TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 39/39 pp1263−1281
DOI: 10.26599/TST.2025.9010091
Volume 31, Number 2, April 2026

Research Article

Drone-Based High-Precision Object Detection in Remote Sensing with Attention-Guided Feature Fusion

Hanxiang Wang, Yanfen Li*, Yuanke Zhang, Junliang Shang, Guangshun Li, Liem Dinh-Tien, L. Minh Dang, Hyoung-Kyu Song, and Hyeonjoon Moon*

Abstract: Small object detection in remote sensing imagery is a challenging task due to the small size of targets, complex background, and low contrast, which makes achieving high precision difficult. To enhance the accuracy of detection, this study proposes a novel oriented object detection model with three significant innovations: Firstly, a lightweight feature extraction network is designed to achieve efficient feature representation at a reduced computational cost, which is particularly effective for the recognition of small targets in remote sensing imagery. Secondly, a Feature-Focused Channel Attention (FFCA) is introduced that enhances the model's ability to focus on small target areas by combining spatial and channel attention, enhancing the model's capacity to capture and represent features more effectively. Lastly, an attention-guided multi-scale feature fusion module is proposed to integrate features from different levels, which substantially boosts the model's ability to accurately detect small-scale objects, especially in remote sensing scenarios with vast fields of view and complex backgrounds. The experimental outcomes validate that our model achieves the best detection performance on two benchmark public datasets for remote sensing imagery, confirming its effectiveness and practicality in remote small object detection tasks.

Key words: remote sensing; oriented object detection; attention mechanism; small object recognition

1 Introduction

Small target detection in the field of computer vision presents a challenging problem, particularly in the analysis of drone or satellite imagery. Factors, such as the small size of targets, low image resolution, and low contrast between targets and background, make detection particularly difficult. Recently, as deep learning technologies have progressed, a multitude of approaches have been introduced by researchers to

• Hanxiang Wang, Yanfen Li, Yuanke Zhang, Junliang Shang, and Guangshun Li are with School of Computer Science, Qufu Normal University, Rizhao 276826, China. E-mail: hanxiang@qfnu.edu.cn; yanfen@qfnu.edu.cn; yuankezhang@qfnu.edu.cn; shangjunliang110@163.com; guangshunli@qfnu.edu.cn.

• Liem Dinh-Tien is with Faculty of Fundamental Sciences, Van Lang University, Ho Chi Minh City 70000, Vietnam. E-mail: liem.dt@vlu.edu.vn.

• L. Minh Dang is with Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam; Faculty of Information Technology, Duy Tan University, Da Nang 550000, Vietnam; and Department of Information and Communication Engineering, and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea. E-mail: minhdl@sejong.ac.kr.

• Hyoung-Kyu Song is with the Department of Information and Communication Engineering, and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea. E-mail: songhk@sejong.ac.kr.

• Hyeonjoon Moon is with Department of Computer Science and Engineering, Sejong University, Seoul 05006, Republic of Korea. E-mail: hmoon@sejong.ac.kr.

∗ To whom correspondence should be addressed.

Manuscript received: 2025-02-05; revised: 2025-04-03; accepted: 2025-05-08

tackle this challenge. For example, feature enhancement techniques[1] have been employed to improve the detection performance of small targets, rotation-invariant methods[2] have been proposed to solve the problem of rotating target detection, and attention mechanisms[3] and depth estimation[4] have been introduced to enhance the accuracy and robustness of small target detection. Rotation target detection algorithms in drone or satellite image analysis hold significant research importance and application value across multiple domains. Primarily, in the military and defense sector, rotation target detection can be utilized for monitoring and identifying enemy military installations, equipment, and activities[5, 6]. Secondly, in civilian fields, this technology can be applied to urban planning, agricultural monitoring, environmental protection, and disaster assessment[7, 8]. For instance, by analyzing drone or satellite images, one can detect and count buildings in cities, crop diseases in farmlands, or the extent of affected areas in disaster zones. Furthermore, rotation target detection technology can also enhance the efficiency of traffic management[9], such as monitoring traffic flow through drones and identifying traffic accidents.

Despite the significant progress made in the field of small target detection within computer vision, there are still numerous challenges in ensuring detection accuracy while enhancing the speed of algorithm execution. Particularly in the application of remote sensing image analysis, small targets, due to their limited pixel occupation in the image, are often difficult to accurately identify and are susceptible to interference from complex background noise[10]. Moreover, to capture a broader field of view, remote sensing images typically have high resolutions, which means a substantial increase in the amount of data that needs to be processed. To improve detection speed, researchers strive to simplify the model structure and reduce computational load. For example, methods such as designing lightweight network architectures[11], applying efficient convolutional operations[12], and optimizing algorithmic processes[13] have been used to expedite the detection process. However, such simplifications may come at the cost of detection accuracy, especially in the feature extraction and classification of small targets. Therefore, the research focus is on developing a lightweight feature extraction network that can maintain high-precision detection capabilities while sustaining rapid processing capabilities.

In the pursuit of enhancing small target detection accuracy, scholars have been committed to developing and refining various advanced feature enhancement techniques. Among these, technologies, such as Feature Pyramid Networks (FPNs)[14] and multi-scale feature fusion[15], are widely applied to strengthen the model’s ability to recognize targets at different scales. By constructing multi-layer feature representations, these methods help the model capture richer and more discriminative image features, thereby improving detection accuracy to a certain extent. However, these deep network structures and complex feature extraction mechanisms also increase the computational burden of the model, especially in real-time remote sensing image analysis tasks. In addition, the direct fusion of multi-scale features can sometimes lead to feature dilution issues[16], which is particularly prominent in small target detection, as the features of small targets are easily lost during the fusion process, affecting the model’s detection precision. Therefore, this study designs a new feature fusion strategy to ensure that the key features of small targets are retained and efficiently integrated with feature information from different layers.

Additionally, the attention mechanism, as a low-cost feature enhancement technique, has been widely applied in computer vision tasks. It emulates the focusing characteristics of human vision, enabling the model to concentrate more on key areas within the image, thereby enhancing feature expressiveness and the model’s recognition accuracy. However, despite the significant performance improvements brought by the attention mechanism, it also has certain limitations. For instance, when dealing with large-scale data or high-dimensional features, the attention mechanism may cause the model to overly rely on certain local features, thus neglecting global contextual information[17]. Moreover, within attention models, the computation of attention weights is a crucial process that determines the extent to which the model focuses on key areas in the image. However, some models use simplified dimensionality reduction techniques in the computation process, which may inappropriately lose spatial information of the features[18]. This unreasonable dimensionality reduction approach may lead to the model’s inability to accurately capture the spatial details of targets, affecting the accuracy of detection.

Based on these issues, a multidimensional feature hybrid attention module has been introduced to further enhance the model's precision.

Considering the above limitations, the key contributions of this research can be summarized as follows.

(1) A high-precision, light weight feature extraction network is designed to address the balance between the model's speed and accuracy.

(2) The development of an efficient attention module that enhances the capability of feature expression, ensuring that the model can more accurately identify and focus on relevant information within the image data.

(3) The introduction of an innovative feature fusion strategy aimed at mitigating the issue of feature dilution during the fusion of multi-scale features.

The structure of the ensuing sections in this paper is outlined below. Section 2 delivers an extensive overview of the literature, focusing on the detection of oriented dense objects. Section 3 outlines the proposed system's flowchart and details the methodology employed. Section 4 details the data information and annotation formats utilized in the study. Section 5 then showcases a series of experiments aimed at highlighting the contributions made in this research. Finally, Section 6 concludes the paper by discussing the current performance and potential avenues for future investigation.

2 Related Work

Deep object detection models are becoming increasingly important in remote sensing data processing, especially in the research fields of computer vision. Due to the high resolution, extensive coverage, and the presence of multi-scale targets typically found in remote sensing imagery, object detection in remote sensing imagery through deep learning encounters numerous difficulties. These include significant variations in target size, complex background environments, and limited computational resources. To address these challenges, researchers have developed a series of deep learning architectures and technologies. For example, complex feature extraction networks, multi-specification feature fusion modules, and flexible attention mechanisms.

Within the object detection models, the backbone network for feature extraction is a crucial component, responsible for extracting key information from raw pixel data. Currently, various lightweight network structures, such as MobileNets[19], ShuffleNets[20], GhostNet[21], Vision Transformers (ViTs)[22], and MultiLayer Perceptron (MLP) architectures[23], have been utilized as feature extractors. These structures not only maintain a fast-processing speed but also achieve considerable recognition accuracy. However, compared to some more complex network structures, these lightweight models may sacrifice some accuracy. Some researchers have attempted to reduce computational costs using methods, such as separable spatial convolution[24] and knowledge distillation[25]. However, these approaches often rely on specific data distributions and may struggle to maintain stable performance when faced with significant variations in data. To better adapt to the characteristics of remote sensing imagery, researchers have also designed various network structures with different depths and widths, such as ResNet[26], Inception[27], EfficientNet[28], YDHNet[29], and ConvNeXt[30], to capture multi-scale and multi-level features within the images. While these network structures can provide more accurate feature representation, they also come with higher computational costs. Therefore, designing a deep feature extraction network that balances model accuracy and speed is essential. This requires us to maintain detection accuracy while also considering the computational efficiency and practical applicability of the model.

Multi-scale feature fusion is a pivotal strategy for enhancing the performance of target detection in remote sensing imagery. Given the variability in target sizes within this type of imagery, features from a single scale often struggle to comprehensively represent all targets. To address this, researchers have introduced various feature fusion networks and strategies, such as FPN[31], Multi-scale Feature Fusion (MFF)[32], and Adaptive Scale Feature Fusion (ASFF)[33]. These approaches amalgamate feature maps from different levels and scales, bolstering the model's ability to detect both small and large targets. Additionally, to further boost the detection performance of rotating small targets in remote sensing imagery, researchers have proposed several variants of the FPN structure. For instance, the Bidirectional FPN (BiFPN)[34] and Path Aggregation Network (PANet)[35] facilitate bidirectional exchange of feature information, allowing for the transfer of information from higher to lower layers and vice versa, enabling a more profound level of feature integration. The Graph FPN (GraphFPN)[36] leverages graph-based structures to effectively construct feature pyramids, enhancing the efficacy of object detection algorithms for multi-scale targets in various imaging contexts.

Furthermore, the High Resolution FPN (HRFPN)[37] focuses on preserving high-resolution information during the feature fusion process, which is particularly critical for small target detection. Despite the advancements of FPN and its derivatives in multi-scale feature fusion, challenges remain regarding the loss of valuable spatial information during the feature transfer and fusion processes. This is especially problematic for the identification of minute targets in satellite imagery, where spatial context is crucial for the precise detection and positioning of targets.

Attention modules play a crucial role in remote sensing image target detection. Highly flexible attention mechanisms allow models to adaptively focus on the most informative parts of the image, thereby improving detection accuracy. In the detection of targets within satellite imagery, attention mechanisms are divided into three main types. Spatial attention models enhance detection performance by highlighting key areas of the image, such as Receptive-Field Attention (RFA)[38] and non-local neural networks[39]. Channel attention models strengthen feature expression capabilities, with Squeeze-and-Excitation Networks (SENet)[40] and Efficient Channel Attention (ECA)[41] being typical examples. Channel-spatial attention models combine the advantages of the former two, such as the Convolutional Block Attention Module (CBAM)[42] and the Coordinate Attention (CA)[43]. These models significantly improve the precision and reliability of identifying small targets in remote sensing images by guiding the network to focus on the most relevant information in different dimensions. However, they still face some challenges and limitations in practical applications. For example, spatial attention models may neglect the global dependency of image features, while channel attention models may not be sufficient to deal with complex spatial context information. In addition, although channel-spatial attention models integrate channel and spatial information, their computational costs are often high, and they may require a large number of parameters to capture subtle feature relationships. Therefore, attention modules need a simple and effective design to reduce computational overhead and capture key channel and spatial features.

3 Methodology

Figure 1 comprehensively illustrates the research framework we have developed for the detection of rotating targets in remote sensing imagery. This framework processes a large-scale dataset encompassing a variety of scenarios, and we have customized the format of the annotated data to suit the needs of our proposed algorithm. Our Faster You Only Look Once (F-YOLO) model, which is based on the YOLOv8[44] architecture, integrates three key

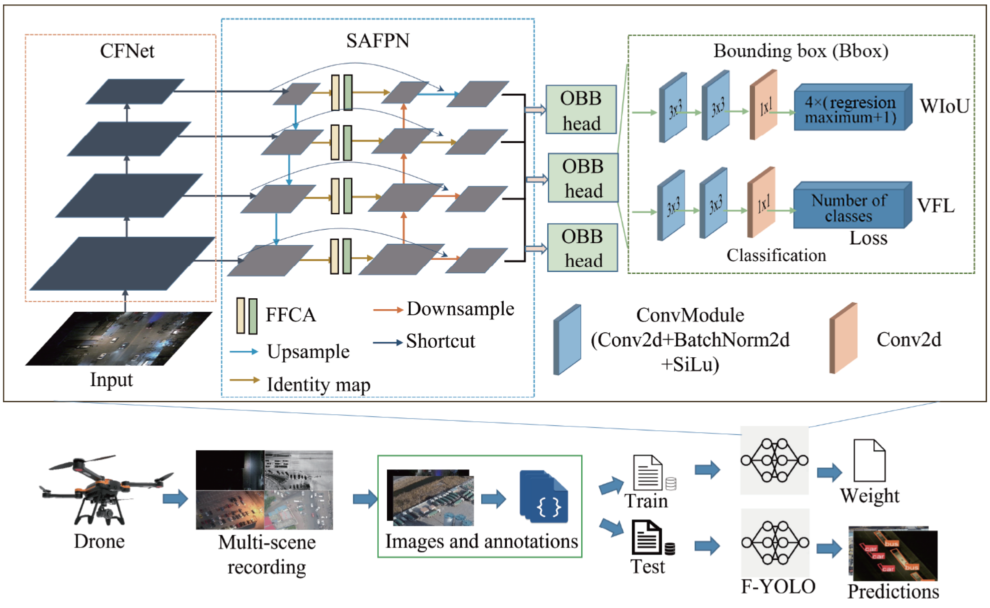


Fig. 1 Diagram of the novel oriented object detection framework (OBB: oriented bounding box dection).

innovations: firstly, a lightweight yet high-precision feature extraction network (as detailed in Section 3.1), designed to maintain detection accuracy while reducing computational load through an elegant network architecture; secondly, a computationally efficient attention module (also detailed in Section 3.1), which enhances the model's feature representation by focusing on critical areas of the image; and finally, an attention-guided feature enhancement module (as detailed in Section 3.2), designed to act as the "neck" of the model, bolstering its ability to detect targets across various scales. In addition, we have made fine-tuned adjustments to the model's loss function (as detailed in Section 3.3) to optimize detection performance. The synergy of these technologies provides an efficient and accurate solution for the detection of rotating targets in remote sensing imagery.

3.1 Feature extraction network

YOLO occupies a pivotal position in the field of object detection, with its unique advantage of being able to directly predict the position and category of objects in an image during a single forward propagation. This characteristic allows YOLO to surpass other object detection methods in terms of processing speed and accuracy. YOLOv8's capabilities are largely attributed to its efficient and robust feature extraction network, C2f-CSPDarkNet[44], which combines Cross Stage Partial (CSP) technology with DenseNet's skip connections, significantly enhancing the network's feature expression and utilization. While C2f-CSPDarkNet excels in feature extraction, it has some limitations. Its extensive use of DenseNet-like skip connections, although beneficial for feature propagation, also increases the model's memory footprint, affecting inference speed and overall efficiency. To tackle these challenges, this study conducted an in-depth exploration and innovative redesign of the feature extraction module for the YOLOv8 model, focusing on precision and speed.

In the research process, we conducted a detailed analysis of the lightweight model FasterNet's[45] structure and operation. FasterNet maintains satisfactory accuracy in detection while diminishing the model's intricacy and parameter volume with its simple and efficient network structure, providing valuable insights for the design of YOLOv8's feature extraction network. Based on a comprehensive consideration of C2f-CSPDarkNet and FasterNet, this study proposes a new feature extraction network architecture. This architecture aims to inherit the powerful feature extraction capabilities of C2f-CSPDarkNet and incorporates the lightweight design philosophy of FasterNet to achieve more efficient memory usage and faster inference speed. Below is a concise introduction to C2f-CSPDarkNet, FasterNet, and the proposed feature extraction network.

3.1.1 C2f-CSPDarkNet

The earlier generations of the YOLO series adopted CSPDarkNet as the backbone network, with the core of this network lying in the CSP structure. This innovative network design aims to optimize the flow of features and the propagation of gradients. The CSP structure splits the input feature map into two parts, one of which is directly passed to the next layer, while the other part undergoes a series of convolutional layers for processing. The processed feature map is then merged with the unprocessed part, a design that helps reduce the vanishing gradient problem and enhances the feature expression capability. Although the backbone network of YOLOv8 also refers to the structure of CSPDarkNet (as shown in Fig. 2), it does not directly adopt the network structure of CSPDarkNet. Instead, YOLOv8 uses the C2f (CSPLayer_2Conv) module to replace some of the cumbersome modules in CSPDarkNet (as marked in the yellow box in Fig. 2).

The C2f structure reduces the count of input channels in each bottleneck, thereby lowering the model's parameter quantity and computational intricacy. This design allows YOLOv8 to maintain high performance while achieving a more lightweight structure. It is worth noting that the C2f module places greater emphasis on preserving spatial information, which is particularly important for object detection tasks that involve handling smaller objects or scenarios with high spatial detail requirements. In comparison, the original module focuses more on capturing contextual information, suitable for dealing with larger objects and complex backgrounds. This architectural refinement is designed to balance the performance and lightweight needs of the model, enabling YOLOv8 to achieve good detection results in various application scenarios.

3.1.2 FasterNet and CFNet (proposed)

FasterNet is an innovative neural network that concentrates on enhancing the operational speed and efficiency of models while maintaining or enhancing

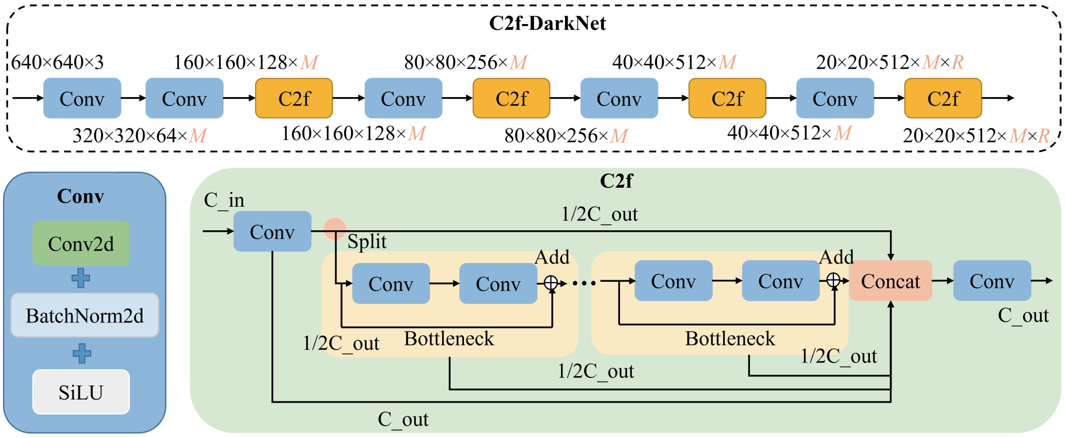
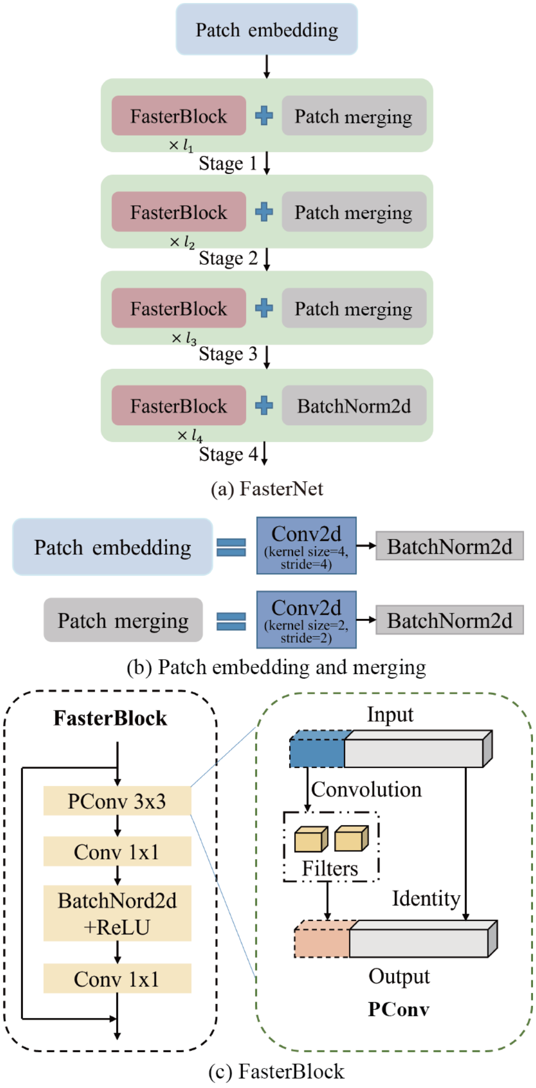


Fig. 2 Structure of C2f-DarkNet. *M* and *R* refer to the width multiple and ratio, respectively. C_in and C_out are the input and output channels, respectively.

performance. Its core innovation lies in the introduction of Partial Convolution (PConv), a novel convolutional operation that significantly reduces computational volume and memory access by only performing convolution on a subset of the input channels, thereby improving overall computational efficiency (as shown in green dotted box in Fig. 3). The network structure of FasterNet consists of multiple levels (as shown in Fig. 3a), each containing embedding layers or merge layers for spatial down-sampling and channel expansion, as well as a series of FasterBlocks that include PConv and Point-Wise Convolution (PWConv, Conv 1 × 1). These blocks utilize an inverted residual block design, effectively reusing input features through shortcut connections to optimize feature propagation. To further reduce latency, FasterNet employs only normalization and activation layers after each intermediate PWConv, and selects GELU or ReLU as the activation function based on the size variant of the model. FasterNet offers a range of model size variants (adjusted by the depth parameter *l*), including tiny, small, medium, and big, to accommodate different computational needs and application scenarios, demonstrating its potential for high-speed operation and efficient performance across various devices.

Although FasterNet has achieved significant results in improving the speed and efficiency of model operation, it still has some potential drawbacks and room for improvement. Due to the design of FasterNet focusing on reducing computational complexity, this can affect the model's transferability on some complex tasks. Additionally, the computational complexity of FasterNet still needs to be further reduced as its



(a) FasterNet

(b) Patch embedding and merging

(c) FasterBlock

Fig. 3 Architecture of FasterNet. It consists of three modules: patch embedding, patch merging, and FasterBlock (l_{1-4} are depth coefficients).

inference speed is much slower compared to C2f-CSPDarkNet. To address the mentioned challenges and improve the backbone's ability to learn multi-scale features, we propose a new architecture called CFNet, which integrates C2f-CSPDarkNet and FasterNet. Its structure is illustrated in Fig. 4. CFNet introduces a novel residual module, CFBlock, which leverages the T-shaped convolution strategies of PConv and PWConv from FasterNet while incorporating an optimized feature splitting and fusion mechanism. Additionally, the Feature-Focused Channel Attention (FFCA) module is designed to enhance CFBlock's multidimensional feature extraction capabilities, further improving the model's efficiency and representational power.

Figure 4 illustrates the specific structure of FFCA (the red dotted box). Initially, the input feature map $U \in \mathbf{R}^{C\times H\times W}$ is processed through two methods: Local Average Pooling (LAP) and Global Average Pooling (GAP), where C denotes the channel count, and H and W denote the height and width, respectively. LAP and GAP can be expressed as follows:

$$\mathrm{LAP}(U) = \frac{1}{N}\sum_{i\in\Omega} U(i) \tag{1}$$

$$\mathrm{GAP}(\mathrm{LAP}(U)) = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathrm{LAP}(U)(h, w) \tag{2}$$

where Ω represents the local receptive field, with N denoting the count of elements it encompasses. The term $\mathrm{LAP}(U)(h, w)$ refers to the local pooling outcome at the specific coordinate (h, w). By integrating both local and global pooling operations, the model effectively captures a comprehensive range of features, from intricate local details to broad global patterns, thereby establishing a robust basis for further feature manipulation and analysis. Subsequently, the features extracted through local and global pooling are subjected to a transformation via a 1D-convolutional layer. This step is instrumental in not only diminishing the complexity of the feature set by reducing its dimensionality but also in distilling the most salient information from the data. Following this transformation, the features derived from the 1D convolution are rearranged through a reshaping process. This reshaping is a pivotal step that ensures the features are appropriately configured to meet the demands of the subsequent stages in the model's architecture and operation.

The feature maps output from the local and global attention branches are combined and added together to compute the attention matrix. FFCA determines the importance of each channel by calculating its attention score,

$$\mathrm{Att} = \sigma(\mathrm{Conv1D}(\mathrm{GAP}(\mathrm{LAP}(U))) + \mathrm{Conv1D}(\mathrm{LAP}(U))) \tag{3}$$

where Att represents the attention score for channels, and $\sigma(\cdot)$ is the activation function that confines the scores within the range of 0 and 1. Ultimately, by

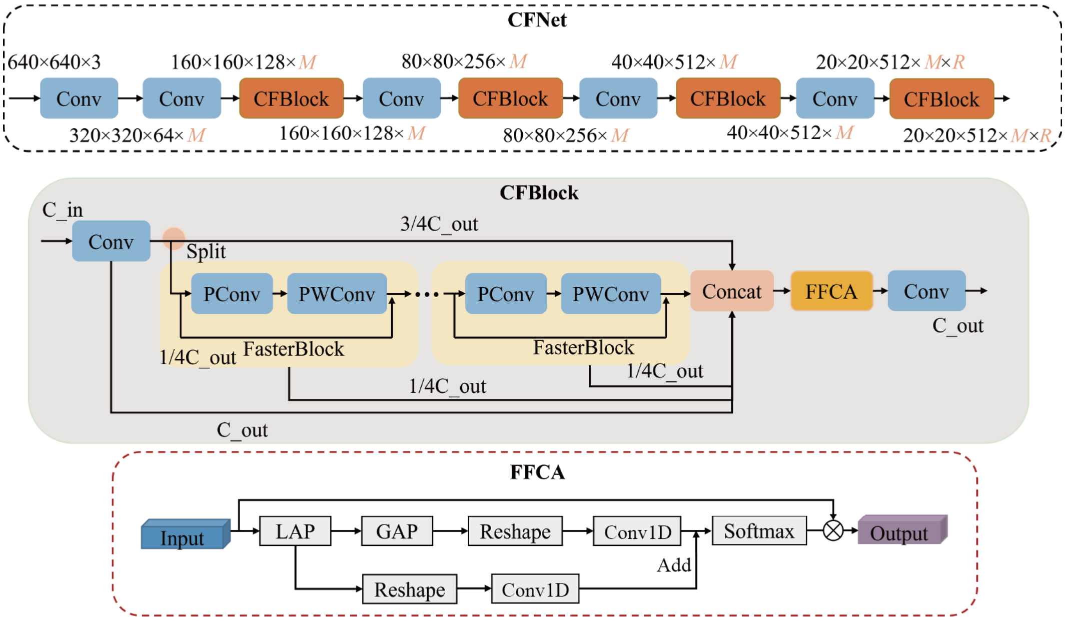


Fig. 4 Architecture of the proposed CFNet.

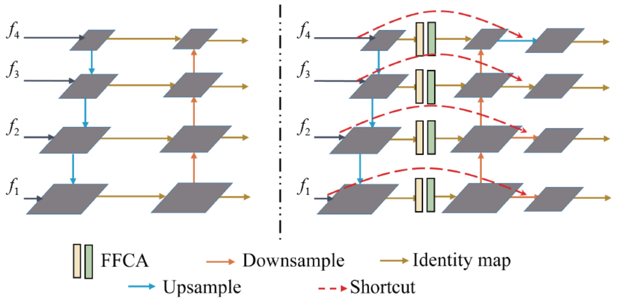
integrating both local and global attention scores, the original feature maps are weighted and merged. This can be achieved in the following equation:

$$\text{FFCA}(U) = \text{Att} \otimes U \tag{4}$$

FFCA adaptively highlights informative channels while suppressing irrelevant ones, enabling more efficient use of network capacity. By jointly leveraging local and global cues, it balances fine-grained details with global context, thus enhancing feature discrimination.

3.2 Semantic augmentation feature pyramid network

The feature fusion mechanism plays an essential role in enhancing the performance of detectors. In the feature maps extracted by the backbone network, low-level features excel at capturing precise target location information but are relatively lacking in semantic context; on the other hand, high-level features are rich in semantic information but are not as precise in detailing localization. YOLOv8 effectively integrates these multi-scale features through the Path Aggregation Feature Pyramid Network (PAFPN)[35], thereby strengthening the network's feature representation and descriptive capabilities. As shown in Fig. 5a, the PAFPN structure employs a dual-path design that goes from top-down (indicated by blue arrows) and bottom-up (indicated by orange arrows) to achieve bidirectional feature enhancement. The top-down path utilizes upsampling to leverage the abundant semantic information from higher levels, while the bottom-up path extracts low-level features that contain detailed target information. The target contour information contained in these low-level features is crucial for the model's accurate target localization. PAFPN achieves an effective complement between high-level semantic information and low-level detailed information in this manner. However, during the fusion of features at different scales, there may be a dilution of semantic information, and the direct merging of features at different scales may lead to a blending effect[15].



(a) PAFPN (b) SAFPN

Fig. 5 Architecture of PAFPN and our proposed SAFPN. The SAFPN integrates the proposed FFCA, a multi-dimensional feature attention module, along with strategic shortcuts into its architecture. This enhancement endows the model with an augmented feature representation capability, particularly excelling in the detection of small targets. f_i denotes the feature map, $i = 1, 2, 3, 4$.

To address the aforementioned challenges, a novel feature fusion network named as the Semantic Augmentation Feature Pyramid Network (SAFPN) is introduced. As shown in Fig. 5b, four different scale features extracted by backbone are input into FFCA. The attention module computes spatial attention and channel attention for the features separately, focusing on specific regions within the feature maps for spatial attention, while evaluating the importance of each channel for channel attention. The feature maps, after being weighted by attention, undergo a series of convolutional operations and non-linear activations to further refine and enhance the features. During the construction of the feature pyramid, as the size of the feature maps decreases, higher-level features, while possessing stronger semantic information, often lose spatial details. Through shortcut connections, the high-resolution features from lower levels can be combined with the semantic features from higher levels, reducing the loss of information. Therefore, the output feature maps of the SAFPN structure are fused with the original input features through a shortcut structure (indicated by the red dashed line in Fig. 5). This process not only retains the rich semantic information from higher levels but also incorporates the precise detail information from lower levels, which is particularly helpful for the detection of small objects. The fusion of multi-level features provides a more abundant and useful feature representation for subsequent detection tasks.

3.3 Loss function

In the design of object detection models, classification loss and bounding box regression loss are two core components. The classification loss is primarily responsible for measuring the discrepancy between the predicted categories and the ground truth, thereby adjusting network parameters to enhance the model's recognition capabilities. However, when dealing with specific datasets captured by drones, there is a large number of easy-to-classify negative samples. These samples may cause the model to focus excessively on easily identifiable negative samples, neglecting the learning of more challenging samples.

To counter the impact of a substantial quantity of simple negative instances, the VariFocal Loss (VFL)[46] is employed in this study to calculate the classification loss. VFL is an advanced loss function utilized within the field of object detection to enhance the model's capability in identifying dense objects. It is an improvement based on focal loss and incorporates Intersection over Union (IoU)-Aware Classification Scores (IACS), a scoring mechanism that combines the likelihood of object existence and localization accuracy. VFL allows the model to perform more precise sorting when dealing with a large number of candidate detection boxes. VFL reduces the influence of negative samples and enhances the importance of positive samples by unequally weighting the positive and negative samples, especially for those hard-to-classify positive samples. This calculation method helps the model to concentrate on learning more critical instances. The computation of VFL for positive and negative samples is represented by Eqs. (5) and (6), respectively,

$$\mathrm{VFL}(p,q)_{\text{positive}} = -q(q\log(p)+(1-q)\log(1-p)) \tag{5}$$

$$\mathrm{VFL}(p,q)_{\text{negative}} = -\alpha p^{\gamma}\log(1-p) \tag{6}$$

where p represents the classification probability predicted by the model, q denotes the IoU score of the object, hyperparameter α is used to balance the weight of negative instances, and the adjustment factor γ is employed to regulate the shape of the loss function, focusing on hard-to-classify positive samples.

In rotated object detection, the design of the bounding box loss function is crucial for model performance. However, when collecting data using drones, low-quality training samples are often generated due to altitude issues. Traditional IoU loss has some problems when dealing with low-quality training samples, as these samples can negatively impact the model's learning because they typically have low IoU scores, leading to overfitting to these samples during training. This study addresses the class imbalance issue by adopting an advanced bounding box regression loss function, Wise-IoU (WIoU)[47]. WIoU introduces the concept of outlier degree (od) to evaluate the quality of anchor boxes. The od is calculated based on the IoU quality metric. High-quality anchor boxes have smaller od values. The core of WIoU lies in the dynamic non-monotonic gradient gain assignment strategy, where the gradient gain g can be calculated based on the o (od) of the anchor box,

$$g(o) = \frac{1}{a+e^{-\delta\cdot(o-T)}} \tag{7}$$

where δ is a hyperparameter that controls the rate of change of the gradient gain, and T is a threshold used to determine the dynamic classification standard for anchor box quality. The WIoU loss function combines the IoU loss with the gradient gain, and its formula is as follows:

$$\mathrm{WIoU} = -\rho\cdot g(o)\cdot L_{\mathrm{IoU}} \tag{8}$$

where ρ is a hyperparameter used to adjust the impact of the gradient gain, and L_{IoU} represents the loss value based on IoU.

4 Oriented Object Detection Dataset

In this study, we utilize two public datasets: DroneVehicle dataset[48] and DOTA dataset[49], each comprising a training set, a validation set, and a test set for training and validating the performance of oriented object detection models. These datasets contain a diverse range of object categories, including vehicles, ships, planes, and storage tanks, with objects appearing at various scales and orientations. Notably, they include a large number of small object instances, making them well-suited for evaluating the model's ability to detect small targets in complex aerial imagery.

The DroneVehicle dataset is crafted to cater to the unique challenges of object detection in aerial imagery, focusing on 5 vehicles (car, truck, bus, van, and freight_car). Comprising 17 990 images designated for the training set and 8980 images for the test set, along with a validation set of 1469 images, it presents a comprehensive range of vehicle types across diverse environmental settings. The dataset enables models to refine their detection capabilities under various real-world conditions, including fluctuating light exposure and instances of occlusion. The training set facilitates the learning and fine-tuning processes of the model, the validation set serves to refine and regulate the learning parameters, and the test set is crucial for assessing the model's ability to generalize from previously unseen data.

The DOTA dataset is a large-scale collection for object detection in aerial imagery, created to train and test models for detecting targets in high-resolution aerial images. Spanning 15 square kilometers, this dataset comprises more than 188 000 labeled instances within 15 object categories, with their abbreviations detailed in Table 1. It is segmented into three subsets: a training subset with around 1411 images, a validation subset containing 458 images, and a test subset with approximately 938 images. The image resolutions vary from 800 pixel × 800 pixel up to 4000 pixel × 4000 pixel. For convenience in data annotation reading and subsequent processing, the dataset annotations are offered in both textual and XML formats. Figure 6 displays a selection of example images from both datasets.

5 Experimental Result

5.1 Experimental environment

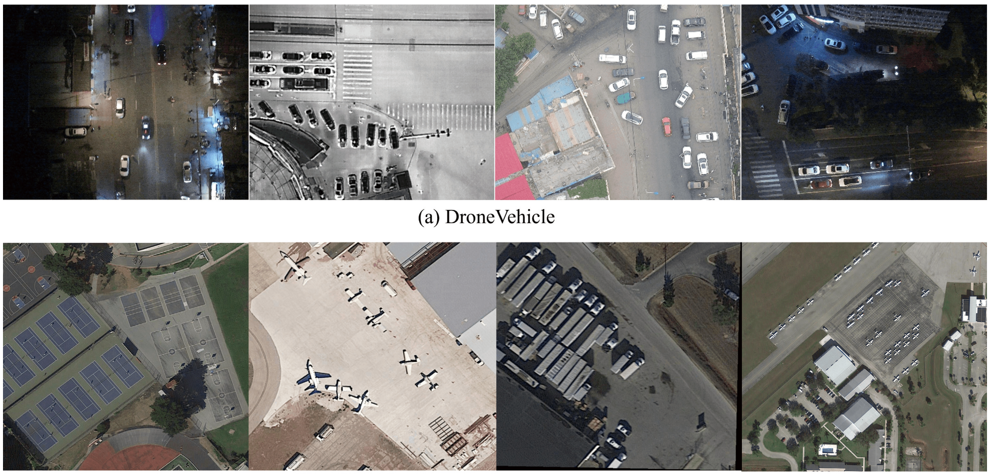
In this research, we have configured a high-performance computing environment, centered around two NVIDIA GeForce RTX 4090 GPUs that provide substantial parallel processing power for deep learning and computer vision tasks. The system is equipped with the latest Intel Core i9 processor for efficiently handling complex computations and coordinating GPU operations. We have also installed high-speed memory and Solid State Drives (SSDs) to accelerate data read/write speeds. On the software front, we are using an optimized 64-bit operating system, CUDA Toolkit 11.8, and the cuDNN library to enhance GPU computations. Development and model training are conducted using Python 3.9 and PyTorch 2.1, while numerical computations are handled by NumPy and SciPy. Data visualization is facilitated by Matplotlib and Seaborn. This combination of hardware and software provides a robust and efficient platform for our research.

5.2 Feature extraction

The first experiment is conducted to stress the excellent performance of the proposed backbone model by making a scientific comparison with other related backbones. Table 2 lists the detailed results of these models based on the YOLOv8 framework and PAFPN neck part. In this section, the performance is evaluated in terms of both efficiency and effectiveness. As for the model’s efficiency, the proposed CFNet (n) is the lightest model with the minimum network parameter

Table 1 Abbreviations for 15 categories in DOTA dataset.

Category	Abbreviation	Category	Abbreviation	Category	Abbreviation
Bridge	BR	Small vehicle	SV	Basketball court	BC
Harbor	HA	Large vehicle	LV	Soccer-ball field	SB
Ship	SH	Baseball diamond	BD	Roundabout	RA
Plane	PL	Ground track field	TF	Swimming pool	SP
Helicopter	HC	Tennis court	TC	Storage tank	ST



(a) DroneVehicle

(b) DOTA

Fig. 6 Examples of two public datasets.

Table 2 Performance comparison of rotational object detection on the DroneVehicle dataset using different backbone models baset on the YOLOv8 framework and PAFPN net part. Note: Metrics include parameter scale (Param.), floating point operations (FLOPs), mean Average Precision (mAP and mAP$_{50}$), and inference speed per image. The bold style represents the best performing data.

Backbone	Param.	FLOPs	mAP (%)	mAP$_{50}$ (%)	Inference (ms)
C2f-CSPDarkNet (n)	3.08×10^{6}	8.4×10^{9}	53.3	74.6	0.32
C2f-CSPDarkNet (s)	1.141×10^{7}	2.94×10^{10}	57.8	79.3	0.64
FasterNet (T2)	1.526×10^{7}	3.73×10^{10}	55.9	77.5	0.80
CFNet (n) (proposed)	**2.38×10^{6}**	**6.7×10^{9}**	54.8	76.2	**0.29**
CFNet (s) (proposed)	8.61×10^{6}	2.26×10^{10}	58.6	80.5	0.58
CFNet (s)+ FFCA (proposed)	8.63×10^{6}	2.34×10^{10}	**59.5**	**81.2**	0.60

scale (2.38 × 10^{6}) and computations (6.7 × 10^{9}). Owing to its light character, CFNet (n) achieves fast inference speed of 0.29 ms per image, which is 0.51 ms faster than the FasterNet (T2). For the model's effectiveness, the presented another CFNet (s) with FFCA obtained the highest mAP of 59.5% and mAP$_{50}$ of 81.2%. That may be because the advantages of FFCA, which enhances feature selection and context awareness. Compared with the initial C2f-CSPDarkNet (n), our model is considerably improved by the mAP of 6.2% and the mAP$_{50}$ of 6.6%. This experiment reflects that the proposed model achieves a 25% reduction in parameter count and a 23% decrease in FLOPs compared to the original model, and the designed backbone has leading feature extraction ability for the drone vehicle data.

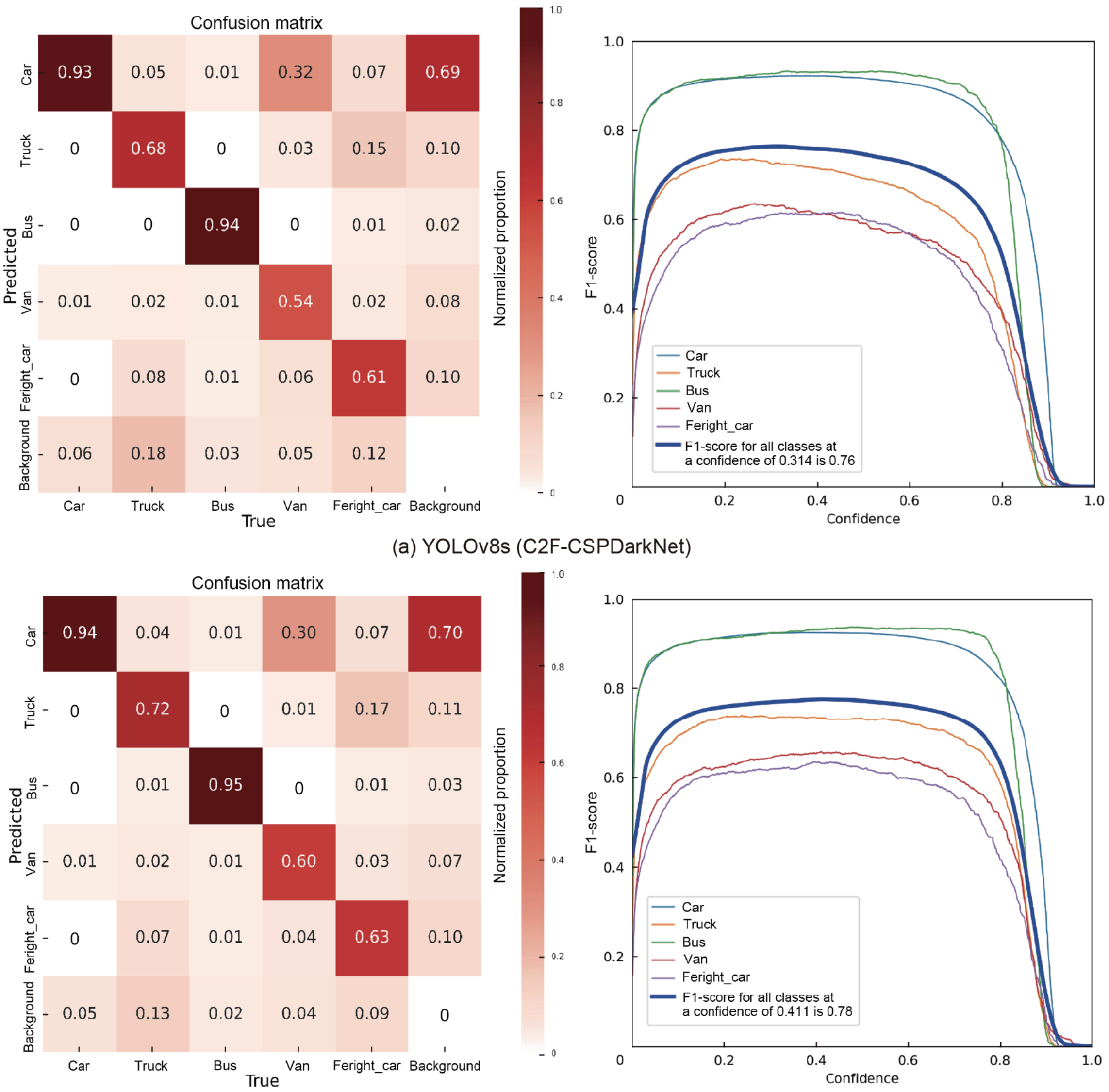
Figure 7 shows the corresponding confusion matrix and F1-confidence curves for 5 classes (car, truck, bus, van, and feright_car). According to the comparison results, the CFNet with FFCA can detect each class with higher precision, especially for the van class. Also, the under F1-confidence curves hve a smoother trend than the upper curve, which indicates the proposed model is more stable in the process of feature extraction.

In addition, the visualized feature extraction results of C2f-CSPDarkNet and CFNet+FFCA are compared to show the advantages of FFCA. Figure 8 illustrates the results of two models in both daytime and night scenes. This experiment suggests the designed model can capture more discriminative features from the input images compared to another experimental model.

5.3 Feature fusion

During the feature fusion process, different neck models are configured with different structural settings. Table 3 presents the object detection performances of the four experimental neck models based on the uniform backbone. The main modified components are listed from the second column to the fourth column, and various evaluation results are shown in the last six columns. Even though the SAFPN with FFCA has the fewest parameters (8.55 × 10^{6}) and computations (2.24 × 10^{10}), its detection effectiveness does not achieve satisfactory expected results. However, the SAFPN with both modifications performs the first place among all the experimental models, which obtains the precision of 0.831, the recall of 0.779, the F1-score of 0.804, and the mAP$_{50}$ of 0.823.

Figure 9 presents a comparative experimental analysis of two models engaged in an object detection task, with a particular focus on their architectural differences in the neck structure. The first model utilizes PAFPN as the neck component, designed to achieve multi-scale feature fusion, while the second model incorporates the proposed SAFPN structure. Figure 9 illustrates the detection outcomes for each model across various scenarios, including those captured under both nocturnal and diurnal lighting conditions. A closer examination reveals that Fig. 9a displays the detection results for F-YOLO with PAFPN, while Fig. 9b corresponds to those of F-YOLO with SAFPN. It is observable that within identical testing environments, the SAFPN model delivers more precise detection in certain instances. For example, in the top set of images, the SAFPN model more accurately identifies the vehicles within the areas encircled by black frames, whereas the PAFPN model exhibits several missed detections and false positives. In the middle set, the SAFPN model also demonstrates enhanced stability in detecting grayscale images. The bottom set of images further illustrates that the SAFPN model's detection accuracy for the vehicles marked with yellow circles surpasses that of the PAFPN model.



Confusion matrix

Predicted \ True	Car	Truck	Bus	Van	Feright_car	Background
Car	0.93	0.05	0.01	0.32	0.07	0.69
Truck	0	0.68	0	0.03	0.15	0.10
Bus	0	0	0.94	0	0.01	0.02
Van	0.01	0.02	0.01	0.54	0.02	0.08
Feright_car	0	0.08	0.01	0.06	0.61	0.10
Background	0.06	0.18	0.03	0.05	0.12	

(a) YOLOv8s (C2F-CSPDarkNet)

Confusion matrix

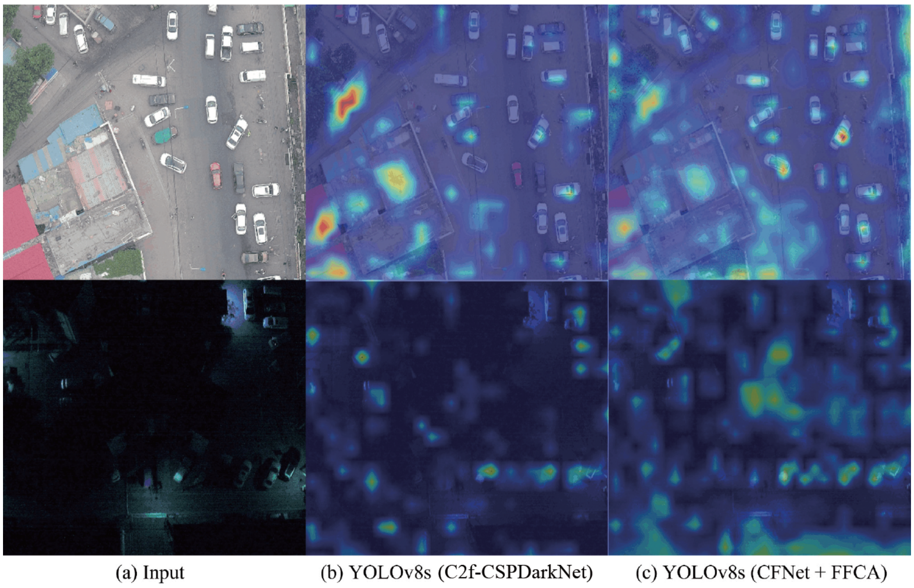
Predicted \ True	Car	Truck	Bus	Van	Feright_car	Background
Car	0.94	0.04	0.01	0.30	0.07	0.70
Truck	0	0.72	0	0.01	0.17	0.11
Bus	0	0.01	0.95	0	0.01	0.03
Van	0.01	0.02	0.01	0.60	0.03	0.07
Feright_car	0	0.07	0.01	0.04	0.63	0.10
Background	0.05	0.13	0.02	0.04	0.09	0

(b) YOLOv8s (CFNet + FFCA)

Fig. 7 Comprehensive performance evaluation of YOLOv8 with two feature extraction models on DroneVehicle dataset: Analysis via confusion matrix and F1-confidence curve.

These comparisons underscore the superior performance and robustness of the SAFPN structure in handling multi-scale feature fusion, providing higher detection accuracy across diverse settings. This suggests that SAFPN possesses distinct advantages in feature selection and preservation of spatial information, contributing to the overall enhancement of the model’s detection capabilities.

To demonstrate the strengths of F-YOLO in small object detection, comparative tests are conducted on another public dataset, DOTA. The experimental outcomes are presented in Fig. 10, which displays a bar chart comparing the performance of the proposed F-YOLO with YOLOv8s, specifically focusing on the Average Precision for each category. In the chart, the blue bars represent the performance of YOLOv8s, while the orange bars denote the performance of F-YOLO. The horizontal axis of the chart is labeled with abbreviations for various categories. The vertical axis indicates the average precision, ranging from 0 to 1. A visual analysis of the chart reveals that F-YOLO exhibits higher average precision in the majority of categories. For instance, in categories like PL, BD, TF, and SV, F-YOLO demonstrates a significantly superior detection accuracy over YOLOv8s. Moreover, in certain categories where the two models show similar performance, F-YOLO maintains a slight edge. F-YOLO shows a higher detection accuracy,



(a) Input (b) YOLOv8s (C2f-CSPDarkNet) (c) YOLOv8s (CFNet + FFCA)

Fig. 8 Comparing feature visualization results: YOLOv8s (C2f-CSPDarkNet) vs. proposed YOLOv8s (CFNet + FFCA).

Table 3 Object detection performance of models with different pyramid structures using the CFNet(s) backbone. "√" represents the use of corresponding module, "×" represents the absence of the module. The bold style represents the best performing data.

PAFPN[35]	SAFPN			Param.	FLOPs	Precision	Recall	F1-score	mAP_{50}
	FFCA	CBAM[42]	Shortcut	(× 10^6)	(× 10^{10})				(%)
√	×	×	×	√	√	0.810	0.756	0.782	81.2
×	√	×	×	**8.55**	**2.24**	0.827	0.763	0.794	81.8
×	×	√	×	8.58	2.26	0.823	0.76	0.79	81.4
×	×	×	√	8.56	2.26	0.822	0.762	0.791	81.5
×	×	×	√	8.59	2.29	**0.831**	**0.779**	**0.804**	**82.3**

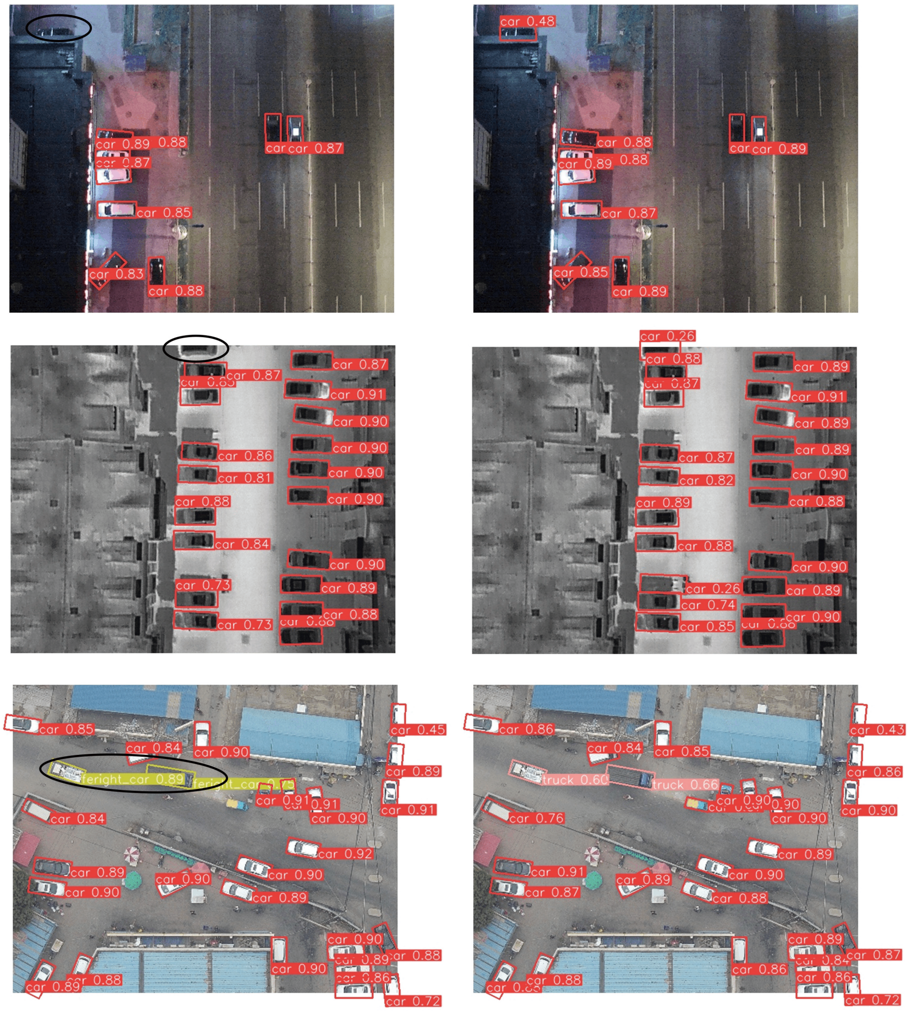
suggesting that the proposed enhancements have effectively improved the model's detection capabilities and robustness. The comparative data illustrate that F-YOLO possesses greater adaptability and precision when dealing with a variety of complex targets.

5.4 Model optimization and comparison

Figure 11 presents an in-depth comparative analysis of the performance between YOLOv8s and the proposed F-YOLO model under various loss functions. In Fig. 11a, F-YOLO demonstrates a consistently higher mAP at an IoU threshold of 0.5, stabilizing towards the end of training and outperforming YOLOv8s. Figure 11b refines this trend, illustrating F-YOLO's superior convergence and ultimate precision across the full range of IoU from 0.50 to 0.95. Figures 11c and 11d break down the training process losses, with Fig. 11c focusing on the box loss. F-YOLO not only shows a rapid decrease in loss but also maintains a lower level in the later stages of training, indicating an advantage in precise bounding box regression. Figure 11d, which concentrates on the classification loss, also shows F-YOLO achieving a quick reduction in loss and sustaining a low loss value in the later training phase, highlighting its high efficiency in object classification.

Figures 11e and 11f focus on performance during the validation phase. Figure 11e indicates that F-YOLO maintains a lower box loss during validation, reflecting its excellent generalization capabilities. The classification loss chart on Fig. 11f corroborates this, demonstrating F-YOLO's stability and accuracy throughout the validation set. Synthesizing the data from Fig. 11, F-YOLO shows an advantage over YOLOv8s on both mAP and loss performance metrics.

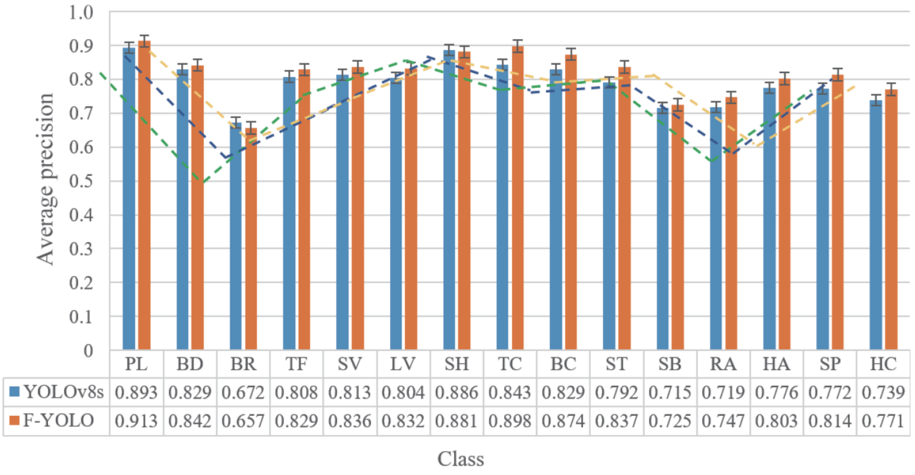
Figure 12 demonstrates the capability of the F-YOLO model in detecting small targets within the DOTA test set. The image displays eight satellite photographs, each highlighting small targets with distinct colored bounding boxes. Upon inspection, it is clear that F-YOLO can accurately identify a variety of



(a) F-YOLO (PAFPN)

(b) F-YOLO (SAFPN)

Fig. 9 Object detection performance of two models in complex environments on DroneVehicle test set. Black circles indicate some missed detections and misidentifications.



	PL	BD	BR	TF	SV	LV	SH	TC	BC	ST	SB	RA	HA	SP	HC
■ YOLOv8s	0.893	0.829	0.672	0.808	0.813	0.804	0.886	0.843	0.829	0.792	0.715	0.719	0.776	0.772	0.739
■ F-YOLO	0.913	0.842	0.657	0.829	0.836	0.832	0.881	0.898	0.874	0.837	0.725	0.747	0.803	0.814	0.771

Fig. 10 Precision comparison of models (YOLOv8s and F-YOLO) on the DOTA test dataset.

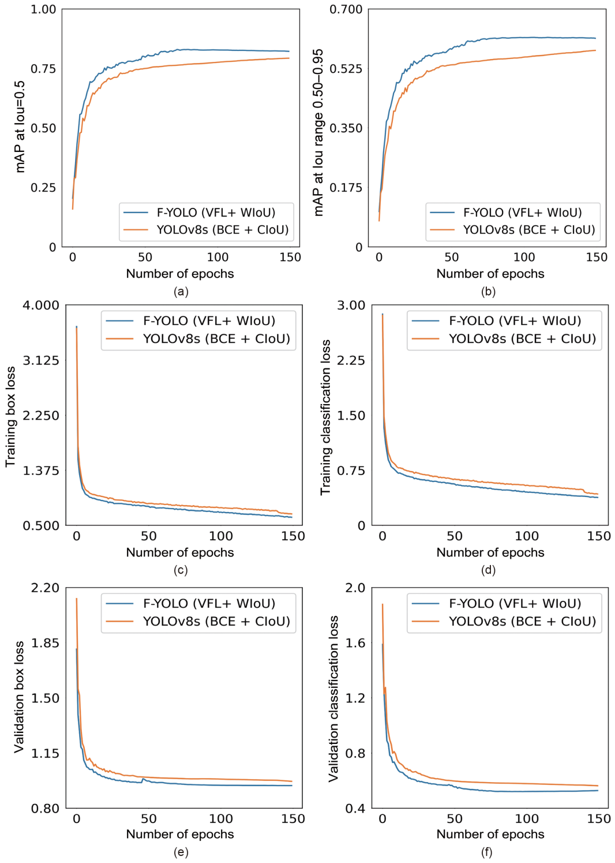


Fig. 11 Comparative analysis of mAP and loss for YOLOv8s and the proposed F-YOLO model utilizing diverse loss functions.

small targets, such as airplanes, ships, and vehicles. Each type of object is marked with a bounding box in a unique color. A color legend at the bottom of the image clearly matches each color to its respective target class. These results underscore F-YOLO's high precision in detecting small targets against complex backdrops, affirming its efficiency and stability in handling high-resolution remote sensing imagery.

To demonstrate the accuracy and efficiency of our proposed model, we conduct an impartial comparison on the DOTA test set, including F-YOLO and several contemporary state-of-the-art models across metrics of inference speed and precision. This evaluation ensured that all models are subjected to identical training configurations, such as model input dimensions, the number of training epochs, and initial learning rates. The results are shown in Table 4, detailing the precision for 15 distinct categories, mAP_{50}, and the Frames processed Per Second (FPS) for inference. The superior results for each category are distinguished in boldface to facilitate easy recognition. The data reveal that F-YOLO obtains the most top rankings, with a pronounced advantage in inference speed. Specifically, F-YOLO attains an ultimate mAP_{50} of 81.73% on the DOTA test set, coupled with a striking processing speed of 364.2 FPS.

6 Conclusion

In this research, we propose a novel small object detection framework designed for efficient oriented object detection in Unmanned Aerial Vehicle (UAV) environments. Unlike conventional approaches, we introduce CFNet, a backbone that redesigns feature extraction by integrating a novel CFBlock, which

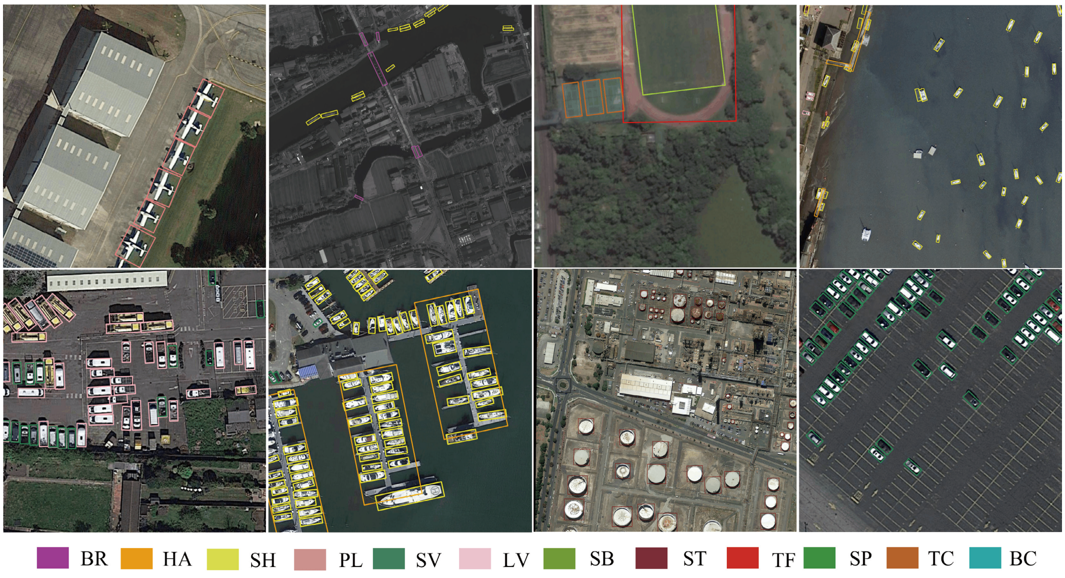


Fig. 12 Small object detection performance of F-YOLO model on DOTA testing set.

Table 4 Performance evaluation of our method against leading techniques on the DOTA test set. The best performance is highlighted in bold.

Method	PL	BD	BR	TF	SV	LV	SH	TC	BC	ST	SB	RA	HA	SP	HC	mAP_{50} (%)	FPS
Mask OBB[50]	89.61	85.09	51.85	72.90	75.28	73.23	85.57	90.37	82.08	85.05	55.73	68.39	71.61	69.87	66.33	74.86	–
ReDet[51]	88.79	82.64	53.97	74.00	78.13	84.06	88.04	90.89	87.78	85.75	61.76	60.39	75.96	68.07	63.59	76.25	–
Oriented RepPoints[52]	87.02	83.17	54.13	71.16	80.18	78.40	87.28	90.90	85.97	86.25	59.90	70.49	73.53	72.27	58.97	75.97	–
Oriented RCNN[53]	89.46	82.12	54.78	70.86	78.93	83.00	88.20	90.90	87.50	84.68	63.97	67.69	74.94	68.84	52.28	75.87	15.0
ORCNN-X[54]	90.58	**88.2**	62.72	80	80.31	83.11	88.06	**91.87**	**87.48**	**87.74**	**72.97**	74.43	**80.77**	80.81	75.12	81.61	16.1
YOLOv8s[44]	89.3	82.9	**67.2**	80.8	81.3	80.4	88.6	84.3	82.9	79.2	71.5	71.9	77.6	77.2	73.9	79.26	350.7
YOLOv11s[55]	90.8	87.6	66	80.5	80.9	79.7	86.4	84.7	86.5	78.7	72.1	70.6	77.5	80.6	71.7	79.62	362.7
YOLOv12s[56]	90	85.4	66.8	79.3	82.6	81.6	**89.1**	84	84.2	80.5	70.7	71.6	76.9	80.8	74.3	79.85	358.9
F-YOLO	**91.3**	84.2	65.7	**82.9**	**83.6**	**83.2**	88.1	89.8	87.4	83.7	72.5	**74.7**	80.3	**81.4**	**77.1**	**81.73**	**364.2**

leverages the T-shaped convolution strategies of PConv and PWConv from FasterNet, while incorporating an optimized feature splitting and fusion mechanism. This design enhances computational efficiency and feature representation, making it particularly suitable for small object detection. To further refine feature selection, we develop the FFCA module, which combines local and global feature aggregation to improve multi-scale feature learning. Additionally, we propose the SAFPN, which reformulates feature fusion by integrating FFCA-based attention mechanisms and strategic shortcut connections, ensuring better semantic retention and mitigating feature dilution. These innovations collectively optimize small object detection by improving both accuracy and computational efficiency. Experimental results validate that F-YOLO surpasses traditional models in balancing precision and speed, demonstrating state-of-the-art performance in UAV-based object detection.

In the future, the focus should intensify on refining the model’s feature extraction techniques, with the aim of developing a learning approach that more closely mirrors biological visual systems to achieve more efficient recognition capabilities.

Acknowledgments

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (No. 62502271), the Young Scientists Fund of the Natural Science Foundation of Shandong Province (No. ZR2025QC630), the Natural Science Foundation of Rizhao City (Nos. RZ2024ZR33 and RZ2024ZR34), the Key Research and Development Program of Shandong Province (No. 2025CXGC010113), the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2020R1A6A1A03038540).

References

[1] S. Liu, P. Chen, and M. Woźniak, Image enhancement-based detection with small infrared targets, *Remote Sens.*, vol. 14, no. 13, p. 3232, 2022.

[2] H. Mo and G. Zhao, RIC-CNN: Rotation-invariant coordinate convolutional neural network, *Pattern Recognit.*, vol. 146, p. 109994, 2024.

[3] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, Deep learning for generic object detection: A survey, *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.

[4] H. M. Wang, H. Y. Lin, and C. C. Chang, Object detection and depth estimation approach based on deep convolutional neural networks, *Sensors*, vol. 21, no. 14, p. 4755, 2021.

[5] A. Hanif, M. Muaz, A. Hasan, and M. Adeel, Micro-Doppler based target recognition with radars: A review, *IEEE Sens. J.*, vol. 22, no. 4, pp. 2948–2961, 2022.

[6] H. Chen, High-precision target detection of remote sensing image based on feature enhancement with 6G technology, *Adv. Multimedia*, vol. 2022, p. 6095308, 2022.

[7] Z. Liu, J. Li, M. Ashraf, M. S. Syam, M. Asif, E. M. Awwad, M. Al-Razgan, and U. A. Bhatti, Remote sensing-enhanced transfer learning approach for agricultural damage and change detection: A deep learning perspective, *Big Data Res.*, vol. 36, p. 100449, 2024.

[8] Y. Liu, T. Zhou, J. Xu, Y. Hong, Q. Pu, and X. Wen, Rotating target detection method of concrete bridge crack based on YOLO v5, *Appl. Sci.*, vol. 13, no. 20, p. 11118, 2023.

[9] J. Zou, H. Zheng, and F. Wang, Real-time target detection system for intelligent vehicles based on multi-source data fusion, *Sensors*, vol. 23, no. 4, p. 1823, 2023.

[10] S. Wang, S. Huang, S. Liu, and Y. Bi, Not just select samples, but exploration: Genetic programming aided remote sensing target detection under deep learning, *Appl. Soft Comput.*, vol. 145, p. 110570, 2023.

[11] C. Deng, D. Jing, Y. Han, Z. Deng, and H. Zhang, Towards feature decoupling for lightweight oriented

object detection in remote sensing images, *Remote Sens.*, vol. 15, no. 15, p. 3801, 2023.

[12] J. Yun, D. Jiang, Y. Liu, Y. Sun, B. Tao, J. Kong, J. Tian, X. Tong, M. Xu, and Z. Fang, Real-time target detection method based on lightweight convolutional neural network, *Front. Bioeng. Biotechnol.*, vol. 10, p. 861286, 2022.

[13] W. Zhang, X. Xia, G. Zhou, J. Du, T. Chen, Z. Zhang, and X. Ma, Research on the identification and detection of field pests in the complex background based on the rotation detection algorithm, *Front. Plant Sci.*, vol. 13, p. 1011499, 2022.

[14] H. Wang, Y. Li, L. M. Dang, and H. Moon, An efficient attention module for instance segmentation network in pest monitoring, *Comput. Electron. Agric.*, vol. 195, p. 106853, 2022.

[15] Y. Li, H. Wang, L. M. Dang, H. K. Song, and H. Moon, Attention-guided multiscale neural network for defect detection in sewer pipelines, *Comput.-Aided Civil Infrastruct. Eng.*, vol. 38, no. 15, pp. 2163–2179, 2023.

[16] L. Chen, H. Liu, J. Mo, D. Zhang, J. Yang, F. Lin, Z. Zheng, and R. Jia, Cross channel aggregation similarity network for salient object detection, *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 8, pp. 2153–2169, 2022.

[17] A. Mewada and R. K. Dewang, SA-ASBA: A hybrid model for aspect-based sentiment analysis using synthetic attention in pre-trained language BERT model with extreme gradient boosting, *J. Supercomput.*, vol. 79, no. 5, pp. 5516–5551, 2023.

[18] Z. Mi, X. Zhang, J. Su, D. Han, and B. Su, Wheat stripe rust grading by deep learning with attention mechanism and images from mobile devices, *Front. Plant Sci.*, vol. 11, p. 558126, 2020.

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, MobileNets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv: 1704.04861, 2017.

[20] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, ShuffleNet V2: Practical guidelines for efficient CNN architecture design, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 116–131.

[21] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, GhostNet: More features from cheap operations, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 1580–1589.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in *Proc. 9th Int. Conf. Learning Representations*, Virtual Event, https://dblp.uni-trier.de/db/conf/iclr/iclr2021.html#DosovitskiyB0WZ21, 2021.

[23] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al., MLP-mixer: An all-MLP architecture for vision, in *Proc. 35th Int. Conf. Neural Information Processing Systems*, Virtual Event, 2021, p. 1857.

[24] H. Zhang, M. Liu, Y. Qi, N. Yang, S. Hu, L. Nie, and W. Zhang, Efficient brain tumor segmentation with lightweight separable spatial convolutional network, *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 20, no. 7, p. 229, 2024.

[25] Y. Qi, W. Zhang, X. Wang, X. You, S. Hu, and J. Chen, Efficient knowledge distillation for brain tumor segmentation, *Appl. Sci.*, vol. 12, no. 23, p. 11980, 2022.

[26] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the inception architecture for computer vision, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2818–2826.

[28] M. Tan and Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in *Proc. 36th Int. Conf. Machine Learning*, Long Beach, CA, USA, 2019, pp. 6105–6114.

[29] W. Liu, L. Zhou, S. Zhang, N. Luo, and M. Xu, A new high-precision and lightweight detection model for illegal construction objects based on deep learning, *Tsinghua Science and Technology*, vol. 29, no. 4, pp. 1002–1022, 2024.

[30] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, pp. 16133–16142.

[31] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, Feature pyramid networks for object detection, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 2117–2125.

[32] J. Zhang, C. Xu, S. Shen, J. Zhu, and P. Zhang, MFF-YOLO: An improved YOLO algorithm based on multi-scale semantic feature fusion, *Tsinghua Science and Technology*, vol. 30, no. 5, pp. 2097–2113, 2025.

[33] S. Liu, D. Huang, and Y. Wang, Learning spatial fusion for single-shot object detection, arXiv preprint arXiv: 1911.09516, 2019.

[34] M. Tan, R. Pang, and Q. V. Le, EfficientDet: Scalable and efficient object detection, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10781–10790.

[35] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, Path aggregation network for instance segmentation, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 8759–8768.

[36] G. Zhao, W. Ge, and Y. Yu, GraphFPN: Graph feature pyramid network for object detection, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision*, Montreal, Canada, 2021, pp. 2763–2772.

[37] S. Wei, H. Su, J. Ming, C. Wang, M. Yan, D. Kumar, J. Shi, and X. Zhang, Precise and robust ship detection for high-resolution SAR imagery based on HR-SDNet, *Remote Sens.*, vol. 12, no. 1, p. 167, 2020.

[38] X. Zhang, C. Liu, D. Yang, T. Song, Y. Ye, K. Li, and Y. Song, RFAConv: Innovating spatial attention and standard convolutional operation. arXiv preprint arXiv: 2304.03198, 2023.

[39] X. Wang, R. Girshick, A. Gupta, and K. He, Non-local neural networks, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7794–7803.

[40] J. Hu, L. Shen, and G. Sun, Squeeze-and-excitation networks, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[41] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 11534–11542.

[42] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, CBAM: Convolutional block attention module, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 3–19.

[43] Q. Hou, D. Zhou, and J. Feng, Coordinate attention for efficient mobile network design, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 13713–13722.

[44] Ultralytics YOLO, https://github.com/ultralytics/ultralytics, 2025.

[45] J. Chen, S. H. Kao, H. He, W. Zhuo, S. Wen, C. H. Lee, and S. H. G. Chan, Run, Don't walk: Chasing higher FLOPS for faster neural networks, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, pp. 12021–12031.

[46] H. Zhang, Y. Wang, F. Dayoub, and N. Sünderhauf, VarifocalNet: An IoU-aware dense object detector, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 8514–8523.

[47] Z. Tong, Y. Chen, Z. Xu, and R. Yu, Wise-IoU: Bounding box regression loss with dynamic focusing mechanism, arXiv preprint arXiv: 2301.10051, 2023.

[48] Y. Sun, B. Cao, P. Zhu, and Q. Hu, Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6700–6713, 2022.

[49] G. S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, DOTA: A large-scale dataset for object detection in aerial images, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 3974–3983.

[50] J. Wang, J. Ding, H. Guo, W. Cheng, T. Pan, and W. Yang, Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images, *Remote Sens.*, vol. 11, no. 24, p. 2930, 2019.

[51] J. Han, J. Ding, N. Xue, and G. S. Xia, ReDet: A rotation-equivariant detector for aerial object detection, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 2786–2795.

[52] W. Li, Y. Chen, K. Hu, and J. Zhu, Oriented RepPoints for aerial object detection, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 1829–1838.

[53] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, Oriented R-CNN for object detection, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision*, Montreal, Canada, 2021, pp. 3520–3529.

[54] Y. Li, H. Wang, L. M. Dang, H. K. Song, and H. Moon, ORCNN-X: Attention-driven multiscale network for detecting small objects in complex aerial scenes, *Remote Sens.*, vol. 15, no. 14, p. 3497, 2023.

[55] R. Khanam and M. Hussain, YOLOv11: An overview of the key architectural enhancements, arXiv preprint arXiv: 2410.17725, 2024.

[56] Y. Tian, Q. Ye, and D. Doermann, YOLOv12: Attention-centric real-time object detectors, arXiv preprint arXiv: 2502.12524, 2025.

Hanxiang Wang received the BEng degree in software engineering from Linyi University, China in 2018, and the PhD degree in computer science from Sejong University, Seoul, Republic of Korea in 2023. Currently, he is a lecturer at School of Computer Science, Qufu Normal University, China. His research interests are in computer vision, natural language processing, and artificial intelligence.

Yanfen Li received the BEng degree in software engineering from Linyi University, China in 2018, and the PhD degree in computer science from Sejong University, Seoul, Republic of Korea in 2023. Currently, she is a lecturer at School of Computer Science, Qufu Normal University, China. Her research interests include image processing, object detection, and image segmentation.

Yuanke Zhang received the PhD degree in computer application technology from Xidian University, Xi'an, China in 2011. He is currently a professor at School of Computer Science, Qufu Normal University, China. His current research interests include computer vision, medical image processing and analysis, and deep learning.

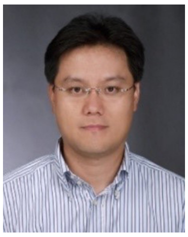
Junliang Shang received the BEng, MEng, and PhD degrees from Xidian University, Xian, China in 2007, 2010, and 2013, respectively. He is currently a professor at School of Computer Science, Qufu Normal University, China. His research interests are bioinformatics, artificial Intelligence, and big data mining.

Guangshun Li received the PhD degree from Harbin Engineering University, China in 2008. He is currently a professor at School of Computer Science, Qufu Normal University, China. He was a visiting scholar at The Hong Kong Polytechnic University in the second half year of 2019. His research interests include networks security, artificial intelligence, and big data. He is a member of IEEE.

Liem Dinh-Tien received the BEng and MEng degrees from Da Lat University, Lam Dong, Vietnam in 1996 and 2010 respectively. He is currently a lecturer at Faculty of Fundamental Sciences, Van Lang University, Vietnam. His research interests are mathematical analysis and optimal, artificial intelligence, and big data mining.

L. Minh Dang received the BEng degree in information systems from University of Information Technology, VNU HCMC, Vietnam in 2016, and the PhD degree in computer science from Sejong University, Seoul, Republic of Korea in 2021. Starting from 2017, he joined Sejong University, Republic of Korea. His current research interests include computer vision, natural language processing, and artificial intelligence.

Hyoung-Kyu Song received the BEng, MEng, and PhD degrees in electronic engineering from Yonsei University, Seoul, Republic of Korea in 1990, 1992, and 1996, respectively. From 1996 to 2000, he was a managerial engineer at Korea Electronics Technology Institute (KETI), Republic of Korea. Since 2000, he has been a professor at Department of Information and Communication Engineering, and Convergence Engineering for Intelligent Drone, Sejong University, Republic of Korea. His research interests include digital and data communications.



Hyeonjoon Moon received the BEng degree in electronics and computer engineering from Korea University, Republic of Korea in 1990, the MEng and the PhD degrees from State University of New York, USA in 1992 and 1999, respectively. He is currently a professor and chairman at Department of Computer Science and Engineering, Sejong University, Republic of Korea. His current research interests include image processing, biometrics, artificial intelligence, and machine learning.