IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 July 2024, accepted 12 August 2024, date of publication 16 August 2024, date of current version 27 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3444606

RESEARCH ARTICLE

An Improved YOLOv8 Detector for Multi-Scale Target Detection in Remote Sensing Images

MIN YUE[1], LIQIANG ZHANG[1], YUJIN ZHANG[2], (Member, IEEE), AND HAIFENG ZHANG[1]
[1]School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China
[2]School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

Corresponding author: Min Yue (yuemincn@sues.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 52375503.

ABSTRACT Target detection via remote sensing is extensively utilized across diverse domains because of its inherent potential value in applications. However, most objects within remote sensing images consist of multi-scale and dense small objects, observed from diverse angles against complex backgrounds, resulting in insufficient detection performance. To enhance the detection accuracy and robustness in detecting multi-scale objects, we present the YOLO-GE algorithm based on you only look once (YOLO). We introduce the ghost convolution hierarchical graph (G-HG) block that combines ghost convolutions and the cross-stage partial (CSP) strategy. This enhancement can efficiently utilize redundant feature maps, broaden the receptive field, and accurately extract multi-scale objects and advanced semantic features in complex backgrounds. By incorporating the G-HG block, we establish the ghost-convolution enhanced hierarchical graph (GE-HGNet) feature extraction backbone, thereby enhancing its ability to capture multi-scale object features and advanced semantic information. Additionally, we develop the E-SimAM attention mechanism using residual techniques to address the low-resolution feature loss limitation, thereby enhancing the precision in identifying low-resolution features against intricate backgrounds. Furthermore, to improve the capability of detecting densely packed small objects, we reconstruct the structure of the neck and add a tiny detection head. This additional tiny detection head is specifically designed to better focus on densely packed small targets, fully leveraging the fine-grained information in shallow feature maps. Extensive experiments conducted on the DIOR, NWPU VHR-10, and VisDrone2019 datasets demonstrate the effectiveness and robustness of our YOLO-GE algorithm. Notably, compared to the state-of-the-art algorithm, our YOLO-GE-n achieves improvements of 20.1% and 22.2% in $mAP_{0.5}$ and $mAP_{0.5:0.95}$ respectively on the VisDrone2019 dataset.

INDEX TERMS YOLO, multi-scale target detection, remote sensing image, attention mechanism.

I. INTRODUCTION

As satellite techniques and object detection methodologies progressively evolve, remote sensing target detection plays a pivotal role in various fields, particularly in applications involving traffic management [1], military monitoring [2], and marine resource management [3]. Its primary objectives in remote sensing images involve automatically identifying and localizing specific targets within images acquired by satellites, aircraft, or spacecraft. These targets can range from buildings, vehicles, water bodies, etc. Nevertheless, due to the diverse shooting angles and high distances within remote sensing images, detecting objects within such images is quite different from traditional target detection in general datasets, such as ImageNet [4], Pascal VOC [5], etc. There are a considerable number of small objects within remote sensing images, including some that are particularly small (less than 10 pixels), as well as instances of high-density clusters. Moreover, the targets vary in scale and the backgrounds are complex in various scenarios, with certain images possibly containing multiple categories of objects or backgrounds at various scales simultaneously [6]. For example, buildings and airplanes often differ significantly in scale from vehicles within remote sensing images. Therefore, it is challenging

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

to extract features of low-resolution and dense small objects, which can seriously hinder further improvements in detection accuracy [7].

Wang et al. [8] proposed the full-scale object detection network (FSoD-Net) to address the issue of scale and category variance of multiple objects. They presented the multi-scale enhancement network (MSE-Net) backbone integrating a Laplace kernel with fewer parallel multiscale convolution layers to enhance tiny object feature extraction. By designing regression layers that cater to a large-scale span, FSoD-Net effectively covers the full range of object sizes found in optical remote sensing imagery. This addresses the challenges posed by the varied scales of objects within large-view-scale scenes. Yu and Ji [9] introduced the spatial-oriented object detector, which considers the unique spatial properties and scales. This approach proposed the scale-attention boosted CNN heatmaps and deformable convolutions to capture multi-scale objects. Peng et al. [10] adopted the coordinate attention (CA) mechanism and the bidirectional feature pyramid network (BiFPN) by improving multi-scale feature fusion to capture the direction and location-aware information across channels.

Although these methods have been validated and applied in practical scenarios, they often contain complex feature fusion mechanisms while yielding only modest improvements in terms of detection accuracy. In addition, they usually overlook the fine-grained information in shallow feature maps. In order to address these issues, we developed YOLO-GE, a more accurate and robust algorithm for multi-scale target detection within remote sensing images. Firstly, We introduce the G-HG block that combines ghost convolutions and the CSP strategy. This enhancement can efficiently utilize redundant feature maps, broaden the receptive field, and accurately extract multi-scale objects and advanced semantic features in complex backgrounds. By incorporating the G-HG block, we establish the GE-HGNet feature extraction backbone, thereby enhancing its ability to capture multi-scale object features and advanced semantic information. Additionally, we develop the E-SimAM attention mechanism using residual techniques to address the low-resolution feature loss limitation, thereby enhancing the precision in identifying low-resolution features against intricate backgrounds. Finally, in order to improve the capability of detecting densely packed small objects, we reconstruct the structure of the neck and add a tiny detection head. This additional tiny detection head is specifically designed to better focus on densely packed small targets, fully leveraging the fine-grained information in shallow feature maps. By organically integrating these enhancements, our proposed YOLO-GE further enhances the extraction and integration mechanism for multi-scale feature information. The main contributions of this article are summarized as follows:

(1) We proposed a novel G-HG block and an improved feature extraction backbone GE-HGNet based on ghost convolution and PP-HGNetv2. The GE-HGNet backbone based on the G-HG block can efficiently utilize redundant feature maps, broaden the receptive field, and accurately extract multi-scale objects and advanced semantic features in complex backgrounds.

(2) In response to the limitation of the SimAM attention mechanism, especially in dealing with low-resolution and multi-scale objects, we improved the SimAM attention mechanism and introduced an enhanced version named the E-SimAM attention mechanism. This approach enhances the most representative information while minimizing the loss of details from lower resolution or weaker characteristics, significantly improving its capability to detect multi-scale objects and accurately localize targets.

(3) To improve the capability of detecting densely packed small objects, we reconstructed the structure of the neck and detection head. By employing this strategy, the FPN+PAN network can capture and retain more intricate features of smaller target objects, while the added small detection heads can fully utilize these features, consequently enhancing the model's capability to recognize small objects from low-resolution images.

The rest of this study is organized as follows. Section II provides a review of CNN-based object detection. Section III details the improvements of our proposed YOLO-GE algorithm. Section IV describes the experimental datasets, accompanied by extensive experiments to evaluate the efficacy of our proposed enhancements and the performance of YOLO-GE. The discussion is presented in section V. Section VI summarizes the research conducted in this study.

II. RELATED WORKS

CNN-based object detection plays a crucial role in computer vision, focusing on recognizing and positioning objects in images or video sequences. This technique has gained immense importance across numerous fields, including surveillance systems, robotics, autonomous vehicles, etc. There are primarily two categories of CNN-based object detection: two-stage and one-stage methods.

The two-stage method divides the process of detecting objects into two distinct phases: initially, a region proposal network suggests a series of potential boxes that may contain the target objects; subsequently, these proposed boxes are subjected to classification and localization to recognize and precisely locate the detected targets. The two-stage methods primarily encompass R-CNN [11], Fast R-CNN [12], Faster R-CNN [13], and Cascade R-CNN [14], among others. Due to the necessity of generating candidate regions before conducting object detection, these methods exhibit high detection accuracy but are less suitable for real-time scenarios.

However, in the one-stage approach, category probabilities and bounding box coordinates are directly regressed without the necessity of generating candidate regions, resulting in faster detection speeds albeit with a potential loss in accuracy. The one-stage approaches mainly consist of the YOLO series algorithms [15], [16], [17], [18], [19], [20], [21], [22],

[23], SSD [24], RetinaNet [25], etc. In 2016, Redmon et al. [15] introduced the YOLO algorithm, which utilizes a single network to directly predict position frames and category probabilities across the entire image. It employs the same fully connected layer for classification and regression, leading to coarse object localization and suboptimal detection performance for small objects. In addressing these challenges, numerous researchers have conducted extensive studies, leading to the emergence of YOLOv2 [16] to YOLOv8 [22] algorithms. Through their persistent efforts, the YOLO series algorithms have been continuously improved, achieving a remarkable balance between speed and precision.

YOLOv8 [22] is the latest state-of-the-art (SOTA) model proposed by Ultralytics, which is an improvement over YOLOv5. Compared to YOLOv5, YOLOv8 utilizes the C2f structure with enhanced gradient flow, replacing all C3 structures within the backbone and neck. Additionally, it substitutes the coupled head with the prevalent decoupled head framework, distinguishing between the classification and detection heads. Furthermore, YOLOv8 transitions from anchor-based to anchor-free methodology, aligning ground truth and predicted boxes through an assigner. Based on the number of network channels and parameters, YOLOv8 can be categorized into five types: n, s, m, l, and x. Due to its advantages in accuracy and complex model structure, YOLOv8 is more suitable for scenarios that require higher detection accuracy.

Despite achieving excellent results on general datasets, the YOLO series approaches for detecting objects still encounter obstacles when utilized in remote sensing imagery, including the detection of multi-scale objects, handling complex backgrounds across diverse scenes, and fulfilling the requirements for applications in real-time surveillance. In order to address these limitations, more recent attention has focused on enhancing the precision in identifying small objects based on the YOLO series algorithms. Zakria et al. [26] introduced the classification setting of the non-maximum suppression threshold and K-means anchor frame scheme based on YOLOV4 to improve detection performance. However, these hyperparameters are fixed and not suitable for arbitrary datasets. When handling datasets with significant scale variations and densely packed small objects, these hyperparameters must be set empirically. Liu et al. [27] presented the YOLO-extract approach inspired by YOLOv5, by employing residual concepts to boost the capacity for extracting features, which incorporated the coordinate attention mechanism and mixed dilated convolution into the model. To accelerate the convergence of the model, Focal-α EIoU was introduced to replace CIoU loss. However, to improve the detection of small and densely packed objects, the algorithm introduces an additional detection head specifically for tiny objects. Nevertheless, it removes two detection heads for medium and large objects, which impairs the ability to capture high-level semantic information and consequently reduces detection performance for medium and large-scale objects. Lin et al. [28] developed the YOLO-DA approach inspired by YOLOv5, which strikes a balance between precision and speed by incorporating an attention module and a streamlined decoupled detection head featuring a CBAM module. Xie et al. [29] introduced the Partial Hybrid Dilated Convolution (PHDC) blocks and the CSP strategy to propose a lightweight detection algorithm CSPPartial-YOLO for Remote Sensing Images. This approach effectively utilizes redundant feature maps and reduces the model's parameter count, while also enlarging the receptive field to detect objects against complex backgrounds. Although YOLO-DA and CSPPartial-YOLO have improved the detection efficiency for remote sensing objects, the enhancements in detection accuracy are not significant. Liu et al. [30] introduced an improved model based on YOLOv8, named YOLO-SSP. This model enhances its detection accuracy by refining the downsampling layers to capture finer details and employing hierarchical pooling operations to derive weights from various spatial locations. Several attempts have been made to enhance the precision of detecting multi-scale objects, but effectively detecting densely packed small targets within remote sensing imagery remains a notable challenge despite multiple attempts.

III. METHOD

The comprehensive architecture of our YOLO-GE algorithm is depicted in Figure 1. The GE-HGNet backbone is primarily composed of HGStem [31], ConvModule, SPPF, and the newly proposed G-HG block. The G-HG block is a novel GhostConv-based module, specifically crafted to substitute C2f for extracting features. The neck utilizes upsampling, C2f, ConvModule, and the E-SimAM for feature fusion. The E-SimAM is a newly enhanced attention mechanism designed to address the constraints inherent in the SimAM approach. Additionally, we reconstruct the structure of both the neck and head by introducing a features fusion structure with a 4x down-sampling rate and an additional detection head.

Similar to the YOLOv8 algorithm, our YOLO-GE algorithm is categorized into four models based on the depth and width scales, named YOLO-GE-n, YOLO-GE-s, YOLO-GE-m, and YOLO-GE-l, respectively. This allows for the selection of an appropriate model according to the application scenario. Table 1 showcases the compound scales of YOLO-GE. In this section, we first detail the G-HG block and GE-HGNet backbone, then describe the E-SimAM attention mechanism, and finally introduce the improvements of the neck and head.

TABLE 1. The compound scales of YOLO-GE.

Model	depth	width	Max channels	Feature channels of the head
YOLO-GE-n	0.33	0.25	1024	[32, 64, 128, 256]
YOLO-GE-s	0.33	0.25	1024	[64, 128, 256, 512]
YOLO-GE-m	0.67	0.5	768	[96, 192, 384, 576]
YOLO-GE-l	1	1	512	[128, 256, 512, 512]

A. GE-HGNet BACKBONE

The main function of the backbone network involves extracting features across multiple scales and advanced semantic

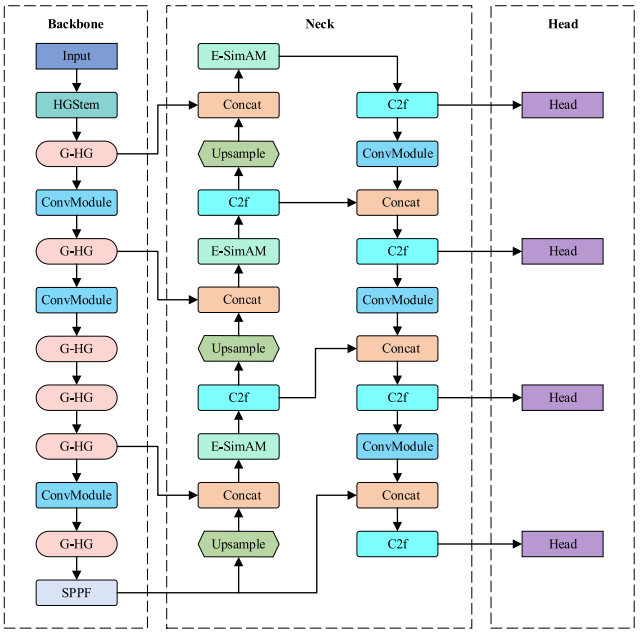


FIGURE 1. The architecture of YOLO-GE.

information for various targets, thereby promoting improved feature fusion in subsequent stages. Within YOLOv8n, the cross-stage partial network (CSPNet) [32] is retained, while the C3 module from YOLOv5 has been replaced with the C2f structure. However, with the continuous stacking of 3×3 convolutions, the parameters and computational cost in the backbone also increase. To tackle this challenge and establish an efficient yet potent backbone for feature extraction, we attempt to replace 3×3 convolution with lightweight convolution.

Ghost convolution is a lightweight module that can capture additional features with a low computational cost [33]. The ghost convolution is executed in a dual-phase process, as illustrated in Figure 2. Firstly, a traditional convolution module is employed to generate an intrinsic feature map with a smaller number of channels. Subsequently, novel ghost feature maps are generated from the obtained intrinsic feature maps through cheap operations. Finally, these two collections of feature maps are merged to produce the ultimate feature maps. Compared to other lightweight convolutions, ghost convolution efficiently exploits the correlations and redundancies among feature maps. Therefore, we attempt to employ ghost convolution as a substitute for traditional convolution for feature extraction.

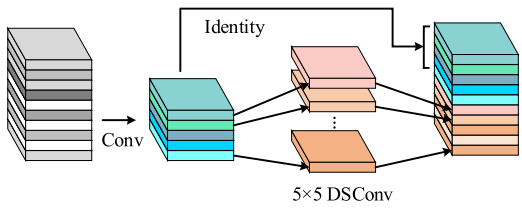


FIGURE 2. The workflow of ghost convolution.

However, directly replacing the backbone's convolutional modules with ghost convolutions may instead weaken its feature extraction capabilities, resulting in a minor reduction in detection accuracy. This occurs because the feature maps provided by the cheap operations of ghost convolutions are not adequately exhaustive. Therefore, it may be necessary to stack more ghost convolutions to increase redundant feature maps and expand the receptive field, enabling effective feature extraction. Inspired by the recent success of RT-DETR [34], we incorporate the HG block [31] from its backbone, as illustrated in Figure 3(a), replacing convolution operations with ghost convolutions. Additionally, to enhance the extraction of advanced semantic features, we propose an enhanced version G-HG block, as depicted in Figure 3(b).

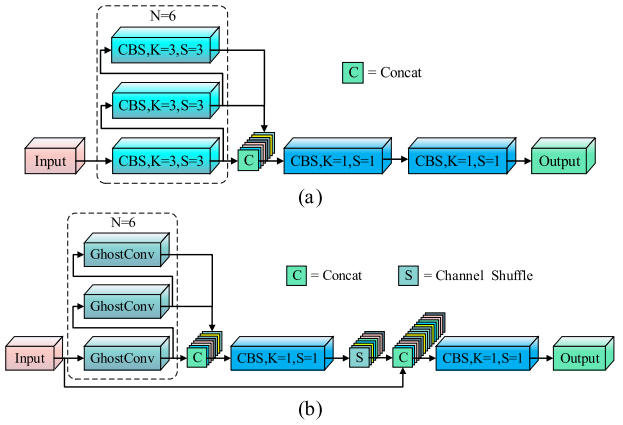


FIGURE 3. The architecture of (a) HG block and (b) our G-HG block.

The HG block employs a hierarchical stacking approach to blend features from diverse convolution layers. This boosts the model's proficiency in processing fine-grained and coarse-grained information, aligning with the earlier strategy of stacking ghost convolutions to enlarge the receptive field for extracting advanced semantic features. After concatenating all the feature maps outputted by ghost convolutions, a hierarchical feature fusion is conducted through two 1×1 convolutional modules, significantly increasing non-linearity while preserving the feature map dimensions. Although the HG block can effectively extract feature map information by stacking convolutional layers to enlarge the receptive field, it becomes challenging to capture advanced semantic information as the network depth increases due to the potential loss of gradient information.

To further boost the effectiveness of the HG block, we employ the CSP [32] strategy and channel shuffle [35] method to restructure the HG block and introduce the G-HG block. Firstly, the feature map generated by the initial 1×1 convolution undergoes channel shuffling. Subsequently, the feature map resulting from channel shuffling is concatenated with the original feature map. Finally, the obtained feature map is refined through the second 1×1 convolution to yield the ultimate feature map. By employing channel shuffling, the features within the channels are further integrated. Through

the CSP strategy, the infusion of gradient flow information is introduced, facilitating the network's learning and feature extraction.

In pursuit of extracting advanced semantic information and multi-scale details more comprehensively, we synthesize the strengths of CSPDarknet53 and HGNet to create GE-HGNet. The GE-HGNet backbone is comprised of a convolution-stacked stem, and four feature extraction stages composed of G-HG blocks and convolutions, along with an SPPF, as shown in Figure 4. Initially, the original image undergoes a 4x down-sampling and channel expansion, laying the foundation for subsequent feature extraction. Subsequently, the obtained feature map undergoes four feature extraction stages composed of G-HG blocks and convolution layers. In this process, the convolutional module is responsible for down-sampling, while the HG block focuses on extracting features across multiple scales. Finally, the SPPF module performs multi-scale fusion on the input feature maps to extract richer advanced semantic information.

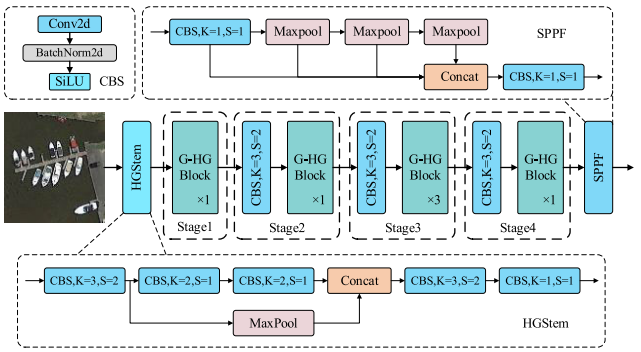


FIGURE 4. The architecture of our GE-HGNet backbone.

B. E-SIMAM ATTENTION MECHANISM

In practical remote sensing scenarios, accurately identifying multi-scale targets becomes a considerable challenge, attributed to the dense distribution of small targets and the complexity of constantly changing backgrounds. To better extract features of these targets, the attention mechanism has garnered significant attention among researchers. Various attention mechanisms proposed in prior research suggest that they can address the limitations of convolutional models [36]. Therefore, there have been numerous attempts to design various attention mechanisms to enhance detection accuracy. Woo et al. [37] introduced the convolutional block attention module (CBAM), which sequentially inferred a 1D channel and 2D spatial attention map to extract features from the channel and spatial dimensions. Misra et al. [38] constructed the triplet attention module that extracts inter-dependencies through rotation operations and residual transformations. Pan et al. [36] presented the ACmix attention module, which combines self-attention and convolution to extract semantic features effectively. Although these attention mechanisms have achieved significant success, the parameters and complexity of the networks have also continuously increased. Yang et al. [39] proposed the SimAM attention module, as illustrated in Figure 5(a), which establishes three-dimensional attention weights for the input feature map by identifying the significance of individual neurons through the optimized energy function. In particular, the mechanism features a simple structure without unnecessary parameters. Aiming to enhance detection precision in capturing small objects and optimize the model for efficiency, we integrate the SimAM module and propose an enhanced version named the E-SimAM attention module as illustrated in Figure 5(b).

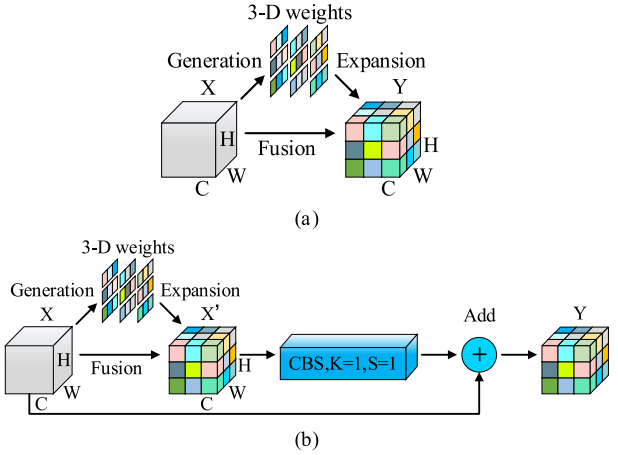


FIGURE 5. The illustration of (a) SimAM and (b) our proposed E-SimAM.

To generate the three-dimensional attention weights, we should calculate the significance of individual neurons within the input feature maps through the optimized energy function. Suppose the input feature maps refer to $X \in \mathbb{R}^{H\times W\times C}$, t, i, and x_i correspond to the target neuron, the index over spatial dimension, and other neurons in a channel of the feature maps

$X.\mathrm{M} = H \times W$ represents the number of all neurons on that channel. The optimized energy function is defined as follows:

$$\mathrm{e}_t = \frac{4(\sigma^2+\lambda)}{(t-\mu)^2+2\sigma^2+2\lambda}, \tag{1}$$

$$\mu = \frac{1}{M}\sum_{i=1}^{M} x_i, \tag{2}$$

$$\sigma^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i-\mu)^2 \tag{3}$$

where e_t represents the lower energy of the neuron t, μ and σ^2 refer to the mean and variance of all neurons. As a result, the significance of individual neurons can be determined by $1/e_t$, and the expression for SimAM can be defined as follows:

$$X' = \mathrm{Sigmoid}(\frac{1}{E}) \odot X \tag{4}$$

where E groups all e_t across channel and spatial dimensions, and X' refers to the output feature maps of SimAM. However, the Sigmoid function is a non-linear function that ensures its output always lies between 0 and 1. Therefore, neurons with

prominent features can be preserved, while those with lower resolution or weaker characteristics in small targets may be susceptible to loss. To address this limitation, we introduced the residual concept [40] to enhance the attention mechanism. To begin with, the obtained feature maps X' undergo further processing through a 1×1 convolutional module to ensure the extraction and utilization of effective features. Subsequently, the original feature maps X are combined with the newly obtained feature maps. This operation not only reinforces dominant features but also preserves features with lower resolution or weaker characteristics, facilitating the detection and localization of small targets. The E-SimAM module can be defined as:

$$Y = X + Conv1 \times 1(X') \tag{5}$$

where the output feature of E-SimAM is denoted by $Y \in \mathbb{R}^{H \times W \times C}$.

To fully leverage the E-SimAM attention mechanism, we place it after the concatenation operation of feature maps of various stages within the backbone network and up-sampling feature maps of the neck. This facilitates the rapid extraction of pivotal details from the fusion of feature maps in the backbone and neck, preparing for the subsequent multi-scale fusion by the C2f module.

C. IMPROVED NECK AND HEAD

In general, deep-layer feature maps encompass more robust semantic features but offer less precise localization information. In contrast, shallow-layer feature maps provide strong positional information with comparatively weaker semantic features. In order to enhance multi-scale feature fusion more effectively, YOLO-GE still employs the FPN [41] and PAN [42] structures from YoLOv8. Firstly, the FPN network integrates the feature maps from the SPPF layer using a top-down approach via up-sampling, enabling fusion with lower-level features. This mechanism enables the transfer of advanced semantic features to lower levels, thereby enriching semantic representation across various scales. Subsequently, PAN further enhances the FPN architecture by incorporating an additional fusion pathway, extending from lower to deeper layers through down-sampling. This facilitates the transmission of robust localization features from lower to higher layers, enhancing the model's capacity to localize across diverse scales. Finally, the feature maps produced by FPN and PAN are inputted into the decoupled heads for classification and localization. The workflow of the neck and head is illustrated in Figure 6.

For object detection within remote sensing imagery, the majority of detected objects are small, with certain objects densely distributed and typically covering only a few pixels. However, the lowest-level feature maps processed by the FPN+PAN network undergo an 8x down-sampling rate. When dealing with low-resolution or small objects, this leads to loss the fine-grained information, consequently reducing the precision in detecting small targets. In addressing this concern, we introduced a feature map with a 4x down-sampling rate from the P2 layer within the backbone. This facilitates the extraction and fusion of features for small targets through the FPN+PAN network. Consequently, we have also incorporated an additional tiny detection head specifically designed for more focus on low-resolution or small targets.

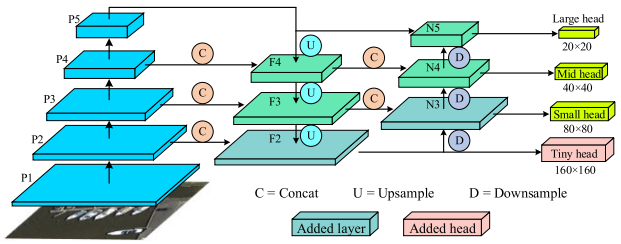


FIGURE 6. The workflow of the neck and head.

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

The experiments were carried out on an Ubuntu 22.04.3 system, Intel®i5-13600KF CPU, and NVIDIA 4090 graphics cards with 24G video memory. The experimental environment was set up with Python 3.10, PyTorch 2.1.1, and Cuda 11.8. During the training stage, we employed the SGD optimizer with a momentum of 0.937 and utilized a batch size of 8. The learning rate was consistently set at 0.01, and the network underwent training for 200 epochs. For a fair experiment comparison, all training and testing processes shared the same set of parameters.

B. DATASET

In order to assess the effectiveness of our YOLO-GE model, we conducted experiments on the extensively employed DIOR dataset [43]. The DIOR dataset serves as a vast benchmark dataset developed by Northwestern Polytechnical University, specifically tailored for detecting objects within remote sensing imagery, featuring 23,463 images and 190,288 instances. The dataset consists of images sized 800×800 pixels, exhibiting spatial resolutions varying from 0.5 m to 30 m. It encompasses a diverse array of 20 object categories, including airplanes, buildings, vehicles, and so on. Compared to other datasets, the DIOR dataset boasts advantages such as large-scale data, diverse instance object sizes, rich image diversity, high inter-class similarity, and significant intra-class differences.

In order to validate the performance capabilities of our proposed model and to ascertain its robustness, we carried out extensive experiments on the NWPU VHR-10 dataset [44], [45]. Published by Northwestern Polytechnical University in 2014, this dataset possesses 650 images with targets and 150 background images, totaling 800 images for spatial object detection. These images manually annotated by experts were extracted from Google Earth and the Vaihingen dataset.

In order to further verify the detection capabilities and robustness of our algorithm, we conducted extensive experiments on the VisDrone2019 dataset [46]. This dataset is collected by the AISKYEYE team from Tianjin University's Machine Learning and Data Mining Laboratory, captured from drone perspectives. It is a horizontal bounding box dataset specifically designed for optical remote sensing object detection. The dataset includes 288 video clips, totaling 261,908 video frames and 10,209 static images. It's worth noting that the data was gathered using various drone platforms under different scenes, weather conditions, and lighting conditions. It also includes special situations such as scene visibility, object categories, and occlusion scenarios.

C. EXPERIMENTAL METRICS

In order to assess our approach's performance, the mean average precision(mAP) is adopted as the evaluation criterion. The precision (P) represents the ratio of accurately predicted samples to total samples, while the recall rate (R) refers to the ratio of correctly predicted samples to all actual positive samples, and they can be formulated as follows:

$$P = \frac{TP}{TP + FP} \times 100\%, \tag{6}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{7}$$

where TP signifies true positive, and FP refers to false positive. FN corresponds to false negative. The average precision (AP) and mean average precision (mAP) are defined below:

$$AP = \int_0^1 P(R)dR, \tag{8}$$

$$mAP_{0.5} = \frac{1}{N}\sum_{i=1}^{N} AP_i, \text{Iou} = 0.5 \tag{9}$$

$$mAP_{0.5:0.95} = \frac{1}{N}\sum_{i=1}^{N} AP_i, \text{Iou} = 0.5 : 0.05 : 0.95 \tag{10}$$

where N signifies the total number of categories.

D. ABLATION STUDY

In order to assess the effectiveness of our proposed YOLO-GE approach, we carried out three sets of ablation experiments, utilizing YOLOv8-n as the baseline for comparison. In the first set of experiments, to evaluate the performance of the backbone improvements made to the GE-HGNet backbone through the integration of the G-HG block, we conducted three experiments. Initially, in the first experiment, YOLOv8-n served as the benchmark for comparative analysis. In the subsequent experiment, we verified the effectiveness of the G-HGNet backbone, which was engineered by substituting the standard convolutions within the HG block with ghost convolutions. Finally, we proceeded to assess the backbone of GE-HGNet, which is developed based on the G-HG block.

Table 2 provides a detailed comparison of model performance utilizing different backbones. The baseline model showcases good performance, with a $mAP_{0.5}$ of 85.7% and a $mAP_{0.5:0.95}$ of 61.4%. Nevertheless, the introduction of G-HGNet has slightly reduced the model's detection accuracy. However, GE-HGNet has realized further improvements upon the G-HG block, matching the baseline's precision at 88.8% while significantly enhancing the recall to 81.1%. In addition, it outperforms both the baseline and G-HGNet in $mAP_{0.5}$ and $mAP_{0.5:0.95}$, achieving 86.2% and 62.4% respectively. This validates the feasibility of improvements made with the G-HG block and the GE-HGNet backbone, particularly in enhancing the feature extraction capabilities of the backbone, along with its outstanding capability in detecting multi-scale objects.

During the second set of experiments, to validate the efficacy of the E-SimAM enhancements, we also carried out three experiments. The initial experiment once again utilized YOLOv8-n as the reference baseline for comparison analysis. In the second experiment, we introduced the SimAM attention mechanism and incorporated it within the model's neck structure to evaluate its performance. For the third experiment, we implemented and tested the model with the refined E-SimAM attention mechanism to validate its efficacy.

Table 3 presents the comparison result of model performance incorporating different attention mechanisms. After incorporating the SimAM attention mechanism, precision slightly improved to 88.9%, yet there was a slight decrease in recall and mAP. This observation confirms the analysis previously discussed, indicating that this attention mechanism tends to overlook small objects with lower resolution or weaker features, leading to a reduction in detection accuracy. However, compared to SimAM, the E-SimAM approach experienced a minor reduction in precision to 88.7%, yet it notably enhanced the recall to 80.7%. Additionally, the model also achieved significant enhancements in $mAP_{0.5}$ and $mAP_{0.5:0.95}$, with 86.3% and 62.3%, respectively. This demonstrates the efficacy of E-SimAM in preserving the detection accuracy with low resolution or subtle characteristics. Despite a slight increase in computational parameters, the improvement in detection performance highlights the potential of E-SimAM in enhancing models for the challenge of detecting small objects.

In the final series of ablation research, we aimed to measure the performance impact of three proposed enhancements, including the GE-HGNet for the backbone, the E-SimAM attention mechanism, and refinements to the fusion modules within both the neck structure and the detection head. Table 4 details the ablation results of various improvements evaluated on the DIOR dataset.

In Table 4, experiment 1 displays the results derived from the initial YOLOv8-n algorithm. In the second experiment, the adoption of the GE-HGNet maintains precision at 88.8%, significantly raises recall to 81.1%, and advances both $mAP_{0.5}$ and $mAP_{0.5:0.95}$ to 86.2% and 62.4%, respectively. Experiment 3 introduces the E-SimAM attention mechanism within the neck structure, resulting in a slight decrease in precision, but a notable increase in recall to 80.7%, along with $mAP_{0.5}$ and $mAP_{0.5:0.95}$ enhanced to 86.3% and 62.3%,

TABLE 2. Performance with different backbones evaluated on the DIOR dataset.

Method	P(%)	R(%)	$mAP_{0.5}$(%)	$mAP_{0.5:0.95}$(%)	Params(M)	GFLOPs(G)
Baseline	88.8	79.6	85.7	61.4	**3.0**	**8.2**
G-HGNet	88.6	79.1	85.6	61.5	3.2	9.6
GE-HGNet	**88.8**	**81.1**	**86.2**	**62.4**	3.4	9.9

TABLE 3. Performance with different attention mechanisms evaluated on the DIOR dataset.

Method	P(%)	R(%)	$mAP_{0.5}$(%)	$mAP_{0.5:0.95}$(%)	Params(M)	GFLOPs(G)
Baseline	88.8	79.6	85.7	61.4	**3.0**	**8.2**
SimAM	**88.9**	79.1	85.3	61.2	**3.0**	**8.2**
E-SimAM	88.7	**80.7**	**86.3**	**62.3**	3.2	9.2

TABLE 4. Performance with different enhancements evaluated on the DIOR dataset.

Id	GE-HGNet	E-SimAM	Head	P(%)	R(%)	$mAP_{0.5}$(%)	$mAP_{0.5:0.95}$(%)	Params(M)	GFLOPs(G)
1				88.8	79.6	85.7	61.4	**3.0**	**8.2**
2	✓			88.8	81.1	86.2	62.4	3.4	9.9
3		✓		88.7	80.7	86.3	62.3	3.2	9.2
4			✓	89.4	79.6	86.5	62.1	2.9	12.5
5	✓	✓		**89.5**	80.4	86.7	63.3	3.5	10.9
6	✓		✓	88.5	81.7	87.2	63.2	3.3	14.2
7		✓	✓	89.1	80.8	86.8	62.9	3.2	14.2
8	✓	✓	✓	88.4	**81.9**	**87.4**	**63.8**	3.5	15.9

respectively. In experiment 4, the fusion modules of the neck structure and detection head were enhanced, improving detection precision to 89.4%, while the recall rate remained constant. Additionally, the $mAP_{0.5}$ and $mAP_{0.5:0.95}$ were also enhanced to 86.5% and 62.1%, respectively. The aforementioned three experiments validate the effectiveness of our proposed improvements, with each improvement notably improving the $mAP_{0.5}$ and $mAP_{0.5:0.95}$.

In order to verify the effect of integrating different enhancements, we conducted experiments 5-8 as shown in Table 4. Experiment 5 integrates the GE-HGNet backbone and the E-SimAM attention mechanism. Experiment 6 then combines the GE-HGNet backbone with the fusion modules of the neck and head. In experiment 7, we integrate the E-SimAM attention mechanism with the fusion modules of the neck and head. The experimental results show that integrating these enhancements in pairs leads to additional advancements in $mAP_{0.5}$ and $mAP_{0.5:0.95}$. Notably, pairwise combinations of these improvements surpass the outcomes achieved by any single improvement, particularly in terms of improvements to both $mAP_{0.5}$ and $mAP_{0.5:0.95}$ metrics. In the final experiment, we integrate all improvements to assess the performance of our ultimate YOLO-GE approach. It is significantly observable that, despite a slight decrease in precision, the model achieved the highest detection accuracy, showcasing recall, $mAP_{0.5}$, and $mAP_{0.5:0.95}$ at 81.9%, 87.4%, and 63.8%, respectively. The results detailed in Table 4 validate the efficacy of our proposed improvements and illustrate that the organic combination of these improvements substantially improves the model's detection capabilities, confirming the rationality and efficacy of the YOLO-GE algorithm.

E. COMPARISON EXPERIMENTS

To further validate the performance and efficacy of our YOLO-GE approach, we conducted comparisons against multiple state-of-the-art methodologies. Firstly, we conducted comparisons of the evaluation metric curves throughout the training process. Then, we proceeded to compare several classic algorithms, including one-stage methods like SSD [24] and the YOLO series [15], [16], [17], [18], [19], [20], [21], [22], [23], alongside two-stage algorithms like Faster R-CNN [13]. Finally, to confirm the robustness of our YOLO-GE algorithm, comparison experiments were also conducted on the NWPU VHR-10 dataset.

Figure 7 illustrates the evaluation curves during training on the DIOR dataset. It's noteworthy that the $mAP_{0.5}$ and $mAP_{0.5:0.95}$ consistently outperform YOLOv8-n, exhibiting a sustained and steady upward trend as iterations increase. Moreover, our approach achieves a higher convergence speed and exhibits a stronger continual learning ability compared to YOLOv8n. Therefore, our YOLO-GE outperforms YOLOv8-n in terms of detecting precision and convergence performance.

Table 5 presents the performance comparison results evaluated from the DIOR dataset. Compared to other algorithms with similar parameter sizes, YOLO-GE exhibits superior performance in terms of detection accuracy. In contrast to lightweight models such as YOLOv5-n, YOLOv6-n,

TABLE 5. Performance comparison results evaluated on the DIOR dataset.

Network	Input size	$mAP_{0.5}$(%)	$mAP_{0.5:0.95}$(%)	Params(M)	GFLOPs(G)	Year
Faster RCNN	300×300	67.7	39.8	137.9	370.2	2015
SSD	300×300	68.0	43.5	26.3	62.7	2016
YOLOv5-n	640×640	84.2	58.8	**2.5**	**7.2**	2020
YOLOX-s	640×640	86.1	60.9	8.9	26.8	2021
YOLOv6-n	640×640	82.6	59.4	4.2	11.9	2022
YOLOv8-n	640×640	85.7	61.4	3.0	8.2	2023
YOLOv8-s	640×640	87.9	64.7	11.1	28.7	2023
YOLOv8-m	640×640	89.1	67.6	25.9	79.1	2023
YOLOv8-l	640×640	89.5	68.7	43.6	165.5	2023
YOLOv8-x	640×640	89.7	69.0	68.2	258.2	2023
YOLO-SSP[30]	640×640	87.3	64.7	28.3	72.3	2024
YOLO-GE-n	640×640	87.4	63.8	3.5	15.9	2024
YOLO-GE-s	640×640	89.0	66.5	13.0	50.7	2024
YOLO-GE-m	640×640	90.3	69.2	34.1	132.1	2024
YOLO-GE-l	640×640	**90.8**	**70.2**	56.3	263.6	2024

and YOLOv8-n, YOLO-GE-n achieves a $mAP_{0.50}$ of 85.7%, showing improvements of 3.8%, 5.8%, and 2.0%, respectively. Compared to the YOLOv8-x model, YOLO-GE-l achieves the best detection accuracy, with improvements of 1.2% and 1.7% in $mAP_{0.5}$ and $mAP_{0.5:0.95}$ respectively, while also reducing parameter size by 21.1%. Compared to the YOLO-SSP [30], YOLO-GE-s achieves 1.9% and 2.8% enhancements in $mAP_{0.5}$ and $mAP_{0.5:0.95}$ respectively, accompanied by 54.1% and 29.9% reduction in parameter size and GFLOPs.

To further compare with the YOLO series of algorithms, we carried out extensive experiments, and the findings are depicted in Figure 8. It is evident that the YOLO-GE algorithm outperforms other YOLO series models in terms of $mAP_{0.5}$ and $mAP_{0.5:0.95}$. Notably, YOLO-GE-l achieves the highest level of detection precision. Although there has been a certain increase in parameter size, YOLO-GE has achieved a significant improvement in detection accuracy.

Furthermore, to further evaluate the performance capabilities and robustness of our model, we conducted extensive experiments on the NWPU VHR-10 dataset. The performance comparison results are presented in Table 6. Our proposed YOLO-GE demonstrates superior performance compared to models with similar parameter sizes. Significantly, YOLO-GE-l maintains the highest detection precision, achieving $mAP_{0.5}$ and $mAP_{0.5:0.95}$ values of 92.3% and 61.0%, respectively.

Finally, we conducted extensive experiments on the more challenging VisDrone2019 dataset. It can be seen from Table 7 that our YOLO-GE-l achieves the highest detection accuracy in terms of P, R, $mAP_{0.5}$, and $mAP_{0.5:0.95}$. Compared to the state-of-the-art algorithms, our YOLO-GE-n achieves improvements of 20.1% and 22.2% in $mAP_{0.5}$ and $mAP_{0.5:0.95}$ respectively on the VisDrone2019 dataset. Compared to TA-YOLO-s [47], our YOLO-GE-s achieves 4.0% and 2.5% enhancements in $mAP_{0.5}$ and $mAP_{0.5:0.95}$ respectively, accompanied by a 6.5% reduction in parameter size. In addition, our YOLO-GE-s improves $mAP_{0.5}$ by 10.0% and 9.0% respectively, compared to BDH-YOLO [48] and PVswin-YOLOv8s [49]. These comparative experiments further validate that our proposed algorithm has significant advantages in terms of detection accuracy.

F. VISUALIZATION

The comparison of some detection examples between YOLO-GE-n and YOLOv8-n on the DIOR dataset is depicted in Figure 9. In airplane detection tasks, both algorithms can identify every instance within remote sensing imagery. However, the confidence scores for detections made by YOLO-GE-n are generally higher than that of YOLOv8-n. When it comes to identifying small targets such as cars, it is worth noting that YOLO-GE-n can accurately detect all instances, whereas YOLOv8-n had one instance of false detection and missed two instances. In the densely packed detection scenarios involving harbors, ships, and vehicles, both algorithms successfully identified all harbors and ships. However, with vehicles, YOLOv8-n missed one vehicle, whereas YOLO-GE-n accurately detected all vehicle targets. This confirms that YOLO-GE-n possesses superior detail feature extraction capabilities, enabling it to detect smaller targets more effectively.

In addition, in order to further verify the effectiveness and robustness of our YOLO-GE approach, we conducted another visualization experiment on the VisDrone2019 dataset. Figure 10 shows some representative detection results on the VisDrone2019-test dataset. From row 1 in Figure 10, it can be observed that our YOLO-GE-m can accurately detect various

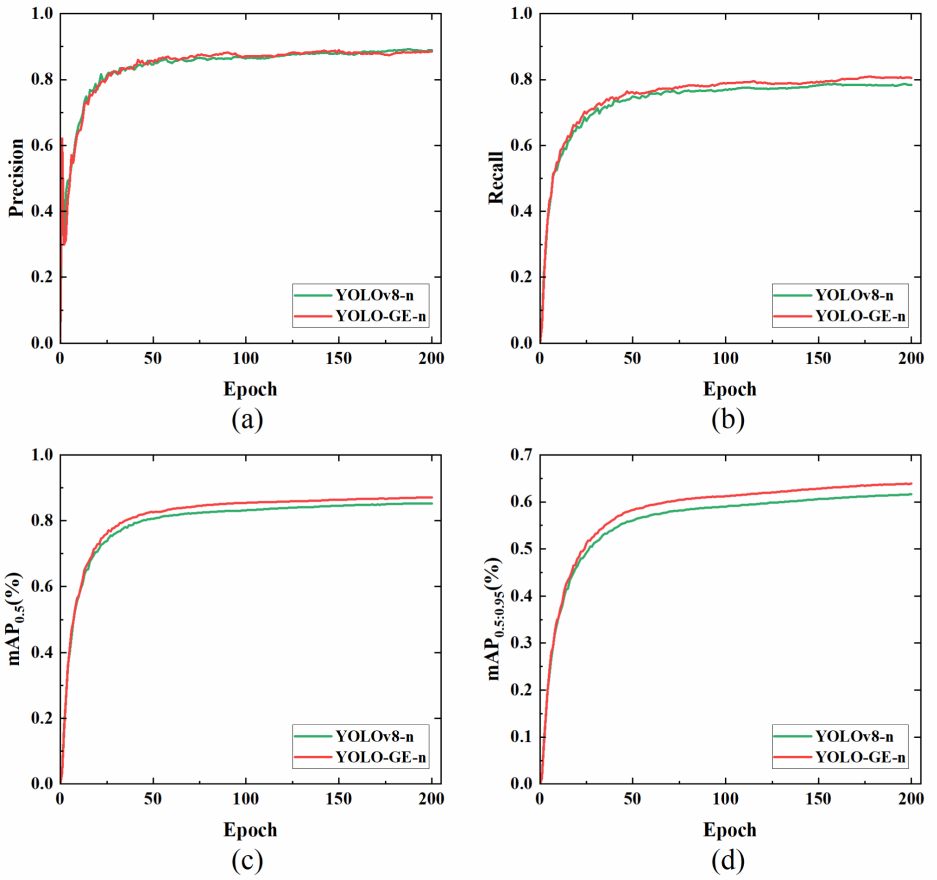


FIGURE 7. Comparison of evaluation indicators during training on the DIOR dataset;(a-d) refer to the comparison curves of precision, recall, mAP0.5, and mAP0.5:0.95 respectively.

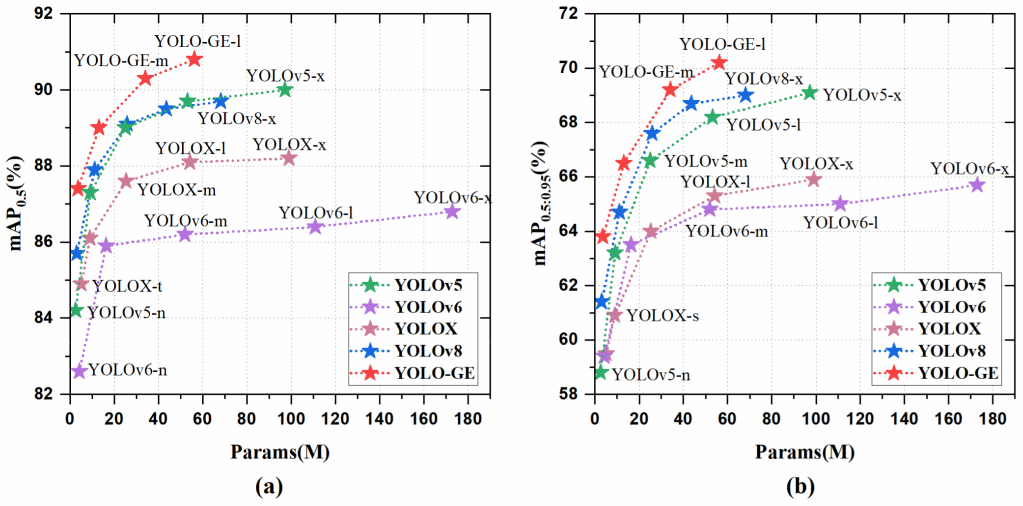
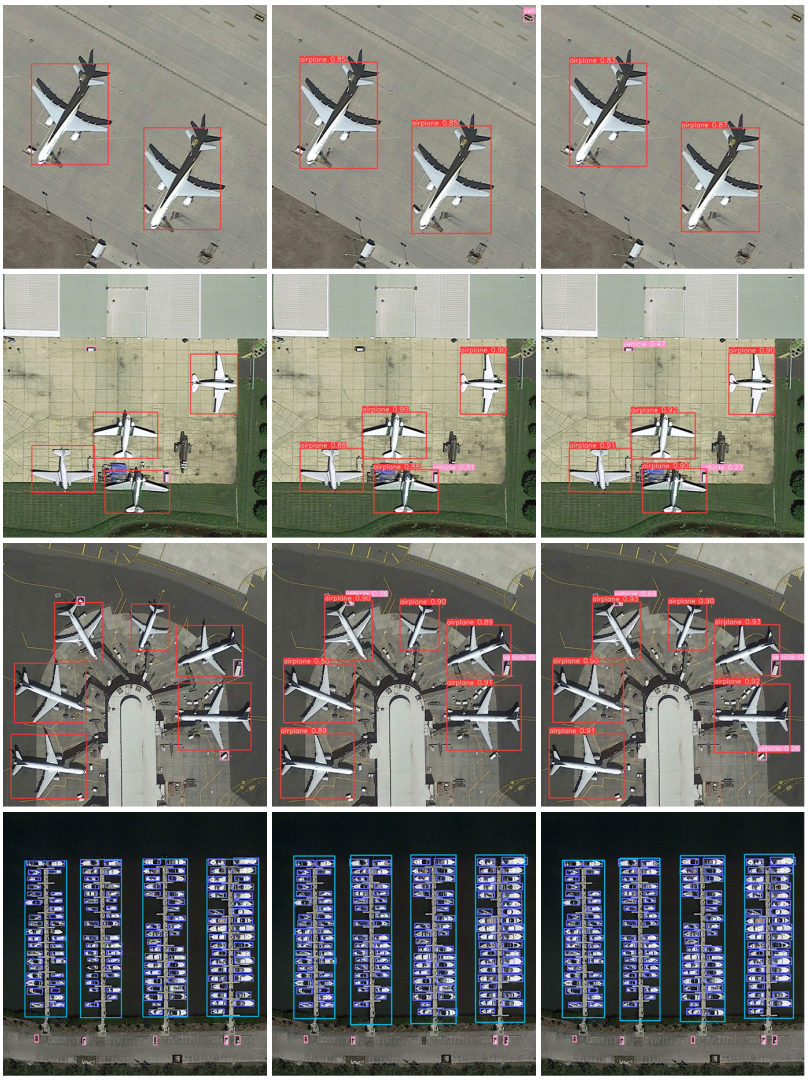


FIGURE 8. Performance comparison of YOLO-GE with YOLO series algorithms. (a) mAP0.5, (b) mAP0.5:0.95.

types of objects, even vehicles that are partially obscured by trees. As can be seen from row 2 and row 3 in Figure 10, our approach can also accurately detect objects of various scales and types even in poor lighting conditions.

V. DISCUSSION

Tables 2 and 3 present the performance of the backbone and attention mechanism enhancements evaluated on the DIOR dataset. The performance results reveal that the G-HGNet



(a) (b) (c)

FIGURE 9. Comparison of some detection examples on the DIOR dataset: (a) ground truth, (b) YOLOv8-n, (c) YOLO-GE-n.

TABLE 6. Performance comparison results evaluated on the NWPU VHR-10 dataset.

Network	Input size	$mAP_{0.5}$(%)	$mAP_{0.5:0.95}$(%)	Params(M)	GFLOPs(G)	Year
Faster RCNN	300×300	77.8	40.5	137.9	370.2	2015
SSD	300×300	70.5	38.4	26.3	62.7	2016
YOLOv5-n	640×640	87.9	54.5	**2.5**	**7.2**	2020
YOLOv5-m	640×640	89.9	58.2	25.1	64.4	2020
YOLOX-s	640×640	88.0	56.5	8.9	26.8	2021
YOLOX-l	640×640	89.7	59.2	54.2	155.8	2021
YOLOv6-s	640×640	89.8	58.3	16.3	44.2	2022
YOLOv6-m	640×640	89.9	57.0	52.0	161.6	2022
YOLOv8-s	640×640	91.0	59.1	11.1	28.7	2023
YOLOv8-l	640×640	91.8	60.8	43.6	165.5	2023
YOLO-GE-n	640×640	91.1	56.5	3.5	15.9	2024
YOLO-GE-s	640×640	91.8	58.9	13.0	50.7	2024
YOLO-GE-m	640×640	92.2	60.3	34.1	132.1	2024
YOLO-GE-l	640×640	**92.3**	**61.0**	56.3	263.6	2024

backbone and SimAM attention mechanism have slightly diminished the model's accuracy. However, the incorporation of GE-HGNet utilizes redundant feature maps and broadens the receptive field, thereby enhancing the extraction

TABLE 7. Performance comparison results evaluated on the VisDrone2019-val dataset.

Network	P(%)	R(%)	$mAP_{0.5}$(%)	$mAP_{0.5:0.95}$(%)	Params(M)	Year
YOLOv8-n	43.9	33.5	33.9	19.4	**3.0**	2023
YOLOv8-s	51.3	39.2	39.9	23.4	11.1	2023
YOLOv8-m	54.5	41.6	43.3	26.2	25.9	2023
YOLOv8-l	57.2	43.2	45.3	27.7	43.6	2023
YOLOv8-x	56.0	44.0	45.6	28.2	68.2	2023
TA-YOLO-n[47]	50.2	38.9	40.1	24.1	3.8	2024
TA-YOLO-s[47]	53.9	44.3	45.4	27.7	13.9	2024
TA-YOLO-m[47]	58.3	46.6	48.8	30.2	29.7	2024
BDH-YOLO[48]	-	41.6	42.9	26.2	9.4	2024
PVswin-YOLOv8s[49]	54.5	41.8	43.3	26.4	21.6	2024
YOLO-GE-n	49.3	41.2	40.7	23.7	3.5	2024
YOLO-GE-s	54.6	46.2	47.2	28.4	13.0	2024
YOLO-GE-m	57.8	49.1	50.4	30.8	34.0	2024
YOLO-GE-l	**59.8**	**50.0**	**52.6**	**32.6**	56.3	2024

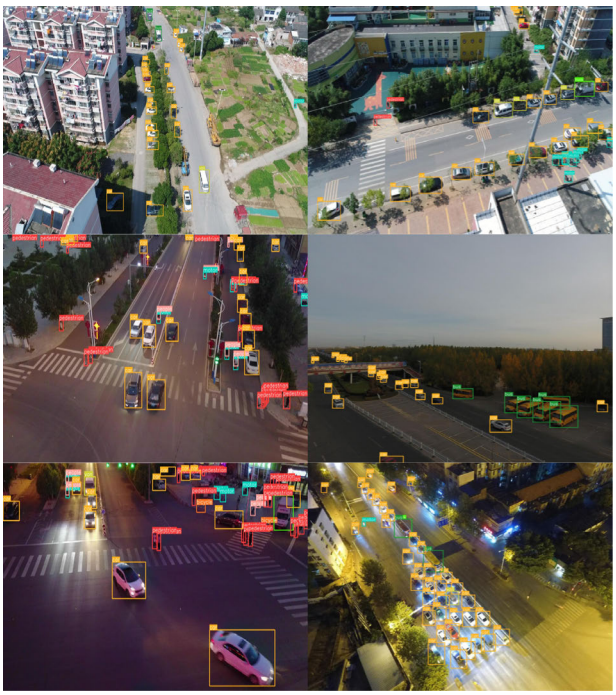


FIGURE 10. Some representative detection results on the VisDrone2019-test dataset.

of multi-scale objects and advanced semantic features in complex backgrounds. Additionally, the E-SimAM attention mechanism employs residual techniques to address the low-resolution feature loss limitation, thereby enhancing the precision in identifying low-resolution features against intricate backgrounds.

Table 4 demonstrates each enhancement contributes to the model's detection accuracy. Notably, when the three improvements are combined, our proposed YOLO-GE achieves the highest detection accuracy, surpassing algorithms that use only a single improvement. This indicates that our proposed YOLO-GE algorithm effectively combines the benefits of all three enhancement methods, thereby significantly improving the model's overall performance.

Tables 5, 6, and 7 illustrate the comparison results between YOLO-GE and other algorithms evaluated on the diverse datasets. The overall performance of our YOLO-GE outperforms all other detection algorithms, even the current state-of-the-art remote sensing target detection algorithms such as BDH-YOLO and PVswin-YOLOv8s. In addition, our approach adapts well to different scenarios, especially excelling in the more challenging VisDrone2019 dataset, where it exhibits a significant improvement in detection accuracy compared to YOLOv8. This may be attributed to our enhanced approach's strong ability to extract features from densely packed small objects, thus improving the detection accuracy.

Figures 9 and 10 show some representative detection results evaluated on DIOR and VisDrone2019 datasets. The results indicate that our YOLO-GE can accurately detect objects of various scales and types even in densely packed detection scenarios or poor lighting conditions.

In summary, extensive experiments across diverse datasets have fully demonstrated the effectiveness and robustness of our proposed enhancements.

VI. CONCLUSION

With the widespread application of remote sensing technology across diverse industries, remote target detection holds crucial strategic significance for aerial remote sensing technology. To address the obstacles of multi-scale object detection within remote sensing and further enhance detection accuracy, we propose YOLO-GE based on YOLOv8. We introduce three proposed enhancements, including the GE-HGNet for the backbone, the E-SimAM attention

mechanism, and refinements to the fusion modules within both the neck structure and the detection head. Through extensive experimental validation on the DIOR, NWPU VHR-10, and VisDrone2019 datasets, we have fully demonstrated the effectiveness and robustness of our suggested enhancements. It has been proven that the integrated application of these improvements significantly boosts YOLO-GE's capability in accurately detecting small objects in remote sensing images, demonstrating both its effectiveness and rationality. In particular, the YOLO-GE-l outperforms current mainstream algorithms in achieving the highest detection precision, while also having fewer model parameters compared to several larger models. Given that YOLO-GE features several models with superior detection accuracy, it allows for the selection of a suitable model based on the application scenario, presenting wide-ranging application potential.

The introduction of the GE-HGNet backbone and E-SimAM attention mechanism in YOLO-GE has improved the model's detection accuracy, but it has also increased the number of model parameters and floating-point operations, thereby increasing the computational burden on the system. Therefore, it is relatively challenging to deploy it on embedded devices with limited computational resources. In future work, we aim to enhance the YOLO-GE algorithm by employing techniques such as network pruning and knowledge distillation to achieve parameter compression, while maintaining detection accuracy, while maintaining detection accuracy. Additionally, we also plan to boost the model's detection performance by improving approaches related to localization or classification loss functions.

REFERENCES

[1] J. Feng, J. Wang, and R. Qin, "Lightweight detection network for arbitrary-oriented vehicles in UAV imagery via precise positional information encoding and bidirectional feature fusion," *Int. J. Remote Sens.*, vol. 44, no. 15, pp. 4529–4558, Aug. 2023, doi: 10.1080/01431161.2023.2197129.

[2] Y. Shen, D. Liu, F. Zhang, and Q. Zhang, "Fast and accurate multi-class geospatial object detection with large-size remote sensing imagery using CNN and truncated NMS," *ISPRS J. Photogramm. Remote Sens.*, vol. 191, pp. 235–249, Sep. 2022, doi: 10.1016/j.isprsjprs.2022.07.019.

[3] H. Zhang, X. Zhang, G. Meng, C. Guo, and Z. Jiang, "Few-shot multi-class ship detection in remote sensing images using attention feature map and multi-relation detector," *Remote Sens.*, vol. 14, no. 12, p. 2790, Jun. 2022, doi: 10.3390/rs14122790.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[5] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015, doi: 10.1007/s11263-014-0733-5.

[6] T. Xie, W. Han, and S. Xu, "YOLO-RS: A more accurate and faster object detection method for remote sensing images," *Remote Sens.*, vol. 15, no. 15, p. 3863, Aug. 2023, doi: 10.3390/rs15153863.

[7] Z. Zhou and Y. Zhu, "KLDet: Detecting tiny objects in remote sensing images via Kullback–Leibler divergence," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4703316, doi: 10.1109/TGRS.2024.3382099.

[8] G. Wang, Y. Zhuang, H. Chen, X. Liu, T. Zhang, L. Li, S. Dong, and Q. Sang, "FSoD-Net: Full-scale object detection from optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602918, doi: 10.1109/TGRS.2021.3064599.

[9] D. Yu and S. Ji, "A new spatial-oriented object detection framework for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4407416, doi: 10.1109/TGRS.2021.3127232.

[10] G. Peng, Z. Yang, S. Wang, and Y. Zhou, "AMFLW-YOLO: A lightweight network for remote sensing image detection based on attention mechanism and multiscale feature fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4600916, doi: 10.1109/TGRS.2023.3327285.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2013, *arXiv:1311.2524*.

[12] R. Girshick, "Fast R-CNN," 2015, *arXiv:1504.08083*.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[14] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," 2017, *arXiv:1712.00726*.

[15] 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, doi: 10.1109/CVPR.2016.91.

[16] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.

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

[18] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[19] *YOLOv5*. Accessed: Mar. 15, 2023. [Online]. Available: https://github.com/ultralytics/yolov5

[20] 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*.

[21] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[22] *YOLOv8*. Accessed: Nov. 15, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

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

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Single shot MultiBox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 21–37.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017, *arXiv:1708.02002*.

[26] Z. Zakria, J. Deng, R. Kumar, M. S. Khokhar, J. Cai, and J. Kumar, "Multiscale and direction target detecting in remote sensing images via modified YOLO-v4," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1039–1048, 2022, doi: 10.1109/JSTARS.2022.3140776.

[27] 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, doi: 10.1109/ACCESS.2023.3233964.

[28] J. Lin, Y. Zhao, S. Wang, and Y. Tang, "YOLO-DA: An efficient YOLO-based detector for remote sensing object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023, doi: 10.1109/lgrs.2023.3303896.

[29] S. Xie, M. Zhou, C. Wang, and S. Huang, "CSPPartial-YOLO: A lightweight YOLO-based method for typical objects detection in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 388–399, 2024, doi: 10.1109/jstars.2023.3329235.

[30] Y. Liu, D. Yang, T. Song, Y. Ye, and X. Zhang, "YOLO-SSP: An object detection model based on pyramid spatial attention and improved downsampling strategy for remote sensing images," *Vis. Comput.*, vol. 40, pp. 1–18, 2024.

[31] PaddlePaddle. (2023). *HGNetv2*. Accessed: Nov. 15, 2023. [Online]. Available: https://github.com/PaddlePaddle/PaddleDetection/blob/develop/ppdet/modeling/backbones/hgnet_v2.py

[32] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "CSPNet: A new backbone that can enhance learning capability of CNN," 2019, *arXiv:1911.11929*.

[33] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," 2019, *arXiv:1911.11907*.

[34] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," 2023, *arXiv:2304.08069*.
[35] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," 2017, *arXiv:1707.01083*.
[36] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," 2021, *arXiv:2111.14556*.
[37] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.
[38] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," 2020, *arXiv:2010.03045*.
[39] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, in Proceedings of Machine Learning Research, vol. 139, Jul. 2021, pp. 11863–11874.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.
[41] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016, *arXiv:1612.03144*.
[42] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," 2018, *arXiv:1803.01534*.
[43] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 01, 2020, doi: 10.1016/j.isprsjprs.2019.11.023.
[44] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016, doi: 10.1016/j.isprsjprs.2016.03.014.
[45] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014, doi: 10.1016/j.isprsjprs.2014.10.002.
[46] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7380–7399, Nov. 2022, doi: 10.1109/TPAMI.2021.3119563.
[47] M. Li, Y. Chen, T. Zhang, and W. Huang, "TA-YOLO: A lightweight small object detection model based on multi-dimensional trans-attention module for remote sensing images," *Complex Intell. Syst.*, vol. 10, pp. 5459–5473, May 2024, doi: 10.1007/s40747-024-01448-6.
[48] J. Sui, D. Chen, X. Zheng, and H. Wang, "A new algorithm for small target detection from the perspective of unmanned aerial vehicles," *IEEE Access*, vol. 12, pp. 29690–29697, 2024, doi: 10.1109/access.2024.3365584.
[49] N. U. A. Tahir, Z. Long, Z. Zhang, M. Asim, and M. ELAffendi, "PVswin-YOLOv8s: UAV-based pedestrian and vehicle detection for traffic management in smart cities using improved YOLOv8," *Drones*, vol. 8, no. 3, p. 84, Feb. 2024, doi: 10.3390/drones8030084.

MIN YUE received the master's degree in bionics and robotics from Beijing Institute of Technology, in 2007. Since 2007, he has been with Shanghai University of Engineering Science. He is currently an Experimentalist with the School of Mechanical and Automotive Engineering. His research interests include computer vision, artificial intelligence, and intelligent control systems.

LIQIANG ZHANG received the Ph.D. degree in mechanical manufacturing and automation from Shanghai Jiao Tong University, Shanghai, China, in 2008. He is currently a Professor with the School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science. His research interests include intelligent manufacturing systems, aviation equipment digital twins, and artificial intelligence.

YUJIN ZHANG (Member, IEEE) received the Ph.D. degree in communication and information systems from Shanghai Jiao Tong University, Shanghai, China, in 2014. He is currently an Associate Professor with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science. His research interests include multimedia forensics, signal processing, artificial intelligence, and pattern recognition.

HAIFENG ZHANG received the bachelor's degree in computer science and technology and the master's degree in computer application technology from Chang'an University, in 2000 and 2004, respectively, and the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, in 2007. From 2017 to 2018, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, Kansas State University. He is currently an Associate Professor with the School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science. His research interests include machine vision technology and applications and image processing. His research results have been published in various domestic and international journals.