ITC 3/55
Information Technology and Control
Vol. 55 / No. 3/ 2026
pp. 1248-1265
DOI 10.5755/j01.itc.55.3.45036

YOLOv8-PSA: A YOLOv8 Enhanced by Partial Self-Attention Mechanism for Urban UAV Detection

Received 2026/05/15 | Accepted after revision 2026/08/01

HOW TO CITE: Zhao, Y., Chen, Y., Liu, Y. (2026) YOLOv8-PSA: A YOLOv8 Enhanced by Partial Self-Attention Mechanism for Urban UAV Detection. *Information Technology and Control*, 55(3), 1248-1265. https://doi.org/10.5755/j01.itc.55.3.45036

YOLOv8-PSA: A YOLOv8 Enhanced by Partial Self-Attention Mechanism for Urban UAV Detection

Yiwei Zhao#

Department of Civil Engineering, Qilu Institute of Technology, Jinan, China, 250200;
e-mail: 241101300301@aliyun.com

Yu Chen#

College of Computer and Information Technology, China Three Gorges University, Yichang, China, 443002;
e-mail: cystuctgu2025@163.com

Yiyao Liu*

Glasgow College, University of Electronic Science and Technology of China, Chengdu, China, 611731; e-mail: 2960078L@student.gla.ac.uk

#Yiwei Zhao and Yu Chen contributed equally to this work, they are both first authors.

Corresponding author: 2960078L@student.gla.ac.uk

Unmanned Aerial Vehicle (UAV) detection in urban environments is critical for airspace management and emergency response. However, dynamic scale variations, complex backgrounds, and real-time requirements hinder existing methods. This research proposes YOLOv8-PSA, which integrates a Partial Self-Attention (PSA) mechanism to capture global contextual dependencies and suppress background noise. A multi-scale feature fusion module with a lightweight C2fCIB block enhances detection accuracy across varying scales. CIoU loss is adopted to optimize bounding box localization. Experiments on an urban UAV dataset demonstrate 97.8% precision, 92.5% recall, and 96.5% mAP50, with consistent improvements over the baseline YOLOv8n. Comparative and ablation studies validate the effectiveness of the proposed modules for practical urban UAV detection.

KEYWORDS: urban UAV detection; YOLOv8; global attention mechanism; multi-scale feature fusion; Urban UAV real-time detection

1. Introduction

With the advancement of flight control technology, Unmanned Aerial Vehicles (UAVs) have been widely utilized in people's daily life. The detection of UAV is of great importance in optimizing the management of airspace, ensuring enhanced aviation safety, and maintaining social order, and enhancing the efficiency of emergency response systems [15,18]. UAV-based sensing and imaging technologies have been extensively deployed for power system inspection and infrastructure monitoring, demonstrating versatile applicability across industrial and disaster scenarios [14, 26].

Existing UAV detection approaches encompass diverse sensing technologies, each tailored to specific operational scenarios—including acoustic sensors, visual imaging sensors, photoelectric detectors, radar systems, and radio frequency (RF) sensing technologies [27, 30]. Radar-based methods boast strong anti-interference capabilities but come with high costs. Infrared-based approaches perform well in nighttime scenarios yet are not only expensive but also susceptible to weather conditions [13]. Owing to the unique characteristics of UAVs—including low-altitude flight, diverse postures, small size, and rapid movements—these methods have notable limitations when it comes to UAV detection in complex urban environments. In recent years, vision-driven detection approaches have drawn substantial research focus for UAV identification, primarily due to their strengths in direct visual data parsing and adaptability to dynamic operational environments. In comparison to other methods, vision-based techniques are more practical and efficient for target detection, as they can directly analyze videos to detect UAVs [17, 43]. Comprehensive surveys have systematically reviewed passive sensor-based UAV detection and tracking methodologies specifically tailored for urban environments [39].

Traditional vision-based detection methods use manual operations such as sliding windows and operators for feature extraction [2]. The deep learning approaches leverages convolutional neural network (CNN) to learn feature extraction from large amounts of data. They have demonstrated remarkable efficacy in computer vision tasks. CNN-based UAV detection approaches can mainly be primarily divided into two-stage approaches (such as the R-CNN series) [6] and one-stage approaches (including YOLO and SSD) [2, 25]. The former has been shown to achieve high detection accuracy but with high computational complexity]. In contrast, the latter has been demonstrated to be faster but has shortcomings in detecting small targets and dense targets. Therefore, two-stage detection is more focused on accuracy, complex environments and small things. For achieving real-time UAV detection, single-stage detection approaches estimate locations and classifications directly from feature maps in one forward pass, eliminating the intermediate region proposal generation step that is integral to two-stage frameworks. Furthermore, in conjunction with the YOLO algorithm and multi-frame motion analysis method, the real-time detection of small UAVs is achieved, thereby enhancing the accuracy and efficiency of detection [16]. Recent advances further integrate YOLO with multi-frame motion analysis to achieve robust real-time detection of small UAVs under dynamic flight conditions [23]. In addition, deformable separation convolution blocks are used to perform feature separation. This method has been shown to significantly reduce the computational burden of convolution operations. By incorporating a channel weighting module, the extraction of feature information is further enhanced. These studies provide significant theoretical underpinnings and technical references for the development of UAV target detection. To conclude, notable advancements have been achieved in the field of UAV target detection over recent years. Nevertheless, there remain several challenges that require further attention:

First, in emergency situations such as fires, earthquakes, floods, UAVs can serve as important tools for rescue and monitoring. Real-time detection enables rapid location and tracking of UAVs allows for timely discovery and the taking of measures to ensure the smooth progress of emergency rescue work. Therefore, detection algorithms are required to have excellent real-time performance [22].

Second, UAVs in urban areas undergo dynamic scale changes during flight due to variations in distance and altitude [13]. Besides, the different flight attitudes of UAVs (hovering, rising, falling, turning) result in

different viewing angles and appearances in the images. For example, top-down and side views are very significantly different. This requires detection algorithms to detect targets with different scales and attitudes to avoid false detection or missing detection.

Third, urban environments pose significant challenges to UAV detection. Complex backgrounds with buildings, vehicles and pedestrians, as well as fluctuating lighting and weather, create noise and occlusion. Lighting changes due to weather and shadows distort UAV visibility, making feature extraction difficult and causing exposure problems. Adverse weather degrades sensor data, while moving urban objects increase false positive and false negative rates, significantly reducing detection accuracy and efficiency [1].

Accordingly, the core objectives of this research are summarized as follows:

1 To design convolutional neural networks to achieve real-time and accurate urban UAV detection.
2 To develop feature pyramid network to realize accurate feature extraction of UAVs with different scales.
3 To propose attention mechanism to realize accurate UAV detection in urban with complicated background.

The subsequent structure of this manuscript is organized in the following manner: Section II delivers a comprehensive literature review encompassing UAV detection methodologies and attention mechanism advancements. This review not only synthesizes state-of-the-art approaches – ranging from traditional vision-based techniques to emerging deep learning frameworks – but also critically dissects existing limitations, such as inadequate handling of dynamic scales in urban clutter or inefficient integration of attention modules with lightweight detectors, thereby laying a theoretical groundwork for the innovations proposed in subsequent sections.

To address above problems, this research adapts and integrates the Partial Self-Attention (PSA) mechanism—originally introduced in recent YOLOv8 variants for general object detection – into a dedicated architecture for robust urban UAV detection (YOLOv8-PSA). To the best of our knowledge, this is the first work that systematically evaluates the efficacy of PSA-enhanced YOLOv8 for UAV detection in complex urban environments, where the PSA module is strategically embedded to capture global contextual dependencies, suppress background noise, and enhance the discriminability of UAV features in cluttered scenes.

2. Literature Review

2.1. UAV Detection

Vision-based methodologies excel in achieving precise UAV detection with superior speed and reduced computational overhead, making them particularly advantageous in resource-constrained scenarios such as urban surveillance and low-latency monitoring [19, 28, 33, 40]. These vision-driven approaches can be broadly categorized into two paradigms: conventional detection techniques and deep learning-enabled frameworks, each characterized by distinct mechanisms for feature extraction and target recognition.

Traditional detection approaches hinge on manually crafted feature extraction operators to identify and detect UAVs, relying heavily on domain expertise to design discriminative features. For instance, Dalal et al. [7] pioneered an object detection framework that extracts Histogram of Oriented Gradients (HOG) features from training samples—capturing local edge distributions and shape nuances—and subsequently trains a Support Vector Machine (SVM) classifier to recognize object patterns in images. While this method demonstrated efficacy in controlled environments, its performance is constrained by sensitivity to variations in UAV poseand lighting conditions, limiting its robustness in cluttered urban settings. Liu et al. [24] proposed a fast aerial UAV detection framework employing improved inter-frame differencing and SVM classification, offering an efficient alternative for resource-constrained scenarios. Similarly, Viola et al. [32] introduced the AdaBoost algorithm for object categorization, whose core innovation lies in cascading weak classifiers—each derived from distinct feature subsets extracted from samples – to forge a high-performance strong classifier. Diao et al. [9] further demonstrated the efficacy of modified local sliding window methods for vision-based industrial inspection, highlighting the adaptability of handcrafted feature extraction in structured environments. This paradigm significantly enhanced real-time detection capabilities,

yet its reliance on hand-engineered features renders it less adaptable to the diverse appearances and dynamic motions of UAVs in complex backgrounds. Dong et al. [10] propose a UAV detection method that combines adaptive inter-frame differencing with morphological filtering for motion-based ROI extraction. It extracts multi-scale HOG features (using 64×128 windows and 8×8 cells), followed by FLD dimension reduction, and employs a polynomial-kernel SVM classifier. Unlu et al. [31] utilized Generic Fourier Descriptors (GFD) to characterize binary silhouettes of drones and birds, achieving scale, rotation, and translation invariant feature representation. Dong et al. [10] design online feature classification module in which Sobel operators extract gradient strength and orientation information, while HOG features quantify local edge distribution characteristics. Then, the dual-threshold strategy (gradient strength threshold and orientation variance threshold) is applied to filter candidate UAV targets with distinct edge patterns.

Deep learning-based approaches build neural networks to learn from UAV image datasets under various conditions, thereby enabling UAV feature extraction and detection. Du et al. [11] introduced the Faster module for the original ELAN architecture in YOLOv7, with the objective of acquiring rich and informative feature representations while maintaining a moderate parameter footprint. The model utilizes partial convolution (PConv) [21, 22], a specialized convolutional mechanism engineered to dynamically select and process a subset of input channels rather than the entire feature map. Additionally, Depth-wise separable convolution (DWConv) layer facilitates inter-channel feature integration across all channels. Recent architectures have designed multi-branch modules for localized spatial feature extraction, cross-channel semantic fusion, and identity skip connections to preserve raw input details [5]. During inference, all branches are reparametrized into a single lightweight 3×3 convolutional layer through linear parameter fusion, eliminating multi-path computational redundancy while retaining learned feature hierarchies. Ye et al. [35] design a dual-branch CNN module with Ghost convolution [13] to extract shallow local features and integrate with an Attention-Enhanced Transformer Block (AETB) for deep modeling. The left branch orchestrates the synthesis of semantic features via an intricate computational pipeline, incorporating a Ghost module—renowned for its lightweight design that mitigates computational redundancy by generating intrinsic features through cheap operations – alongside 3×3 depthwise convolution, and 1×1 convolution integrated with batch normalization (BN) to stabilize training dynamics and prevent internal covariate shift. In contrast, the middle branch deploys 3×3 depthwise convolution in tandem with a Ghost module, followed by SiLU activation to augment the model's capacity for capturing complex, non-linear patterns, thereby strengthening its expressive power in distinguishing subtle UAV features amid cluttered urban backgrounds. Concurrently, attention-enhanced tracking algorithms have been developed to mitigate occlusion challenges in low-altitude complex backgrounds [34].

Fundamentally, while conventional detection methodologies can attain precise UAV detection across heterogeneous scenarios, they remain inherently constrained by their heavy dependence on manually engineered feature extraction operators, thereby exhibiting compromised generalization across unseen environments. In contrast, deep learning-based paradigms demonstrate superior accuracy and robustness by leveraging large-scale UAV datasets for knowledge induction, with their performance predominantly predicated on the architectural ingenuity of neural network configurations.

2.2. Attention Mechanism

Empirically, attention mechanisms have been validated to significantly bolster the accuracy and robustness of UAV detection, particularly in navigating complex scenarios characterized by cluttered backgrounds and dynamic target scales. Rooted in the analogy of human visual perception – wherein the visual system inherently prioritizes salient regions while suppressing irrelevant stimuli – these mechanisms are transfiguring the landscape of computer vision, with their transformative impact being especially pronounced in UAV detection tasks.

Along the dimension of feature processing, attention mechanisms can be taxonomized into three primary paradigms: spatial attention, channel attention, and hybrid attention. Spatial attention mechanisms operate by adaptively weighting spatial locales, hon-

ing in on critical regions such as UAV contours amid urban clutter; channel attention, by contrast, recalibrates feature channels to amplify discriminative cues (e.g., texture or motion signatures) while attenuating redundant information; hybrid attention, as an integrated framework, synergizes the strengths of both, enabling simultaneous refinement of spatial and channel-wise features to tackle the multifaceted challenges of real-world UAV detection.

Channel attention mechanisms operate by modeling interdependencies between feature channels. Through global context aggregation and channel-wise recalibration, these methods achieve effective feature selection with high parameter efficiency. For instance, channel attention mechanisms have been widely adopted to augment representational capacity within the channel dimension [3]. The integration of such attention mechanism serves to augment the representational capacity of targets within the channel dimension. Concurrently, through accentuating target-related features, suppressing background interference, and fine-tuning network parameters, it demonstrates distinct superiority in image classification tasks. While channel attention mechanisms enhance feature discriminability by weighing critical channels, their standalone use suffers from inadequate spatial localization for small targets like UAVs due to unmodeled positional relationships and feature dilution in deep layers.

Spatial attention mechanisms model inter-pixel dependencies across spatial dimensions, enabling adaptive weighting of critical regions. Following the paradigm of Non-local Networks [36], spatial attention has been widely adopted to augment CNNs with global context modeling capabilities. For instance, SCINet [12] strategically embeds a spatial attention module within its inaugural ResBlock, which is specifically engineered to zero in on the structural contours and morphological details of UAVs in infrared imagery – effectively mitigating interference from cluttered backgrounds such as thermal noise, overlapping objects, or uneven radiation in complex urban or low-light scenarios. Despite such merits, spatial attention mechanisms are inherently confined to spatial-wise feature recalibration: they prioritize refining pixel-level spatial distributions while overlooking the saliency of channel-specific features, which encode critical semantic cues like texture variations or spectral signatures unique to UAVs. This inherent imbalance culminates in pronounced constraints on detection accuracy, particularly when tasked with identifying small-scale targets (e.g., micro-UAVs) in highly intricate environments – such as dense urban canyons or foliage-rich areas – where the dearth of channel-wise feature enhancement exacerbates the ambiguity between faint target signals and background clutter.

Hybrid attention frameworks—encompassing classical mechanisms such as CBAM [37] and recently advanced variants like DynamicViT [29]—integrate both channel and spatial attention mechanisms, thereby demonstrating superior efficacy in intricate vision tasks (e.g., dense multi-target detection, low-visibility image parsing, or cluttered scene segmentation). Unlike standalone channel or spatial attention mechanisms, which operate in relative isolation, hybrid attention frameworks excel by synergistically optimizing cross-modal interactions between channel-wise semantic cues (e.g., texture or spectral signatures) and spatial-wise structural details (e.g., contours or positional relationships). This joint enhancement markedly boosts the robustness of UAV detection, particularly in scenarios plagued by complex backgrounds (e.g., overlapping buildings, dynamic foliage) and diminutive targets (e.g., micro-UAVs), where the fused attention mechanism effectively mitigates ambiguity between faint target signals and environmental clutter.

Attention mechanisms augment the accuracy and robustness of UAV detection by dynamically adapting to intricate UAV operational contexts – encompassing not only forested regions but also urban canyons, low-light environments, and scenarios with heavy occlusion from dynamic objects like moving vehicles or foliage. In forested areas, for instance, these mechanisms can intelligently zero in on critical regions or discriminative features (such as the UAV's propeller motion, thermal signatures, or unique silhouette edges) while suppressing irrelevant noise from dense tree canopies, overlapping branches, or fluctuating shadows.

2.3. Research Gaps and Motivations

Despite notable strides in deep learning-driven UAV detection – encompassing breakthroughs in lightweight model architectures, multi-scale fea-

ture fusion, and real-time inference optimization – persistent research lacunae endure, with particular prominence in tackling challenges inherent to urban UAV detection scenarios. While hybrid attention mechanisms (e.g., CBAM) have shown promise in enhancing feature representation, their effective integration with lightweight detection frameworks remains underexplored. Although recent studies have attempted to embed hybrid attention modules into lightweight YOLO variants for UAV and remote sensing tasks [20, 27, 38], these efforts primarily focus on generic object detection scenarios without tailoring the synergy between attention mechanisms and efficient feature fusion modules to address urban-specific challenges such as dynamic scale variations and severe background clutter. The critical importance of domain-specific adaptation has been underscored by the emergence of industrial foundation models for intelligent manufacturing [42], yet analogous principled adaptation remains underexplored in urban UAV detection. As a result, existing approaches rarely optimize the joint deployment of attention mechanisms and lightweight network structures for urban UAV detection.

Future research ought to concentrate on developing more efficient attention mechanisms, aiming to achieve precise detection of multi-scale UAVs in complex environments. Specifically, leveraging the strengths of channel attention (inter-channel correlations), spatial attention (pixel-wise dependencies), and hybrid attention (synergistically combining multiple operational domains) to achieve composite feature enhancement that aligns with the demands of urban UAV detection.

3. Methodology

3.1. Yolov8 with Enhanced Partial Self-attention

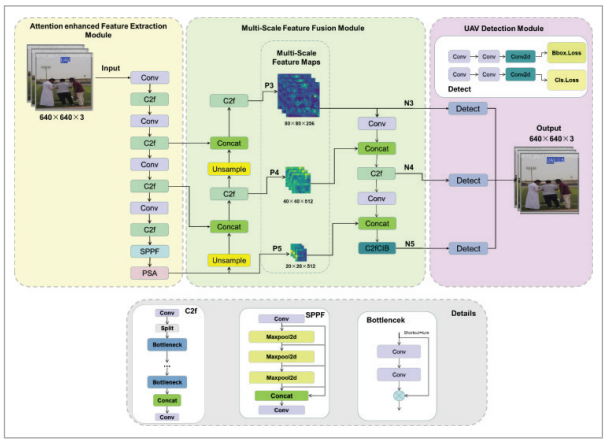
In this study, the YOLOv8-PSA is designed for UAV detection in urban environments, its detailed network structure is shown Figure 1.

YOLOv8-PSA network utilizes 640×640×3 images as input and outputs images with bounding boxes. The attention-augmented feature extraction block progressively derives image feature maps that traverse a hierarchical spectrum, ranging from low-level representations – such as edge contours, texture gradients, and pixel-level spatial correlations – to high-level semantic features, including UAV structural patterns, contextual attribute associations, and target-background discrimination cues. This incremental derivation process, modulated by the embedded attention mechanism, ensures that critical features are dynamically emphasized at each stage, avoiding information loss during the transition from primitive visual cues to abstract semantic representations. The shallower layers specialize in extracting features such as edges and contours, whereas the deeper layers focus on capturing high-level semantic features associated with object categories. The multi-scale feature fusion module leverages the Feature Pyramid Network (FPN) to integrate multi-scale feature maps, consolidating semantic and detailed information from diverse hierarchical levels. This integration enhances the network's adaptability and robustness when confronting targets of varying scales. The UAV detection module forecasts the bounding boxes and categories of targets based on these fused feature maps. Specifically, the bounding box delineates the spatial position and dimensional attributes of the target, while the category probability quantifies the confidence in the classification outcome. Together, these components enable precise object detection.

The detailed architecture of the PSA module and its integration into the feature extraction backbone are elaborated in Section 3.2, while the multi-scale feature fusion strategy and the lightweight C2fCIB module are presented in Section 3.3.

Figure 1

The structure of the YOLOv8-PSA network.



3.2. Attention-enhanced Feature Extraction Module

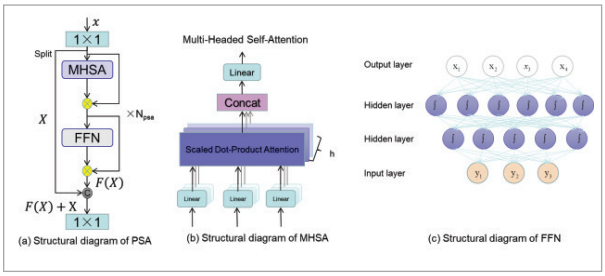
The attention-enhanced feature extraction module is tasked with deriving multi-scale feature maps across three distinct spatial resolutions with 80×80, 40×40, 20×20. To enhance the feature representation of UAV, the Partial Self-Attention (PSA) mechanism is incorporated into the YOLOv8-PSA architecture. The PSA embodies polarization strategy to enable the network to focus on distinct feature subspaces, thereby fostering a more comprehensive feature representation. Unlike full self-attention, which processes all spatial positions and incurs quadratic computational overhead, the proposed Partial Self-Attention (PSA) mechanism adopts a feature segmentation strategy. It splits the feature map into two branches along the channel dimension: one branch is processed by Multi-Headed Self-Attention (MHSA) and Feed-Forward Network (FFN), while the other bypasses these operations via a residual connection. This partial design preserves raw feature information, reduces redundant computations, and improves efficiency without sacrificing global contextual modeling. The PSA consists of three primary components including Feature Segmentation (FS), Multi-Headed Self-Attention (MHSA), and Feed-Forward Network (FFN), the network structure is shown Figure 2.

The FS processes the input feature maps x. Initially, this feature map undergoes processing via a 1×1 convolutional layer to adjust its channel dimensionality, perform feature transformation, generating transformed feature map X. Then, this transformed feature map is partitioned by the channel dimension into two components. One part directly propagates feature map X to the output referring to residual block. The other part is processed by the subsequent MHSA and FFN to obtain feature map $F(X)$. In the end, the two portions are concatenated to form ultimate output feature map $F(X)+X$, as shown in Figure 2(a).

The MHSA is applied to capture long-range dependencies within feature maps. It leverages self-attention mechanism to realize projections of Query (Q), Key (K), and Value (V). The outputs from all heads are subsequently concatenated along the channel dimension and fused via a linear layer to generate the final output. The attention weights are computed as follows:

Figure 2

The detailed diagram of PSA: (a) Structural diagram of PSA, (b) Structural diagram of MHSA, and (c) Structural diagram of FFN.



$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where d_k represents dimensionality of key vectors. The MHSA module consists of multiple parallel attention heads, each processing different subspaces to learn complex semantic features and capturing diverse dependency patterns, Additionally, residual connection is used to facilitate the preservation of original feature information while integrating the enhanced representations. The structure of MHSA structure is shown as Figure 2.

The output of the MHSA module is then fed into a FFN, see Fig.2(b). It comprises two fully connected layers containing non-linear activation function, enabling further feature transformation and enhancement, as shown in Fig.2(c).

$$FFN(x) = FC_2(\sigma(FC_1(x))), \quad (2)$$

where FC_1 and FC_2 denote the first and second fully connected layers, and σ represents the activation function (e.g., ReLU). The output of the FFN is combined with the second component from Feature Segmentation through element-wise summation, implementing a residual connection:

$$Y = FFN\left(MHSA\left(X_1'\right)\right) + X_2'. \quad (3)$$

The entire PSA process, including Feature Segmentation, MHSA, FFN, and Feature Fusion, can be repeated N times (denoted as "×N PSA" in Figure 3) to iteratively refine the feature representation. Ulti-

mately, a 1×1 convolutional layer is deployed to fuse the aggregated features, calibrate channel dimensions, and integrate the attention-refined feature maps derived from preceding stages. This multi-fold operation consolidates discriminative cues across spatial scales and semantic levels, ultimately yielding the PSA module's final output

3.3. Multi-scale Feature Fusion and UAV Detection Modules

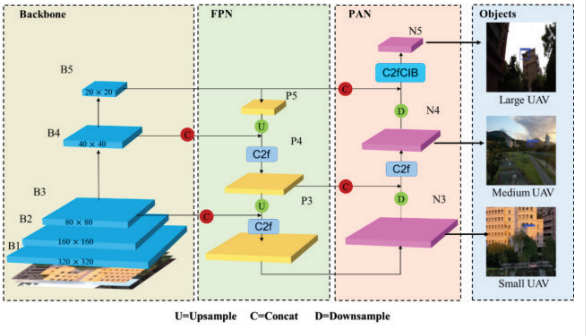
The multi-scale feature fusion module is meticulously designed to fully harness the potential of feature maps encompassing multi-level information, aiming to address the challenge of detecting UAVs across varying scales. Specifically, large-scale feature maps, which are rich in fine-grained details such as subtle edges, texture variations, and local contour information, are strategically employed for the precise detection of small UAVs. These detailed features play a critical role in distinguishing small targets from complex backgrounds, ensuring that even diminutive UAVs – often easily obscured or overlooked – can be accurately identified.

In contrast, small-scale feature maps, characterized by their abundance of high-level semantic information, including category-related attributes and contextual associations, are dedicated to the detection of large UAVs. Such semantic-rich features enable the network to grasp the overall characteristics and contextual relevance of larger targets, enhancing the accuracy of classification and localization for these relatively prominent UAVs. By rationally allocating feature maps of different scales to corresponding target sizes, the module effectively bridges the gap in detection performance caused by scale discrepancies, thereby significantly boosting the network's ability to detect UAVs of varying dimensions.

To achieve this multi-scale feature fusion, the network integrates both FPN and PAN are adopted for multi-scale feature fusion. Five scale feature maps B1 to B5 are extracted for UAV detection. Both FPN and PAN work together to integrate B3, B4 and B5 feature maps and generate N5, N4 and N3 feature maps for more precise UAV detection. The structure of feature pyramid network of YOLOV8 is shown in Figure 3.

The top-down pathway starts with B5 (P5), which is upsampled and concatenated with B4 to form

Figure 3
The multi-scale feature pyramid network of YOLOv8 across different object scales (Large, Medium, and Small UAVs).



P4, merging high-level semantics with mid-level details. P4 is further concatenated with B3 to generate P3, optimized for small-object detection by retaining fine-grained spatial information. The bottom-up pathway in YOLOv8 uses downsampling from P3 (N3), concatenating with P4/P5 to generate N4/N5. N4 fuses mid-level semantics for large-object detection, while N5 emphasizes deep semantics for medium objects. Bidirectional FPN-PANet fusion ensures detection layers (P3-N5) integrate spatial and semantic features, enhancing scale adaptability.

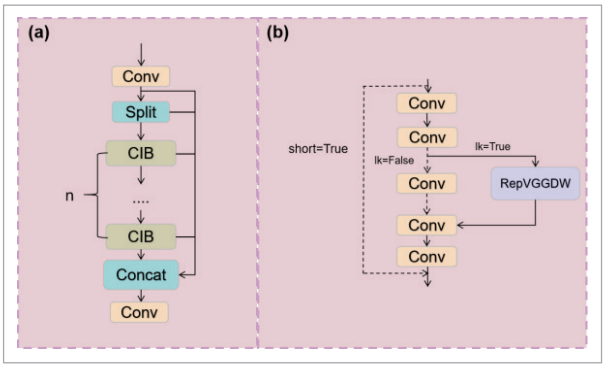
The C2fCIB (Cross-Scale Feature Fusion and Channel Information Bottleneck) module integrates pointwise convolution and depthwise convolution to realize a lightweight design. This integration enables efficient processing of high-dimensional data while preserving semantic information through inter-channel information interaction and local spatial information extraction. By doing so, it optimizes the quality of feature maps and empowers the network to more accurately capture key details of targets. The processing workflow of e C2fCIB module is illustrated in Fig. 4. To quantitatively evaluate the efficiency of the proposed C2fCIB module, we compared its computational cost with the standard C2f module under identical input conditions (384 channels, 20×20 feature map, n=1). The standard C2f module in YOLOv8n requires 0.493M parameters and 0.198G FLOPs, whereas the proposed C2fCIB reduces these to 0.283M parameters and 0.114G FLOPs, achieving a 42.7% reduction in parameters and a 42.2% reduction in FLOPs. This significant reduction is primarily attributed to the depthwise sep-

arable convolutions and channel splitting strategies within the CIB sub-modules, enabling a lightweight design that preserves semantic integrity while maintaining real-time detection capability.

The module initially employs a 1×1 convolutional layer to refine semantic features while preserving the spatial information. Subsequently, feature map is partitioned along channel dimension into two equal segments, each with dimensions $H\times W\times(C/2)$. One segment is then processed sequentially by two CIB (Channel-wise Information Bottleneck) sub-modules. Each CIB module outputs a feature map of size $H\times W\times(C/2)$. The output of the second CIB module is concatenated with the unprocessed feature segment, resulting in a feature map of size $H\times W\times[C/2\times(n+2)]$. Finally, a 1×1 convolutional layer is applied to this concatenated feature map to project it back to the original channel dimension, producing the final output feature map of size $H\times W\times C$. Within the CIB sub-modules, diverse convolutional operations are employed, including 3×3 convolutions, 1×1 convolutions, and RepVGGDW, enabling flexible feature extraction and adaptation based on specific parameter configurations. Each convolutional operation is succeeded by Batch Normalization (BatchNorm2d) and the SiLU activation function. This sequential processing not only facilitates stable training by mitigating internal covariate shift and accelerating convergence but also strengthens the module's non-linear representation capabilities—enabling it to model complex feature relationships and capture intricate patterns within the input data more effectively.

Figure 4

C2FCIB detailed diagram. (a) Structural diagram of C2FCIB, and (b)Structural diagram of CIB.



3.4. Loss Function

While YOLOv8n originally adopts a combination of Distribution Focal Loss (DFL) and CIoU loss, this study employs CIoU as the sole bounding box regression loss to simplify optimization and enhance convergence stability. Compared to EIoU and SIoU, CIoU balances overlap area, center distance, and aspect ratio without introducing additional hyperparameters, which is critical for maintaining generalization across diverse UAV scales. Furthermore, experiments showed that adding DFL did not yield significant gains on our dataset, motivating our streamlined design.

CIoU holistically incorporates multiple critical factors: the overlapping ratio of bounding boxes, discrepancies in central coordinates, and aspect ratio harmony. This comprehensive consideration not only elevates the precision of target detection by refining the localization accuracy of predicted boxes but also expedites the convergence rate of model training through more effective gradient propagation. The mathematical formulations for CIoU Loss and its associated parameters are presented below:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b_p,b_t)}{c^2} + \alpha\upsilon \tag{4}$$

$$\rho^2(b,b_t) = (x_p - x_t)^2 + (y_p - y_t)^2 \tag{5}$$

$$\alpha = \frac{\upsilon}{1-IoU+\upsilon} \tag{6}$$

$$\upsilon = \frac{4}{\pi^2}\left(arctan\left(\frac{w_t}{h_t}\right) - arctan\left(\frac{w_p}{h_p}\right)\right)^2 \tag{7}$$

$$c = \sqrt{(x_{max} - x_{min})^2 + (y_{max} - y_{min})^2}\,. \tag{8}$$

where w_p and h_p are the width and height of prediction boxes, and w_t and h_t are width and height of real boxes.

The coordinates of bounding box's center point are denoted as (x_p,y_p), while those of ground-truth bounding box's center point are (x_t,y_t). x_{max}, x_{min} represent the maximum and minimum coordinate values on the x-axis. y_{max}, y_{min} correspond to the maximum and minimum coordinate values on the y-axis. These

parameters collectively define the spatial boundaries and positional relationships between the predicted and real bounding boxes, serving as foundational metrics for calculating geometric constraints in loss functions such as CIoU.

4. Experiment

4.1. Experiment Setting

The detailed experimental environment, encompassing both hardware and software specifications, are listed in Table 1. Regarding the training setup, the total number of training iterations was configured to 450, with each iteration processing a pre-defined mini-batch of annotated UAV images to update model parameters. Concurrently, an early stopping mechanism was incorporated to prevent overfitting, featuring a patience parameter set to 100 epochs. Validation is performed after every training epoch. The patience of 100 epochs is intentionally set to allow sufficient room for performance plateaus, given the total of 450 epochs. This configuration ensures that early stopping does not interrupt training prematurely unless no improvement is observed over an extended period (approximately 22% of total iterations), which is considered indicative of convergence saturation. During the training process, the batch size was fixed at 16, and the input images were resized to an initial resolution of 640×640. The initial learning rate was set to 0.01. These hyperparameters follow the default training configuration recommended by the Ultralytics YOLOv8 framework [16], which has been widely adopted in recent object detection studies [41]. The epoch number was extended to 450 to ensure full convergence on the UAV dataset, while the patience of 100 was set to prevent overfitting given the dataset size. To monitor the generalization capability of model in the real time, validation was performed iteratively during training. For each round of validation, the segmentation strategy adhered to the default split of the validation set.

In this experiment, a UAV detection dataset is established for model training and testing. This dataset is collected from the public dataset 'Urban-UAV-Detection' (or alternative name) released on the Roboflow platform (https://github.com/mugheessarwar-awan/UETT4K-Anti-UAV?tab=readme-ov-file),

Table 1
Model training environment.

Environment Name	to configure
CPU	16 vCPU Intel(R) Xeon(R) Platinum 8481C
GPU	RTX 4090D(24GB) * 1
Python	Python 3.12(ubuntu22.04)
PyTorch	PyTorch 2.5.1
CUDA	CUDA 12.4

which is organized to be used primarily for the task of UAV detection in complex environments. The images in the dataset include urban UAV under different angles, altitudes and lighting conditions, which effectively ensures the applicability of UAV detection. The dataset encompasses a cumulative total of 5,275 annotated images, which are stratified into training, validation, and testing subsets adopting a 7:2:1 split ratio.

To conduct an objective and comprehensive assessment of the model's performance, three core metrics – Precision (P), Recall (R), and Mean Average Precision (mAP) – are employed. These metrics effectively quantify the model's accuracy and comprehensiveness in handling classification tasks, thereby furnishing a critical foundation for evaluating its overall performance. A detailed elaboration of each metric is as follows:

Precision connotes the fraction of model-predicted positive-class instances that are truly positive. This metric encapsulates the model's capability to minimize false alarms within its positive predictions, serving as a critical indicator of discriminative rigor in distinguishing genuine targets from background or irrelevant classes. It is computed using the formula:

$$P = \frac{TP}{TP+FP}, \tag{9}$$

where, TP denotes the quantity of positive-class samples that are accurately recognized and appropriately categorized by the model, whereas FP designates the number of negative-class samples that are erroneously predicated as positive. Mathematically, Precision is calculated as:

Recall, also known as sensitivity, quantifies the model's capacity to identify all actual positive samples within the dataset. It reflects the completeness of the model in capturing true positive instances, measuring how many of the real positive cases are successfully retrieved by the model. Mathematically, it is calculated as:

$$R = \frac{TP}{TP+FN}, \tag{10}$$

FN specifically refers to the count of positive samples that are erroneously categorized as negative, representing instances where the model fails to recognize actual targets. A higher recall value directly signifies that the model can detect positive samples more comprehensively, minimizing the number of missed true positives—an attribute of particular significance in UAV detection, where overlooking targets may lead to critical oversights.

mAP serves as a holistic metric to evaluate the model's capacity to sustain high precision across varying recall thresholds. Unlike single-point precision or recall, mAP integrates performance across the entire spectrum of recall values, offering a more robust assessment of the model's overall effectiveness. A higher mAP value indicates that the model can maintain elevated precision even as recall increases, reflecting superior balance between accuracy and completeness in target detection. Its calculation formula is as follows:

$$mAP = \frac{1}{k}\sum_{r=0}^{k}(R_n - R_{n-1})P(R_n). \tag{11}$$

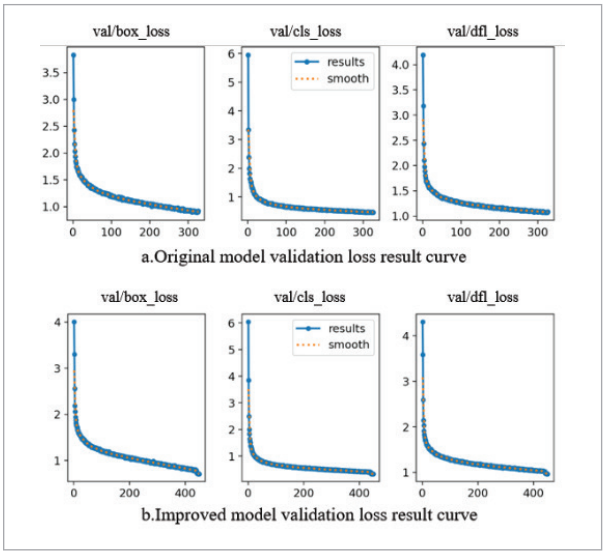
4.2. Performance Evaluation

This section shows the experiment results including the loss and mAP during the model training, model performance evaluation and detection results, and comparison between the YOLOv8-PSA and the original YOLOv8n.

4.2.1. Model Training

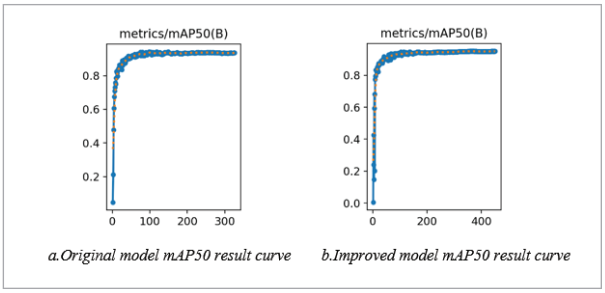
The loss values of the baseline YOLOv8n (including box loss, classification loss, and distribution focal loss) and the proposed YOLOv8-PSA (including only box loss and classification loss, as DFL was omitted) were calculated to monitor the training process; detailed results are shown in Fig. 5. For YOLOv8-PSA, the box_loss decreases from about 4.0 to about 1.5 after 300 iterations, and the cls_loss decreases from about 12.5 to about 2.0. Comparitively, for YOLOv8n, the box_loss decreases from about 4.5 to about 2.5 after 400 iterations, the cls_loss decreases from about 13.5 to about 3.5, the dfl_loss decreases from about 4.5 to about 2.5. It demonstrates that YOLOv8-PSA outperforms the original YOLOv8n model in terms of bounding-box regression generalization, classification performance.

Figure 5
Loss curve during the model training.



The mAP results of YOLOv8-PSA and YOLOv8n are also evaluated, as shown in Fig. 6. In terms of mAP50, the initial value of the original model is about 0.6, and it rises to about 0.8 after 300 iterations; the initial value of YOLOv8-PSA is about 0.65, and it reaches about 0.85 after 200 iterations, and is close to 0.9 after 300 iterations, which rises faster and converges to a higher value in the end. This indicates that YOLOv8-PSA is better than the original model in terms of detection accuracy. In summary, the YOLOv8-PSA is better in terms of both the rise speed and the final convergence value, which indicates that the improvement measures effectively enhance the detection accuracy and generalization, making it adaptable to target detection under different IoU thresholds, and more accurately identifying and localizing the target objects.

Figure 6
mAP50 curve during the model training.



a.Original model mAP50 result curve *b.Improved model mAP50 result curve*

4.2.2. Model Evaluation

The detection performance of proposed YOLOv8-PSA is also quantitative evaluated across different UAV sizes: small, medium, and large, as well as overall. The metrics used for this analysis are precision, recall, mAP50 (mAP at an IoU threshold of 0.5). The results of multi-scale UAV detection metrics are shown in Table 2.

Table 2
Detection performance of YOLOv8-PSA for multi-scale UAV.

	precision	Recall	mAP50
small	0.974	0.891	0.935
Medium	0.968	0.946	0.969
Large	0.966	0.966	0.985
overall	0.978	0.925	0.965

The relatively lower recall for small UAVs (0.891) compared to large ones (0.966) can be attributed to their limited pixel footprint in the input images (640×640), which reduces the spatial details available for feature extraction. Small UAVs are also more susceptible to occlusion and background clutter in urban scenes, leading to occasional missed detections. In contrast, large UAVs occupy more pixels, providing richer semantic and structural cues, thereby improving recall. A more detailed analysis of each scale is provided below.

For small UAVs, it achieves a high precision of 0.974, a recall of 0.891 and mAP50 score was 0.935. In the detection of medium UAVs, the model demonstrated a precision of 0.968, a strong recall of 0.946 and mAP50 score was 0.969. This suggests a balanced and robust performance for medium-sized targets. For large UAVs, the model yielded a precision of 0.966 and a recall of 0.966, indicating consistent identification and capture of these targets. The mAP50 score was the highest among the categories at 0.985. This highlights the effectiveness of model in detecting larger, more prominent UAVs.

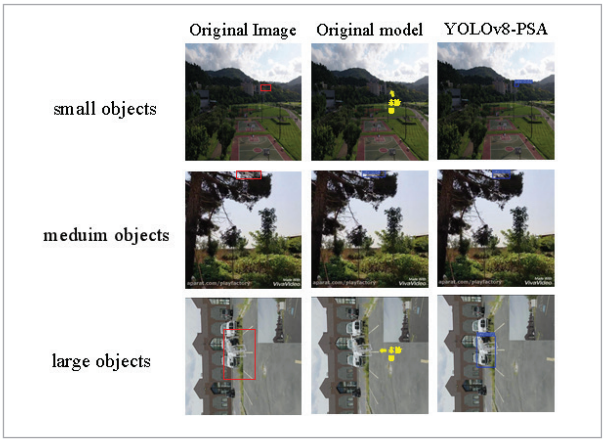
Collectively, the model attained a precision of 0.978, a recall of 0.925, and an overall mean Average Precision at 50% IoU (mAP50) of 0.965. These consolidated metrics underscore the model's superior accuracy and robust efficacy in detecting UAVs spanning diverse scales within the dataset. The consistently high precision across all categories underscores the model's reliability in its positive predictions.

4.2.3. Detection Results

This section evaluates the performance of the YOLOv8-PSA model in detecting multi-scale UAVs against complex urban scenarios through visualized detection results. Sample images were selected from the test set to visually compare the efficacy of YOLOv8-PSA model against the original model, as illustrated in Figure 7.

As illustrated in Figure 7, key differences between YOLOv8-PSA and the original YOLOv8n are annotated with colored boxes and arrows. Red boxes indicate the ground-truth UAV positions. For the detection of small UAV in complex urban background, the original model fails to detect the target, whereas YOLOv8-PSA correctly identifies it. For medium UAV with partial occlusion, both models detect the target, but

Figure 7
Detection Results of proposed YOLOv8-PSA for multi-scale UAVs.



YOLOv8-PSA achieves a higher confidence score compared to the original model. For large UAV with 90 rotation, the original model misses the target, while YOLOv8-PSA successfully localizes it. These comparisons demonstrate that the proposed model is more robust to scale variation, occlusion, and orientation changes in urban UAV detection.

By capturing long-distance dependencies and solving the deficiencies of traditional convolution in capturing detailed features, the PSA attention mechanism enables the model to filter out irrelevant background information and locate the target object more accurately. This leads to a reduction in both false positives and false negatives, even in visually cluttered environments. Furthermore, the integration of the C2fCIB module contributes to a lightweight design and efficient processing of high-dimensional data, preserving semantic information and optimizing feature map quality to better capture key target details. The CIoU loss function also plays a role by improving target localization accuracy.

In summary, the proposed enhancements, particularly the global attention mechanism, allow our model to achieve more accurate and reliable UAV detection in complex urban environments. It can successfully identify targets that the original model might miss or misclassify, demonstrating superior performance in handling variations in UAV scale, pose, and challenging background conditions. This lays a solid foundation for practical application in real-world urban surveillance and airspace management.

4.3. Comparative Study

In order to verify the superiority of the improved YOLOv8-PSA in the UAV detection performance, YOLOv8n and its various improved algorithms are selected to be compared and analyzed on the dataset used in this experiment. The experiment results are shown in Table 3.

After comparative analysis, the improved model achieves competitive performance gains across the three key metrics of P, R, and mAP50, with consistent margins over the baseline YOLOv8n and other variants, which are 0.04, 2.7, and 2.1 percentage points higher than the optimal values of the suboptimal algorithm (YOLOv8n+GIoU+ShuffleAttention), respectively. Compared with the base YOLOv8n model and the other three models (YOLOv8n+CIoU+GAM, YOLOv8n+DIoU+GAM, YOLOv8n+DIoU+SE), the improved model improves by 2.9, 2.4, 1, and 0.9 percentage points on the P metric, 2.5, 2.3, 2.6, and 3 percentage points on the R metric, and 2.6, 2.7, 2.6, and 1.8 percentage points on the mAP50 index, respectively.

It is worth noting that the current comparative study focuses on YOLOv8-based variants to isolate the effect of our proposed modules. However, recent state-of-the-art detectors such as YOLOv9 and RT-DETR have demonstrated strong performance in object detection tasks. Future work will include these models as benchmarks to further validate the superiority and generalizability of YOLOv8-PSA.

4.4. Ablation Study

A total of 8 different module combinations is designed in the experiment, and each module and its combination is gradually introduced based on the base model (Model 1), and their effects on precision (P), recall (R) and mAP50 are analyzed, as shown in Table 4.

Table 3
Comparative experiments with other detection approaches.

Method	P	R	mAP50
YOLOv8n	0.949	0.9	0.939
YOLOv8n+CIoU+GAM	0.954	0.902	0.938
YOLOv8n+DIoU+GAM	0.968	0.899	0.939
YOLOv8n+DIoU+SE	0.969	0.895	0.947
YOLOv8n+GIoU+ShuffleAttention	0.974	0.898	0.944
Modified model	0.978	0.925	0.965

Table 4
Ablation experiment.

model	CIoU	PSA	C2fCIB	P	R	mAP50	Latency(ms)
1	—	—	—	0.949	0.9	0.939	2
2	√	—	—	0.973	0.908	0.944	0.8
3	—	—	√	0.943	0.904	0.949	3.1
4	√	—	√	0.973	0.908	0.944	1.1
5	—	√	—	0.971	0.897	0.941	2.6
6	√	√	—	0.969	0.913	0.948	1.7
7	—	√	√	0.966	0.912	0.95	1.2
8	√	√	√	0.978	0.925	0.965	1.4

Model 1 is the baseline YOLOv8n with default settings, achieving 0.949 precision, 0.900 recall, 0.939 mAP50, and an inference latency of 2.0ms on an NVIDIA RTX 4090 GPU with batch size 1 and 640×640 input resolution.

Model 2 introduces only the CIoU loss. Compared to the baseline, precision improves to 0.973, recall to 0.908, and mAP50 to 0.944, while latency is significantly reduced to 0.8ms. This demonstrates that CIoU not only enhances localization accuracy but also accelerates convergence and reduces inference time.

Model 3 adds only the C2fCIB module. Precision slightly drops to 0.943, whereas recall increases to 0.904 and mAP50 to 0.949. The latency increases to 3.1ms due to the additional depthwise separable convolutions and channel splitting operations. The improved recall suggests that C2fCIB enhances sensitivity to low-level and mid-level features, but the increased feature diversity may introduce false positives from background elements, causing a marginal precision drop.

Model 4 combines CIoU and C2fCIB. It achieves the same accuracy as Model 2, with a latency of 1.1ms. The feature extraction advantage of C2fCIB may not be fully exploited when paired with CIoU, possibly due to mutual constraints.

Model 5 introduces only the PSA mechanism. Precision decreases to 0.971, recall to 0.897, and mAP50 to 0.941, with a latency of 2.6ms. PSA focuses on global contextual dependencies but may over emphasize certain features, leading to slight bias in localization and identification.

Model 6 combines CIoU and PSA. Precision is 0.969, recall increases to 0.913, mAP50 to 0.948, and latency is 1.7ms. This combination balances localization precision and feature selection, improving recall without significantly harming precision.

Model 7 combines PSA and C2fCIB. Precision is 0.966, recall 0.912, mAP50 0.950, and latency 1.2ms. PSA and C2fCIB synergistically enhance feature extraction and selection, effectively mining target features and boosting recall, though the altered feature representation causes a slight precision decrease.

Model 8 integrates all three modules. It achieves the best overall performance: precision 0.978, recall 0.925, mAP50 0.965, with a latency of 1.4ms. This latency is lower than the baseline (2.0ms) and remains well within real time requirements. The synergy of CIoU, PSA, and C2fCIB enables significant gains in both precision and recall, confirming the effectiveness of the proposed enhancements for practical urban UAV detection.

While statistical significance testing was not performed due to computational constraints, the consistent margin of improvement across precision, recall, and mAP50 suggests robust performance gains attributable to the proposed modules.

4.5. Limitations and Failure Cases

Although YOLOv8-PSA demonstrates superior performance in most urban UAV detection scenarios, several limitations remain. First, under extreme low-light conditions (e.g., nighttime or heavy shad-

ow), the model suffers from degraded recall due to insufficient texture details and low signal-to-noise ratio, occasionally missing small UAVs that are barely visible. Second, in cases of severe occlusion—where UAVs are partially or fully hidden behind buildings, vehicles, or dense foliage—the detection rate drops significantly, as the PSA mechanism may fail to capture adequate global cues when the target is largely invisible. Third, adverse weather conditions such as rain, fog, or haze degrade image quality and introduce scattering noise, leading to increased false negatives. These failure cases are exemplified in the test set by the relatively lower recall for small UAVs (0.891) compared to large ones (0.966), indicating that extremely small or heavily obscured targets remain challenging. Future work will address these limitations through multi-modal sensor fusion and low-light image enhancement.

5. Conclusion

In this study, a partial self-attention mechanism-enhanced YOLOv8 is proposed to address the challenges of UAV detection in complex urban environments. The key contributions and findings are summarized as follows:

1. **PSA mechanism.** We integrate the Partial Self-Attention (PSA) mechanism into YOLOv8 to enhance the model's ability to capture global contextual dependencies and suppress background noise. By partitioning feature maps, applying multi-headed self-attention, and fusing features via residual connections, PSA effectively focuses on critical UAV features (e.g., shape, contours) in cluttered urban scenes, reducing false positives and missed detections.
2. **C2fCIB multi-scale fusion.** We design a multi-scale feature fusion module combining FPN and PAN with the C2fCIB block, which strengthens adaptability to dynamic UAV scales. C2fCIB, through lightweight design and cross-channel information exchange, optimizes feature extraction efficiency while preserving fine-grained details, enabling accurate detection of small, medium, and large UAVs.
3. **CIoU loss optimization.** We adopt the CIoU loss function to improve bounding box localization by jointly considering the overlapping ratio, center distance, and aspect ratio harmony. This accelerates model convergence and enhances positioning precision during training.

Experimental results on a urban UAV dataset demonstrate that YOLOv8-PSA demonstrates measurable and consistent improvements over the original YOLOv8n and other variants. It achieves an overall precision of 0.978, recall of 0.925, and mAP50 of 0.965, with exceptional performance on small UAVs (mAP50 = 0.935) and large UAVs (mAP50 = 0.985). Ablation studies confirm that the synergistic effect of PSA, C2fCIB, and CIoU is critical: their combination yields a 2.6% improvement in mAP50 compared to the baseline YOLOv8n.

In summary, the YOLOv8-PSA model achieves a favorable detection accuracy and real-time performance, rendering it well-suited for practical applications such as urban UAV surveillance, airspace management, and emergency response scenarios. Its ability to efficiently capture key target details while maintaining rapid inference speeds ensures reliable operation in dynamic and complex environments, addressing the core demands of real-world UAV monitoring tasks.

For future work, two primary directions will be prioritized: First, to reduce the computational overhead of the PSA module while maintaining detection accuracy, we will investigate efficient attention variants that replace dense self-attention with sparse or linear-complexity approximations. Specifically, sparse attention patterns with fixed local-global factorization (e.g., Sparse Transformers [4]) and low-rank projection methods (e.g., Linformer [35]) will be explored to reduce the quadratic complexity of self-attention. Additionally, IO-aware exact attention kernels (e.g., FlashAttention [8]) will be considered to accelerate inference on resource-constrained edge devices deployed in urban surveillance systems. Second, the dataset will be expanded to include more samples captured under different weathers. By incorporating such challenging scenarios, the model's robustness and generalization capability will be strengthened, enabling it to maintain stable performance across a wider range of practical environments and thus extending its applicability to more complex real-world scenarios. Furthermore, the complete source code

and pre-trained weights of the proposed YOLOv8-PSA model will be released and made open-source on GitHub to facilitate reproducibility and future collaborative research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

Data Sharing Agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Bayomi, N., Fernandez, J. E. Eyes in the Sky: Drones Applications in the Built Environment Under Climate Change Challenges. Drones, 2023, 7(10), 637. https://doi.org/10.3390/drones7100637
2. Chen, C., Zheng, Z., Xu, T., Guo, S., Feng, S., Yao, W., Lan, Y. YOLO-Based UAV Technology: A Review of the Research and Its Applications. Drones, 2023, 7(3), 190. https://doi.org/10.3390/drones7030190
3. Cheng, D., Meng, G., Cheng, G., Pan, C. SeNet: Structured Edge Network for Sea-Land Segmentation. IEEE Geoscience and Remote Sensing Letters, 2016, 14(2), 247-251. https://doi.org/10.1109/LGRS.2016.2637439
4. Child, R., Gray, S., Radford, A., Sutskever, I. Generating Long Sequences with Sparse Transformers. arXiv Preprint arXiv:1904.10509, 2019. https://arxiv.org/abs/1904.10509
5. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 1251-1258. https://openaccess.thecvf.com/content_cvpr_2017/html/Chollet_Xception_Deep_Learning_CVPR_2017_paper.html https://doi.org/10.1109/CVPR.2017.195
6. Cui, J., Zhang, X., Zhang, J., Han, Y., Ai, H., Dong, C., Liu, H. Weed Identification in Soybean Seedling Stage Based on UAV Images and Faster R-CNN. Computers and Electronics in Agriculture, 2024, 227, 109533. https://doi.org/10.1016/j.compag.2024.109533
7. Dalal, N., Triggs, B. Histograms of Oriented Gradients for Human Detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, 1, 886-893. https://doi.org/10.1109/CVPR.2005.177
8. Dao, T., Fu, D. Y., Ermon, S., Rudra, A., Ré, C. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. Advances in Neural Information Processing Systems, 2022, 35, 16344-16359. https://doi.org/10.52202/068431-1189
9. Diao, Y., Cheng, W., Du, R., Wang, Y., Zhang, J. Vision-Based Detection of Container Lock Holes Using a Modified Local Sliding Window Method. EURASIP Journal on Image and Video Processing, 2019, 2019(1), 69. https://doi.org/10.1186/s13640-019-0472-1
10. Dong, Q., Zou, Q. Visual UAV Detection Method with Online Feature Classification. 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2017. https://doi.org/10.1109/ITNEC.2017.8284767
11. Du, Y., Wu, T., Dai, Z., Xie, H., Hu, C., Wei, S. F-YOLOv7: Fast and Robust Real-Time UAV Detection. Computing, 2025, 107(1), 50. https://doi.org/10.1007/s00607-024-01406-7
12. Fang, H., Ding, L., Wang, X., Chang, Y., Yan, L., Liu, L., Fang, J. SCINet: Spatial and Contrast Interactive Super-Resolution Assisted Infrared UAV Target Detection. IEEE Transactions on Geoscience and Remote Sensing, 2024. https://doi.org/10.1109/TGRS.2024.3471786
13. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C. GhostNet: More Features from Cheap Operations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 1580-1589. https://doi.org/10.1109/CVPR42600.2020.00165
14. He, Y., Liu, Z., Guo, Y., Zhu, Q., Fang, Y., Yin, Y., Liu, Z. UAV-Based Sensing and Imaging Technologies for Power System Detection, Monitoring and Inspection: A Review. Nondestructive Testing and Evaluation, 2024, 1-68. https://doi.org/10.1080/10589759.2024.2421938
15. Hildmann, H., Kovacs, E. Using Unmanned Aerial Vehicles (UAVs) as Mobile Sensing Platforms (MSPs)

for Disaster Response, Civil Security and Public Safety. Drones, 2019, 3(3), 59. https://doi.org/10.3390/drones3030059

16. Jocher, G., Chaurasia, A., Qiu, J. Ultralytics YOLOv8 (Version 8.0.0) [Computer Software], 2023. https://github.com/ultralytics/ultralytics
17. Kaleem, Z., Orakzai, F. A., Ishaq, W., Latif, K., Zhao, J., Jamalipour, A. Emerging Trends in UAVs: From Placement, Semantic Communications to Generative AI for Mission-Critical Networks. IEEE Transactions on Consumer Electronics, 2024. https://doi.org/10.1109/TCE.2024.3434971
18. Khan, A., Gupta, S., Gupta, S. K. Emerging UAV Technology for Disaster Detection, Mitigation, Response, and Preparedness. Journal of Field Robotics, 2022, 39(6), 905-955. https://doi.org/10.1002/rob.22075
19. Khan, M. A., Menouar, H., Eldeeb, A., Abu-Dayya, A., Salim, F. D. On the Detection of Unauthorized Drones-Techniques and Future Perspectives: A Review. IEEE Sensors Journal, 2022, 22(12), 11439-11455. https://doi.org/10.1109/JSEN.2022.3171293
20. Li, M., Chen, Y., Zhang, T., Huang, W. TA-YOLO: A Lightweight Small Object Detection Model Based on Multi-Dimensional Trans-Attention Module for Remote Sensing Images. Complex & Intelligent Systems, 2024, 10(4), 5459-5473. https://doi.org/10.1007/s40747-024-01448-6
21. Liu, G., Dundar, A., Shih, K. J., Wang, T. C., Reda, F. A., Sapra, K., Yu, Z., Yang, X., Tao, A., Catanzaro, B. Partial Convolution for Padding, Inpainting, and Image Synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(5), 6096-6110. https://doi.org/10.1109/TPAMI.2022.3209702
22. Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., Catanzaro, B. Image Inpainting for Irregular Holes Using Partial Convolutions. Proceedings of the European Conference on Computer Vision (ECCV), 2018, 85-100. https://doi.org/10.1007/978-3-030-01252-6_6
23. Liu, J., Plotegher, L., Roura, E., De Souza Junior, C., He, S. Real-Time Detection for Small UAVs: Combining YOLO and Multi-Frame Motion Analysis. IEEE Transactions on Aerospace and Electronic Systems, 2025. https://doi.org/10.1109/TAES.2025.3577592
24. Liu, X., Li, S., Zhang, B., Wang, Y., Xu, F. Fast Aerial UAV Detection Using Improved Inter-Frame Difference and SVM. Journal of Physics: Conference Series, 2019, 1187(3), 032082. https://doi.org/10.1088/1742-6596/1187/3/032082
25. Liu, X., Li, Y., Shuang, F., Gao, F., Zhou, X., Chen, X. ISSD: Improved SSD for Insulator and Spacer Online Detection Based on UAV System. Sensors, 2020, 20(23), 6961. https://doi.org/10.3390/s20236961
26. Lyu, M., Zhao, Y., Huang, C., Huang, H. Unmanned Aerial Vehicles for Search and Rescue: A Survey. Remote Sensing, 2023, 15(13), 3266. https://doi.org/10.3390/rs15133266
27. Ni, J., Zhu, S., Tang, G., Ke, C., Wang, T. A Small-Object Detection Model Based on Improved YOLOv8s for UAV Image Scenarios. Remote Sensing, 2024, 16(13), 2465. https://doi.org/10.3390/rs16132465
28. Rahman, M. H., Sejan, M. A. S., Aziz, M. A., Tabassum, R., Baik, J. I., Song, H. K. A Comprehensive Survey of Unmanned Aerial Vehicles Detection and Classification Using Machine Learning Approach: Challenges, Solutions, and Future Directions. Remote Sensing, 2024, 16(5), 879. https://doi.org/10.3390/rs16050879
29. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C. J. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. Advances in Neural Information Processing Systems, 2021, 34, 13937-13949. https://doi.org/10.48550/arXiv.2106.02034
30. Shelekhov, A., Afanasiev, A., Shelekhova, E., Kobzev, A., Tel'minov, A., Molchunov, A., Poplevina, O. Low-Altitude Sensing of Urban Atmospheric Turbulence with UAV. Drones, 2022, 6(3), 61. https://doi.org/10.3390/drones6030061
31. Unlu, E., Zenou, E., Riviere, N. Using Shape Descriptors for UAV Detection. Electronic Imaging 2017, 2018, 1-1. https://hal.science/hal-01740282/ https://doi.org/10.2352/ISSN.2470-1173.2018.09.SRV-128
32. Viola, P., Jones, M. Rapid Object Detection Using a Boosted Cascade of Simple Features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), 2001, 1, I-I. https://doi.org/10.1109/CVPR.2001.990517
33. Wang, B., Li, Q., Mao, Q., Wang, J., Chen, C. L. P., Shangguan, A., Zhang, H. A Survey on Vision-Based Anti Unmanned Aerial Vehicles Methods. Drones, 2024, 8(9), 518. https://doi.org/10.3390/drones8090518
34. Wang, C., Shi, Z., Meng, L., Wang, J., Wang, T., Gao, Q., Wang, E. Anti-Occlusion UAV Tracking Algorithm with a Low-Altitude Complex Background by Integrating Attention Mechanism. Drones, 2022, 6(6), 149. https://doi.org/10.3390/drones6060149

35. Wang, S., Li, B., Khabsa, M., Fang, H., Ma, H. Linformer: Self-Attention with Linear Complexity. arXiv Preprint arXiv:2006.04768, 2020. https://arxiv.org/abs/2006.04768

36. Wang, X., Girshick, R., Gupta, A., He, K. Non-Local Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 7794-7803. https://doi.org/10.1109/CVPR.2018.00813

37. Woo, S., Park, J., Lee, J. Y., Kweon, I. S. CBAM: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision (ECCV), 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

38. Xiong, X., He, M., Li, T., Zheng, G., Xu, W., Fan, X., Zhang, Y. Adaptive Feature Fusion and Improved Attention Mechanism-Based Small Object Detection for UAV Target Tracking. IEEE Internet of Things Journal, 2024, 11(12), 21239-21249. https://doi.org/10.1109/JIOT.2024.3367415

39. Yan, X., Fu, T., Lin, H., Xuan, F., Huang, Y., Cao, Y., Hu, H., Liu, P. UAV Detection and Tracking in Urban Environments Using Passive Sensors: A Survey. Applied Sciences, 2023, 13(20), 11320. https://doi.org/10.3390/app132011320

40. Yasmeen, A., Daescu, O. Recent Research Progress on Ground-to-Air Vision-Based Anti-UAV Detection and Tracking Methodologies: A Review. Drones, 2025, 9(1), 58. https://doi.org/10.3390/drones9010058

41. Ye, T., Zhang, J., Li, Y., Zhang, X., Zhao, Z., Li, Z. CT-Net: An Efficient Network for Low-Altitude Object Detection Based on Convolution and Transformer. IEEE Transactions on Instrumentation and Measurement, 2022, 71, 1-12. https://doi.org/10.1109/TIM.2022.3165838

42. Zhao, S., Liu, S., Jiang, Y., Zhao, B., Lv, Y., Zhang, J., Wang, L., Zhong, R. Y. Industrial Foundation Models (IFMs) for Intelligent Manufacturing: A Systematic Review. Journal of Manufacturing Systems, 2025, 82, 420-448. https://doi.org/10.1016/j.jmsy.2025.06.011

43. Zhao, S., Zhong, R. Y., Jiang, Y., Besklubova, S., Tao, J., Yin, L. Hierarchical Spatial Attention-Based Cross-Scale Detection Network for Digital Works Supervision System (DWSS). Computers & Industrial Engineering, 2024, 192, 110220. https://doi.org/10.1016/j.cie.2024.110220

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).