IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 September 2023, accepted 30 October 2023, date of publication 2 November 2023,
date of current version 16 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3329713

RESEARCH ARTICLE

YOLO-UAV: Object Detection Method of Unmanned Aerial Vehicle Imagery Based on Efficient Multi-Scale Feature Fusion

CHENGJI MA[1], YANYUN FU[2], DEYONG WANG[3], RUI GUO[3], XUEYI ZHAO[3], AND JIAN FANG[3]

[1]School of Information Science and Engineering (School of Cyberspace Security), Xinjiang University, Ürümqi 830017, China
[2]Beijing Academy of Science and Technology, Beijing 100035, China
[3]Key Laboratory of Big Data of Xinjiang Social Security Risk, Xinjiang Lianhaichuangzhi Information Technology Company Ltd., Ürümqi 830011, China

Corresponding author: Yanyun Fu (fyyun163@163.com)

This work was supported in part by the National Natural Science Foundation of China: Intelligent Perception and Real-time Simulation and Deduction Technology for Urban Emergency Management Events under Grant U20B2060, in part by the Key Laboratory of Big Data of Xinjiang Social Security Risk Prevention and Control, and in part by the Autonomous Region High-Level Talent Introduction Project: Research on Unmanned Aerial Vehicle Information Reconnaissance and Big Data Analysis Technology for Public Safety.

ABSTRACT As Unmanned Aerial Vehicle (UAV) remote sensing technology progresses, the utilization of deep learning in UAV imagery object detection has become more prevalent. However, detecting small targets in complex backgrounds and distinguishing dense targets remains a major challenge. To address these issues and improve object detection efficiency, this study proposes an UAV imagery object detection method called YOLO-UAV by optimizing YOLOv5. YOLO-UAV first reconstructs the backbone and feature fusion networks by simplifying the network structure and reducing computational burden. The employment of a Dense_CSPDarknet53 backbone network, fashioned via the incorporation of dense connections, facilitates the extraction of latent image information through the recurrent utilization of features. In the Neck structure, an efficient feature fusion block with structural re-parameterization and ELAN strategies is integrated to effectively reduce interference from complex background noise while extracting more accurate and rich features. In addition, by proposing GS-Decoupled Head, this approach diminishes the parameter count of the decoupled head without compromising accuracy. It also separates classification tasks from regression tasks to lessen the influence of task disparities on prediction bias. To tackle the discrepancy between positive and negative samples in bounding box regression tasks, this study introduces a new loss function, Focal-ECIoU, capable of expediting network convergence and improve model positioning ability. Experimental findings from the public VisDrone2019 dataset indicate that YOLO-UAV outperforms other advanced object detection methods in comprehensive performance. Compared with the baseline model YOLOv5s, YOLO-UAV increased mAP0.5 from 35.1% to 46.7%, while mAP0.5:0.95 increased from 19.1% to 27.4%. For small-scale targets, AP_{small} increased from 10.2% to 17.3%. The experiment proves that YOLO-UAV performs well in improving object detection accuracy and has strong generalization ability, satisfying the practical requirements of UAV imagery object detection tasks.

INDEX TERMS UAV imagery, object detection, YOLO-UAV, VisDrone2019.

I. INTRODUCTION

With the progress in onboard sensor technology and computational capabilities, UAVs have become indispensable for acquiring high-resolution remote sensing imagery. Capturing imagery from an aerial viewpoint, UAVs offer a fresh perspective for a range of industries. Especially when combined with deep learning technology, the application value of UAVs in various fields has been greatly expanded. The combination of UAV imagery and deep learning has

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

considerably influenced the domain of identifying and categorizing objects.

UAV small object detection, a pivotal technology in UAV image processing, can identify small, morphologically intricate objects and is hard to distinguish by color. This technique confronts challenges due to the intricate backgrounds and substantial detail found in images captured by UAVs. Traditional detection methods typically rely on manually-designed features, sensitive to illumination, angles, and occlusions, with limited capacity to handle complex backgrounds. Despite their decent performance in simpler environments, these methods could lead to false and missed detections in more complex scenarios. In contrast, deep learning methods compensate for these shortcomings, demonstrating excellent performance in complex environments. The domain of UAV small-object detection has seen a significant application of deep learning methods, which now represent key solutions. Within these developments, Convolutional Neural Networks (CNNs) stand out as the dominant model structure. Addressing the issue of identifying small-scale objects in UAV imagery, advancements in deep learning have contributed significantly to research breakthroughs. [1]. Currently, object detection algorithms can primarily be divided into two categories: two-stage detectors and single-stage detectors.

The two-stage detectors include algorithms such as R-CNN (Region Convolutional Neural Networks) [2], Faster R-CNN [3], Mask R-CNN [4], Cascade R-CNN [5], Libra R-CNN [6], etc. These algorithms typically exhibit higher accuracy compared to the single-stage methods, particularly in detecting small targets and managing complex backgrounds, enabling more accurate object identification. Moreover, two-stage methods initially generate proposal regions, then perform classification and regression on these regions to distinctly segregate the targets from the background. However, due to the involvement of two computational stages, these methods tend to have slower processing speeds and require higher computational resources. In contrast, single-stage detectors like the YOLO (You Only Look Once) series [7], SSD (Single Shot MultiBox Detector) [8], RetinaNet [9], CenterNet [10], etc., are characterized by their fast processing speeds and robust real-time performance. Nevertheless, these methods often exhibit lower accuracy, especially when detecting small targets and managing complex backgrounds, which may lead to false positives or missed detections. Additionally, single-stage algorithms need to perform classification and regression for all potential positions, potentially leading to class imbalance issues.

While these object detectors have performed well, they often focus on general scene detection rather than explicitly addressing the challenges of UAV imagery. During UAV flights, environmental lighting, weather conditions, and sensor noise can introduce instability in the image quality of the dataset. In UAV-captured images, objects are often captured from different perspectives and distances. Thus, directly performing object detection at the same scale can result in significant errors and missed detections. Additionally, UAV imagery datasets often contain densely packed and small objects, along with scenes where objects are heavily occluded, making the object features less prominent. Furthermore, details of small objects may be lost as a result of downsampling during image processing, causing it to be difficult for the network to gather enough details for accurate identification. Due to these challenges, current detection methods face difficulties in precisely localizing and detecting objects in UAV imagery. There is still a lot that can be done to address these issues.

Centered on detecting small objects in UAV imagery, this research utilizes YOLOv5s, a prevalent detection approach from the YOLO family. Our research refines the YOLOv5s detection approach to bolster its detection efficacy, making it increasingly suitable for identifying targets in UAV images. We have proposed an object detection method based on efficient multi-scale feature fusion termed as YOLO-UAV. The YOLO-UAV strategy is designed to tackle the difficulties unique to UAV imagery, consequently enhancing the accuracy of the model's detection in this domain.

The following are the primary improvement strategies discussed in this paper:

1) This study proposes an improved YOLOv5 network for object detection in UAV imagery. We simplified the network architecture by eliminating the 20×20 large object detection head, subsequently reducing redundant computations and model parameters, while also streamlining the network. Furthermore, by introducing a 160×160 detection head, we enhanced the model's sensitivity towards smaller objects. To efficiently merge features of different scales, we employed the BiFPN architecture. These modifications not only enhanced the detection accuracy of the model but also lowered its complexity and computational requirements.
2) This study proposes a Dense_CSPDarknet53 backbone network that utilizes feature reuse to leverage the latent information within the network. By combining and connecting feature maps that are learnt in various layers, the network enhances feature diversity and reduces feature loss.
3) The efficient feature fusion block is integrated into the multi-scale feature fusion process. This allows for effective extraction and learning of both local and global features in the feature maps, resulting in more affluent and more accurate feature representations.
4) This study proposes an efficient and simple decoupled head called GS-Decoupled Head. It decouples the feature channels for classification and regression tasks, reducing prediction biases resulting from differences between tasks. This strengthens the model's localization and regression capabilities while reducing the parameter and computational overhead of the decoupled head module while maintaining accuracy.

5) This study proposes Focal-ECIoU loss as a remedy for the problem of sample imbalance in bounding box regression assignments. This optimisation decreases the effect of low-quality samples on the performance of the model and focuses the regression processing on great anchor boxes, thereby accelerating model convergence.

II. RELATED WORKS

Object detection stands as a pivotal task in the realm of computer vision. In conventional object detection approaches, the design and selection of features are substantially contingent upon prior conditions, limiting their accuracy, objectivity, robustness, and generalizability. Additionally, most traditional methods predominantly employ a sliding window strategy, culminating in extended computational time, diminished efficiency, intricate processing, and compromised precision. With the ascent of computational prowess and the evolution of dataset scales, it has become evident that traditional techniques are no longer adept at meeting contemporary demands. Consequently, deep learning-based object detection methodologies, distinguished by their superior detection performance, have garnered substantial attention from the research community.

Influenced by the altitude of UAV flight, UAV imagery tends to encompass a larger number of small objects compared to traditional ground-based imagery. These objects often manifest irregular orientations and distributions, frequently encountering challenges such as clustering and occlusions. Moreover, imagery captured from UAVs at varying flight positions exhibits notable differences in background, illumination, weather conditions, and topography. Potential image blurring and noise induced by slight jitters of onboard cameras exacerbate the difficulty in object identification. A singular object, when viewed from disparate angles, can manifest a myriad of forms, sizes, and textures. These intricacies augment the complexities of object detection, rendering conventional deep learning detection methodologies less than optimal for UAV imagery. To better cater to these unique characteristics of UAV imagery, researchers have proactively introduced a plethora of innovative strategies to refine existing detection techniques.

Wang et al. [11] enhanced the Faster R-CNN for improved small object detection. They expanded the output feature maps within the main network to emphasize the texture features of minor objects. Additionally, considering the histogram distribution of objects in training data, they incorporated additional anchor boxes and fine-tuned their parameters.

Huang et al. [12] proposed an object detection approach based on the Cascade R-CNN. This method subdivides the detection head for different object categories, allowing for enhanced extraction of edge frames and precise adjustments to them. This ensures a more accurate region of interest, enhancing the reliability of the detection results.

Lin et al. [13] proposed the ECascade-RCNN object detection network. This network comprises the Trident-FPN backbone, RPN, and a cascaded dual-head detector. Furthermore, based on the size distribution of targets in UAV imagery, the anchor boxes in the RPN were re-clustered to obtain more refined parameters.

Liu et al. [14] introduced the CBSSD method. Building on the foundation of VGG-16, CBSSD incorporated the ResNet-50 network as an auxiliary backbone, which enhanced feature extraction capabilities and facilitated the retention of richer semantic information. The CBSSD model boasts higher recognition rates and lower false detection rates, maintaining commendable detection performance even under low-light conditions.

Gao et al. [15] proposed a single-stage detector tailored for UAV imagery. It adopts the anchor-free concept from FOCS, which provides a more rational judgment of positive and negative samples. By employing a matching score map strategy, the detector effectively leverages similar information from the feature maps. Additionally, the use of the Soft-NMS method alleviates the miss detection issues caused by dense arrangements, proving beneficial for the detection of slanting objects.

In the recent years, the YOLO detection approach, known for its superior speed and precision, has seen extensive use in the realm of identifying objects within UAV imagery [16].

Jawaharlalnehru et al. [17] addressed challenges in UAV imagery object detection such as low multi-scale object localization accuracy, slow detection speed, and missed detections by proposing an enhanced YOLOv2 algorithm. To tailor anchor box parameters to the specific detection task, they re-clustered their custom aerial detection dataset. During the network's training process, they altered the model's input size every ten iterations, enhancing robustness to images of varying scales.

Sahin and Ozer [18] explored how modifications to the YOLO architecture influence the detection efficiency of tiny objects within UAV imagery. Building upon the foundation of YOLOv3, they introduced the YOLODrone object detection method. This approach expanded the original three different scale output layers to five. Such modifications aid in acquiring more positional information, enhancing the localization performance for small objects

Cheng [19] tackled challenges arising from camera vibrations in UAV aerial shots, inconsistent lighting exposure, and transmission noise. They introduced an enhanced YOLO variant which incorporates various data augmentation strategies like affine shifts, Gaussian blur, and grayscale conversion. This amplifies the preprocessing efficiency of the YOLOv4 framework and mitigates training challenges due to data scarcity.

Shen et al. [20] proposed the CA-YOLO model based on YOLOv5. CA-YOLO, by implementing the CA (Coordinate Attention) module and Spatial Pyramid Pooling, alongside an optimized anchor box method and loss function, enhances

the detection capability of multi-scale objects and inference speed.

Koay et al. [21] introduced the YOLO-RTUAV model. YOLO-RTUAV builds upon YOLOv4-Tiny and reduces suppression errors using DIoU-NMS, reducing missed detections. It also utilizes 1 × 1 convolutions to reduce model complexity. However, YOLO-RTUAV is predominantly designed to detect small targets in UAV imagery and may not be suitable for detecting objects of various sizes.

Wang et al. [22] presented SPB-YOLO, a streamlined detector tailored for UAV imagery. By incorporating a custom Strip Bottleneck module, it boosts the detection of objects across various scales. Furthermore, utilizing a feature map upsampling technique inspired by the PAN (Path Aggregation Network), it elevates its efficiency in dense object detection tasks specific to UAV visuals.

Huang et al. [23] proposed a new model, TCA-YOLOv5m. TCA-YOLOv5m enhances the accuracy of feature extraction and detection of tiny objects by utilizing the transformer algorithm and CA module. Additionally, with the integration of the PAN and an additional detection layer, it bolsters the feature representation and the ability to capture targets.

The progress in the YOLO series and its variants have shown considerable potential for identifying small objects within UAV imagery, underlining the continuous efforts to refine identification techniques in this specific domain.

Liang et al. [24] proposed an enhanced Sparse R-CNN approach, integrating coordinate attention blocks with the ResNeSt architecture. They further constructed a feature pyramid to reinforce the backbone network, resulting in improved object detection accuracy. To address the challenges in complex scenarios, they introduced novel data augmentation techniques. Moreover, the inclusion of Self-Adaptive Augmentation (SAA) and Detection-Time Augmentation (DTA) modules bolstered the robustness of the model. This strategy offers a promising avenue for UAV object detection, especially when dealing with UAV imagery against intricate backdrops.

Liang et al. [25] proposed an object detector termed DetectFormer, which leverages category-augmented transformers. By utilizing a Class Decoder that synergizes proposed category information with the Global Extract Encoder (GEE), enhanced category sensitivity and detection performance are achieved. To cater to varied scenarios, data augmentation techniques were employed, and attention mechanisms were incorporated within the network backbone to capture spatial features and directional information across channels. DetectFormer presents an array of innovative strategies for UAV object detection, significantly elevating detection accuracy in challenging environments.

Overall, the myriad of techniques and methodologies emerging in the realm of object detection in recent years, coupled with the latest advancements in this domain, have ushered in substantial innovations and possibilities for the field of UAV imagery object detection. These novel techniques not only significantly elevate detection accuracy and real-time capabilities but also bolster the adaptability of drones in diverse and intricate scenarios.

III. METHODS

A. YOLO-UAV MODEL

In this study, the central challenges we address are the complex backgrounds, minute target sizes, and significant target occlusions present in UAV imagery. Current object detection methods often fall short when confronted with these specific issues. To better address these challenges, we propose enhancements to the YOLOv5s model.

YOLOv5s [26] is highly regarded in the realm of object detection algorithms due to its simplicity, computational efficiency, and outstanding detection performance, striking an effective balance between detection precision and speed. During its training process, YOLOv5s conducts a meticulous statistical analysis of the object size distribution in the dataset, thereby autonomously gauging the most fitting anchor sizes. Moreover, YOLOv5s continues the YOLO series' multi-scale detection scheme, performing detections on multiple feature map scales to identify targets of various sizes. As such, in this research, YOLOv5s was chosen as the baseline model to enhance the performance in detecting targets within UAV imagery.

First and foremost, we simplified the network architecture by eliminating the 20×20 large object detection head, thereby reducing redundant computations and model parameters. Concurrently, we introduced a 160 × 160 small object detection head, enhancing the model's detection capability for smaller targets. Furthermore, to effectively fuse features across varying scales, we adopted the BiFPN architecture. At the crux of feature extraction, we introduced the Dense_CSPDarknet53 backbone network, which capitalizes on feature reuse to harness latent information within the network. We also incorporated an efficient feature fusion block during the multi-scale feature fusion process, enabling the model to adeptly extract and learn both local and global features from the feature maps. To mitigate prediction biases caused by discrepancies between tasks and to fortify the model's localization and regression capabilities, all while decreasing the parameter count of the decoupled head module, we introduced the GS-Decoupled Head. Lastly, to address the sample imbalance issue in bounding box regression tasks, we proposed the Focal-ECIoU loss. This optimization diminishes the influence of low-quality samples on model performance, subsequently accelerating model convergence. Fig. 1 illustrates the YOLO-UAV structure.

B. IMPROVEMENTS IN NETWORK STRUCTURE FOR OBJECT DETECTION IN UAV IMAGERY

1) SIMPLIFIED NETWORK STRUCTURE

In CNNs, lower-level feature maps contain a great deal of information regarding location but less information related to semantics. In contrast, higher-level feature maps contain less

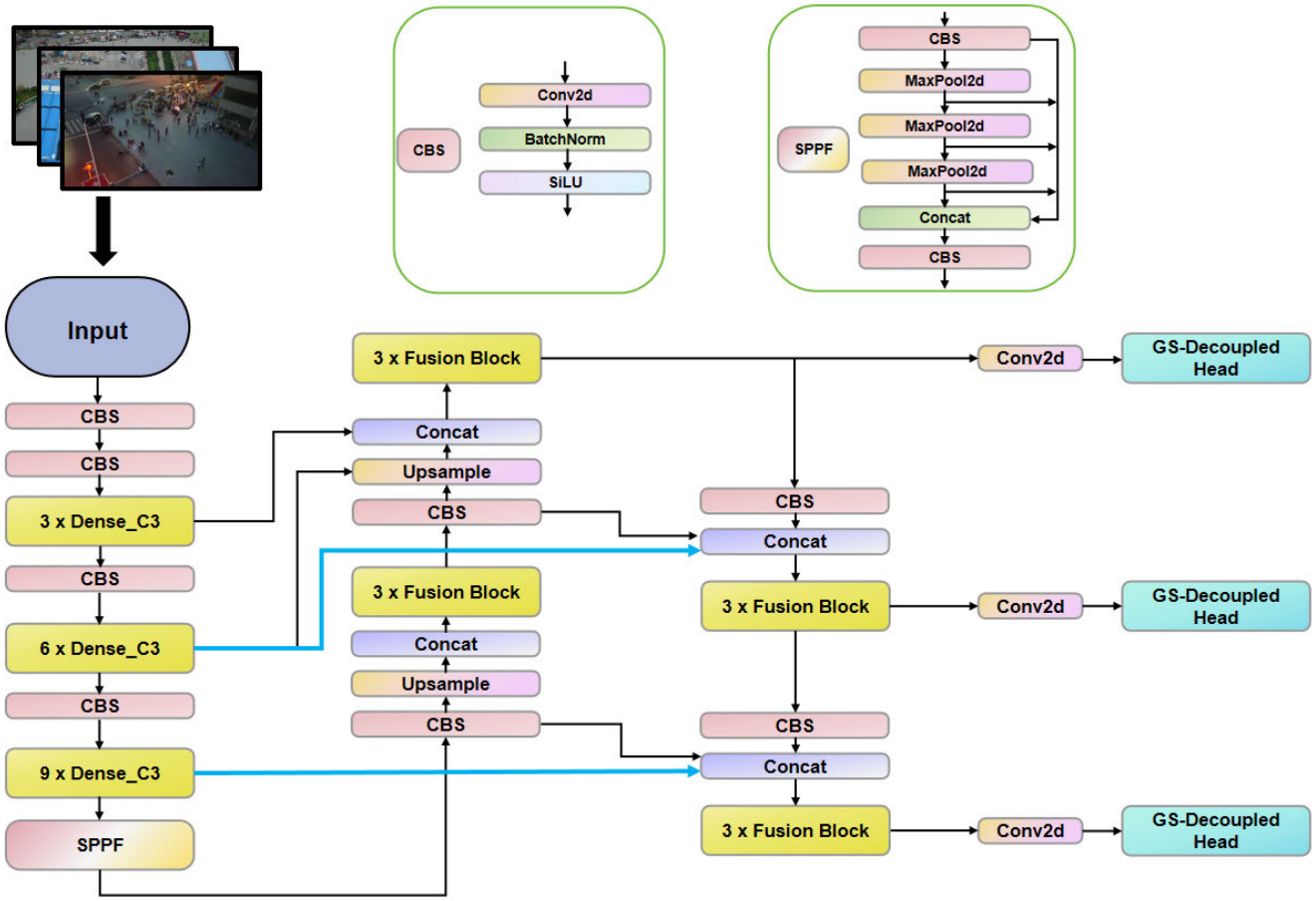


FIGURE 1. The network structure of YOLO-UAV.

information regarding positions but a wealth of information related to semantics. There are only three detection layers in the YOLOv5, and when the input image size is 640 × 640, the network downsamples by factors of 8, 16, and 32, respectively. The respective sizes of the detection layer feature maps for detecting targets of various sizes are 80×80, 40×40, and 20×20. In practical scenarios, objects within images exhibit diversity in pixel occupancy, size, clarity, and position; hence the mapping to images encompasses objects of varying scales. Given that the target sizes in UAV images are predominantly small, in YOLOv5, the layer responsible for large object detection results from the image being downsampled by a scale of 32. When the size of an object is less than 32 pixels, the model may only perceive a single point or might entirely miss the object. Moreover, the 20×20 feature layer can result in some degree of semantic loss, impacting feature fusion within the network and causing the model to overlook the target, thereby impeding network detection. Consequently, the 20×20 large object detection layer in YOLOv5 is considered redundant for detecting small-sized objects. Furthermore, in multi-scale feature fusion, the information lost cannot be recovered through upsampling due to the irreversible nature of downsampling. This results in the model being adversely affected by lower-resolution feature layers, leading to decreased detection accuracy.

Based on the conclusions mentioned above, We have modified the YOLOv5 architecture in our study to lessen the loss of valuable semantic details. The large object detection head of 20×20, along with its corresponding feature extraction and fusion layers, are eliminated. The model then solely utilises the feature maps from the operations of downsampling by factors of 8 and 16 for object detection tasks. The refined network architecture design minimizes the impact of downsampling on capturing object traits, improve the network's capacity to acquire detailed characteristics, eliminates a substantial amount of redundant computation, significantly reduces the model's parameter count, and shrinks the network structure, ensuring accuracy while alleviating computational bottlenecks.

2) ADD A SMALL OBJECT DETECTION HEAD

The YOLOv5 network possesses a large receptive field for mid-to-high-level features, primarily focusing on representing abstract semantic information. However, its capability to represent objects' location and detailed information is relatively weak. In feature extraction for small objects,

selecting an appropriate receptive field size or considering a multi-scale receptive field can effectively retain more local feature information of the small objects. The objects that must be detected in UAV imagery are minuscule. The original multi-scale detection structure is prone to missing such objects. Therefore, the inclusion of a smaller detection head is essential to strengthen the network's aptitude for identifying small targets.

To optimize the perceptual range of the network and boost its ability to identify small objects without raising the input resolution, a new 160 × 160 detection head was introduced in this study. This architecture enables the network to retain vital information when extracting features from small objects. In the network's second layer, we integrated an output feature map, designated as P2, and connected it to the feature fusion network. When the input image dimensions are 640 × 640, the P2 feature map dimensions are 160 × 160. Each unit of the P2 feature map corresponds to a perceptual area of 4 × 4 in the input image. This feature map size has a smaller perceptual range. This characteristic brings richer positional data about the target, which is pivotal for enhancing the detection of minuscule objects and providing valuable input for other layers in the process of feature fusion. The newly introduced detection head focuses on low-level features, thereby increasing sensitivity to small objects.

3) EFFICIENT MULTI-SCALE FEATURE FUSION NETWORK

The Neck structure utilises a combination of FPN (Feature Pyramid Network) and PAN. The original Neck structure of YOLOv5 is shown in Fig. 2. FPN [27] delivers high-level semantic characteristics, whereas PAN [28] delivers low-level position information towards deeper layers. The Neck performs upsampling on features acquired from the backbone network followed by downsampling fusion, enabling each feature to contain more information. However, the P2 section undergoes only 4× downsampling, introducing significant noise. To elevate its ability to extract features, an improvement is employed to the Neck.

To further optimize the ability to detect objects, ensuring that they possess both comprehensive location information and full semantic information, the structure of BiFPN (Bidirectional Feature Pyramid Network) [29] is adopted to improve the existing feature fusion structure. Fig. 3 illustrates the BiFPN structure. The original design of BiFPN incorporated a learnable-weight topology, which facilitated efficient information flow across different scales, thereby enhancing the model's performance. However, despite potential performance improvements under certain circumstances when introducing learnable weights, it concurrently escalates the computational complexity of the model and the quantity of its parameters. This increase in turn slows down the model's operation speed and amplifies the difficulty of training the model. Additionally, inappropriate initialization or setting of the learning rate may detrimentally affect

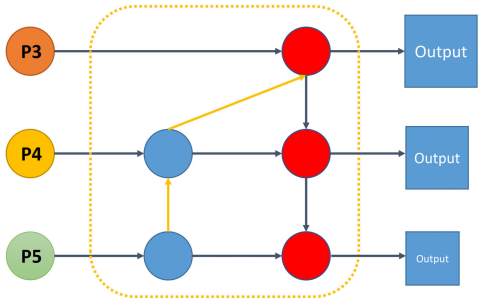


FIGURE 2. The network structure of the original Neck.

the stability of model training. For instance, the weights could dramatically increase or decrease to zero, causing the model to struggle with convergence. Therefore, our design did not adopt the approach of introducing weights, but instead retained the concept of BiFPN in constructing fusion channels. Fig. 4 illustrates the improved Neck structure.

Upon integrating P2, an upsampling node is added to the Neck to acquire a 160 × 160 feature map. The 160 × 160 feature map obtained from upsampling the P3 layer concatenates with the backbone network's P2 feature map. In the P3 and P4 layers, skip connections to the output nodes are added, concatenating the initial feature from the backbone with the top-down indirect feature and the downsampling output feature to enrich the semantic data of the feature maps without adding excessive complexity. After improving the network structure, we finally obtained YOLOv5s-stru. The improved Neck can fuse deep feature layers with larger receptive fields and shallow feature layers with stronger position information to allow the model to acquire stronger position features from shallow feature layers, allowing in-depth features to perform more accurate fine-grained detection. The improved multi-scale feature fusion structure increases complexity without increasing parameter and computational complexity. However, each feature map can fuse more information, which considerably strengthens the network detection precision and lowers the rate of missed detection. Fig. 5 illustrates the YOLOv5s-stru structure.

UAV imagery features images with diverse backgrounds, from urban scenes to rural settings. This variety results in objects of different sizes and looks. Our tailored skip connections help preserve spatial hierarchies, guaranteeing accurate detection of both large and small objects. UAV imagery is taken from different altitudes and angles, leading to scale and perspective variations. The integrated skip connections facilitate multi-scale feature learning, making the model robust against such variations. Due to the bird's-eye view in many UAV imagery, object occlusions and overlaps are common. With the assistance of our skip connections, our model can better differentiate overlapping objects and identify partially occluded objects by effectively utilizing features from different layers. As UAV imagery can be captured at different times of the day and under varied

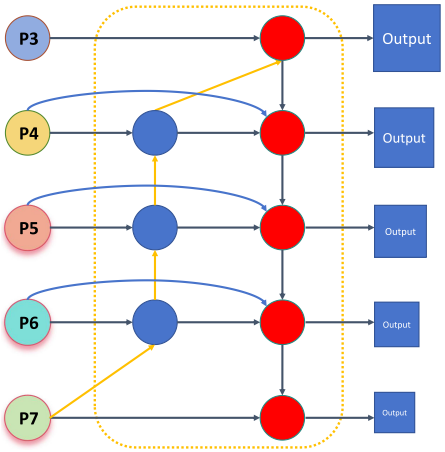


FIGURE 3. The network structure of BiFPN. The backbone network preceding the original BiFPN structure is EfficientNet, hence it has five inputs, ranging from P3 to P7.

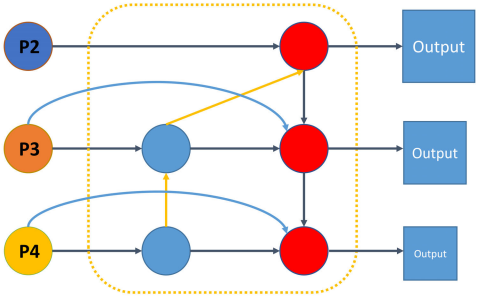


FIGURE 4. The network structure of the improved Neck.

weather conditions, they often exhibit stark lighting contrasts and shadows. Our tailored skip connections ensure that features affected by such lighting conditions are enhanced, leading to improved detection accuracy.

C. EFFICIENT DENSE CONNECTION FEATURE EXTRACTION NETWORK

YOLOv5 is a high-performance object detection model that employs the CSPDarknet53 feature extraction network as its backbone. The CSPDarknet53 connection method creates a direct link between the front feature layer and the back feature layer. The formula is shown in Equation (1)

$$x_n = H_n(x_{n-1}) + x_{n-1} \tag{1}$$

Here, x_n represents the outcome from the nth layer, whereas $H_n(\cdot)$ stands for the nonlinear transformation operation.

This connection method effectively solves the vanishing gradient problem. However, as the network grows in depth during object detection, crucial information about features may be lost during convolution and downsampling process.

To address the previously mentioned issues, Huang et al. [30] introduced DenseNet, employing a dense connection technique within its Dense Block. Unlike traditional methods, this block doesn't use element-wise addition for shortcut connections between a specific layer and its antecedent. Instead, it connects a particular layer densely with all previous layers. It performs channel-wise concatenation of feature maps through skip connections. Fig. 6 depicts the dense connection mechanism of the Dense Block. Forward feedback connections are established between every layer within the Dense block and the rest. The formula is shown in Equation (2)

$$x_n = H_n([x_0, x_1, \cdots, x_{n-1}]) \tag{2}$$

Here, $[x_0, x_1, \cdots, x_{n-1}]$ represents the combination of feature maps derived from the network's 0 to $n-1$ layers, and x_n indicates that the n layer essential information regarding features from every previous $x_0, x_1, \cdots, x_{n-1}$ layers.

In DenseNet, every layer receives feature maps derived from the that came before layers. Compared to the CSPDarknet53 network, the DenseNet network's connection method retains vital feature maps and can repeatedly reuse critical information regarding features, resulting in more abundant and diverse features. It makes the network better at extracting vital information and effectively solves the problem of vital information getting lost in the CSPDarknet53 network.

In this study, we employ the dense connection structure from the DenseNet network. Specifically, we select the C3 module in the CSPDarknet53 backbone network, and within its BottleNeck structure, we integrate two 3×3 convolutions. Moreover, we establish dense connections between each layer within the BottleNeck and its preceding and subsequent layers. Each layer retrieves additional input from all previous layers and transmits its feature mapping to all subsequent layers, resulting in the creation of a novel Dense_C3 module. A comparison illustration of the C3 and Dense_C3 structures is depicted in Fig. 7. As a further step, we replaced all C3 modules in the main network with Dense_C3, forming a new backbone network, the Dense_CSPDarknet53. The structure of the Dense_CSPDarknet53 network is displayed in Fig. 8.

D. EFFICIENT FEATURE FUSION BLOCK BASED ON ELAN AND STRUCTURAL RE-PARAMETRIZATION

In YOLOv5, the Neck combines shallower features information with deeper semantic features information in order to extract features and merge feature maps derived from various phases. This contributes to the enhancement of what the model can do to identify characteristics across different scales. The C3 is one of the vital sections in the Neck section. The primary purpose of the C3 module is to establish cross-stage connections to improve its perception and accuracy. Although the C3 in the Neck of YOLOv5 has yielded positive results, the C3 module's high complexity level and extensive parameters require more storage space and slow down the model's computation speed. When CNNs increase in depth, there is a critical problem: when input or gradient information passes through many layers, it may disappear or overinflate.

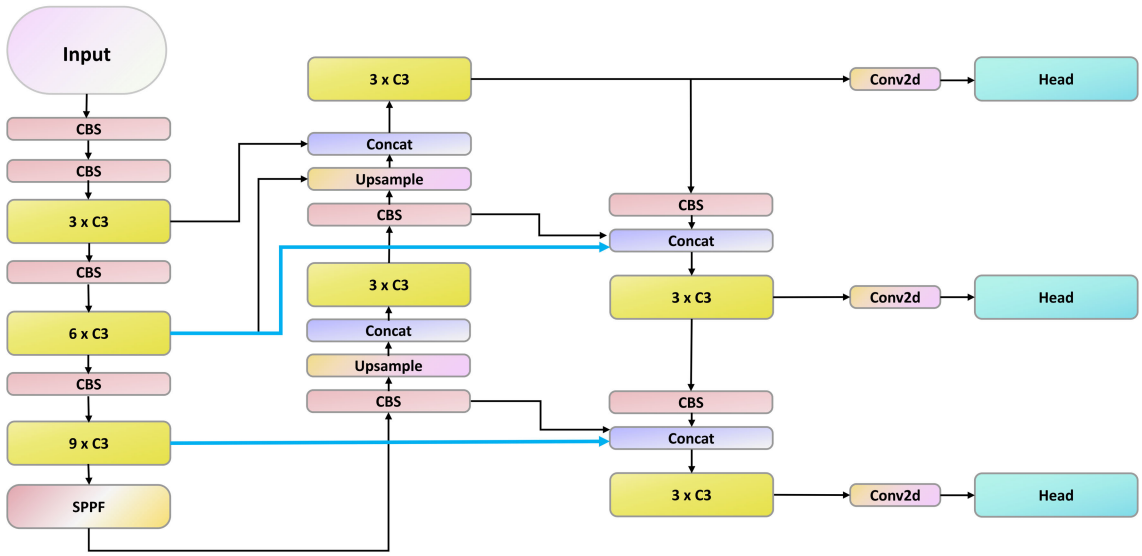


FIGURE 5. The network structure of YOLOv5s-stru.

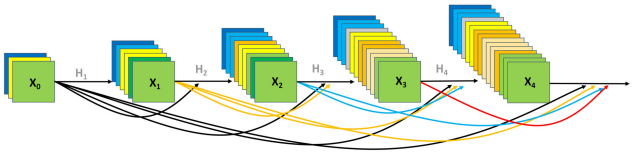


FIGURE 6. The dense connection mechanism of the Dense Block.

We studied standard techniques to augment the learning capacity of CNNs, such as DenseNet [30], VoVNet [31], CSPVoVNet [32], and CSPNet [33]. According to the study, if CNNs have shorter connections within layers near the layers that provide inputs and outputs, the network can be considerably deepened and trained with greater effectiveness. The ELAN (Efficient Layer Aggregation Networks) in YOLOv7 [33] incorporates the idea of segmenting gradient flow from VoVNet and CSPNet to get detailed gradient flow information. Using the stack structure in the calculation block, the gradient length of all networks is optimally optimised. The ELAN structure is displayed in Fig. 9.

Compared to the C3 structure in YOLOv5, ELAN indeed increases the network's layer count, enhancing the model's training accuracy and generalization capability. However, adding multiple nonlinear layers increases the network depth, resulting in substantial computational overhead, and may lead to a decline in network performance. Complex multi-branch structures, while capable of achieving relatively high accuracy, tend to reduce the model's inference speed and memory utilization. In RepVGG [34], propose a structural re-parameterization method that decouples the multi-branch topology at the phase of training from the normal structure at the phase of inference in order to accomplish a more favourable balance between accuracy and speed. By incorporating the structural re-parameterization

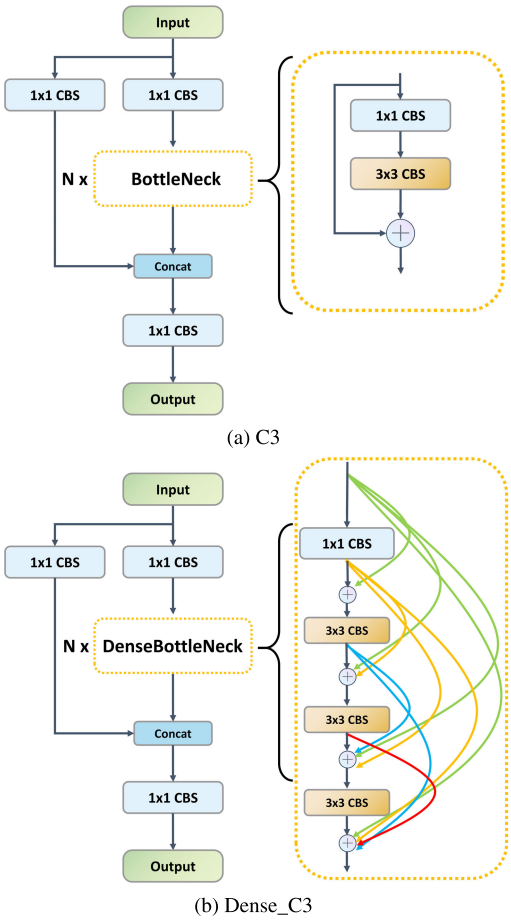


FIGURE 7. The structure of the C3 and Dense_C3. The (a) is the C3 structure, and the (b) is the Dense_C3 structure with dense connection mechanism.

technique, we decouple the standard structure during the inference phase from the multi-branch structure during the

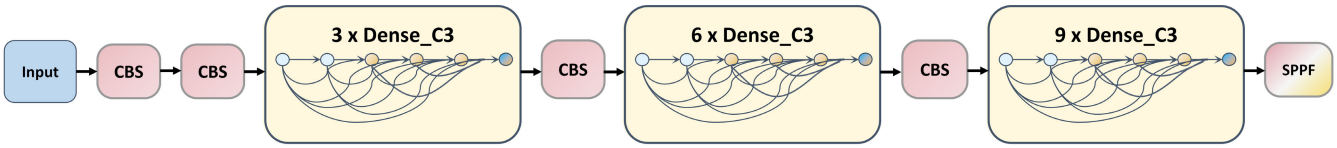


FIGURE 8. The network structure of Dense_CSPDarknet53.

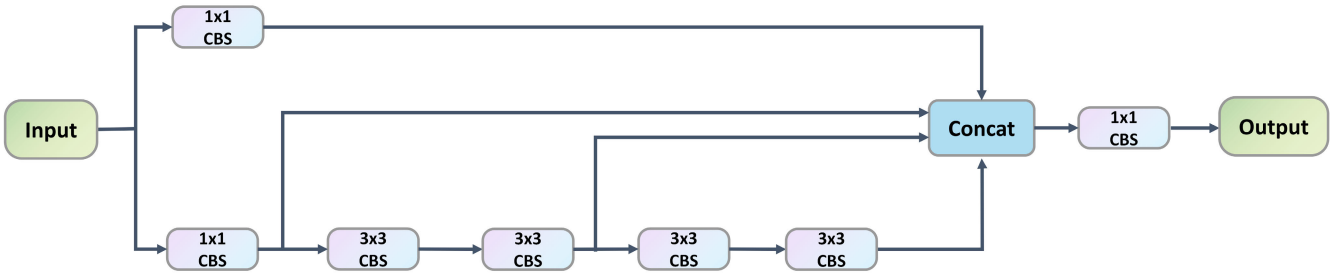


FIGURE 9. The structure of ELAN.

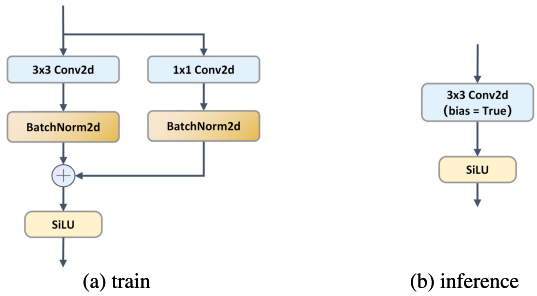


FIGURE 10. Schematic diagram of RepConv. The (a) represents the network structure employed during the training of RepConv, and the (b) represents the structure utilized during the inference of RepConv.

training phase. While the multi-branch structure during training can lead to higher accuracy, the inference is equivalently transformed into a single-branch structure through the structural re-parameterization technique, resulting in faster inference speeds while maintaining consistent accuracy. Fig. 10 shows the transition of RepConv between training and inferred states. In the training state, with additional 1×1 convolutions, RepConv can guarantee accuracy during training. The re-parameterized structure can be equivalently translated to the inferred state in the inferred state.

In the feature fusion block of DAMO-YOLO [35], the CSPNet connection is employed to replace the original feature fusion based on 3×3 convolution. Then, CSPNet is upgraded by combining the re-parameterization technique and the connection method of ELAN. The idea of segmenting gradient flow is introduced into the feature fusion block of DAMO-YOLO so as to enhance the inadequate efficacy of the node stacking operation, improve the model's accuracy without increasing more computation, and optimize the feature fusion. We replaced all C3 sections in Neck regarding DAMO-YOLO's feature fusion block so as to enhance its feature fusion capability. The input feature map for the feature fusion block of DAMO-YOLO, along the channel dimension, has been separated into two sections, each of which is adjusted by a 1×1 convolution layer. The lower branch depicts the idea of an ELAN module, which is comprised of multiple 3×3 RepConv and 3×3 Conv. Finally, the resultant outputs of each of the branches are concatenated by channel number and output to the next layer for processing. Based on CSPNet, structural re-parameterization technique, ELAN, and other strategies, the feature fusion block enhances the capabilities of feature fusion. It can obtain more rich and accurate feature representation so as to the network to more effectively learn small object features, thereby improving small object detection accuracy. Fig. 11 displays the efficient feature fusion block structure based on ELAN and structural re-parameterization techniques.

E. EFFICIENT AND SIMPLE DECOUPLED HEAD

There is a disparity among regression as well as classification assignments in object detection. The contradiction between classification and regression is essentially the contradiction between the invariance and identity of convolutional translation, scale, and identity. The classification task hopes that the category information remains unchanged after the target state is translated, rotated, illuminated, and scaled, that is, the invariance of translation and scale. For the regression task, the change of target state should be reflected on the feature and then accurately regressed to the position, that is, the identity of translation and scale. Object detection tasks are more challenging in the intricate background of UAV imagery and the influence of objects of multi-class. The detection head of YOLOv5 incorporates a coupled architecture. By sharing parameters in the final detection head, the classification and

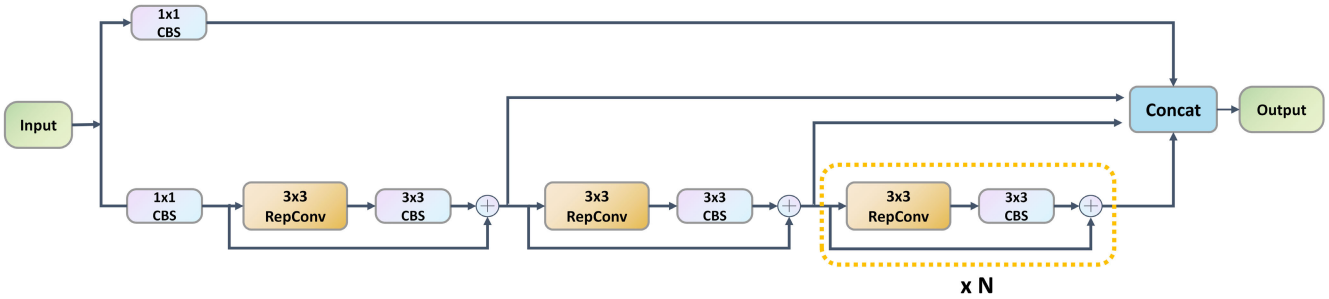


FIGURE 11. The structure of efficient feature fusion block.

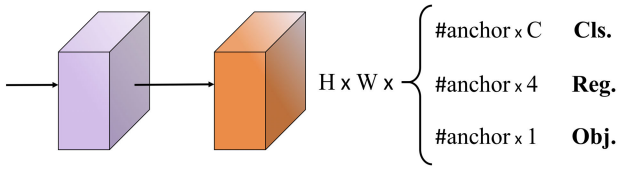


FIGURE 12. The structure of the coupled head.

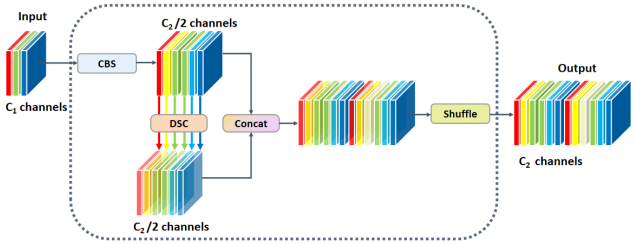


FIGURE 13. The structure of the GSConv module.

regression assignments are strongly coupled. Fig. 12 displays the coupled head of YOLOv5.

YOLOX [36] has proved that decoupling classification and regression tasks and introducing two additional 3×3 convolutional layers in each branch can better the network's capacity for detection. As a result of the outstanding capability of the decoupled head, the decoupled head and its modified version have been applied to a number of subsequent variants in the YOLO series. For example, YOLO-Extract [37] directly replaces YOLOv5's coupled head with YOLOX's decoupled head; YOLOv6 [38], building on the decoupled head from YOLOX, adopts a hybrid channels method to redesign a better decoupled head structure. It maintains accuracy while reducing latency, alleviating the additional delay cost brought about by the 3×3 convolution in the decoupled head. However, the decoupled head of YOLOX adds multiple additional convolutional layers. The standard convolution process, involving multiplication and addition operations between the convolution kernels at each position with every location of the input feature map, results in substantial computational expense. This procedure results in a substantial increase in the total amount of module parameters.

This paper replaces the traditional 3×3 convolution in the original decoupled head with GSConv. GSConv [39] is a mixed convolution of SC (Standard Convolution), DSC (Depth-Wise Separable Convolution), and Shuffle. Fig. 13 depicts the GSConv structure. By means of the Shuffle operation, GSConv infiltrates the data produced by SC into every portion of the data produced by DSC. This approach allows the data from SC to be completely mixed into the output of DSC. Shuffle is a uniform mixing strategy, which uniformly exchanges local feature information on different channels to make the information from SC wholly mixed into the output of DSC. The SC maximises the preservation of concealed connections between channels, whereas the DSC completely eliminates these connections. GSConv retains these connections as much as possible. Replacing SC operation with GSConv reduces the computational burden of convolution calculations while producing outputs that are as close to SC as feasible.

After considering the balance between parameter, computation, and accuracy, this paper redesigns an efficient and simple decoupled head, the GS-Decoupled Head. The GSConv can maintain the accuracy improvement brought by the decoupled head module while reducing its parameter volume. In this paper, the proposed GS-Decoupled Head separates the localization and classification functions into distinct feature channels for object detection tasks. These channels are individually responsible for bounding box coordinate adjustment and object categorization. The GS-Decoupled Head reduces the dimensionality of the input features via a 1×1 convolution, curtailing the parameter generation. Subsequently, it bifurcates the output features into two pathways. The first pathway focuses on classification, employing two 3×3 GSConvs for feature extraction and a 1×1 SC to adjust the feature channel dimensions to align with the target classes. The secondary pathway, in charge of regression, utilizes a pair of 3×3 GSConvs for feature extraction. After the secondary pathway feature extraction, the feature map is partitioned into two parts: one for the prediction of bounding box attributes such as centre coordinates, height, and width, and the other for determining the confidence score of the object to obtain the Intersection over Union for

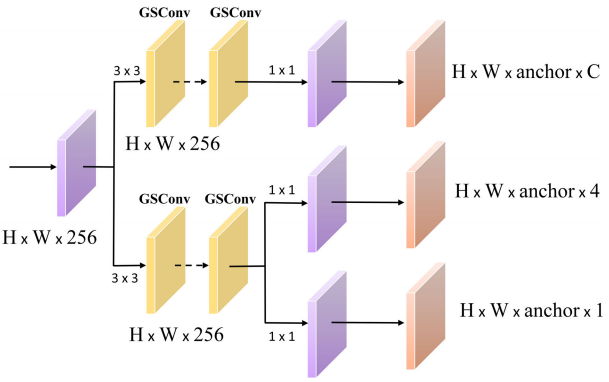


FIGURE 14. The structure of the GS-Decoupled head.

the actual and predicted boxes. Unlike the coupled head, which incorporates various information directly into a single feature map, the GS-Decoupled Head is capable of reducing conflicts among the different information regarding features needed for multiple tasks, thereby enhancing localization and regression capabilities. Simultaneously, the decoupled head, through depth and breadth operations, can well preserve the information in each channel, reducing the prediction bias caused by differences between tasks and thereby enhancing the precision of model detection. Fig. 14 depicts the structure of the GS-Decoupled Head.

F. FOCAL EFFICIENT COMPLETE INTERSECTION OVER UNION (FOCAL-ECIOU) REGRESSION LOSS FUNCTION

Object detection predicts the location of targets in images via bounding box regression, and early works in object detection employed IoU (Intersection over Union) [40] as the localization loss. However, IoU loss encounters a problem of vanishing gradients when the predicted box isn't overlapping the actual box, resulting in delayed convergence and less precise detectors. This has spurred several improved designs based on IoU loss, including GIoU (Generalized-IoU) [41], DIoU (Distance-IoU) [42] and CIoU [42]. In the detection of objects, bounding box regression is crucial for determining the efficacy of target localization.

In the context of bounding box regression, YOLOv5 utilizes the CIoU as its distinctive strategy for the computation of loss. CIoU incorporates three significant geometric factors into the target box regression function: overlap area, center point distance, and aspect ratio. The expression for CIoU shown in Equation (3).

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b^{gt}, b)}{c^2} + \alpha v. \tag{3}$$

$$IoU = \frac{A \cap B}{A \cup B} \tag{4}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{5}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{6}$$

Here, A and B respectively represent the predicted box and the actual box, b represents the center point of the predicted box, IoU represents the Intersection over Union of the predicted and the actual boxes, $\frac{\rho^2(b^{gt},b)}{c^2}$ represents the distance among the centre points of the predicted and actual boxes, α represents used for balancing the proportion and is a tunable parameter, v signifies the consistency parameter between the aspect ratios of the predicted and actual boxes.

CIoU is an enhancement of DIoU with adding an aspect ratio parameter v, which measures the consistency between the predicted box and actual boxes. CIoU can accelerate the regression pace of the predicted box to some degree. Nonetheless, once the aspect ratio of the predicted and actual boxes exhibits a linear ratio during the regression process of the predicted box, the relative ratio penalty item added in CIoU no longer functions. As inferred from the gradient formulas of w and h for the predicted box, when one value increases, the other must decrease; they cannot both increase or decrease simultaneously.

To resolve this matter, Zhang et al. [43] proposed the EIoU (Efficient-IoU), a solution which divides the height-to-width ratio based on the CIoU and leverages its penalty term to separate the aspect ratio impact factor, enabling separate calculations for the height and width of both the target box and the anchor box. EIoU is divided into three sections: overlap loss, center distance loss, and height-width loss. The initial two sections of EIoU extend the CIoU methods, the height-width loss aims to directly minimize the discrepancies in height and width between the predicted box and actual boxes, enabling quicker convergence and superior localisation results. The expression for EIoU shown in Equation (7).

$$L_{EIoU} = 1 - IOU + \frac{\rho^2(b^{gt}, b)}{c^2} + \frac{\rho^2(h^{gt}, h)}{c_h^2} + \frac{\rho^2(w^{gt}, w)}{c_w^2} \tag{7}$$

Here, $\frac{\rho^2(h^{gt},h)}{c_h^2}$ represents the height difference among the predicted and actual boxes, $\frac{\rho^2(w^{gt},w)}{c_w^2}$ represents the width difference among the predicted and actual boxes.

When dealing with a box with a distant edge, the computation of EIoU can slow down and convergence isn't achieved in advance. To resolve this matter, Chen et al. [44] proposed the ECIoU (Efficient Complete-IoU), that is able to improve the adjustment of the predicted box and speed up its regression convergence rate. The ECIoU is based on the combination of CIoU and EIoU loss functions. First, CIoU adjusts the predicted box's aspect ratio till it converges to an acceptable range; then, EIoU adjusts each of the sides till it converges to the correct value. The expression for ECIoU is shown in Equation (8).

$$L_{ECIoU} = 1 - IoU + \alpha v + \frac{\rho^2(b^{gt}, b)}{c^2} + \frac{\rho^2(h^{gt}, h)}{c_h^2} + \frac{\rho^2(w^{gt}, w)}{c_w^2} \tag{8}$$

The majority of loss functions disregard the imbalance between positive and negative samples. Specifically, a majority of predicted box with a small overlap area with the actual box dominate in the final bounding box optimization. Considering there's also a sample imbalance issue during bounding box regression, within an image, high-quality anchor boxes with minimal regression errors are significantly outnumbered by low-quality samples with substantial errors. These low-quality samples can result in excessive gradients that negatively effect the training procedure. To solve these issues, we have made adjustments to the ECIoU loss and, combined with Focal Loss [9], proposed a novel Focal-ECIoU Loss. From the perspective of gradients, we separate between high-quality and low-quality anchor boxes. In particular, it decreases the contribution of a multitude of anchor boxes that have fewer overlaps with the ground truth box to the optimisation of bounding box regression. This approach ensures the regression process focuses more on high-quality anchor boxes.

Due to the influence of surrounding environmental factors, UAV images can have variable quality. In some cases, a part of the target may be easily overlooked due to factors like viewing angles and lighting, resulting in these kinds of targets being significantly less numerous than other categories. Different flying heights and shooting areas of the UAV can lead to variations in the number and distribution of captured object types. Moreover, UAV imagery contains a great deal of background content, resulting in a lack of balance between the content in the foreground and the background. These factors may cause the imbalance between positive and negative examples. As a result, the trained model may be able to recognize easy-to-classify samples but perform poorly with difficult samples.

To address this matter and better improve model performance, we use FocalL1 Loss to set different gradients. The FocalL1 Loss formula is shown in Equation (9). FocalL1 loss, according to β, is capable of enhancing the gradients' value for inliers while reducing it for outliers. When β is larger, it demands fewer regression errors from the inliers and rapidly diminishes the value of the outliers' gradients. We assign higher gradients to areas with larger error rates, thus placing greater emphasis on the recognition of difficult-to-classify samples. This can mitigate the effect of low-quality samples on the model's performance. By integrating ECIoU and FocalL1, we obtain the final Focal-ECIoU loss. The Focal-ECIOU loss formula is shown in Equation (10).

$$L_{f(x)} = \begin{cases} -\dfrac{\alpha x^2[2\ln(\beta x) - 1]}{4}, \\ \qquad 0 < x \leqslant 1; \dfrac{1}{e} \leqslant \beta \leqslant 1 \\ \\ -\alpha \ln(\beta) x + C, \\ \qquad x > 1; \dfrac{1}{e} \leqslant \beta \leqslant 1 \end{cases} \tag{9}$$

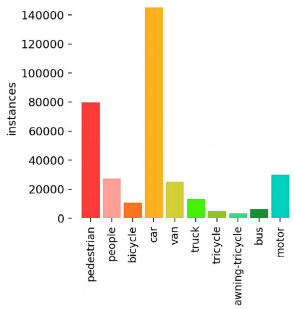


FIGURE 15. Category sample distribution statistics of VisDrone2019 dataset.

$$L_{\text{Focal}-\text{ECIoU}} = IoU^{\gamma} L_{\text{ECIoU}} \tag{10}$$

Here, x denotes the disparity between the true value and the predicted value, e denotes the base of the natural logarithm, β denotes used to control the curvature of the curve, C denotes a constant, and γ denotes a parameter used to control the degree to which outliers are suppressed.

As the IoU increases, the associated loss also intensifies, serving as an effective weighting mechanism. By assigning a greater loss to superior regression targets, this approach aids in enhancing the precision of the regression.

IV. EXPERIMENTS AND RESULTS

A. DATASET

This study uses the VisDrone2019 [45] public dataset. It is a vast image library tailored for UAV vision research, offering high-quality UAV visuals along with their corresponding ground truth annotations, sourced from varied times, locations, and weather conditions. The VisDrone dataset is collected using professional UAV photography techniques to ensure high image resolution, quality, and clarity, and each image is accurately annotated with objects, encompassing a broad range of scenes, lighting, weather, and seasonal changes. VisDrone2019 has ten detection categories. The distribution of category samples in the VisDrone2019 dataset is uneven. The category distribution sample statistics are shown in Fig. 15.

Fig. 16 represents a distribution chart which illustrates the dimensions of all label sizes in the training dataset. In this graph, the vertical axis denotes the height of the label box, while the horizontal axis corresponds to its width. Observations reveal a concentration of points in the lower left quadrant, suggesting a predominance of smaller objects within the VisDrone2019 dataset. This trend mirrors the real-world applications of drones, aligning well with the research context and issues addressed in this paper.

B. EXPERIMENTAL ENVIRONMENT AND TRAINING PARAMETER SETTINGS

This experiment used an Intel Core ™i7-12700K CPU processor with 64G of running memory and Ubuntu 22.04 LTS

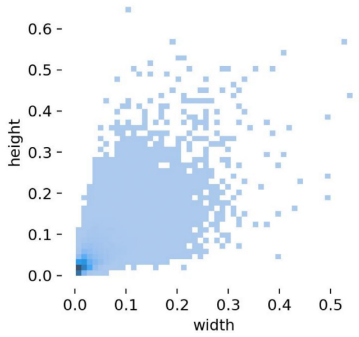


FIGURE 16. Size distribution of all labels in the training set. The horizontal axis (x-axis) represents the relative width of the object compared to the entire image width, while the vertical axis (y-axis) represents the relative height of the object compared to the image's total height.

TABLE 1. Hyperparameter settings.

Hyperparameter	Value
Epochs	300
Warmup epochs	3
Image size	640
batch_size	32
Learning_rate	0.01
Weight_decay	0.0005
Momentum	0.937
Optimizer	SGD

operating system. The experiment was performed using an NVIDIA RTX2080Ti GPU with 12G of video memory for the parallel acceleration of the network. PyCharm was used as the development platform. Training, validating, and testing were performed under the same hyperparameters. The Hyparameter settings are shown in Table 1.

C. EVALUATION INDICATORS

In this study, we aim to conduct a comprehensive analysis of our model, focusing on its precision and complexity. FLOPs (Floating-point Operations Per Second), indicating floating-point operations per second, gauge the computational complexity during model training. Params (Parameters), denoting the count of parameters in the model, quantify the computational memory resources utilized.

Precision (P) denotes the fraction of accurately predicted positive samples relative to all actual positive samples. The formula for Precision is displayed in the following Equation(11).

$$P = \frac{TP}{TP + FP} \tag{11}$$

Recall (R) represents the proportion of correctly identified positive samples out of all predicted samples. The formula for Recall is provided in the subsequent Equation(12).

$$R = \frac{TP}{TP + FN} \tag{12}$$

We hope that both precision and recall are high, but the two indicators are contradictory and cannot achieve both high. Therefore, based on the balance point between them, we choose an appropriate threshold to define a new indicator: F1-Score (F1). F1 is the harmonic mean of P and R, serving as a combined metric to optimize and maintain a balance between their performances. The formula for the F1 is presented in the upcoming Equation(13).

$$F1 = \frac{2P * R}{P + R} \tag{13}$$

mAP (Mean Average Precision) denotes the mean value of average precision across all individual categories. The mAP formula is shown in Equation(14).

$$mAP = \frac{1}{c}\sum_{i=1}^{c}\left(\int_0^1 P(R)dR\right)_i \tag{14}$$

Here, TP (true positives) signifies the count of samples correctly identified as positive. FP (false positives) refers to the quantity of samples wrongly classified as positive, though they are negative. FN (false negatives) refers to the count of instances incorrectly labeled as negative when they are in fact positive.

D. COMPARISON EXPERIMENTS

1) COMPARISON EXPERIMENT OF IMPROVED DECOUPLED HEAD

To establish the accuracy and efficacy of the GS-Decoupled Head put forth in this study, we designed a comparative experiment against YOLOX Decoupled Head. The baseline model for this experiment was the YOLOv5s-stru equipped with a Coupled Head. This was sequentially replaced first by YOLOX Decoupled Head and then by the GS-Decoupled Head developed in this research. To ensure an unbiased and valid experiment, we conducted a full cycle of retraining, validating, and testing for all the models under scrutiny. The experimental results are displayed in Table 2.

The comparison experimental results demonstrate that both Model A, which utilizes a YOLOX decoupled head, and Model B, employing a GS-Decoupled Head, substantially outperform the baseline model in performance metrics such as mAP0.5, mAP0.5:0.95, mAP0.75, and F1 Score, evidencing the significant optimization effect of the decoupled head. The performance of the GS-Decoupled Head closely mirrors that of the YOLOX decoupled head across the three mAP indicators. Although the F1-Score of Model B is slightly lower than that of Model A, the marginal difference of only 0.006 indicates that their F1 performance is nearly identical. In processing small and medium-scale targets, both models exhibit comparable performance; however, when dealing with large-scale targets, Model B, with the GS-Decoupled Head, has a slight edge, corroborating the decoupled head's

TABLE 2. Performance comparison of the detector with different decoupled heads.

Model	mAP0.5 (%)	mAP0.5:0.95 (%)	mAP0.75 (%)	F1	AP_{small} (%)	AP_{medium} (%)	AP_{large} (%)	Params (M)
Baseline(Coupled Head)	40.2	22.0	21.1	0.440	12.9	29.8	34.0	2.12
A(Decoupled Head)	43.9	25.3	25.4	0.472	16.0	32.8	37.4	9.33
B(GS-Decoupled Head)	43.8	25.1	24.8	0.466	15.8	32.6	38.2	5.83

advantage in handling targets of different scales, particularly large-scale ones. Notably, this advantage is preserved even when GSConv is applied to reduce computational complexity.

While Models A and B are nearly equivalent in terms of key performance indicators, Model B (5.8M) reduces parameter volume by approximately 37% compared to Model A (9.3M). This implies that Model B requires less storage space and computational resources, thereby enhancing operational efficiency on resource-limited devices.

In conclusion, the experimental results validate that our GS-Decoupled Head, while maintaining performance comparable to the YOLOX Decoupled Head, significantly reduces the model parameter quantity, boosts model efficiency, and meets practical application requirements. This attests to the efficacy and efficiency of the GS-Decoupled Head.

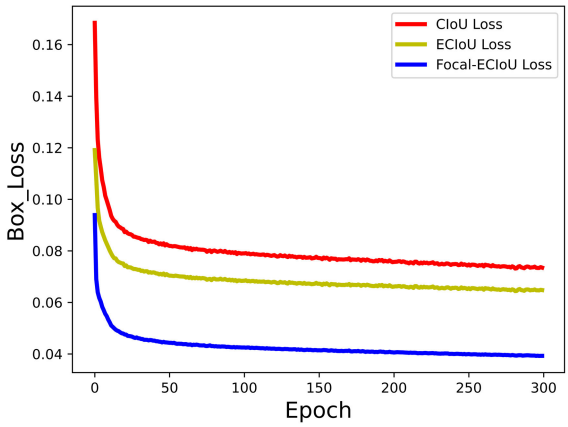


FIGURE 17. The loss comparison curve for CIoU, ECIoU, and Focal-ECIoU.

2) COMPARISON EXPERIMENT OF IMPROVED LOSS FUNCTION

To substantiate the efficacy of the proposed Focal-ECIoU for the task of UAV imagery object detection, we conceived a comparison experiment encompassing CIoU, ECIoU, and Focal-ECIoU. The baseline performance was gauged using the YOLOv5s-stru model that employs CIoU Loss as the loss function. This was successively replaced with ECIoU Loss and the proposed Focal-ECIoU Loss for this research. To ensure the experiment's fairness and legitimacy, we conducted a full retraining, validation, and testing cycle for all the models in consideration. The experimental results are displayed in Fig. 17 and Table 3.

Upon analyzing the comparative experimental results, we find that Model B has shown an enhancement relative to both the Baseline and Model A in mAP0.5, elevating by 2.7% and 1%, respectively. This indicates that Focal-ECIoU Loss can augment the comprehensive performance of the model. Under stricter evaluation metrics such as mAP0.5:0.95 and mAP0.75, Model B continues to outperform the Baseline and Model A, signifying that Focal-ECIoU Loss maintains high performance even when higher overlaps between predicted and true bounding boxes are required. In terms of the F1-Score evaluation metric, Model B also improved compared to the Baseline and Model A, suggesting that Focal-ECIoU Loss achieves a satisfactory balance between recall and precision. When recognizing objects of different sizes, Focal-ECIoU Loss likewise exhibits superiority, especially in the identification of small objects, where Model B exhibits the most significant improvement compared to the Baseline and Model A, with the AP_{small} increasing by 4.7% and 3%, respectively.

The curve of the loss function can illustrate the model's behavior during the training phase and directly signal the rate of its convergence. The bounding box regression loss, a measure employed in object detection to gauge the divergence between the predicted and actual boxes, serves as an indicator of the model's proficiency in the bounding box regression task. This study compared models' bounding box regression losses using CIoU, ECIoU, and Focal-ECIoU respectively. Upon analysis of the loss function curves, one can note that the bounding box regression losses for CIoU, ECIoU, and Focal-ECIoU all exhibit a downward trend as the number of training iterations growth. However, Focal-ECIoU Loss converges the fastest, with the loss dropping below 0.045 after 50 epochs. At the end of model training, the losses for CIoU, ECIoU, and Focal-ECIoU were 0.073, 0.065, and 0.039, respectively, with Focal-ECIoU Loss's loss value being closest to 0.

Comparative experimental results indicate that Focal-ECIoU Loss outperforms both CIoU Loss and ECIoU Loss in performance enhancement. It significantly improves the model's localization capability. Focal-ECIoU Loss is characterized by a swift convergence rate and smaller concluding loss values. Demonstrating strong stability and adaptability, Focal-ECIoU Loss tackles the imbalance between samples classified as positive and those classified as

TABLE 3. Performance comparison of the detector with different loss functions.

Model	mAP0.5 (%)	mAP0.5:0.95 (%)	mAP0.75 (%)	F1	AP_{small} (%)	AP_{medium} (%)	AP_{large} (%)
Baseline(CloU)	40.2	22.0	21.1	0.440	12.9	29.8	34.0
A(ECIoU)	40.9	22.2	21.0	0.441	13.1	30.1	33.8
B(Focal-ECIoU)	41.3	22.6	21.6	0.449	13.5	30.5	34.1

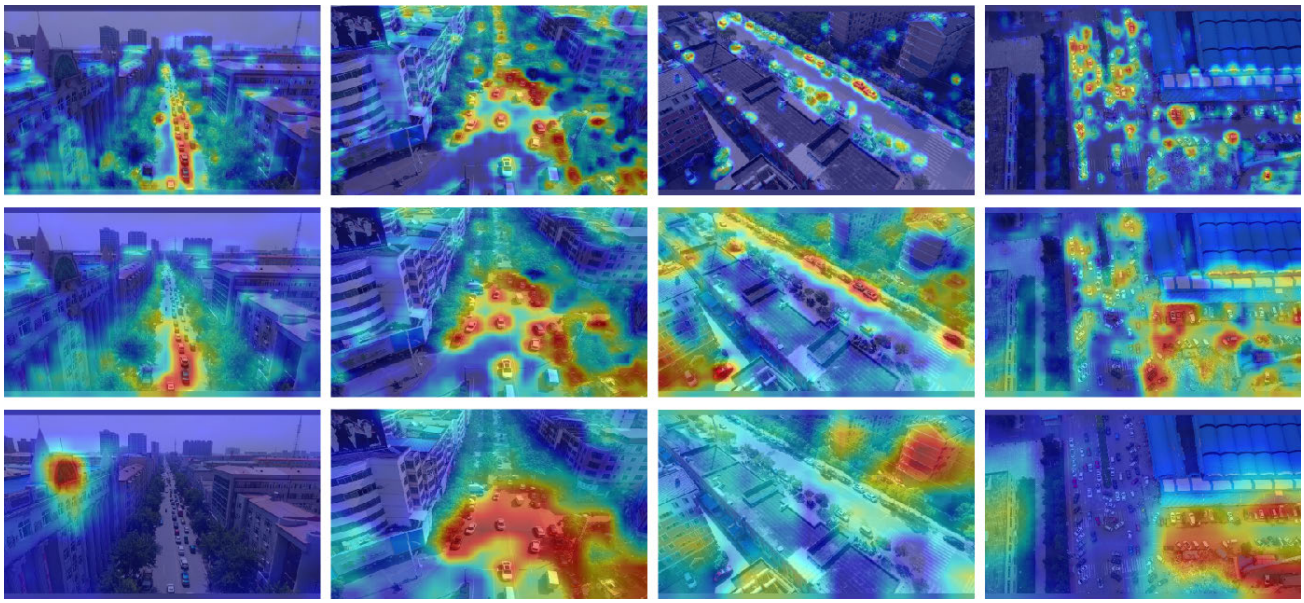


FIGURE 18. The attention of different detection heads to the small objects.

negative, proficiently differentiating complex backgrounds. This ultimately significantly boosts the model's performance.

E. ABLATION STUDY

1) ABLATION STUDY ON THE IMPROVED NETWORK STRUCTURES

We evaluate the effect of numerous network structure optimisation methods on the performance of object detection algorithms under identical experimental conditions through ablation experiments of the improved network structure. In identical conditions, YOLOv5s is used as the baseline model in the experiments,its results functioning as a standard. The ablation studies on the enhanced network structure are conducted by progressively incorporating the network improvement strategies proposed in this research to the base model of YOLOv5s. The experimental results are displayed in Fig. 18, Fig. 19, and Table 4.

As shown in Fig. 18, the first, second, and third rows each respectively illustrate the attention given to small objects by the detection heads of sizes 80 × 80, 40 × 40, and 20 × 20. Comparing the attention of each detection head on small objects, we found that the 20 × 20 resolution detection head has a lot of noise in its attention to object features, covering more background and having lower attention to small objects, even completely ignoring small objects in UAV imagery. Only detection heads with resolutions of 80×80 and 40×40 are capable of identifying these small objects in UAV imagery.

Fig. 19 shows the attention of the 160×160 detection head on small objects. We observe that the 160 × 160 detection head exhibits a high degree of attention to small targets within the UAV images. Consequently, it can derive small object features from UAV imagery more efficiently.

Through the analysis of ablation experiment results, we found that upon the removal of the 20 × 20 large target detection head and the related feature extraction and fusion layers in Model A, which incorporated YOLOv5s, the accuracy decreased marginally. However, the model's parameters and computation significantly reduced by approximately 71% and 26%, respectively. This indicates that while the elimination of these layers might marginally decrease accuracy, it significantly reduces the model's computational complexity and memory overhead, thus rendering the model more lightweight and efficient.

Model B, built upon Model A, incorporated a 160 × 160 small target detection head, leading to a notable enhancement in the model's performance metrics, particularly in the detection performance of small targets,

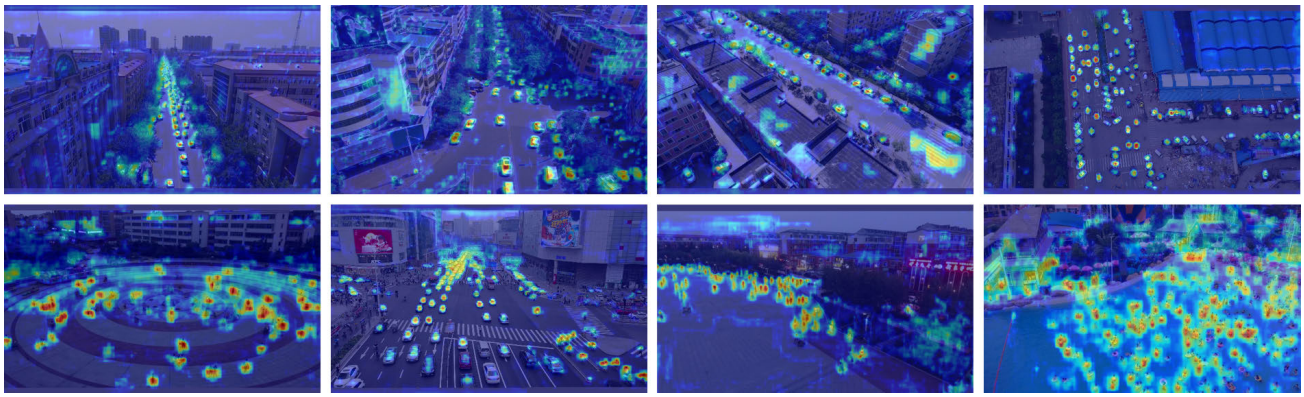


FIGURE 19. The attention of 160 × 160 detection head to small objects.

TABLE 4. Results of the ablation study on the improved network structure.

Model	mAP0.5 (%)	mAP0.5:0.95 (%)	F1	AP_{small} (%)	AP_{medium} (%)	AP_{large} (%)	Params (M)	FLOPs (G)
YOLOv5s	35.1	19.1	0.396	10.2	26.6	35.4	7.04	15.8
A	34.6	18.3	0.392	9.90	25.9	29.3	2.02	11.6
B	39.8	20.6	0.429	12.9	29.2	31.0	2.04	14.1
YOLOv5s-stru	40.2	22.0	0.440	12.9	29.8	34.0	2.12	14.5

TABLE 5. Results of the ablation study on the improved strategies.

Model	Structural Improvement	Dense CSP-Darknet53	Fusion Block	GS-Decoupled Head	Focal-ECIoU	mAP0.5 (%)	mAP0.5:0.95 (%)	AP_{small} (%)	AP_{medium} (%)	AP_{large} (%)
Baseline						35.1	19.1	10.2	26.6	35.4
A	✓					40.2	22.0	12.9	29.8	34.0
B	✓	✓				41.0	22.5	13.5	29.9	34.8
C	✓	✓	✓			42.9	24.7	15.1	33.1	37.6
D	✓	✓	✓	✓		45.7	26.6	16.9	35.0	39.4
YOLO-UAV	✓	✓	✓	✓	✓	46.7	27.4	17.3	36.0	39.7

✓ indicates the selected module.

which improved by approximately 30%. Concurrently, the increase in the number of parameters and computation due to this modification was minimal. This suggests that the addition of the small target prediction layer significantly improves performance while having a minimal impact on model complexity.

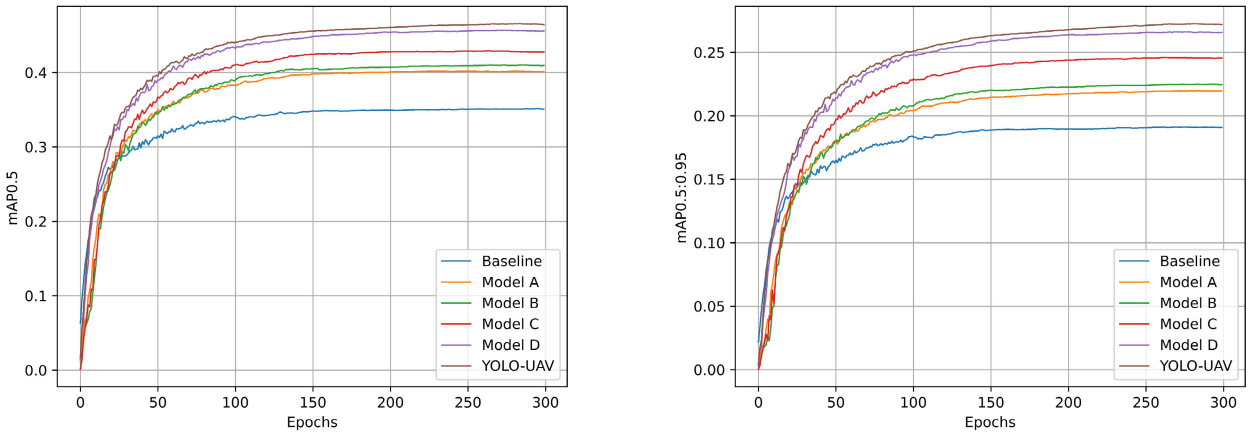
YOLOv5s-stru, derived from Model B, underwent improvements in the Neck structure. This change had a negligible impact on the model's parameters and computational complexity, yet it further enhanced the model's performance, particularly in the detection performance of medium and large targets.

Compared to the baseline model, YOLOv5s, the improved YOLOv5s-stru model demonstrated enhancements in all performance metrics. Moreover, it exhibited improved performance across targets of various sizes, especially in the detection performance of small and medium targets. Simultaneously, the model's parameters and computation exhibited a significant decrease.

The results of the ablation experiments reveal that our improvements to the network structure of YOLOv5s are effective. Our enhancement strategies significantly reduced the model's complexity and memory requirements while improving performance.

2) ABLATION STUDY ON THE IMPROVED STRATEGIES

To substantiate the efficacy of the enhanced strategies proposed in this study, we executed ablation experiments, examining the influence of various strategies on the object detection algorithm's performance under identical experimental parameters. Using YOLOv5s as the baseline model under the same test conditions, we progressively incorporated



(a) metrics/mAP0.5 (b) metrics/mAP0.5:0.95

FIGURE 20. Comparison of experimental mAP0.5 and mAP0.5:0.95 for ablation. The (a) is the comparison of experimental mAP0.5 for ablation, and the (b) is the comparison of experimental mAP0.5:0.95 for ablation.

TABLE 6. AP and mAP0.5 comparison of different target detection methods on VisDrone2019 dataset.

Method	Backbone	Target class(AP/%)										mAP0.5/%
		Pede-strian	Person	Bicycle	Car	Van	Truck	Tricycle	Awning-tricycle	Bus	Motor	
Faster R-CNN [46]	ResNet-50	21.4	15.6	6.7	51.7	29.5	19.0	13.1	7.7	31.4	20.7	21.7
Faster R-CNN [46]	ResNet-101	20.9	14.8	7.3	51.0	29.7	19.5	14.0	8.8	30.5	21.2	21.8
Cascade R-CNN [46]	ResNet-50	22.2	14.8	7.6	54.6	31.5	21.6	14.8	8.6	34.9	21.4	23.2
RetinaNet [46]	ResNet-50	13.0	7.9	1.4	45.5	19.9	11.5	6.3	4.2	17.8	11.8	13.9
CenterNet [47]	Hourglass-104	22.6	20.6	14.6	59.7	24.0	21.3	20.1	17.4	37.9	23.7	26.2
DMNet [48]	ResNet-50	28.5	20.4	15.9	56.8	37.9	30.1	22.6	14.0	47.1	29.2	30.3
CDNet [48]	ResNeXt-101	35.6	19.2	13.8	55.8	42.1	38.2	33.0	**25.4**	49.5	29.3	34.2
HR-Cascade++ [48]	HRNet-W40	32.6	17.3	11.1	54.7	42.4	35.3	32.7	24.1	46.5	28.2	32.5
MSC-CenterNet [48]	Hourglass-104	33.7	15.2	12.1	55.2	40.5	34.1	29.2	21.6	42.2	27.5	31.1
DBAI-Det [49]	ResNeXt-101	36.7	12.8	14.7	47.4	38.0	41.4	23.4	16.9	31.9	16.6	28.0
FPN [50]	ResNet-50	33.0	25.8	13.9	69.4	40.0	34.3	27.4	13.4	49.1	37.6	35.6
YOLOv3 [50]	Darknet-53	18.1	9.9	2.0	56.6	17.5	17.6	6.7	2.9	32.4	17.0	17.1
SlimYOLOv3 [51]	Darknet-53(SPP3-50)	17.4	9.3	2.4	55.7	18.3	16.9	9.1	3.0	26.9	17.0	17.6
YOLOv3-LITE [52]	Darknet-53(DSC)	34.5	23.4	7.9	70.8	31.3	21.9	15.3	6.2	40.9	32.7	28.5
YOLOv4 [53]	CSPDarknet	24.8	12.6	8.6	64.3	22.4	22.7	11.4	7.6	44.3	21.7	30.7
Modified YOLOv4 [53]	CSPDarknet	28.2	15.9	5.8	65.7	25.2	26.1	13.8	8.1	40.2	26.1	32.5
YOLOv5s [26]	CSPDarknet53(C3)	40.8	32.6	13.6	74.6	37.6	32.8	21.9	12.5	44.9	40.0	35.1
MSA-YOLO [54]	CSPDarknet(MSAU)	33.4	17.3	11.2	76.8	41.5	41.4	14.8	18.4	60.9	31.0	34.7
YOLOv8 [55]	CSPDarknet(C2f)	50.2	39.7	21.3	74.8	**50.5**	**46.2**	**33.3**	22.1	**67.4**	45.3	45.1
YOLO-UAV	Dense_CSPDarknet53	**55.7**	**45.3**	**21.4**	**84.8**	49.4	42.3	32.0	19.1	63.5	**53.1**	**46.7**

Bold text indicates optimal results.

the improved strategies proposed in our study. During the training process, we used pre-trained weights for all models, and training started directly from 0. The experimental results are displayed in Fig. 20 and Table 5.

Through the analysis of ablation experiment results, we found that in Model A, through network structure adjustments, the model achieved significant improvements across all performance indicators. In particular, the AP_{small} of small-scale targets increased the most, by 26.47%. This suggests that the improvements to the network structure enable the model to extract and utilize image feature information more effectively, especially the extraction of detailed features.

In Model B, the addition of the Dense_C3 module improved the target detection performance across all scales. This indicates that the Dense_C3 module can effectively promote feature propagation, enhance the richness of features, and reduce feature loss.

(a) Detection effect in lighting variations.

(b) Detection effect in complex backgrounds.

(c) Detection effect in object occlusion.

(d) Detection effect in blurry image.

FIGURE 21. YOLO-UAV detection effect in different scenarios.

In Model C, the incorporation of the Fusion Block led to improvements in all performance metrics. This suggests that the Fusion Block, by merging features of different levels, enables the model to better utilize both global and local information.

In Model D, the inclusion of the GS-Decoupled Head led to further improvements in model performance, particularly for small-scale targets, whose AP_{small} increased most significantly by 11.92%. This is because the GS-Decoupled Head decouples classification and regression tasks, allowing each task to be independently optimized, thereby enhancing the model's capacity to identify small targets.

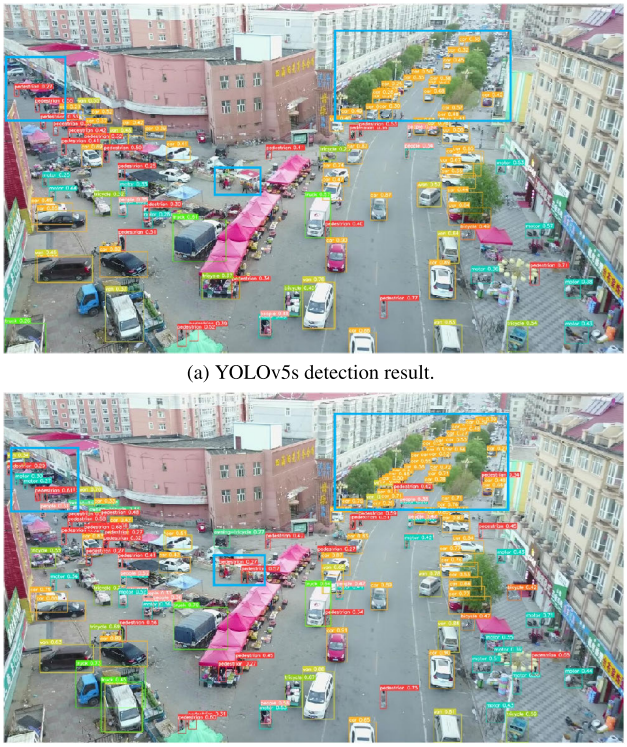
In the final YOLO-UAV model, the introduction of the Focal-ECIoU loss function led to improvements across all performance metrics. This demonstrates the advantage of the Focal-ECIoU loss function in dealing with the imbalance between samples classified as positive and those classified as negative, and optimizing the precision of target positioning.

When all improved strategies were added to the model, compared to the baseline model, the YOLO-UAV achieved optimal performance across all performance metrics, with mAP0.5 and mAP0.5:0.95 improving by 33.05% and 43.46%, respectively. The model's recognition ability for targets of all sizes also significantly improved, especially for small targets, whose AP_{small} increased from 10.2% to 17.3%, an improvement of 69.61%. The results of the ablation experiments demonstrate that each improved strategy effectively improved the model's target detection performance, while also showing that the effects of these improvement strategies are cumulative.

F. COMPARISON WITH STATE-OF-THE-ART METHODS

To establish the efficacy of YOLO-UAV in identifying a variety of targets in UAV imagery, we performed a comparative study on the VisDrone2019 dataset, juxtaposing it with multiple cutting-edge UAV image detection techniques. The outcomes of this comparative analysis with advanced methods are displayed in the Table 6.

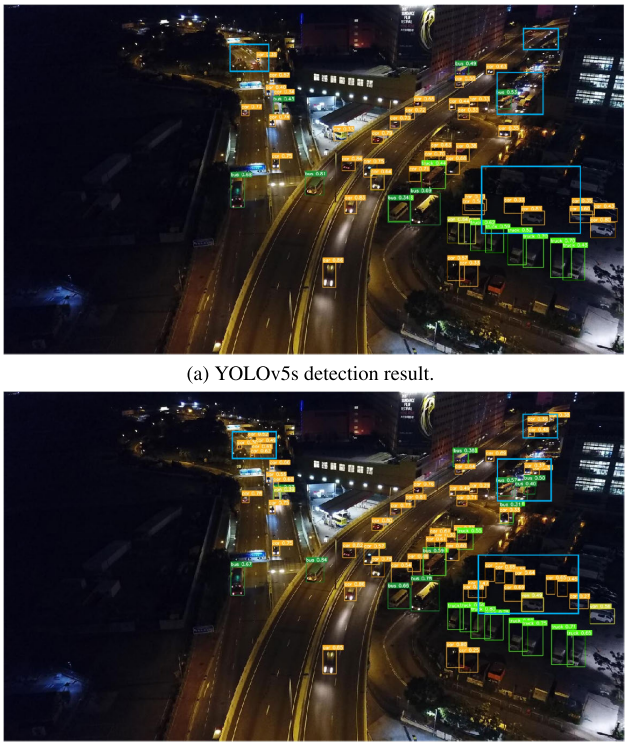
YOLO-UAV outperformed other leading-edge techniques, elevating the mAP0.5 from 45.1% to 46.7%, marking an improvement of 3.6% over the second best, YOLOv8. Compared to other advanced methods, YOLO-UAV obtained the best detection performance in target categories such as Pedestrian, Person, Bicycle, Car, and Motor, with AP for these categories reaching 55.7%, 45.3%, 21.4%, 84.8%, and 53.1% respectively, demonstrating YOLO-UAV excellent



(a) YOLOv5s detection result.

(b) YOLO-UAV detection result.

FIGURE 22. Comparison of detection results between YOLOv5s and YOLO-UAV in complex background environments.



(a) YOLOv5s detection result.

(b) YOLO-UAV detection result.

FIGURE 23. Comparison of detection results between YOLOv5s and YOLO-UAV in low-light nighttime environments.

generalizability and detection accuracy. However, in the Tricycle and Awning-tricycle categories, the performance of YOLO-UAV is not very good. This suggests that while YOLO-UAV performs well in detecting common categories, it underperforms in some uncommon or complex categories due to insufficient training data or a specific model design that cannot fully capture the features of these categories.

YOLO-UAV has a clear advantage in detection accuracy in the Pedestrian, Person, and Motor categories, which have the smallest objects. In these categories, the AP improved by 11.0%, 14.1%, and 17.2%, respectively, compared to the second-best network, YOLOv8. YOLO-UAV also excels in detecting larger categories such as Car, Van, Bus, and Truck, with AP reaching up to 84.8% for Car, 63.5% for Bus, 49.4% for Van, and 42.3% for Truck. For categories with fewer instances, like Bicycle, YOLO-UAV also shows the best detection performance, compared to the second-best YOLOv8. Across all categories, YOLO-UAV performed exceptionally well, especially in categories where other models underperformed, such as Bicycle and Truck, maintaining high accuracy. This demonstrates that YOLO-UAV can sufficiently extract feature information when the number of object instances is low. This experimental evidence shows that YOLO-UAV performs well in UAV image object detection tasks even when object instances are small.

Compared to the original algorithm, YOLO-UAV can capture richer object features, thereby notably augmenting the feature extraction efficacy for small entities and boosting the network's detection precision. This highlights that the YOLO-UAV, as introduced in this paper, possesses noteworthy strengths in dealing with UAV image object detection tasks.

G. METHOD EFFECTIVENESS ANALYSIS

In order to substantiate the real-world applicability and detection performance of YOLO-UAV, we performed evaluations using representative and difficult images drawn from the VisDrone2019 test set. The results derived from these detection tests are illustrated in the Fig. 21. YOLO-UAV exhibits superior detection performance in UAV images with various shooting angles, lighting changes, object occlusions, and complex and densely distributed backgrounds. It can detect more small objects and distant objects, effectively suppress interference from image background noise, and selectively extract important feature information beneficial for UAV imagery object detection assignments.

To delve deeper into the effect variations between the baseline YOLOv5s and the YOLO-UAV in managing UAV imagery object detection tasks, we randomly chose images depicting small object situations set against diverse environmental backgrounds from the VisDrone2019 test set for assessment, and performed a visual comparative study.

By comparing Fig. 22a and Fig. 22b, we found that in complex background environments where the features such as color, texture, and shape of target objects are similar to those of the background, the baseline model

YOLOv5s incorrectly identifies the background as target objects. Also, the dense arrangement of targets in the image causes many small and distant objects to be obscured by other objects, resulting in undetectable areas for small and distant objects, resulting to a large number of missed detection. The YOLO-UAV model strengthens the network's feature extraction ability in small object areas, separating useful information for UAV image object detection from a multitude of feature data. It exhibits strong anti-interference capabilities when dealing with complex background information and can accurately identify objects.

By comparing Fig. 23a and Fig. 23b, we found that in low-light conditions at night, the edges and details of target objects are difficult to discern. Noise and interference in the background increase and the contrast between target objects and the background decreases, causing the outlines and features of target objects to become blurred. This leads to some missed targets in the baseline model YOLOv5s. The YOLO-UAV model, however, uses multi-scale feature fusion to enable the model to learn strong location features from shallow features. This process allows deep features to conduct more precise fine-grained detection, thereby reducing noise interference and effectively improving the detection of missed objects in low-light conditions at night.

Overall, when dealing with UAV imagery object detection tasks, the YOLO-UAV has more obvious advantages compared to the YOLOv5s. The YOLO-UAV is less affected by external conditions such as lighting and still performs well in nighttime conditions. It has stronger detection capabilities for small objects, distant objects, objects in complex backgrounds, and densely arranged objects. It can effectively avoid missed detections and false detections, demonstrating excellent generalization capabilities that can meet the practical task requirements.

V. CONCLUSION

In this study, We have proposed an object detection method based on efficient multi-scale feature fusion. The method is aimed at tackling the complex issues of intricate backdrops, diminutive object sizes, and significant target concealment that are prevalent in UAV images. These challenges often hinder the effectiveness of existing object detectors in target feature extraction and achieving superior detection accuracy. To tackle these issues, we put forth a re-engineered design for both the feature extraction and fusion networks. This design intends to minimize downsampling losses during target feature extraction, streamline the network architecture, and enhance multi-scale feature fusion efficacy. By adding a 160 $\times$ 160 small object detection head, We improve the ability of the networks to extract detailed feature information, particularly concerning small objects. To address the issue of losing feature information when extracting features from small objects, we introduce the Dense_C3 module. This module integrates dense connections within the C3 module of the backbone network, forming the Dense_CSPDarknet backbone network. This network capitalizes on feature reuse to extract implicit network information, increases diversity in the inputs of subsequent layers, and improves feature extraction capability for small objects. To achieve richer and more accurate feature representations, reduce interference from complex background noise, and enable better learning of small object features, we introduce an efficient feature fusion block in the Neck section of the network. This module incorporates various strategies such as structural re-parameterization and ELAN. Moreover, we redesigned a simple and efficient GS-Decoupled Head to decouple the classification and regression tasks, effectively avoiding conflicts arising from different feature information requirements of these tasks and reducing prediction biases caused by task differences. Additionally, we propose a redesigned Focal-ECIoU Loss to tackle the imbalance amid positive and negative samples. This loss function effectively distinguishes complex backgrounds, improves the model's regression and localization capabilities, accelerates network convergence, and enhances model performance. Findings from experimental studies from the VisDrone2019 dataset indicate that our proposed YOLO-UAV significantly improves object detection accuracy compared to various advanced object detection methods. It exhibits excellent generalization ability and fulfills the requirements of practical UAV imagery object detection tasks.

However, despite the achievements of YOLO-UAV, our research has some limitations. For instance, although our GS-Decoupled Head design drastically reduces parameters and computational complexity compared to the decoupled head in YOLOX, it may still impact real-time performance. Additionally, there is room for improvement in optimizing specific categories within our model. As for prospective research directions, we plan to explore the design of lightweight and efficient network models to fulfill real-time detection requirements better. We will also delve deeper into optimizing specific categories, including collecting more targeted data, conducting in-depth feature engineering, and making more detailed adjustments and optimizations to the model structure.

REFERENCES

[1] P. Mittal, R. Singh, and A. Sharma, "Deep learning-based object detection in low-altitude UAV datasets: A survey," *Image Vis. Comput.*, vol. 104, Dec. 2020, Art. no. 104046, doi: 10.1016/j.imavis.2020.104046.

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

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

[4] K. He, G. Gkioxari, P. Dolr, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Jun. 2017, pp. 2961–2969.

[5] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6154–6162.

[6] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 821–830.

[7] J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO: From YOLOv1 and beyond," 2023, *arXiv:2304.00501*.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 21–37.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[10] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.

[11] M. Wang, X. Luo, X. Wang, and X. Tian, "Research on vehicle detection based on faster R-CNN for UAV images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 1177–1180.

[12] H. Huang, L. Li, and H. Ma, "An improved cascade R-CNN-based target detection algorithm for UAV aerial images," in *Proc. 7th Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2022, pp. 232–237.

[13] Q. Lin, Y. Ding, H. Xu, W. Lin, J. Li, and X. Xie, "ECascade-RCNN: Enhanced cascade RCNN for multi-scale object detection in UAV images," in *Proc. 7th Int. Conf. Autom., Robot. Appl. (ICARA)*, Feb. 2021, pp. 268–272.

[14] W. Liu, J. Qiang, X. Li, P. Guan, and Y. Du, "UAV image small object detection based on composite backbone network," *Mobile Inf. Syst.*, vol. 2022, pp. 1–11, Apr. 2022, doi: 10.1155/2022/7319529.

[15] Y. Gao, R. Hou, Q. Gao, and Y. Hou, "A fast and accurate few-shot detector for objects with fewer pixels in drone image," *Electronics*, vol. 10, no. 7, p. 783, Mar. 2021, doi: 10.3390/electronics10070783.

[16] C. Chen, Z. Zheng, T. Xu, S. Guo, S. Feng, W. Yao, and Y. Lan, "YOLO-based UAV technology: A review of the research and its applications," *Drones*, vol. 7, no. 3, p. 190, Mar. 2023, doi: 10.3390/drones7030190.

[17] A. Jawaharlalnehru, T. Sambandham, V. Sekar, D. Ravikumar, V. Loganathan, R. Kannadasan, A. A. Khan, C. Wechtaisong, M. A. Haq, A. Alhussen, and Z. S. Alzamil, "Target object detection from unmanned aerial vehicle (UAV) images based on improved YOLO algorithm," *Electronics*, vol. 11, no. 15, p. 2343, Jul. 2022, doi: 10.3390/electronics11152343.

[18] O. Sahin and S. Ozer, "YOLODrone: Improved YOLO architecture for object detection in drone images," in *Proc. 44th Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2021, pp. 361–365.

[19] Y. Cheng, "Detection of power line insulator based on enhanced YOLO model," in *Proc. IEEE Asia–Pacific Conf. Image Process., Electron. Comput. (IPEC)*, Apr. 2022, pp. 626–632.

[20] L. Shen, B. Lang, and Z. Song, "CA-YOLO: Model optimization for remote sensing image object detection," *IEEE Access*, vol. 11, pp. 64769–64781, 2023, doi: 10.1109/ACCESS.2023.3290480.

[21] H. V. Koay, J. H. Chuah, C.-O. Chow, Y.-L. Chang, and K. K. Yong, "YOLO-RTUAV: Towards real-time vehicle detection through aerial images with low-cost edge devices," *Remote Sens.*, vol. 13, no. 21, p. 4196, Oct. 2021, doi: 10.3390/rs13214196.

[22] X. Wang, W. Li, W. Guo, and K. Cao, "SPB-YOLO: An efficient real-time detector for unmanned aerial vehicle images," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Apr. 2021, pp. 099–104.

[23] M. Huang, Y. Zhang, and Y. Chen, "Small target detection model in aerial images based on TCA-YOLOv5m," *IEEE Access*, vol. 11, pp. 3352–3366, 2023, doi: 10.1109/ACCESS.2022.3232293.

[24] T. Liang, H. Bao, W. Pan, and F. Pan, "Traffic sign detection via improved sparse R-CNN for autonomous vehicles," *J. Adv. Transp.*, vol. 2022, pp. 1–16, Mar. 2022, doi: 10.1155/2022/3825532.

[25] T. Liang, H. Bao, W. Pan, X. Fan, and H. Li, "DetectFormer: Category-assisted transformer for traffic scene object detection," *Sensors*, vol. 22, no. 13, p. 4833, Jun. 2022, doi: 10.3390/s22134833.

[26] G. Jocher. *YOLOv5*. Ultralytics. [Online]. Available: https://github.com/ultralytics/yolov5

[27] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2117–2125.

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

[29] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 10781–10790.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4700–4708.

[31] Y. Lee, J.-W. Hwang, S. Lee, Y. Bae, and J. Park, "An energy and GPU-computation efficient backbone network for real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019.

[32] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[33] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 390–391.

[34] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style ConvNets great again," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2021, pp. 13733–13742.

[35] X. Xu, Y. Jiang, W. Chen, Y. Huang, Y. Zhang, and X. Sun, "DAMO-YOLO: A report on real-time object detection design," 2022, *arXiv:2211.15444*.

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

[37] 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.

[38] 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*.

[39] H. Li, J. Li, H. Wei, Z. Liu, Z. Zhan, and Q. Ren, "Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles," 2022, *arXiv:2206.02424*.

[40] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 516–520.

[41] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 658–666.

[42] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12993–13000.

[43] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, Sep. 2022.

[44] X. Chen, Q. Lian, X. Chen, and J. Shang, "Surface crack detection method for coal rock based on improved YOLOv5," *Appl. Sci.*, vol. 12, no. 19, p. 9695, Sep. 2022, doi: 10.3390/app12199695.

[45] 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.

[46] W. Yu, T. Yang, and C. Chen, "Towards resolving the challenge of long-tail distribution in UAV images for object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3257–3266.

[47] B. M. Albaba and S. Ozer, "SyNet: An ensemble network for object detection in UAV images," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 10227–10234.

[48] D. Du, L. Wen, and P. Zhu, "VisDrone-DET2020: The vision meets drone object detection in image challenge results," in *Computer Vision—ECCV*, A. Bartoli and A. Fusiello, Eds. Cham, Switzerland: Springer, 2020, pp. 692–712.

[49] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang, and L. Bo, "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Seoul, Korea (South), 2019, pp. 213–226.

[50] Y. Wan, Y. Zhong, Y. Huang, Y. Han, Y. Cui, Q. Yang, Z. Li, Z. Yuan, and Q. Li, "ARSD: An adaptive region selection object detection framework for UAV images," *Drones*, vol. 6, no. 9, p. 228, Aug. 2022, doi: 10.3390/drones6090228.

[51] P. Zhang, Y. Zhong, and X. Li, "SlimYOLOv3: Narrower, faster and better for real-time UAV applications," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 27–28.
[52] H. Zhao, Y. Zhou, L. Zhang, Y. Peng, X. Hu, H. Peng, and X. Cai, "Mixed YOLOv3-LITE: A lightweight real-time object detection method," *Sensors*, vol. 20, no. 7, p. 1861, Mar. 2020, doi: 10.3390/s20071861.
[53] S. Ali, A. Siddique, H. F. Ateş, and B. K. Güntürk, "Improved YOLOv4 for aerial object detection," in *Proc. 29th IEEE Conf. Signal Process. Commun. Appl. (SIU)*, Jun. 2021, pp. 1–4.
[54] G. T. Mao, T. M. Deng, and N. J. Yu, "Object detection in UAV images based on multi-scale split attention," *Acta Aeronaut. Astronaut. Sin*, vol. 43, no. 12, 2022, Art. no. 326738.
[55] U. G. Jocher. (Jun. 1, 2023). *YOLOv8*. [Online]. Available: https://github.com/ultralytics/ultralytics

CHENGJI MA received the B.Eng. degree from the Qingdao University of Technology, Qingdao, China, in 2021. He is currently pursuing the master's degree with the School of Information Science and Engineering (School of Cyberspace Security), Xinjiang University. His current research interests include computer vision and deep learning.

YANYUN FU received the Ph.D. degree in safety science and engineering from the University of Science and Technology of China, Hefei, China, in 2017. She was a Postdoctoral Researcher with the Institute for Public Safety Research, Tsinghua University, from 2017 to 2020. She is currently with the Beijing Academy of Science and Technology. Her current research interests include public safety and big data.

DEYONG WANG received the Ph.D. degree in safety science and engineering from the University of Science and Technology of China, Hefei, China, in 2013. He is currently a Professor (Senior Engineer) with the Key Laboratory of Big Data of Xinjiang Social Security Risk, Xinjiang Lianhaichuangzhi Information Technology Company Ltd., Ürümqi, China. His current research interests include public safety, big data, and artificial intelligence.

RUI GUO received the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2019. He is currently a Senior Engineer with the Key Laboratory of Big Data of Xinjiang Social Security Risk, Xinjiang Lianhaichuangzhi Information Technology Company Ltd., Ürümqi, China. His current research interests include computer vision, deep learning, and artificial intelligence.

XUEYI ZHAO received the Ph.D. degree in information and communication engineering from Zhejiang University, China, in 2015. He is currently a Senior Engineer with the Key Laboratory of Big Data of Xinjiang Social Security Risk, Xinjiang Lianhaichuangzhi Information Technology Company Ltd., Ürümqi, China. His current research interests include modeling, information retrieval, and parallel computing.

JIAN FANG received the B.S. degree in law from the Nanjing Army Command College, China, in 2011. He is currently a Senior Engineer with the Key Laboratory of Big Data of Xinjiang Social Security Risk, Xinjiang Lianhaichuangzhi Information Technology Company Ltd., Ürümqi, China. His current research interests include public safety, social governance, and emergency management.