experimental results demonstrate that the customized model outperforms traditional approaches across various evaluation metrics, including recall, precision, F_1-score, and the receiver operating characteristic (ROC) curve, validating its efficacy and innovation in small object detection within radar imagery. The results indicate a substantial improvement in accuracy compared to conventional methods such as image segmentation and standard convolutional neural networks.

Key words: YOLO; radar images; object detection; machine learning

CLC number: TP751 **Document code:** A **Article ID:** 1005-1120(2024)06-0710-15

0 Introduction

In recent years, the rapid development of various types of lightweight commercial unmanned aerial vehicles (UAVs) has posed unprecedented threats to airspace security worldwide. Small air targets, characterized by erratic flight patterns and lack of control, have increasingly become significant challenges for air defense target identification. Radar image-based target detection plays a crucial role in addressing these challenges by enabling the detection and tracking of such small air targets. The ability to accurately detect and track objects in radar imagery is essential across numerous applications, including defense, surveillance, and security[1]. However, detecting small-sized targets remains particularly difficult due to the presence of objects of varying sizes and densely arranged patterns in radar images.

In the field of computer science, object detection has evolved from traditional methods, such as image segmentation, to more sophisticated algo-

*Corresponding author, E-mail address: jiangzhenyu00@yeah.net.

How to cite this article: JIANG Zhenyu, LI Xiaodong, DU Chen, et al. YOLO-v8 with multidimensional attention and upsampling fusion for small air target detection in radar images[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2024, 41(6):710-724.

http://dx.doi.org/10.16356/j.1005-1120.2024.06.004

rithms that leverage deep learning techniques[2]. Among these, the YOLO model has emerged as one of the most popular real-time detection models. YOLO frames the object detection task as a regression problem rather than a classification problem, enabling the model to predict both bounding boxes and class probabilities for objects in a single pass[3]. Object detection has wide-ranging applications across fields like biology, medicine, materials science, architecture, and the arts. However, this flexibility implies that no single model can address all detection problems, as each specific application may present unique challenges. Critical factors influencing in object detection include background interference, object occlusion, computational resource constraints, and data bias[4].

In this study, we present an enhanced version of the YOLO-v8 model, tailored to improve the detection of small objects in radar images. A dataset of radar images, each containing up to three objects of varying sizes, was used for model development and evaluation. To begin, local histogram equalization was applied to the original images to enhance contrast and detail[5]. Subsequently, we replaced the convolutional layers in the backbone network with convolutional kernels based on a multidimensional attention mechanism and a parallel processing strategy, enabling to improve the feature information fusion[6]. Furthermore, an upsampling layer was added to the model's head, which was fused and convolved with outputs from the shallow network in the backbone[7]. We also introduced an additional shallow network detection head. Finally, the loss function was modified to incorporate both focal intersection over union (FIoU) and scaled intersection over union (SIoU), which were applied through a weighting strategy[8].

To evaluate the performance of the customized YOLO-v8 model, we compared it against a basic image segmentation algorithm[9] and the standard convolutional neural network VGG-16[10], typically used for classification tasks. The experimental results demonstrated that the customized YOLO-v8 model achieved an overall weighted accuracy of 86% in detecting small objects in radar images. In contrast, the standard convolutional neural network (CNN) achieved only 34% accuracy, while the image segmentation algorithm performed poorly. Notably, the YOLO-v8 model identified all larger objects with near-perfect accuracy, indicating that the custom modifications did not compromise the model's standard predictive power. Fig.1 illustrates the

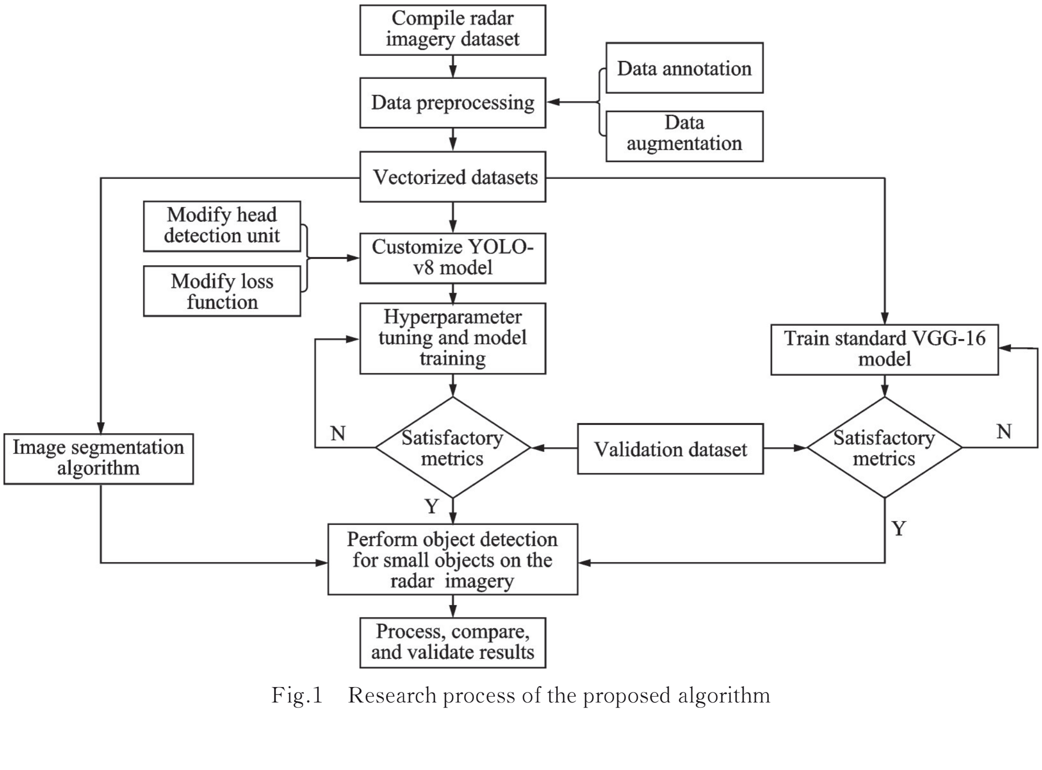

Compile radar imagery dataset
Data preprocessing
Data annotation
Data augmentation
Vectorized datasets
Modify head detection unit
Modify loss function
Customize YOLO-v8 model
Hyperparameter tuning and model training
Train standard VGG-16 model
N
Satisfactory metrics
Validation dataset
Satisfactory metrics
N
Image segmentation algorithm
Y
Y
Perform object detection for small objects on the radar imagery
Process, compare, and validate results

Fig.1 Research process of the proposed algorithm

process of the algorithm used to generate the results of this research. This includes data preprocessing, model customization, hyperparameter tuning and training, as well as iterative optimization through validation metrics, ultimately achieving small object detection and result validation.

1 Methods

1.1 YOLO

YOLO-v8 is an object detection model and the latest version in the YOLO series[3]. Built on the Darknet framework, YOLO-v8 primarily employs CNNs to detect objects within images. Its architecture leverages Darknet-53 as the backbone network and integrates feature fusion methods such as feature pyramid networks (FPNs) and path aggregation networks (PANs). To further enhance the model's receptive field and feature representation capabilities, it incorporates modules like the spatial pyramid pooling (SPP) and the segment anything model (SAM). YOLO-v8 also adopts multi-scale training and testing strategies, along with new loss functions based on intersection over union (IoU), improving both accuracy and robustness.

The key attributes of the YOLO-v8 model include efficiency, innovation, ease of use, and comprehensiveness[11]. It is optimized to run on various hardware platforms, including central processing units (CPUs) and graphics processing units (GPUs), achieving excellent performance in both speed and accuracy. YOLO-v8 builds on the design advantages of previous YOLO versions and has been enhanced to maintain its simplicity and ease of use in engineering applications. In preprocessing, the model has been optimized for tasks such as letterbox resizing (with configurable scaling and padding modes, primarily to resize images to 640 pixel × 640 pixel), and conversion to formats like RGB and CHW.

YOLO-v8's architecture is divided into three primary components: the backbone, the neck (optional), and the head[3]. In the backbone, two consecutive 3×3 convolutions reduce the resolution by a factor of four, maintaining a small receptive field while enriching the model's gradient flow through cross-layer connections. The head, the final layer of the model, adapts depending on the specific detection task, enabling flexibility for various use cases.

The YOLO-v8 model has been pre-trained on the COCO dataset[12-13] and is available in five different pre-trained model sizes: n, s, m, l, and x, catering to diverse hardware and performance requirements. The larger models, l and x, are designed to enhance accuracy while reducing the number of parameters. For the present study, the YOLO-v8-m model was employed with a batch size of eight images.

Figs.2—4 depict the modified YOLO-v8 architecture used in this work. Fig.2 extracts features through the backbone network and the multi-scale fusion of the neck, and finally outputs object detection results through the detection head. Fig.3 introduces the ODConv2d module into the backbone network to optimize feature extraction and adds multi-scale feature fusion to the neck, the detection head outputs the target detection results. In Fig.4, the in-

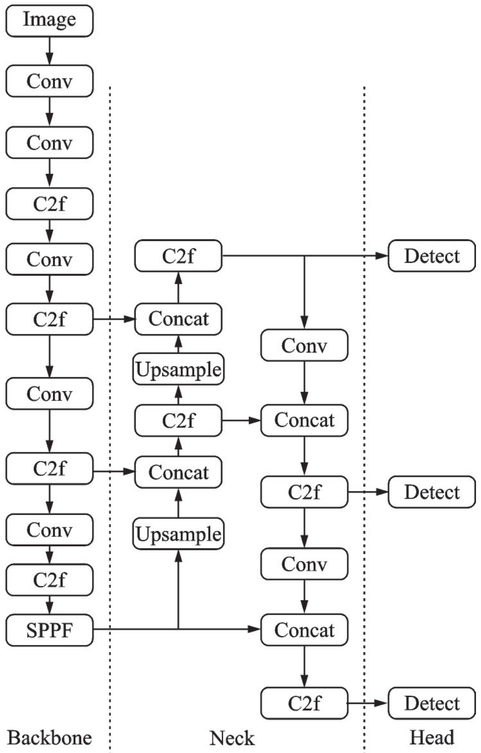


Fig.2 Original YOLO-v8 model architecture

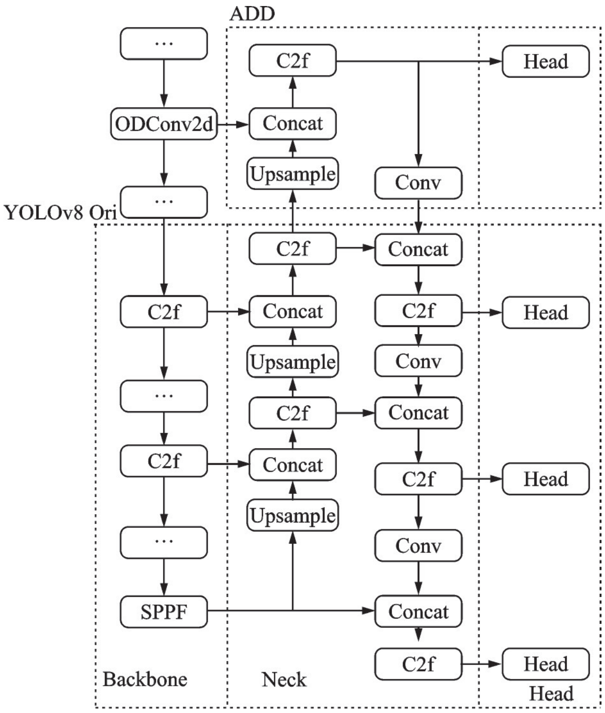


Fig.3 A comprehensive presentation of the modifications made to the YOLO-v8 algorithm

put image is subjected to feature extraction by the backbone network (ODConv2d, SPPF) and enhanced by PAN, and the detection head is combined with SIoU and FIoU to optimize target localization, ultimately outputting the annotation results. In the backbone section, we replaced a standard convolutional layer with an omni-dimensional convolutional layer (ODConv2d), which introduces multidimensional attention[6]. We then selected the YOLO-v8-p2 variant, specifically designed for small object detection. This version of the model differs from the default YOLO-v8 by adding an additional upsampling layer for feature fusion in the neck section, as well as an extra detection head in the layer. Furthermore, the loss function was tailored for each detection head to prioritize the small object detection.

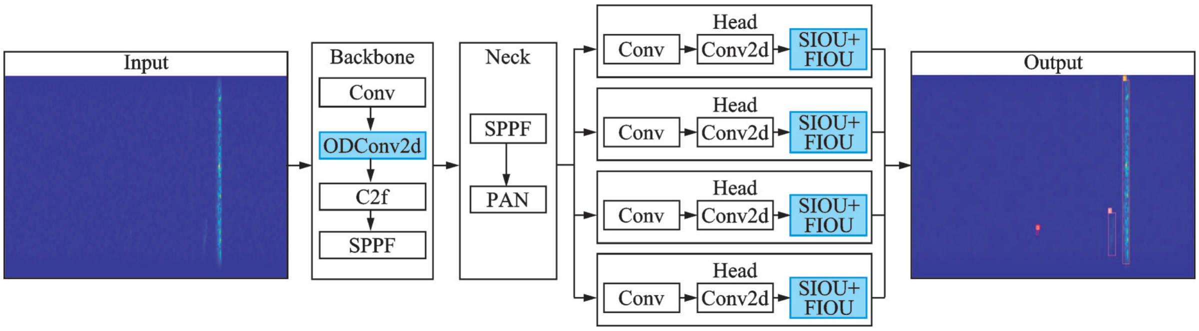


Fig.4 A simplified presentation of the modifications made to the YOLO-v8 algorithm

1.2 Synthetic aperture radar imagery

Synthetic aperture radar (SAR) is a technique that provides high-resolution imaging even under conditions where traditional optical methods are impractical, such as in low-visibility environments[14]. Owing to its flexibility, SAR has become a widely used and reliable tool across various fields, including exploration and surveillance. The combination of polarizations and frequency bands plays a critical role in SAR data processing. In theory, an increased number of polarizations enhances the ability to distinguish between different types of land features[15].

YOLO models have been proven to be effective in SAR imagery, particularly in applications such as ship detection[16]. SAR imagery is especially valuable because it captures high-resolution images regardless of weather conditions or time of day, making it ideal for maritime surveillance and other remote sensing tasks.

The dataset used in this project consists of 668 radar images, which are employed for the detection and classification of target objects. Fig.5 illustrates the original radar image. As illustrated in Fig.6, the target objects are categorized into three distinct classes: 0 (smallest), 1 (medium size), and 3 (largest). Several challenges are evident in this dataset. Firstly, the boundaries of the target objects are often difficult to discern by the human eye, which can lead to misjudgments. Secondly, the size disparity between target class 0 and class 3 is significant, posing challenges for image recognition, particularly in tasks like image segmentation. The dataset contains a total of 1 526 target objects. The images are in standard JPEG format, and the sample distribution across the categories is relatively balanced, with no significant class imbalance. To address any potential issues with class imbalance during training, the focal loss was applied.

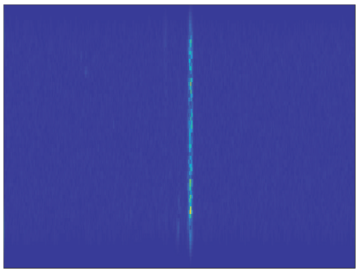


Fig.5 Original data image

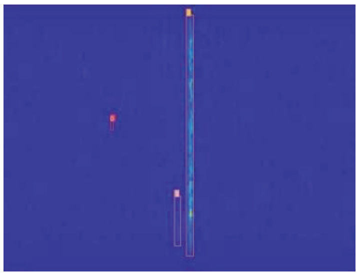


Fig.6 Original data image with target object

1.3 Setup of the enhanced YOLO algorithm

The dataset was partitioned into training, validation, and test sets, with respective ratios of 0.8∶0.1∶0.1. To evaluate the model's accuracy, the IoU metric was employed[17], calculated as

$$\mathrm{IoU}=\frac{A\cap B}{A\cup B}$$

where the numerator represents the area of overlap between the predicted bounding box (A) and the ground truth (GT) bounding box (B), while the denominator represents the area of their union. This metric provides a robust measure of the accuracy of object detection models by quantifying the extent of overlap between predicted and actual bounding boxes.

During the training process, it is crucial to establish a threshold value for IoU[18]. This threshold determines the criteria for a true detection. When the ratio of the intersection area of the predicted bounding box and the GT bounding box to their union area exceeds the threshold, it is classified as a true detection; otherwise, it is considered a false detection. In this project, the IoU threshold was set to 0.8, which is sufficiently stringent to ensure high detection quality. This setting strikes a balance between precision and recall, thereby enhancing the overall performance of the object detection model.

For the backbone network, as shown in the comparison between Fig.2 and Fig.3, we replaced the second convolutional layer with ODConv2d, introducing a multi-dimensional attention mechanism and parallel strategies to enhance the model's feature learning capabilities. ODConv2d is a dynamic convolution design that leverages multi-dimensional attention across four dimensions (channel, filter, spatial, and kernel) and incorporates squeeze-and-excitation (SE) mechanisms. By employing SE attention mechanisms across these dimensions, ODConv2d enables dynamic behavior that improves the accuracy of both lightweight and large CNNs, while maintaining efficient inference speed.

The integration of the multi-dimensional attention mechanism into YOLO-v8 aims at improving accuracy in small object detection, while preserving YOLO's efficiency and speed advantages. By introducing this mechanism, the model selectively focuses on key parts of the image, ensuring efficient feature extraction, particularly in small object detection. The attention mechanism helps to reduce information loss, enabling the model to better capture fine-grained details, which is crucial for detecting small objects.

Small objects, which are often overlooked or affected by background interference, benefit from the network's focused attention, improving detection rates and reducing distractions from irrelevant background noise. This focused attention enhances small object detection rates by preventing the model from being distracted by irrelevant background noise. The mechanism also allows the model to more effectively focus on spatial locations and important features of small objects, addressing YOLO's limitations in detecting small targets.

Moreover, small objects may overlap with larger ones in an image, which presents additional challenges for detection. Although YOLO-v8 incorporates multi-scale feature fusion, combining this with the multi-dimensional attention mechanism further strengthens the model' s ability to distribute attention across scales. This ensures that targets at different scales receive the necessary attention, preventing small objects from being overshadowed or ignored by larger ones. While the attention mechanism introduces some computational overhead, we specifically replaced the second-layer convolution operation, avoiding the increase in parameters typically seen in traditional convolutional networks. This allows for efficient integration with YOLO' s architecture, preventing over-complication and controlling the computational load to ensure that inference speed remains unaffected.

For the neck and head networks, as demonstrated in the comparison between Fig.2 and Fig.3, we employed the YOLO-v8 variant YOLO-v8-p2, which is tailored for small object detection. Unlike the default YOLO-v8 model, YOLO-v8-p2 introduces an additional upsampling layer for feature fusion in the "Neck" section, along with an extra detection head within that layer, further enhancing the model's ability to detect small objects.

For the loss function, we initially employed the SIoU, shown as

$$L_{\mathrm{SIoU}} = 1 - L_{\mathrm{IoU}} + \frac{\Delta + \Omega}{2}$$

where L_{SIoU} represents the SIoU loss, which takes into account the shape difference between the GT and predicted detection boxes, specifically their aspect ratio (width-to-height ratio); Ω the shape loss, which considers the shape difference between the GT and predicted detection boxes; Δ the distance loss, which evaluates the distance between the GT and predicted detection boxes, measured by the area of the minimum enclosing rectangle between the two boxes; L_{IoU} the IoU loss, which calculates the IoU between the two boxes, and by using 1-IoU, it emphasizes the non-overlapping part of the predicted bounding box. Finally, focal loss is applied to the SIoU loss to address class imbalance, ensuring the model focuses on harder-to-classify samples. This improves the training speed, inference accuracy, and generalization ability of the model.

SIoU extends the traditional IoU by introducing three additional loss components: Angle loss, distance loss, and shape loss. The angle loss accounts for the vector angle between the ground-truth bounding box and the predicted bounding box, addressing any directional mismatch between the two boxes. The distance loss evaluates the spatial discrepancy by focusing on the minimum circumscribed rectangular area encompassing both bounding boxes. The shape loss evaluates the shape similarity between the ground-truth bounding box and the predicted bounding box, focusing on the ratio of their lengths to widths.

Subsequently, we incorporated the standard IoU loss to enhance the stability and reliability of the model[19]. These four loss components: Angle loss, distance loss, shape loss, and standard IoU loss, were combined using a weighted averaging approach. This fusion of loss components helps ensure a more comprehensive evaluation of the model' s performance.

To address potential class imbalance issues, we further integrated the focal-IoU loss, shown as

$$L_{\mathrm{Focal\_SIoU}} = L_{\mathrm{SIoU}} + \left(\frac{A_{\mathrm{inter}}}{A_{\mathrm{union}} + \varepsilon}\right)\gamma$$

where γ is a hyperparameter in the focal loss that controls the model's focus on "hard samples" (i.e., samples with a low IoU); ε a very small constant used to prevent division by zero errors and numerical instability; A_{inter} the intersection area; and A_{union} the union area.

Focal-IoU loss specifically targets hard-to-detect objects that are often overshadowed by dominant classes. This modification enhances the model' s ability to focus on small objects, which is crucial for the task at hand.

Finally, an additional coefficient, given by $C - w \times h$, was added to the loss value calculated by the final Focal-SIoU loss function. This coefficient is designed to adjust the loss based on the width (w) and height (h) of the predicted bounding box, detailed as

$$\mathrm{Loss} = (C - w \times h) L_{\mathrm{Focal\_SIoU}}$$

where C is a positive integer greater than all $w \times h$, and $w \times h$ denotes the area of the bounding box. Consequently, the larger the area of the bounding box is, the smaller the loss is, and the smaller the area of the bounding box is, the larger the loss is. This adjustment encourages the model to focus more on objects with smaller areas, thereby improving detection accuracy for small objects.

This approach is designed to direct the model's attention toward challenging-to-distinguish small objects, while maintaining a balance between training speed, accuracy, and model generalization. By emphasizing small objects, the model becomes more effective when detecting them, which is crucial for tasks like small object detection in radar images.

1.4 Setup of the VGG-16 algorithm

VGG-16 is a deep convolutional neural network consisting of 16 layers: 13 convolutional layers, five max-pooling layers, and three fully connected layers. It consistently uses 3×3 convolution filters across the network[10]. Studies indicate that deeper networks, like VGG-16, achieve higher accuracy than shallower architectures. VGG-16 was originally trained on the ImageNet dataset, which contains over 1.2 million images categorized into 1 000 classes[20]. The representations learned by VGG-16 exhibit strong generalizability, allowing it to achieve state-of-the-art performance in image recognition tasks. Due to its ability to extract hierarchical features, VGG-16 is often used as a backbone network for object detection models[21]. The convolutional layers of VGG-16 capture various levels of abstraction, which are crucial for detecting objects of different sizes and shapes.

While VGG-16 provides strong baseline performance, it has been surpassed by more recent architectures, such as ResNet and EfficientNet, in terms of both accuracy and efficiency[22]. Nevertheless, VGG-16 remains a popular choice due to its simplicity and ease of implementation.

For this project, we employed the same train-test-validation dataset split ratios to train a standard VGG-16 CNN. To adapt the model for object detection, its output was modified to predict bounding boxes, transforming the task from classification to regression. During the backpropagation step, we used the root mean squared error (RMSE) in combination with the median absolute error (MedAE) as the loss function. Upon completion of the training, the IoU metric was employed to validate the predicted bounding boxes against the ground-truth bounding boxes in the validation set.

1.5 Image segmentation algorithm for radar images

Image segmentation is a subfield of computer vision that involves dividing an image into multiple regions or segments, where each segment corresponds to an object or a part of an object[23]. The primary objective is to identify object-specific pixels by analyzing similarities among neighboring pixels and regions. Such algorithms are often utilized for object detection due to their ability to directly extract features from images.

A classic and enduring algorithm in this field is the gray-level co-occurrence matrix (GLCM), which calculates the frequency of specific pixel value pairs occurring at defined distances and directions[24]. By analyzing pixel pairs, GLCM extracts crucial texture information, making it a robust tool for texture analysis. With its capability to compute across multiple spatial directions, GLCM offers flexibility and enables comprehensive texture feature extraction.

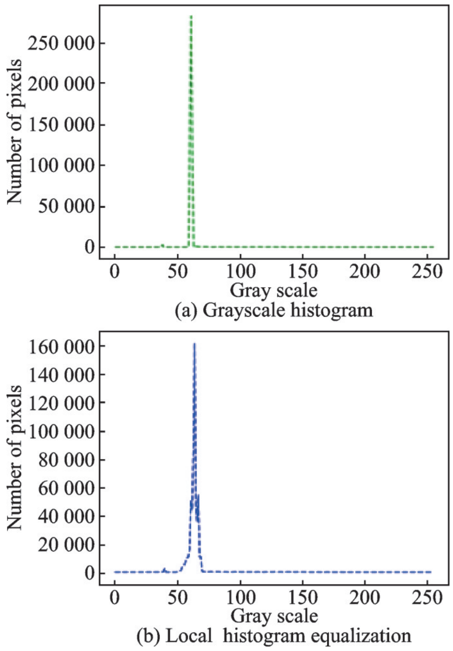
GLCM can be computed in various spatial directions, offering significant flexibility in capturing different types of texture information. The features derived from the GLCM include contrast, correlation, energy, and homogeneity.

In this work, we utilized the GLCM algorithm to segment radar images, effectively distinguishing up to three objects from the background and noise. First, we converted the radar images to true grayscale, ensuring uniformity in pixel intensity values for consistent analysis. Subsequently, a range of threshold values was used to segment and isolate the objects of interest. The GLCM algorithm is especially effective for radar image segmentation because it captures essential spatial relationships between pixels, a key factor for accurate texture analysis[25].

2 Results

2.1 Image preprocessing

Fig.7 presents the histogram visualizations of the dataset both before and after the application of the local histogram equalization technique, which enhance the texture of the relevant objects in the images.



(a) Grayscale histogram

(b) Local histogram equalization

Fig.7 Pixel intensity distribution before and after local histogram equalization technique

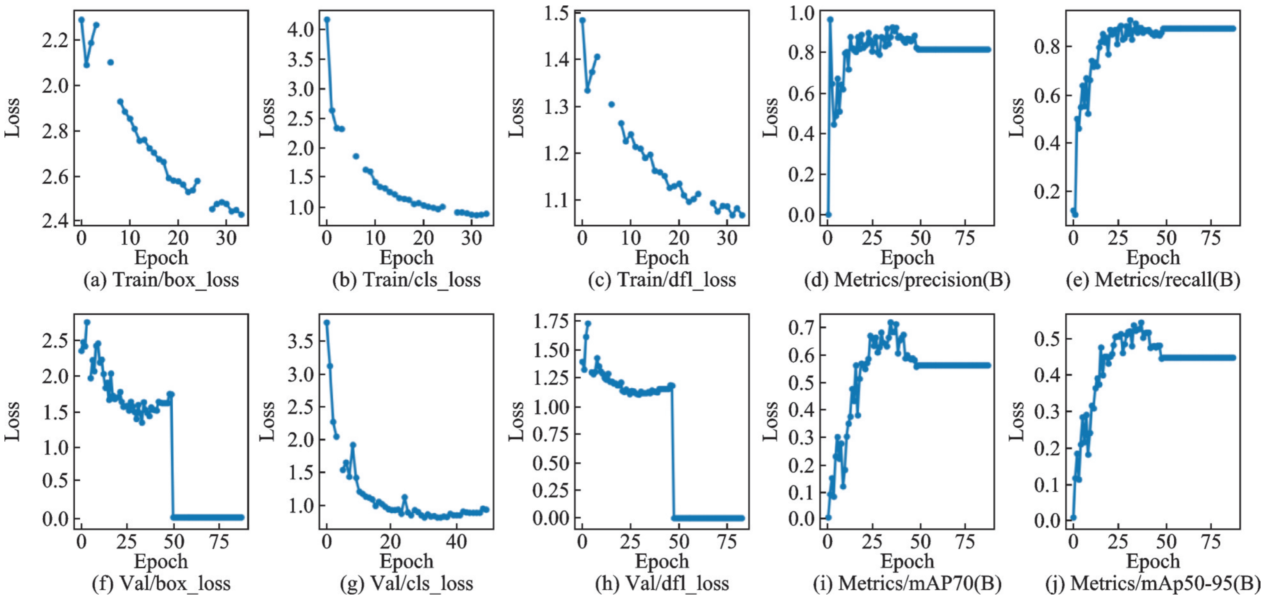
2.2 Ablation study

We conducted ablation experiments using the default YOLO-v11 model. Table 1 provides a comparative analysis of the performance of three distinct models in the context of small object detection within radar imagery, utilizing mean average precision (mAP70) as the benchmark metric. This metric, which stands for mAP at 70% IoU, is a pivotal standard for assessing the efficacy of object detection models, particularly in scenarios involving the identification of smaller targets. As shown in Table 1, the enhanced YOLO-v8 model demonstrates superior performance in detecting small objects in radar images compared to both the default YOLO-v11 model and the YOLO-v8 model with complete IoU (CIoU). Specifically, the enhanced YOLO-v8 model achieved an increase of 0.144 in mAP70 accuracy for small objects and an improvement of 0.095 in mAP70 accuracy across all classes, highlighting the significant impact of SIoU in small object detection.

Fig.8 illustrates the training stage metrics for the YOLO-v8 model with CIoU. During training, it is evident that the loss curve for YOLO-v8 with CI-

Table 1 mAP70 comparison between the default YOLO-v11 model and the enhanced YOLO-v8 model on preprocessed data

Model	mAP70	
	Smaller object	All classes
Default YOLO-v11	0.362	0.770
YOLO-v8 with CIoU	0.310	0.706
Enhanced YOLO-v8	0.454	0.801



(a) Train/box_loss (b) Train/cls_loss (c) Train/dfl_loss (d) Metrics/precision(B) (e) Metrics/recall(B)

(f) Val/box_loss (g) Val/cls_loss (h) Val/dfl_loss (i) Metrics/mAP70(B) (j) Metrics/mAp50-95(B)

Fig.8 YOLO-v8 with CIoU training stage metrics

oU fluctuates significantly, indicating potential issues such as gradient explosion. This instability could impede model convergence. In contrast, the enhanced YOLO-v8 model exhibits a much smoother and more stable loss reduction during training, suggesting that it effectively controls gradient updates and mitigates the risk of gradient explosion. This enhanced stability accelerates convergence, enabling the model to achieve optimal performance more efficiently.

Figs.9 and 10 compare the prediction results of both models on the same dataset. Fig.9 shows the confusion matrix displaying the predictions made by the model across the entire radar image dataset. YOLO-v8 with CIoU exhibits multiple misdetections due to overlapping bounding boxes, where several targets are incorrectly detected as a single object. In contrast, the enhanced YOLO-v8 model shows almost no such misdetections. This improvement highlights the effectiveness of SIoU, which reduces overlapping bounding boxes and misdetections compared to CIoU, thus enhancing both the accuracy and robustness of the model. By optimizing the bounding box regression process, SIoU enables the model to better handle object boundaries and positions, improving detection performance, especially in challenging scenarios where stability and reliability are critical.

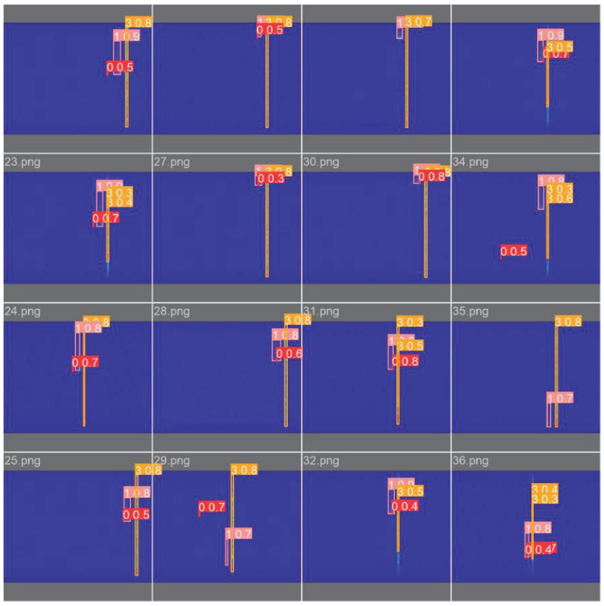


Fig.9 YOLO-v8 with CIoU prediction results

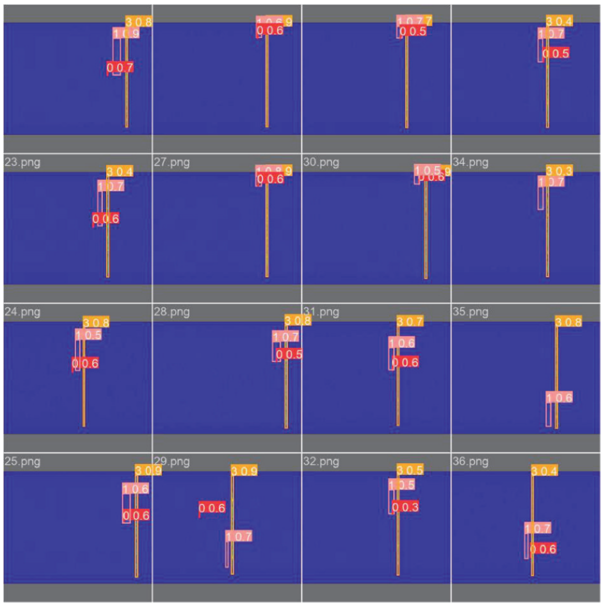


Fig.10 Enhanced YOLO-v8 prediction results

When compared to the default YOLO-v11 model, the enhanced YOLO-v8 model achieved a 0.092 increase in mAP70 accuracy for small objects and a 0.031 increase in mAP70 accuracy across all classes. In terms of model inference speed, the enhanced YOLO-v8 model is 2.1 ms slower per image than the default YOLO-v11 model, which we consider to be within an acceptable range.

Additionally, ablation studies were conducted to evaluate the impact of local histogram equalization, a data preprocessing technique discussed in Section 2.1. The datasets, with and without local histogram equalization, were compared using object segmentation tasks performed with the enhanced YOLO-v8 model. The results, shown in Table 2, the mAP performance of the enhanced YOLO-v8 model was compared when using and not using local histogram equalization, where a more stringent range from 50% to 95% (mAP50-95) is used for validation. The results showed that equalization im-

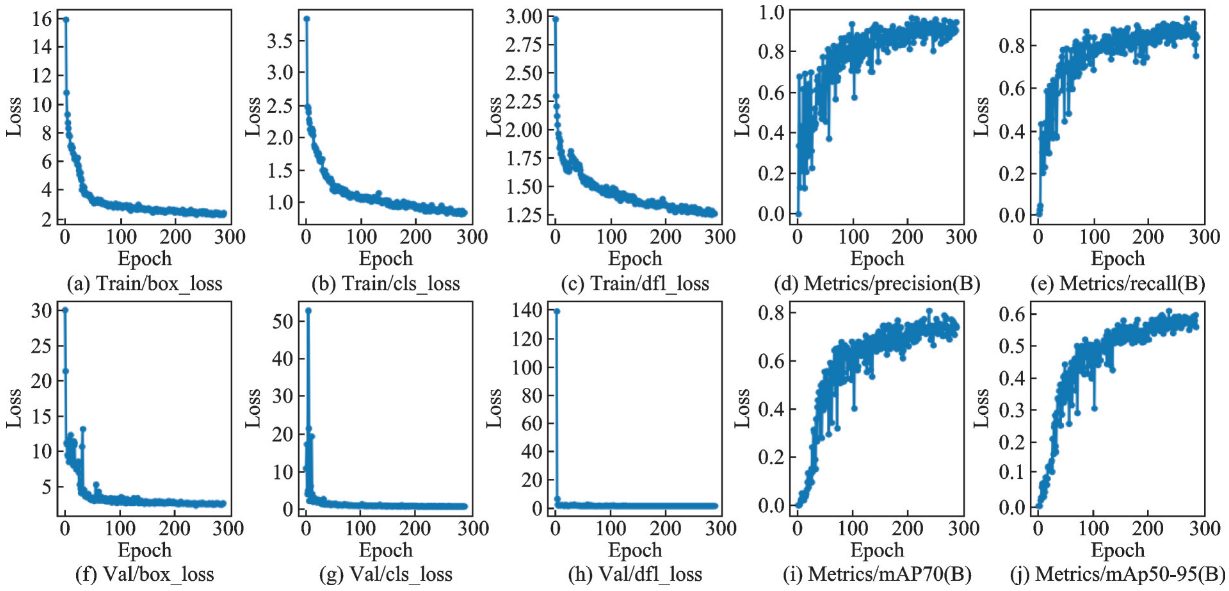
Table 2 mAP comparison of the enhanced YOLO-v8 model on preprocessed data with and without local histogram equalization

Preprocessed data	mAP50-95			
	Smaller object	Medium-size object	Larger object	All categories
Original	0.298	0.592	0.766	0.552
Use local histogram equalization	0.381	0.647	0.795	0.607

proved the mAP for small, medium, large, and all categories, increasing the overall mAP from 0.552 to 0.607, demonstrating a significant improvement in the mAP metric across all target sizes. This highlights the critical role of local histogram equalization in radar image preprocessing. By enhancing local contrast, this technique reveals more details in previously dark or uniformly bright areas, helping the model recognize object edges and textures more effectively. Moreover, it reduces the impact of low-frequency noise, allowing the model to focus on salient features during training and minimizing background interference.

2.3 Training and evaluation of the enhanced YOLO algorithm

After 289 epochs, the training process was stopped, and the best performance metrics were achieved at epoch 239. The model weights from this epoch were saved and used to make predictions on the validation dataset. Fig.11 summarizes the training process for both the training and validation datasets.



(a) Train/box_loss (b) Train/cls_loss (c) Train/dfl_loss (d) Metrics/precision(B) (e) Metrics/recall(B)

(f) Val/box_loss (g) Val/cls_loss (h) Val/dfl_loss (i) Metrics/mAP70(B) (j) Metrics/mAp50-95(B)

Fig.11 YOLO-v8 with CIoU training and validation stage metrics

The bounding box loss measures how well the predicted bounding boxes match the GT bounding boxes, while the classification loss evaluates the accuracy of class predictions for each detected object. The distribution focal loss, used for bounding box regression, is designed to improve the precision of predicted bounding boxes. Other key evaluation metrics include precision: The ratio of true positive predictions to the total number of positive predictions, and recall: The ratio of true positive predictions to the total number of actual positive instances.

Upon completion of the training process, we proceeded to the testing phase using a dataset consisting of 74 images. The results of the testing phase are presented in Table 3, which includes detailed metrics for the three object classes in the radar images.

Table 3 Enhanced YOLO-v8 testing stage metrics

Class	Instance	Precision	Recall	mAP70	mAP50-95
0	56	0.842	0.804	0.454	0.381
1	48	0.919	0.979	0.973	0.647
3	65	0.945	0.846	0.975	0.795
Weighted average	—	0.902	0.876	0.801	0.607

Additional results from the training and evaluation of the YOLO-v8 model are provided below. Fig.12 shows the confusion matrix displaying the predictions made by the model across the entire radar image dataset. In Fig.12, the last column represents the background. It refers to instances where non-background objects are incorrectly predicted as background, leading to missed detections of non-background objects. The last row represents the background as well. It refers to cases where background areas are incorrectly predicted as non-back-

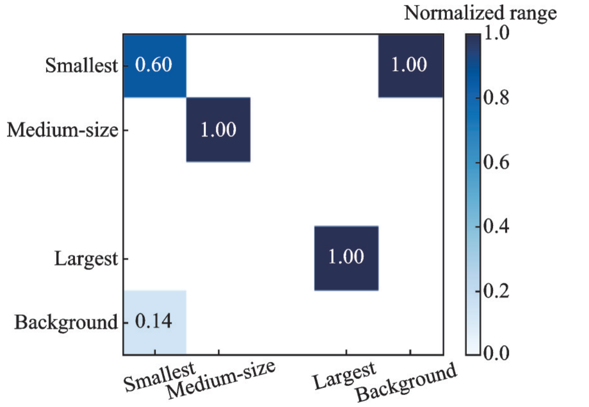


Fig.12 Confusion matrix from predictions made by the enhanced YOLO-v8 model in the evaluation stage

ground, resulting in false detections of objects that do not actually exist.And the progress bar is the normalized range in columns, whose values in the grid are the ratios of the quantity to the total number of a column. Fig.13 shows F_1-confidence curves, which illustrate how the F_1-score varies with different confidence thresholds. The legend "All classes 0.95 at 0.496" indicates that when the confidence threshold is 0.496, the average F_1-score across all classes reaches its maximum value of 0.95. Fig.14 shows recall-confidence curves, demonstrating how recall changes with different confidence thresholds. The legend "All classes 1.00 at 0.000" indicates that when the confidence threshold is 0.000, the average recall across all classes reaches its maximum value of 1.00. Fig.15 shows precision-confidence curves, depicting how precision changes with varying confidence thresholds. The legend "All classes 1.00 at 0.545" indicates that when the confidence threshold is 0.545, the average precision across all classes reaches its maximum value of 1.00. Fig.16 shows

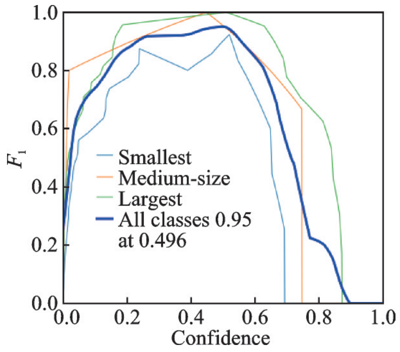


Fig.13 F_1-confidence curves for the enhanced YOLO-v8 in the training stage

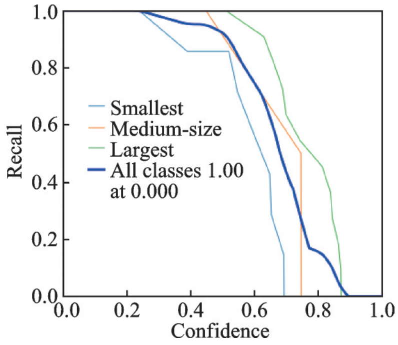


Fig.14 Recall-confidence curves for the enhanced YOLO-v8 in the training stage

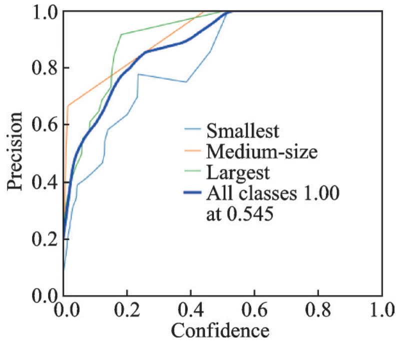


Fig.15 Precision-confidence curves for the enhanced YOLO-v8 in the training stage

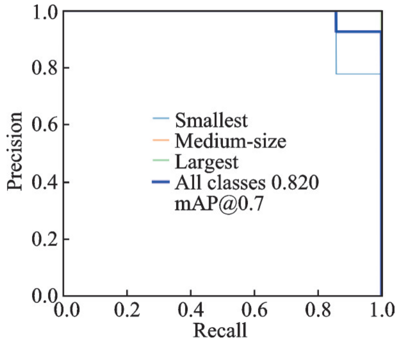


Fig.16 Precision-recall curves for the enhanced YOLO-v8 in the training stage

precision-recall curves, illustrating the trade-off between precision and recall at varying confidence thresholds. The legend "All classes 0.820 mAP@0.7" indicates that when mAP is set to 0.7, the average precision-recall value across all classes reaches its maximum value of 0.820.

2.4 Training and evaluation of the VGG-16 algorithm

The training and validation losses for the VGG-16 model are shown in Fig.17. The model's loss function was specifically designed to predict bound-

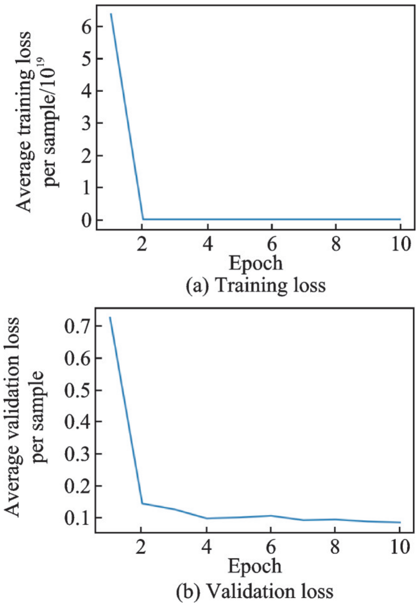


Fig.17 Train and validation loss for the VGG-16 in the training stage

ing boxes through a regression task. For evaluation, we set a threshold based on the IoU between the ground-truth and predicted bounding boxes, with a value of 80% (IoU = 0.8). This threshold allowed us to classify a detection as positive when the IoU exceeded 0.8.

By applying this threshold, we effectively converted the regression task into a classification problem, enabling the construction of a confusion matrix-like graph to better visualize the results. Fig.18 illustrates the VGG-16 model's predictions for objects labeled as 0, 1, and 3, showing only the count of true positives and true negatives under the 80% IoU threshold.

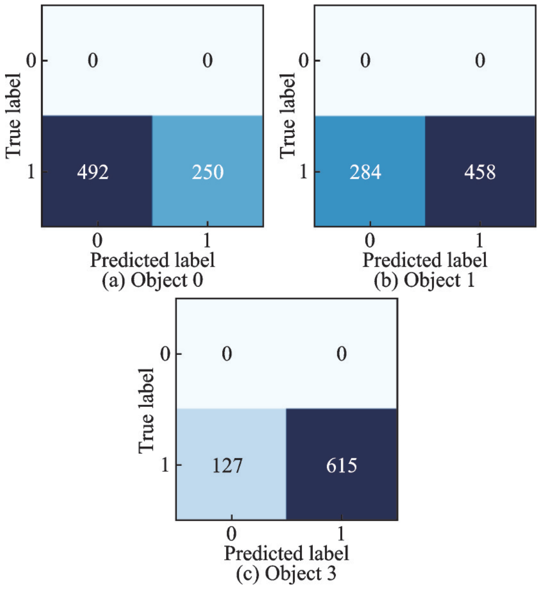


Fig.18 True positives and true negatives from predictions made by the VGG-16 in the evaluation stage

2.5 Image segmentation for radar images

We applied GLCM to a subset of randomly selected images from the dataset. For each image, we manually adjusted the gray threshold value to optimize object detection. Additionally, we determined an average optimal gray threshold value of 93, which was then used for batch processing to segment the entire dataset. This approach generally produced satisfactory results. However, the accuracy was lower when handling small objects and images with noisy backgrounds.

Fig.19 shows three randomly selected image segmentation results as examples. The top row displays the original images, while the bottom row shows the corresponding segmentation results obtained using the GLCM algorithm.

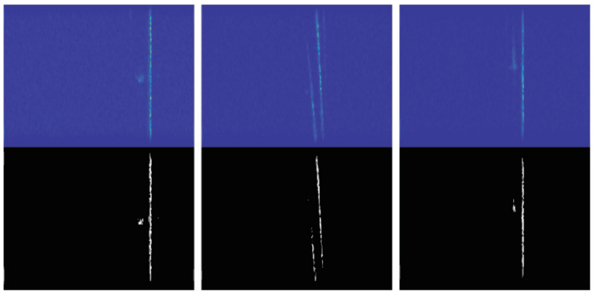


Fig.19 Original and segmented images using the GLCM algorithm

3 Discussion

During the training phase of the enhanced YOLO-v8 model, we observed that after 200 epochs, the improvement in model accuracy became less significant compared to earlier stages. This trend is evident in the metrics shown in Fig.5, indicating that even the most advanced models today cannot achieve perfect solutions for object detection tasks. Nonetheless, the accuracy attained is impressive. As shown in Fig.6, the model successfully detected backgrounds, medium-sized objects, and large objects in radar imagery. However, the primary focus of this study was on the detection of small-sized objects, where the model achieved over 84% weighted accuracy, as detailed in Table 1.

Numerous studies emphasize the challenges posed by speckle noise in radar images for object detection[26-27]. Noise is a well-known issue in computer vision, but effective solutions remain limited. Our model successfully identified all objects except for the smallest ones, which were frequently mistaken for noise. This gives us confidence that future advancements in techniques and algorithms will bring us closer to achieving perfect detection in such tasks.

In evaluating the VGG-16 model, we tested various threshold values for the IoU metric. The results, presented using a stringent threshold value of 0.8, showed that the VGG-16 model performed poorly at this threshold, with an average small-object detection accuracy of 33.7%. The model performed better for medium-sized objects, achieving 61.9% accuracy, and much better for large objects, with an accuracy of 82.8%.

By lowering the threshold value to 0.6, the VGG-16 model' s overall accuracy per class improved to acceptable ranges. Furthermore, using a threshold of 0.4 resulted in near-perfect accuracy across all classes. This suggests that while the VGG-16 model can accurately predict the presence or absence of objects and provide a reasonable bounding box approximation, IoU is a stringent metric that ensures state-of-the-art performance in object detection[28].

After applying the GLCM algorithm to the radar image dataset, we found that image segmentation provided a simple and effective solution to the problem. Image segmentation has been shown to be a valuable technique for SAR imagery, as demonstrated in empirical classification tasks[29]. Other segmentation methods, such as semantic segmentation, have also proven feasible for automation in similar projects[30]. Since image segmentation is not a machine learning technique, its implementation presents a much easier learning curve[31]. As a result, professionals from various fields can utilize image segmentation algorithms without requiring the expertise needed for complex deep learning models like YOLO.

4 Conclusions

This study improved small object detection in radar imagery by customizing YOLO-v8 with algorithm and data modifications. Local histogram equalization enhanced contrast, while a multi-dimensional attention mechanism increased precision in noisy environments. An upsampling layer fused high-level and low-level information, improving performance in complex scenarios.

Two loss functions, FIoU and SIoU, were compared to optimize the model. FIoU addressed class imbalance, while SIoU improved bounding box alignment and convergence speed. Combining the two was most effective for the noisy radar dataset.

YOLO-v8 outperformed VGG-16 in small object detection, showing better precision and robustness. It also surpassed the GLCM algorithm in accuracy, scalability, and real-time analysis, making it more reliable for small object detection in noisy radar images.

References

[1] HELLER M, PETROV N, YAROVOY A. A novel approach to vehicle pose estimation using automotive radar[EB/OL]. (2021-07-20)[2024-06-30]. https://arxiv.org/abs/2107.09607v1.

[2] SUN Y, SUN Z, CHEN W. The evolution of object detection methods[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108458.

[3] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2020.

[4] LI Z, DONG Y, SHEN L, et al. Development and challenges of object detection: A survey[J]. Neurocomputing, 2024, 598: 128102.

[5] LAI Y R, TSAI P C, YAO C Y, et al. Improved local histogram equalization with gradient-based weighting process for edge preservation[J]. Multimedia Tools and Applications, 2017, 76(1): 1585-1613.

[6] LI C, ZHOU A J, YAO A B. Omni-dimensional dynamic convolution[EB/OL]. (2022-09-16). http://arxiv.org/abs/2209.07947.

[7] SWATHI Y, CHALLA M. YOLOv8: Advancements and innovations in object detection[C]//Pro-

ceedings of International Conference on Smart Computing and Communication. Singapore: Springer, 2024: 1-13.

[8] ZHANG H, ZHANG S. Focaler-IoU: More focused intersection over union loss[EB/OL]. (2024-01-19)[2024-06-30]. https://arxiv.org/abs/2401.10525.

[9] YU Y, WANG C, FU Q, et al. Techniques and challenges of image segmentation: A review[J]. Electronics, 2023, 12(5): 1199.

[10] SIMONYAN K. Very deep convolutional networks for large-scale image recognition[EB/OL].(2015-04-10)[2024-06-30]. https://arxiv.org/abs/1409.1556.

[11] WAN D, LU R, HU B, et al. YOLO-MIF: Improved YOLOv8 with multi-information fusion for object detection in gray-scale images[J]. Advanced Engineering Informatics, 2024, 62: 102709.

[12] TERVEN J, CORDOVA-ESPARZA D. A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS[J]. Mach Learn Knowl Extr, 2023, 5: 1680-1716.

[13] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//Proceedings of Computer Vision—ECCV 2014: 13th European Conference. Zurich, Switzerland: Springer International Publishing, 2014: 740-755.

[14] BARAHA S, SAHOO A K. Synthetic aperture radar image and its despeckling using variational methods: A review of recent trends[J]. Signal Processing, 2023, 212: 109156.

[15] NOVRESIANDI D A, SETIYOKO A, ARIEF R. Synthetic aperture radar (SAR) data applications for tropical peatlands monitoring activities: An overview[J]. Remote Sensing Applications: Society and Environment, 2023, 29: 100893.

[16] PASSAH A, SUR S N, ABRAHAM A, et al. Synthetic aperture radar image analysis based on deep learning: A review of a decade of research[J]. Engineering Applications of Artificial Intelligence, 2023, 123: 106305.

[17] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 658-666.

[18] ZOU H, WANG Z. An enhanced object de tection network for ship target detection in sar images[J]. The Journal of Supercomputing, 2024, 80(12): 17377-17399.

[19] WU K, ZHANG Z, CHEN Z, et al. Object-enhanced YOLO networks for synthetic aperture radar ship detection[J]. Remote Sensing, 2024, 16(6): 1001.

[20] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115: 211-252.

[21] SU J, WANG H. Fine-tuning and efficient VGG16 transfer learning fault diagnosis method for rolling bearing[C]//Proceedings of IncoME-Ⅵ and TEPEN 2021: Performance Engineering and Maintenance Engineering. Cham: Springer International Publishing, 2022: 453-461.

[22] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2019: 6105-6114.

[23] MINAEE S, BOYKOV Y, PORIKLI F, et al. Image segmentation using deep learning: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(7): 3523-3542.

[24] SEBASTIAN V B, UNNIKRISHNAN A, BALAKRISHNAN K. Gray level co-occurrence matrices: Generalisation and some new features[J]. (2012-05-22)[2024-06-30]. https://arxiv.org/abs/1205.4831.

[25] JAMES J, HEDDALLIKAR A, CHOUDHARI P, et al. Analysis of features in SAR imagery using GLCM segmentation algorithm[M]//Data Science: Theory, Algorithms, and Applications. [S.l.]: Springer, 2021: 253-266.

[26] DEVAPAL D, HASHNA N, APARNA V, et al. Object detection from SAR images based on curvelet despeckling[J]. Materials Today: Proceedings, 2019(11): 1102-1116.

[27] RAJ N, SETHUNADH R, APARNA P. Object detection in SAR image based on Bandlet transform[J]. Journal of Visual Communication and Image Representation, 2016(40): 376-383.

[28] RADKE K L, KORS M, MULLER-LUTZ A, et al. Adaptive IoU thresholding for improving small object detection: A proof-of-concept study of hand erosions classification of patients with rheumatic arthritis on X-ray images[J]. Diagnostics, 2022,13(1): 104.

[29] WANG Z, WANG Z, QIU X, et al. Global polarimetric synthetic aperture radar image segmentation with data augmentation and hybrid architecture model[J]. Remote Sensing, 2024, 16(2): 380.

[30] ZHENG N R, YANG Z A, YANG H, et al. Semi-automatic annotation method for semantic segmentation of synthetic aperture radar images[C]//Proceedings of China High Resolution Earth Observation Conference. [S.l.]: Springer, 2022: 95-101.

[31] SHUKLA A K, DWIVEDI S K, CHANDRA G, et al. Deep learning-based suppression of speckle-noise in synthetic aperture radar (SAR) images: A comprehensive review[C]//Proceedings of the International Conference on Cognitive and Intelligent Computing. [S.l.]: Springer, 2023.

Acknowledgements This work was supported by the National Natural Science Foundation of China Joint Fund (No.U21B2028), the National Key R&D Program of China (No. 2021YFC 2100100), and the Shanghai Science and Technology Project (Nos.21JC1403400, 23JC1402300).

Author Prof. **JIANG Zhenyu** received his Ph. D. degree from National University of Defense Technology in 2007 and currently serves as a professor in College of Aerospace Science at National University of Defense Technology. Over the years, he has led numerous projects and made significant contributions to the field of aerospace technology. His current research interests focus on advanced methods in aircraft target recognition and online identification, emphasizing the integration of cutting-edge technologies such as artificial intelligence, machine learning, and real-time data processing to enhance system performance and reliability.

Author contributions Prof. **JIANG Zhenyu** described the application requirements, wrote a portion of manuscripts and provided the funds to support the research. Prof. **LI Jinjin** provided the datasets, designed the project frame work, supervised the project development, and provided the funds and hardware to support the research. Mr. **LI Xiaodong** analyzed, interpreted, and annotated the results. Researcher **DU Chen** performed the YOLO-v8 customization and training, also analyzed, interpreted, and annotated the process and results. Mr. **CHEN An** helped with the data curation, visualization and proofreading of the manuscript. Dr. **HAN Yanqiang** helped with the data curation, visualization and proofreading of the manuscript. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: ZHANG Huangqun)

基于多维注意和上采样融合的YOLO-v8雷达图像空中小目标检测

江振宇[1]，李晓东[1]，杜 晨[2,3]，陈 安[2,3]，韩彦强[2,3]，李金金[2,3]

(1.国防科技大学空天科学学院，长沙410073，中国；2.上海交通大学先进微纳制造技术国家重点实验室，上海200240，中国；3.上海交通大学电子信息与电气工程学院微纳电子系，上海200240，中国)

摘要：提出了一种提高YOLO-v8雷达图像中小目标检测性能的创新方法。首先，对原始图像应用局部直方图均衡化技术，显著增强对比度和细节表示。然后，通过结合基于多维注意力机制和并行处理策略的卷积核增强YOLO-v8骨干网络实现更有效的特征信息融合。在模型上部添加了一个上采样层，与浅层网络输出融合，设计一个专门为小物体检测量身定制的检测头，从而进一步提高了精度。此外，对损失函数进行了修改，将局部交并比(Intersection over union, IoU)与尺度IoU结合使用，从而提高了模型的性能。引入了加权策略，有效地提高了小目标的检测精度。实验结果表明，定制模型在各种评估指标上优于传统方法，包括召回率、精确度、F_1评分和受试者特征(Receiver operating characteristic, ROC)曲线，验证了其在雷达图像中小目标检测方面的有效性和创新性。结果表明，与图像分割和标准卷积神经网络等传统方法相比，所提方法准确性有了显著提高。

关键词：YOLO；雷达图像；目标检测；机器学习