IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 May 2024, accepted 11 June 2024, date of publication 14 June 2024, date of current version 21 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3414426

RESEARCH ARTICLE

YOLOX-CA: A Remote Sensing Object Detection Model Based on Contextual Feature Enhancement and Attention Mechanism

CHAO WU AND ZHIYONG ZENG
College of Computer and Cyber Security, Fujian Normal University, Fuzhou, Fujian 350117, China
Digit Fujian Internet-of-Things Laboratory of Environmental Monitoring, Fujian Normal University, Fuzhou, Fujian 350117, China

Corresponding author: Zhiyong Zeng (zzyong@fjnu.edu.cn)

This work was supported in part by the Natural Science Foundation of Fujian Province under Grant 2022J01187 and in part by the Guiding Project of Fujian Science and Technology Plan under Grant 2021Y0011.

ABSTRACT Compared to natural images, remote sensing images have the characteristics of high spatial resolution, large target scale variation, dense target distribution, and complex background. Consequently, there are challenges with insufficient detection accuracy and the inability to identify target locations accurately. Therefore, this paper introduces the YOLOX-CA algorithm, based on the YOLOX model, to address these challenges in remote sensing object detection. Firstly, the YOLOX-CA algorithm optimizes the feature extraction network of the YOLOX model. This optimization employs large-kernel depthwise separable convolution in the backbone network to enhance feature extraction capabilities, comprehensively and accurately capturing information features. Secondly, the ACmix attention mechanism is introduced into the backbone network to identify crucial features, enhance feature extraction capability, and expedite network convergence. Lastly, a Contextual Feature Enhancement (CFE) module is constructed and employed in the upsampling process of feature fusion, aiming to augment the model's awareness of context. Experimental results on the large-scale DIOR dataset for remote sensing object detection demonstrate performance enhancements over the baseline model, with increases of 2.7% in mAP, 1.1% in mAP@0.5, and 2.2% in Recall. The findings from the test dataset suggest that the proposed YOLOX-CA method is applicable and practical for remote sensing object detection, improving detection accuracy while mitigating instances of target omission.

INDEX TERMS Remote sensing images, object detection, attention mechanism, feature enhancement, YOLOX.

I. INTRODUCTION

Remote sensing image target detection refers to identifying and localizing points of interest in images captured from high altitudes or space using remote sensing technology, such as buildings, vehicles, aircraft, and ships. This research topic has significant practical value and challenges involving multiple fields, including computer vision, image processing, machine learning, and deep learning. It is widely applied in civilian and military domains, including land use [1], urban planning [2], disaster monitoring [3], and military reconnaissance [4]. With the development of high-resolution remote sensing satellites, large-scale, high-quality remote sensing image data are continuously emerging, presenting significant opportunities and challenges for remote sensing target detection tasks. However, remote sensing images are typically captured from high altitudes, resulting in smaller target sizes and susceptibility to various factors such as weather conditions, lighting, sea state, sensor parameters, etc. Additionally, in remote sensing images, targets like airplanes,

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif.

cars, etc., are often densely arranged, making it difficult to separate them from the surrounding background, leading to more challenging feature extraction and lower detection accuracy.

Traditional object detection methods generally involve three steps: region proposal, feature extraction, and classifier classification. Firstly, potential regions containing objects are selected from the image using methods like sliding windows or selective search. Secondly, feature vectors are extracted for each candidate region. Finally, a trained classifier is used to classify each candidate region and determine whether it contains an object, providing the object category. However, traditional methods suffer from high computational complexity and low efficiency. The rapid development of artificial intelligence technology has led to its widespread adoption in industries such as manufacturing, medicine [5], and biology [6]. In particular, the advancement of deep learning technology has revolutionized various image analysis tasks, with the most representative being Convolutional Neural Networks (CNNs), it has been widely applied in various scenarios, including image recognition, speech recognition, and natural language processing. In the field of object detection, deep learning has become a mainstream method.

Currently, mainstream target detection algorithms can be broadly classified into two categories: typical two-stage algorithms include R-CNN [7], Fast R-CNN [8], Faster R-CNN [9] and Mask R-CNN [10]. Compared to one-stage target detection methods, two-stage methods generally achieve higher detection accuracy but are slower, and they may lose spatial information about the overall scene or context of objects within the image. Typical one-stage algorithms include the single-shot multibox detector(SSD) [11], the YOLO (You Only Look Once) [12] series, etc. One-stage algorithms exhibit average accuracy but have a faster detection speed.

Scholars have extensively researched object detection for remote sensing images. Li et al. [13] proposed an adaptive attention mechanism to enhance the interaction of features at different scales within the Feature Pyramid Network [14] (FPN), aiming to improve the detection performance of small and dense targets in remote sensing images. However, the structure of the adaptive attention mechanism is overly complex, and its parameter size is enormous, leading to a decrease in detection speed. Zhang et al. [15] introduced shallow information with an attention mechanism before the feature fusion in YOLOv3 [16]. This was done to reduce background interference and enhance the network's representational capacity. Cheng et al. [17] proposed a multi-feature fusion and attention network based on YOLOX [18]. This approach involves fusing multiple branch convolutions and attention mechanisms to enhance feature extraction for objects of different sizes. Li et al. [19] proposed enhancing the expressive power of the network model by modifying the BottleNeckCSP structure. Liu et al. [20] proposed the YOLO-extract method based on the YOLOv5 [21] approach. By incorporating an ensemble of dilated convolutional structures, they enhanced the model's capability to extract features and positional information of objects at different scales. This resulted in reduced computational complexity and accelerated convergence speed. In summary, deep learning methods have shown great value in the field of object detection in remote sensing images and have achieved significant advancements. However, despite continuous improvements and enhancements in algorithms, the unique challenges posed by remote sensing images have not been fully addressed, further research is still needed to improve the detection accuracy.

To further enhance the accuracy of object detection in remote sensing images and address the performance degradation caused by large variations in target scales, dense objects, and complex backgrounds, this paper proposes a novel remote sensing image object detection algorithm based on the YOLOX framework. The algorithm aims to effectively tackle the challenges of object detection in remote sensing images, particularly by providing effective solutions to detect targets with different scales. The significant contributions of this paper are summarized as follows:

(1) We propose a novel YOLOX-CA algorithm for remote sensing image object detection. We have improved the basic building blocks of the backbone network, leading to a significant increase in the effective receptive field. This improvement strengthens the algorithm's ability to extract features from remote sensing images.

(2) By introducing the novel ACmix [22] mixed attention module, we enhance the network's sensitivity to small object detection, thereby improving the accuracy of small object detection.

(3) We introduce the Contextual Feature Enrichment Module, which innovatively enhances the network's perception of targets at different scales. This module enables the model to capture critical information in remote sensing images more accurately.

(4) We conduct ablation and comparison experiments on the DIOR dataset to evaluate the YOLOX-CA algorithm. Compared to existing algorithms, the YOLOX-CA algorithm significantly improves detection accuracy.

The rest of this paper is organized as follows: Section II introduces attention mechanisms and multiscale feature fusion. In Section III, the YOLOX-CA algorithm is described in detail. Section IV presents experimental datasets, evaluation metrics, parameter settings, and the results of ablation experiments and comparison experiments. The proposed algorithm is summarized, and future work is looked forward to in Section V.

II. RELATED WORK

A. ATTENTION MECHANISM

Attention mechanisms have been proven effective in enhancing network performance, and in recent years, they have received significant attention from researchers. The Convolutional Block Attention Module (CBAM) [23] improves object detection performance by calculating channel and spatial attention at different scales. The Squeeze-and-Excitation Network (SENet) [24] optimizes feature maps by learning the importance of each channel. The Global Attention Mechanism (GAM) [25] enhances deep neural network performance by reducing information diffusion and strengthening global interactions. The Coordinate Attention (CA) [26] mechanism embeds position information into channel attention, allowing lightweight networks to perform attention over more significant regions. Wu et al. [27] proposed the Spatial Attention-Guided Upsampling network (SAGU-Net) that utilizes spatial attention to guide cost volume and disparity map upsampling, aiming to enhance the accuracy and speed of stereo matching by emphasizing important spatial information. Inspired by the CoAtNet [28] that combines the advantages of transformers and convolutions, this paper introduces a novel attention mechanism called ACmix. By integrating global information from self-attention and local information from convolutions, the ACmix attention mechanism captures critical features in images and adapts well to different scenarios and tasks. Experimental results demonstrate that ACmix performs exceptionally well in remote sensing image object detection tasks.

B. MULTISCALE FEATURE FUSION

Multiscale feature fusion plays a crucial role in improving the performance of object detection algorithms. By combining features from different scales, the detection model can capture both fine-grained details and high-level semantic information, leading to better object localization and classification. Lin et al. [14] proposed the classic Feature Pyramid Network (FPN), which integrates features from deep to shallow layers. To enhance the effectiveness of information transmission from lower to higher layers in FPN, PANet (Path Aggregation Network) [29] introduced a further connection from bottom to top. Tan et al. proposed the Bidirectional Feature Pyramid Network (BiFPN) [30], which can bi-directionally fuse features and improve the effectiveness of PANet's connection method. Quan et al. introduced a Centralized Feature Pyramid module [31] to optimize global information and fully utilize information at the same scale. Shi et al. [32] proposed a scene categorization model that utilizes deep visually sensitive features. Li et al. [33] proposed a video super-resolution method that combines non-local and multi-scale features to improve the performance of video super-resolution. By fusing deep features from different convolutional layers, the model achieves improved classification performance in complex indoor scenes. However, these methods mainly focus on feature fusion and transmission, while the utilization of global information to optimize detection performance has not been fully considered. To address this issue, we propose a Context Feature Enhancement module to capture contextual information in the scene, thereby improving the model's ability to recognize and localize objects.

III. MODEL INTRODUCTION AND IMPROVEMENT

A. YOLOX MODEL INTRODUCTION

YOLOX is a one-stage object detection algorithm with six versions: n, tiny, s, m, l, and x, each having different network widths and depths. The YOLOX algorithm model comprises four components: input, backbone extraction network, neck, and detection head. In practical scenarios, edge devices typically provide limited computational power. Therefore, detection models should achieve as accurate results as possible with minimal computational cost. Compared to models like YOLOX-m, the YOLOX-s model has fewer parameters, making it more suitable for scenarios with high real-time requirements. Compared to models like YOLOX-nano, YOLOX-s maintains a certain model size, resulting in higher detection accuracy. However, due to the complex backgrounds and numerous small objects in remote sensing images, directly using it to detect objects will result in missed and false detections. Therefore, based on the analysis of YOLOX-s, this paper proposes the YOLOX-CA algorithm.

B. OVERALL STRUCTURE OF YOLOX-CA MODEL

The overall structure of the YOLOX-CA model consists of three main components: the backbone network, the neck network, and the detection network. The model structure diagram of YOLOX-CA is illustrated in Fig.1.

The backbone network is primarily used to extract features from the input image, employing the Darknet53 network structure with a Spatial Pyramid Pooling Fast (SPPF) structure. We replaced the original CSPLayer in the backbone network with CSPNLayr based on the basic building unit CSPNextBlock and introduced the Channel Attention module. In the last layer of the backbone network, we utilized the ACmix attention mechanism to enhance the capability to capture crucial information.

The neck network is mainly responsible for fusing features from the three layers obtained by the backbone network. It adopts the PANet model. In the neck network, we introduced a Contextual Feature Enhancement (CFE) module, which was applied in the downsampling process of the neck network to enrich feature representation. These improvements aim to enhance the performance of remote sensing object detection, enabling the model to better adapt to various complex scenarios.

The detection network performs classification and regression on the feature maps passed from the neck layer. The YOLOX algorithm uses an anchor-free decoupled detection head in this layer to address conflicts between classification and regression tasks in object detection, thereby improving the detection speed of the target detection network. Subsequent sections will introduce each core component in detail.

C. CSPNLAYER MODULE

In object detection, it is necessary to consider high-accuracy detection results, real-time performance, and computational

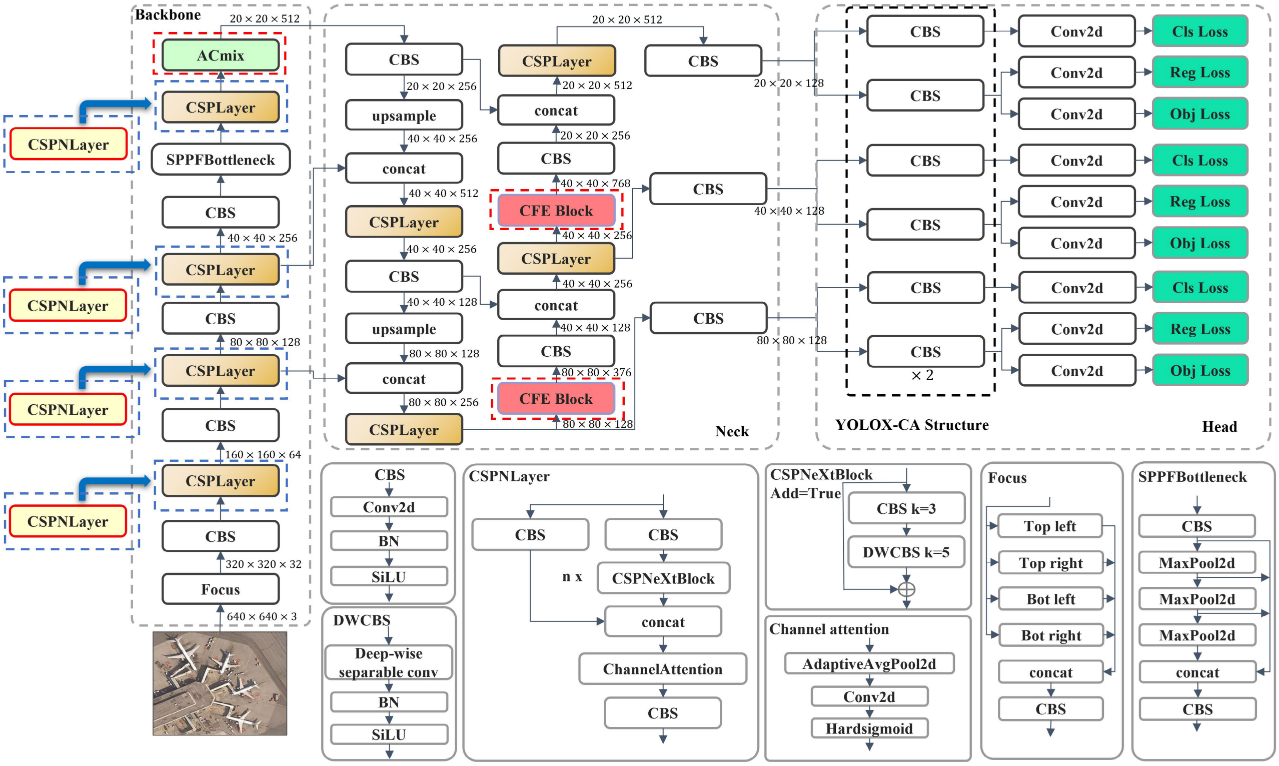

FIGURE 1. YOLOX-CA model structure diagram.

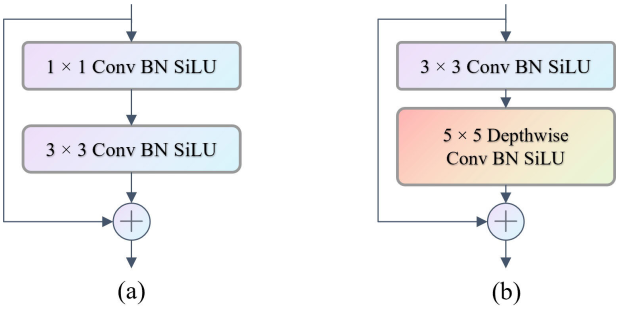

FIGURE 2. Convolution structure. (a) Basic block. (b) CSPNextBlock.

resources simultaneously to reduce redundant expressions and computational resources. This paper introduces CSPN-Layer to improve the performance of the model. The structure of this module is shown in Fig.1, consisting of three CBS (Conv2d, BatchNorm, SiLU activation function), n CSPNextBlocks, and a Channel Attention module.

Inspired by RTMDet [34], CSPNLayer uses CSP-NextBlock as the basic block, as shown in Fig.2 (b) which has a larger receptive field than the traditional CSPLayer and can learn more features in a single convolution module. Compared to the original Basic Block (Fig.2 (a)), the CSPNextBlock is simpler and more efficient. In contrast to Transformer models, the computational complexity of using the CSPNextBlock is lower as it focuses only on local regions, making it suitable for large-scale data processing. Moreover, the use of large-kernel depthwise separable convolutions aids in extracting more shape features, which are crucial for target recognition requiring high shape information. The Channel Attention module consists of AdaptiveAvgPool2d layer, Conv2d layer, and Hardsigmoid activation function, allowing the detection network to concentrate on meaningful features and reduce the interference of non-critical feature information.

D. ACMIX ATTENTION MECHANISM

In remote sensing images, the similarity between targets and the background, as well as the mutual interference among targets, pose challenges for target detection. To address this, this paper introduces ACmix, a hybrid convolutional and self-attention mechanism.

ACmix Attention Mechanism aims to reduce redundant information, extract crucial features, enhance the feature representation of neural networks, and improve accuracy and robustness. The structure of the ACmix Attention Mechanism is illustrated in Fig.3. In the first stage, the input tensor with dimensions H×W×C undergoes three 1×1 convolutions for projection and is reshaped into N, resulting in an intermediate feature set containing 3N feature maps. The second stage consists of self-attention and convolution attention modules. In the self-attention module, ACmix aggregates the intermediate features into N groups, each

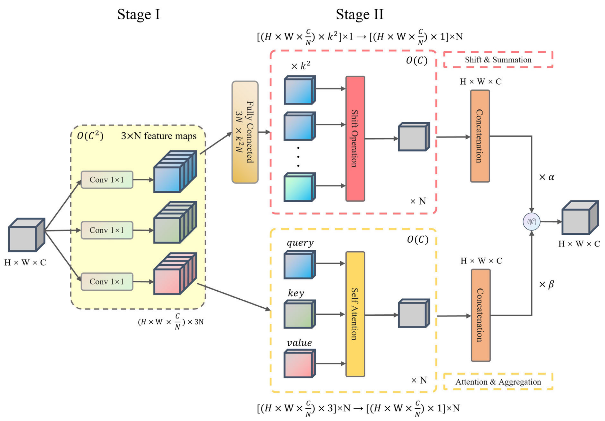

FIGURE 3. The structure of ACmix.

comprising three feature maps originating from the three 1 × 1 convolutions. These feature maps represent query (Q), key (K), and value (V), following the conventional multi-head self-attention model, as shown in formula (1).

$$g_{ij} = \overset{N}{\underset{l=1}{\|}} (\sum\nolimits_{a,b\in N_k(i,j)} A(Q_{ij}^{(l)}, K_{ab}^{(l)}) V_{ab}^{(l)}) \tag{1}$$

g_{ij} represents the projected tensor corresponding to the pixel (i, j), $||$ denotes the concatenation of outputs from N attention heads, $N_k(i, j)$ represents the local region with (i, j) as the center and a spatial range of k for pixel centers, and $A(Q_{ij}^{(l)}, K_{ab}^{(l)})$ is the corresponding attention weight about the inner features $N_k(i, j)$. The magnitude of the value depends on the similarity between the query and key. If the similarity is higher, the assigned weight is correspondingly larger, and vice versa.

In traditional convolutional modules, convolutions using K as the kernel are performed. As shown in formulas (2), (3), these convolutions are connected through lightweight fully connected layers to obtain k^2 feature maps. These feature maps are generated by moving and aggregating features, processing input features in a convolutional manner, and collecting information from local receptive fields, similar to traditional receptive fields.

$$g_{i,j}^{(p,q)} = Shift(\tilde{g}_{ij}^{(p,q)}, p - [\frac{k}{2}, q - [\frac{k}{2}]]) \tag{2}$$

$$g_{ij} = \sum\nolimits_{p,q} g_{ij}^{(p,q)} \tag{3}$$

Finally, the two paths of the traditional convolutional module and the self-attention module are summed, with weights controlled by two learnable scalars, α and β, as shown in formula (4). These parameters are optimized by the mechanism to obtain the final feature representation.

$$F_{\text{out}} = \alpha F_{att} + \beta F_{conv} \tag{4}$$

F_{out} represents the final output of the path, while F_{att} and F_{conv} correspond to the outputs of the self-attention and convolutional attention branches, respectively. The learnable scalar parameters α and β are initialized to 1. The ACmix attention module effectively utilizes both local and global information, enhancing the neural network's feature representation.

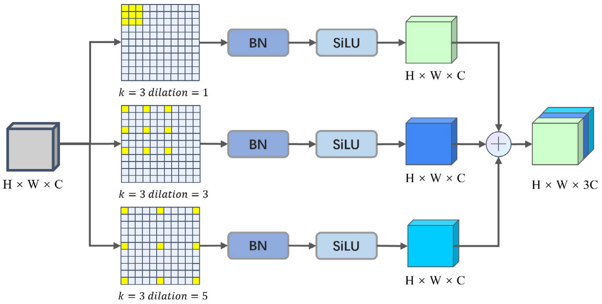

FIGURE 4. Contextual feature enrichment module.

E. CONTEXTUAL FEATURE ENHANCEMENT MODULE

Taking inspiration from the intricate visual systems of humans and animals, renowned for their adeptness at concurrently processing visual information across various scales, our eyes and brains afford us a holistic understanding of our surroundings, enabling sophisticated visual decision-making. Harnessing the insights gained from this multi-scale perception mechanism observed in natural systems, we endeavor to emulate it within the domain of computer vision. Hence, we introduce a novel approach named the Contextual Feature Enhancement(CFE) Module.

As shown in Fig.4, we apply dilated convolutions to the feature map F with dilation rates $d \in \{1, 3, 5\}$ and a 3 × 3 convolution kernel. Following each convolution operation, we introduce the Batch Normalization (BN) layer and the Sigmoid-weighted Linear Unit (SiLU) activation function.

Batch Normalization contributes to stabilizing the training process of the model, accelerating the convergence speed of the network, and making the training process more manageable. Meanwhile, SiLU, as a non-linear activation function, introduces non-linear characteristics to the network, facilitating the model in learning more complex data distributions and features. This is crucial for enhancing the model's representational capacity.

Finally, we perform feature fusion on the three obtained new feature maps, as shown in formula (5).

$$F_{concat} = Concat(F_{1,BN,SiLU}, F_{3,BN,SiLU}, F_{5,BN,SiLU}) \tag{5}$$

where "Concat" denotes the concatenation operation. In other words, the three obtained feature maps are concatenated along the channel dimension, forming a more diverse and rich feature representation. This helps enhance the model's perceptual ability, feature expression, and robustness, thereby improving the handling of contextual information.

After connecting the output feature maps from dilated convolutions with different dilation rates, we use the CBS module (Conv, BN, SiLU) to reduce the channel count back to the same as the input. This is done to decrease the computational cost of subsequent operations while maintaining sufficient expressive power, ensuring that the module's output seamlessly connects with subsequent network layers. This approach helps maintain a well-connected and stable training of the entire neural network.

F. LOSS FUNCTION

The model training is constrained by three different types of loss functions: category prediction loss (L_{cls}), confidence regression loss (L_{obj}), and bounding box regression loss (L_{reg}). The category prediction loss is employed to evaluate the model's accuracy in classifying the target category, aiding the model in learning to classify the target correctly. The confidence regression loss is used to measure the accuracy of the model in determining the presence or absence of the target. Both of these losses utilize the binary cross-entropy loss function (BCEWithLogitsLoss). The calculation method is shown in formulas (6), (7):

$$l(x, y) = L = (l_1, \ldots, l_N)^T \tag{6}$$

$$l_n = -w_n[y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n))] \tag{7}$$

x, y is the input tensor; N is the Batch Size; n is the number of labels predicted in each batch; σ is the Sigmoid nonlinear activation function.

The bounding box regression loss employs the Intersection over Union loss (IoU loss) to measure the model's accuracy in localizing the target position. The calculation method is shown in formula (8).

$$L_{IoU} = 1 - \frac{|b \bigcap b^{gt}|}{|b \bigcup b^{gt}|} \tag{8}$$

The final expression for the loss function is shown in formula (9):

$$L = w_1 L_{reg} + w_2 L_{cls} + w_3 L_{obj} \tag{9}$$

The weights for each loss function are denoted as w_1, w_2, w_3. In this context, the specific values assigned are $w_1 = 5.0$, $w_2 = 1.0$, and $w_3 = 1.0$.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL DATA

To validate the effectiveness of the YOLOX-CA algorithm in detecting remote sensing images, experiments were conducted using the DIOR dataset released by Northwestern Polytechnical University to assess and test the algorithm's performance. This dataset encompasses 20 object classes, including airplanes, airports, baseball fields, basketball courts, bridges, chimneys, dams, highway service areas, highway toll booths, golf courses, athletic fields, harbors, overpasses, boats, stadiums, storage tanks, tennis courts, train stations, vehicles, and windmills. The dataset comprises 23,463 images and 192,472 instances, divided into training, testing, and validation sets. Specifically, the training and validation sets, containing a total of 11,725 images, were utilized for training, while the test set, comprising 11,738 images, was used for evaluation. The dataset is characterized by its large scale in terms of target categories, target instance quantities, and overall image count. It exhibits significant variations in imaging conditions, weather, seasons, and image quality, presenting high inter-class similarity and intra-class diversity. This diversity allows the dataset to represent a wide range of scenarios typical in remote sensing object detection tasks.

TABLE 1. Training parameters.

Name	Parameters
Optimizer	SGD
Learning Rate	0.01
Momentum	0.9
Nesterov	True
Weight decay	0.0005
Learning rate scheduler	CosineAnnealingLR,begin=5,min_lr_ratio=0.05
Batch size	8
Epoch	300
Mosaic	img_scale=(640,640)
RandomAffine	scaling_ratio_range=(0.1,2.0),border=(-320,-320)
RandomFlip	prob=0.5

B. EVALUATION INDICATORS

The evaluation metrics employed in this paper include Precision, Recall, AP (Average Precision), mAP (mean Average Precision), Params (number of parameters), FLOPs (number of floating point operations) and FPS (frames per second), where AP is the area under the Precision-Recall curve, and mAP is the average of AP across all categories, Recall represents the proportion of true positive targets that the model can correctly detect. The mAP and Recall are defined as formula (10), (11).

$$AP = \int_0^1 P(r)dr \tag{10}$$

$$\mathrm{m}AP = \frac{1}{n} \sum\nolimits_{i=1}^{n} AP_i \tag{11}$$

Here, $P(r)$ represents the maximum precision at a recall of r, and dr represents the change in the Recall. n denotes the number of categories, with $n = 20$ in this context, and AP_i represents the average precision for the i-th category.

C. EXPERIMENTAL SETUP

The model training and performance evaluation experiments were conducted on a GPU server with the following hardware configuration: NVIDIA Tesla P100-PCIE-16GB graphics card, Ubuntu 16.04 operating system, Python 3.8.17, PyTorch 1.10.0, and the MMYOLO toolbox [35]. The toolbox version used was mmdet 3.1.0 and mmyolo 0.6.0. For fair comparison, the experimental settings employed uniform training parameters, as follows: the optimizer used standard SGD with a learning rate of 0.01, momentum of 0.9, Nesterov functionality set to True, and a weight decay coefficient of 0.0005. The learning rate scheduler utilized a cosine

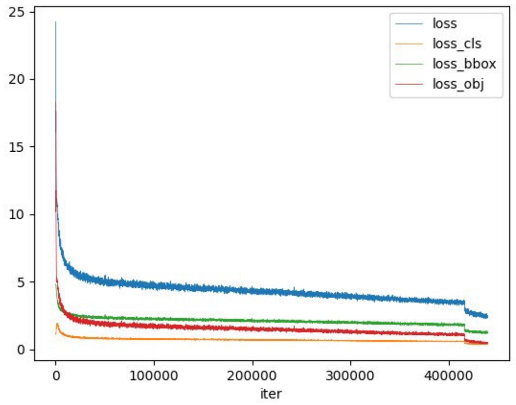

FIGURE 5. Loss function change curve.

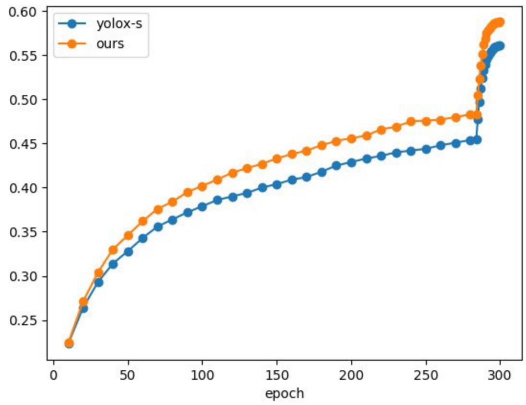

FIGURE 6. mAP change curve.

annealing learning rate strategy, with the minimum learning rate being 5% of the current learning rate, and a quadratic equation was applied for warm-up during the first 5 epochs, followed by a fixed minimum learning rate for the last 15 epochs. The batch size was set to 8, and the total number of iterations was 300. As shown in Table 1.

During model training, the Mosaic data augmentation strategy was employed. Four random images were selected, scaled, and concatenated to create a new image. Random geometric transformations, such as translation, rotation, and scaling, were then applied to the new image to obtain an augmented image of size 640 pixels × 640 pixels. Subsequently, photometric distortion was used to randomly adjust the brightness, contrast, saturation, and hue of the image, resulting in the final augmented image of size 640 pixels × 640 pixels.

D. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental process in this paper is as follows: the study is divided into two stages, training and testing. In the training stage, the model is trained for 300 epochs using the training set and validation set. Subsequently, the trained model is evaluated using the test set. Fig.5 illustrates the change curves of total loss, category prediction loss, bounding box regression loss, and object existence probability loss during the training process. These loss functions reflect the model learning process and performance. As training progresses, these losses gradually decrease, indicating that the improved model becomes more accurate in object detection and classification. Fig.6 presents the evaluation results of average precision (mAP) on the test dataset for both the baseline model and the improved model. Overall, the improved model exhibits better performance improvement during training compared to the baseline model, achieving a higher level of accuracy.

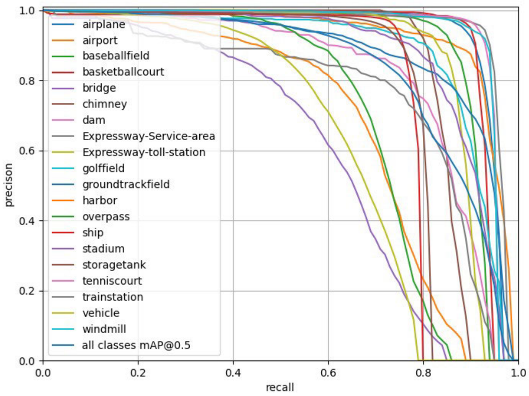

FIGURE 7. Precision and recall curve.

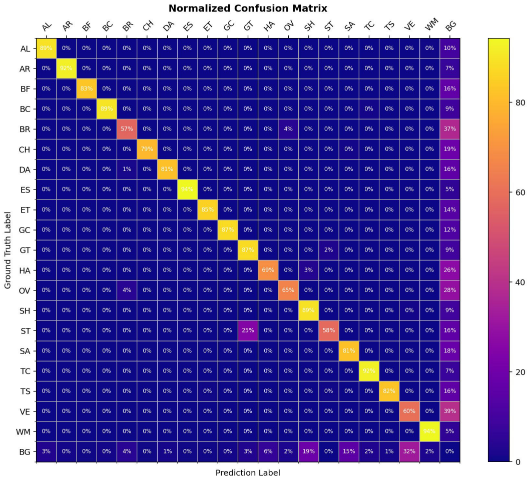

FIGURE 8. Confusion matrix.

The precision and recall curves of the proposed model are depicted in Fig.7. The curves illustrate the variation in accuracy as Recall increases. From the graph, it can be observed that the precision and recall curves for different categories of the proposed model are close to the upper-right corner, indicating high precision and recall rates. The large area under the precision and recall curves suggests the excellent performance of our model. Furthermore, the smooth curves indicate a relatively stable relationship between the model's Recall and precision.

TABLE 2. Ablation study on DIOR dataset.

CFE	ACmix	CSPNLayer	mAP	mAP@0.5	Recall	Params	FLOPs	FPS
			56.1	82.2	64.5	**8.945M**	**13.339G**	**17.5**
✓			57.6	82.4	65.9	12.634M	19.953G	17.0
	✓		57.0	82.5	65.4	9.775M	13.676G	17.4
		✓	57.2	82.9	65.3	9.317M	13.417G	16.4
✓	✓		58.4	82.9	66.7	13.464M	20.289G	16.0
✓	✓	✓	**58.8**	**83.3**	**66.7**	13.836M	20.367G	15.6

TABLE 3. Comparative experiments based on the DIOR dataset.

Method	Backbone	mAP@0.5	AL	AR	BF	BC	BR	CH	DA	ES	ET	GC	GT	HA	OV	SH	ST	SA	TC	TS	VE	WM
SSD [11]	VGG-16	58.6	59.5	72.7	72.4	75.7	29.7	65.8	56.6	63.5	53.1	65.3	68.6	49.4	48.1	59.2	61.0	46.6	76.3	55.1	27.4	65.7
YOLOv3 [16]	Darknet-53	57.1	72.2	29.2	74.0	78.6	31.2	69.7	26.9	48.6	54.4	31.1	61.1	44.9	49.7	87.4	70.6	68.7	87.3	29.4	48.3	78.7
Faster R-CNN [9]	ResNet-101	65.1	54.0	74.5	63.3	80.7	44.8	72.5	60.0	75.6	62.3	76.0	76.8	46.4	57.2	71.8	68.3	53.8	81.1	59.5	43.1	81.2
Mask R-CNN [10]	ResNet-101	65.2	53.9	76.6	63.2	80.9	40.2	72.5	60.4	76.3	62.5	76.0	75.9	46.5	57.4	71.8	68.3	53.7	81.0	62.3	43.0	81.0
RetinaNet [36]	ResNet-101	66.1	53.3	77.0	69.3	85.0	44.1	73.2	62.4	78.6	62.8	78.6	76.6	49.9	59.6	71.1	68.4	45.8	81.3	55.2	44.4	85.5
YOLOv5-s [21]	CSP-Darknet	78.7	84.3	88.5	85.6	91.3	56.1	80.7	69.8	92.7	84.5	78.8	85.5	64.3	67.4	77.3	80.3	74.7	92.6	68.4	59.2	92.3
YOLOv6-s [37]	EfficientRep	79.8	80.3	91.5	86.9	90.8	56.3	79.8	76.1	93.7	82.1	86.5	84.9	68.5	67.1	78.5	82.4	74.5	92.6	71.6	59.0	91.8
PP-YOLOE+-s [38]	CSPResNet	80.1	92.0	88.7	89.3	91.0	53.8	79.2	76.7	90.7	81.6	87.3	87.1	65.3	67.4	78.1	86.3	75.5	93.1	72.8	55.1	91.3
RTMDet-s [34]	CSPNeXt	80.8	89.7	92.3	87.1	**92.9**	54.1	80.9	78.0	**95.2**	78.3	85.6	87.5	**71.0**	69.3	77.6	82.7	75.6	93.4	76.8	58.8	89.0
CATNet [39]	ResNet-50	80.6	89.3	92.0	84.7	92.2	58.3	**84.9**	80.8	93.2	80.2	85.8	**88.3**	68.8	68.8	**84.1**	82.0	70.8	92.2	**76.9**	46.7	91.1
YOLOX-s [18]	CSP-Darknet	82.2	92.9	90.4	**90.2**	92.2	58.4	82.6	76.8	94.6	87.7	84.0	86.5	68.2	69.5	77.6	**89.1**	77.9	**94.5**	73.5	63.5	93.7
YOLOX-CA	CSP-Darknet	**83.3**	**93.6**	**92.5**	89.8	92.2	**60.5**	84.6	**81.9**	94.8	**87.7**	**88.3**	88.1	68.5	**70.8**	77.2	87.5	**78.6**	93.9	76.8	**64.5**	**93.8**

The confusion matrix of the YOLOX-CA algorithm tested on the DIOR dataset is illustrated in Fig.8. It showcases the improved model's precise predictions and the interrelationships among the 20 categories in the dataset. The class abbreviations are as follows: AL - airplane, AR - airport, BF - baseball field, BC - basketball court, BR - bridge, CH - chimney, DA - dam, ES - expressway service area, ET - expressway toll station, GC - golf field, GT - ground track field, HA - harbor, OV - overpass, SH - ship, ST - stadium, SA - storage tank, TC - tennis court, TS - train station, VE - vehicle, WM - windmill, and BG - background.

The confusion matrix is presented in a rectangular form, where rows represent true labels and columns represent predicted categories. The data on the diagonal indicates the proportion of correct predictions, while the off-diagonal data represents cases where the model incorrectly predicted one category as another. As shown in Fig.8, it is evident that the improved algorithm performs well in most classes, with prediction accuracies exceeding 80% for classes such as airplanes, airports, and baseball fields. However, the model underperforms on some categories. For instance, due to the complexity of the background, the model has relatively low prediction accuracy for categories such as bridge, harbor, and overpass. Additionally, because of the dense distribution of targets, the model tends to miss detections for the vehicle category. Moreover, due to the similarity between classes, the model sometimes misclassifies stadium as ground track field. These findings indicate that there is still room for improvement in certain aspects of the enhanced algorithm, necessitating further research and refinement in the future.

To validate the effectiveness and reliability of the proposed YOLOX-CA algorithm, we conducted ablation experiments by selectively removing certain components to assess their impact on the experimental results. Table 2 presents the results of these ablation experiments conducted on the DIOR dataset. The detection results of the original YOLOX-s algorithm are listed as the baseline for comparison in the ablation experiments.

After adding CSPNLayer, attention mechanisms, and contextual feature enhancement modules to the network, although the parameter count and FLOPs increased slightly, improvements were observed in mAP, mAP@0.5, and Recall. The last row presents the improvement achieved by simultaneously adding all three methods. After the improvements, mAP, mAP@0.5, and Recall reached 58.8%, 83.3%, and 66.7%, respectively, representing increases of 2.7%, 1.1%, and 2.2%. These data results demonstrate that the YOLOX-CA algorithm effectively enhances detection accuracy and reduces the probability of target omissions. It is important to acknowledge that these improvements come at the cost of a slight decrease in FPS due to the increased computational complexity introduced by the enhancement modules. Nonetheless, the algorithm still maintains a satisfactory FPS performance, making it suitable for real-time object detection tasks in remote sensing applications.

To further validate the effectiveness of the YOLOX-CA algorithm, we conducted several comparative experiments. We compared the proposed model with a series of models, using AP and mAP@0.5 metrics for 20 classes of objects as evaluation criteria. The experimental comparison results are shown in Table 3.

The comparison results show that our model exhibits significant improvement compared to other models. Specifically, there is a notable enhancement in the detection accuracy of objects such as airplanes, airports, dams, golf courses, ground tracks, and train stations. There are also varying

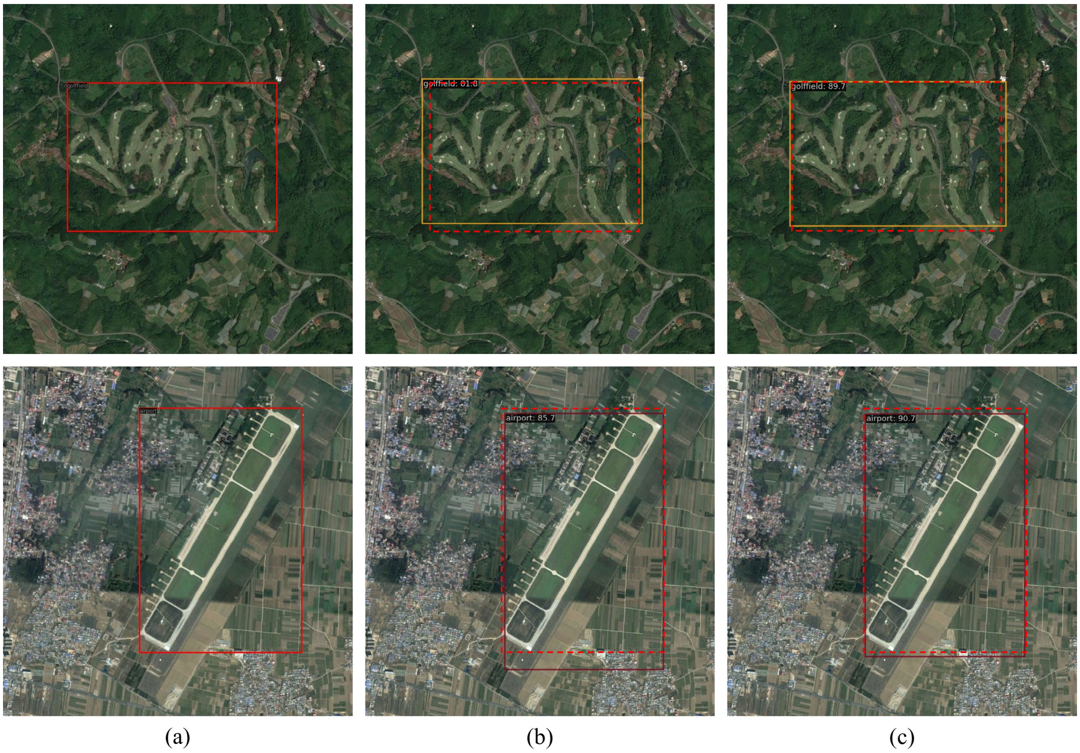

FIGURE 9. Comparison results of YOLOX-s and YOLOX-CA.

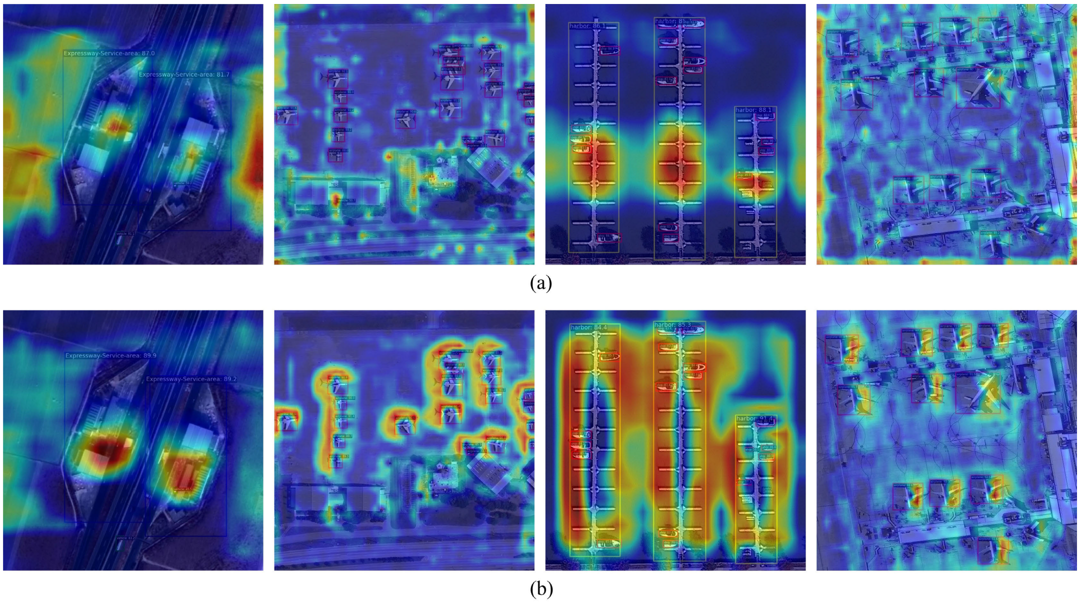

FIGURE 10. Thermal map visualization results of YOLOX-s and YOLOX-CA.

degrees of improvement in the detection accuracy of other objects.

To validate the feasibility of the improved model, we selected multiple images from the test dataset for visual analysis. Fig.9 illustrates the detection results of the baseline model and the improved model in complex backgrounds. Fig.9(a) represents the ground truth bounding boxes for the test images, Fig.9(b) represents the detection results of YOLOX-s, and Fig.9(c) represents the detection results of YOLOX-CA algorithm. It is evident that the improved model

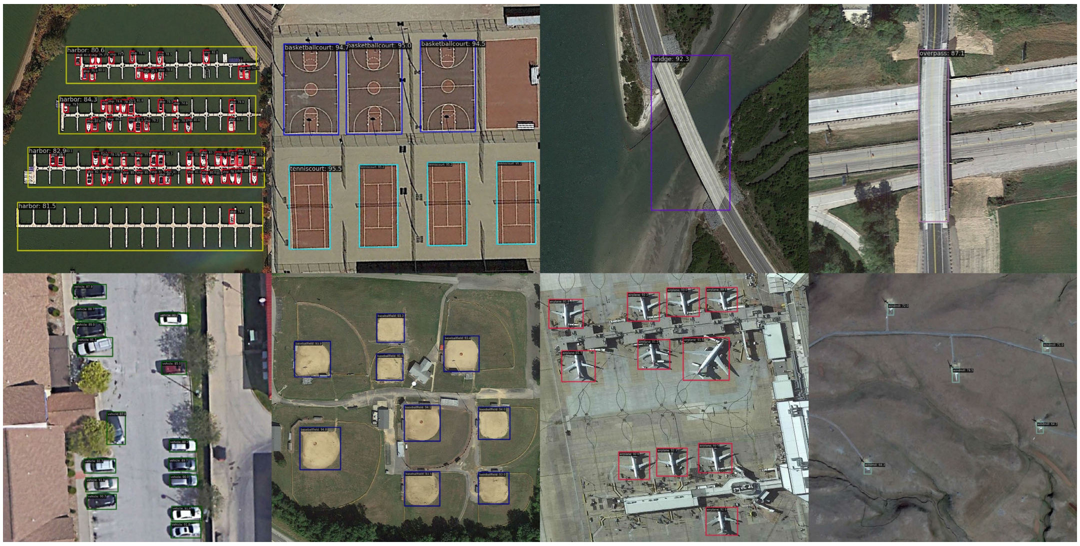

FIGURE 11. Test results of YOLOX-CA on DIOR dataset.

can more accurately locate targets such as golf courses and airports, which align better with the real-world scenario.

Fig.10(a) and (b) respectively illustrate the visual feature maps of the YOLOX-s algorithm and the YOLOX-CA algorithm in different scenes. Through Fig.10, it can be observed that the YOLOX-CA algorithm focuses more on target categories, which helps enhance the accuracy and robustness of target detection while reducing sensitivity to background interference. This implies that in complex scenes, the YOLOX-CA algorithm can more accurately identify and locate targets without being affected by the surrounding environment.

Fig.11 showcases the detection results of the YOLOX-CA algorithm on different categories of targets. It can be seen that targets from various categories are accurately detected and located without significant omissions or false detections. This indicates that the improved model exhibits good generality and stability in remote sensing image target detection tasks and is capable of adapting to various object categories.

V. CONCLUSION

In this work, we conducted an in-depth analysis of the characteristics of optical remote sensing images. We optimized the model results by focusing on feature extraction and feature fusion. We proposed the YOLOX-CA algorithm, a novel framework that introduces large kernel depth-wise separable convolution, ACmix attention mechanism, and context feature enhancement module. This framework improves both the detection accuracy and the speed of the model. Experimental results demonstrate that our method significantly enhances the ability to overcome interference factors such as complex backgrounds and small, dense targets in remote sensing images, thereby improving object detection performance. While our method performs well in handling complex backgrounds and small, dense targets, it may experience a decline in detection performance under extreme conditions, such as extremely low illumination or high occlusion. This could potentially require more computational resources and time. Furthermore, it may exhibit certain preferences for specific types of targets or scenes. In future research, we will attempt to embed more lightweight modules and residual frameworks into the YOLOX-CA object detection algorithm to reduce network size and improve detection accuracy.

REFERENCES

[1] A. Halefom, A. Teshome, E. Sisay, D. Khare, M. Dananto, L. Singh, and D. Tadesse, "Applications of remote sensing and GIS in land use/land cover change detection: A case study of Woreta Zuria Watershed, Ethiopia," *Appl. Res. J. Geographic Inf. Syst.*, vol. 1, no. 1, pp. 1–9, 2018.

[2] H. Shi, L. Chen, F.-K. Bi, H. Chen, and Y. Yu, "Accurate urban area detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1948–1952, Sep. 2015.

[3] O. Ghorbanzadeh, H. Shahabi, A. Crivellari, S. Homayouni, T. Blaschke, and P. Ghamisi, "Landslide detection using deep learning and object-based image analysis," *Landslides*, vol. 19, no. 4, pp. 929–939, Apr. 2022.

[4] H. Liu, Y. Yu, S. Liu, and W. Wang, "A military object detection model of UAV reconnaissance image and feature visualization," *Appl. Sci.*, vol. 12, no. 23, p. 12236, Nov. 2022.

[5] H. Tang and Z. Hu, "Research on medical image classification based on machine learning," *IEEE Access*, vol. 8, pp. 93145–93154, 2020.

[6] J. Bao, X. Liu, Z. Xiang, and G. Wei, "Multi-objective optimization algorithm and preference multi-objective decision-making based on artificial intelligence biological immune system," *IEEE Access*, vol. 8, pp. 160221–160230, 2020.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[8] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–16.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[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. 14th Eur. Conf.* Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 21–37.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[13] W. Li, K. Liu, L. Zhang, and F. Cheng, "Object detection based on an adaptive attention mechanism," *Sci. Rep.*, vol. 10, no. 1, p. 11307, Jul. 2020.

[14] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[15] L.-G. Zhang, L. Wang, M. Jin, X.-S. Geng, and Q. Shen, "Small object detection in remote sensing images based on attention mechanism and multi-scale feature fusion," *Int. J. Remote Sens.*, vol. 43, no. 9, pp. 3280–3297, May 2022.

[16] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[17] Y. Cheng, W. Wang, W. Zhang, L. Yang, J. Wang, H. Ni, T. Guan, J. He, Y. Gu, and N. N. Tran, "A multi-feature fusion and attention network for multi-scale object detection in remote sensing images," *Remote Sens.*, vol. 15, no. 8, p. 2096, Apr. 2023.

[18] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[19] Z. Li, A. Namiki, S. Suzuki, Q. Wang, T. Zhang, and W. Wang, "Application of low-altitude UAV remote sensing image object detection based on improved YOLOv5," *Appl. Sci.*, vol. 12, no. 16, p. 8314, Aug. 2022.

[20] Z. Liu, Y. Gao, Q. Du, M. Chen, and W. Lv, "YOLO-extract: Improved YOLOv5 for aircraft object detection in remote sensing images," *IEEE Access*, vol. 11, pp. 1742–1751, 2023.

[21] G. Jocher et al., "Ultralytics/YOLOv5: V5.0–YOLOv5-P6 1280 models, AWS, Supervise.Ly and YouTube integrations," *Zenodo*, vol. 11, Apr. 2021.

[22] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 805–815.

[23] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[25] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," 2021, *arXiv:2112.05561*.

[26] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.

[27] Z. Wu, H. Zhu, L. He, Q. Zhao, J. Shi, and W. Wu, "Real-time stereo matching with high accuracy via spatial attention-guided upsampling," *Int. J. Speech Technol.*, vol. 53, no. 20, pp. 24253–24274, Oct. 2023.

[28] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 3965–3977.

[29] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[30] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[31] Y. Quan, D. Zhang, L. Zhang, and J. Tang, "Centralized feature pyramid for object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 4341–4354, 2023.

[32] J. Shi, H. Zhu, S. Yu, W. Wu, and H. Shi, "Scene categorization model using deep visually sensitive features," *IEEE Access*, vol. 7, pp. 45230–45239, 2019.

[33] Y. Li, H. Zhu, Q. Hou, J. Wang, and W. Wu, "Video super-resolution using multi-scale and non-local feature fusion," *Electronics*, vol. 11, no. 9, p. 1499, May 2022.

[34] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, "RTMDet: An empirical study of designing real-time object detectors," 2022, *arXiv:2212.07784*.

[35] M. Contributors. (2022). *MMYOLO: OpenMMLab YOLO Series Toolbox and Benchmark*. [Online]. Available: https://github.com/open-mmlab/mmyolo

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[37] 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*.

[38] S. Xu, X. Wang, W. Lv, Q. Chang, C. Cui, K. Deng, G. Wang, Q. Dang, S. Wei, Y. Du, and B. Lai, "PP-YOLOE: An evolved version of YOLO," 2022, *arXiv:2203.16250*.

[39] Y. Liu, H. Li, C. Hu, S. Luo, Y. Luo, and C. W. Chen, "Learning to aggregate multi-scale context for instance segmentation in remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2024.

CHAO WU was born in Jiangxi, China, in 2000. He received the B.S. degree in computer science and technology from the East China University of Technology, Nanchang, China, in 2022. He is currently pursuing the master's degree with the College of Computer and Cyber Security, Fujian Normal University.

His current research interest includes target detection in remote sensing images.

ZHIYONG ZENG received the B.S. degree in chemistry from Gannan Normal University, Ganzhou, in 1987, the M.S. degree in chemistry from Fuzhou University, Fuzhou, in 1998, and the Ph.D. degree in computer application technology from Xidian University, Xi'an, in 2006. From 2007 to 2008, he was an Assistant Professor at the Faculty of Software, Fujian Normal University. Since 2008, he has been an Associate Professor with the College of Computer and Cyber Security, Fujian Normal University. He is the author of more than 50 journal articles and more than five inventions. He holds one patents. His research interests include digital image processes and applications, computer vision, artificial intelligence, and machine learning.