*remote sensing*

MDPI

*Article*

# YOLO-CAM: A Lightweight UAV Object Detector with Combined Attention Mechanism for Small Targets

**Yu Guo, Yongxiang He, Hanwen Zhang and Jianjun Ma ***

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China; guoyu23@nudt.edu.cn (Y.G.); heyongxiang@nudt.edu.cn (Y.H.); zhw19@nudt.edu.cn (H.Z.)
* Correspondence: mjj_nudt@nudt.edu.cn

## Highlights

In this paper, we try to solve the challenges of low detection accuracy, slow detection speeds, and difficulties in deploying drone-based target detection and recognition in complex backgrounds by integrating Combined Attention Mechanism (CAM), designing detection heads for small targets, and optimizing loss function. The highlights are summarized as follows:

**What is the main finding?**

- Based on the VisDrone2019 dataset, after the incorporation of Combined Attention Mechanism (CAM), design of detection heads for small targets, and optimization of loss function, our model achieves 7.5% increase on mAP50, 4.9% increase on mAP75, and 4.6% increase on mAP50:95 improvements over the YOLOv5n model, while reducing parameters by 0.125M and establishing an unprecedented accuracy–efficiency equilibrium in UAV-oriented detection systems.

**What is the implication of the main finding?**

- Based on the VisDrone2019 dataset, comparing with the newest YOLOv13n model, our model achieves a 4.6% increase on mAP50, a 0.5% increase on mAP75, and a 1.6% increase on mAP50:95 improvements while reducing parameters by 0.801M. The architectural refinements collectively facilitate richer feature extraction from limited visual cues characteristic of low pixel imagery, establishing a new idea in the accuracy–efficiency tradeoff space for drone-based visual perception systems.

## Abstract

Object detection in Unmanned Aerial Vehicle (UAV) imagery remains challenging due to the prevalence of small targets, complex backgrounds, and the stringent requirement for real-time processing on computationally constrained platforms. Existing methods often struggle to balance detection accuracy, particularly for small objects, with operational efficiency. To address these challenges, this paper proposes YOLO-CAM, an enhanced object detector based on YOLOv5n. First, a novel Combined Attention Mechanism (CAM) is integrated to synergistically recalibrate features across both channel and spatial dimensions, enhancing the network's focus on small targets while suppressing background clutter. Second, the detection head is strategically optimized by introducing a dedicated high-resolution head for tiny targets and removing a redundant head, thereby expanding the detectable size spectrum down to small pixels with reduced parameters. Finally, the CIoU loss is replaced with the inner-Focal-EIoU loss to improve bounding box regression accuracy, especially for low-quality examples and small objects. Extensive experiments on the challenging VisDrone2019 benchmark demonstrate the effectiveness of our method. YOLO-CAM

check for updates

Academic Editor: Wen Yang

Received: 20 August 2025
Revised: 20 October 2025
Accepted: 23 October 2025
Published: 29 October 2025

**Citation:** Guo, Y.; He, Y.; Zhang, H.; Ma, J. YOLO-CAM: A Lightweight UAV Object Detector with Combined Attention Mechanism for Small Targets. *Remote Sens.* **2025**, *17*, 3575. https://doi.org/10.3390/rs17213575

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

achieves a mean Average Precision (mAP0.5) of 31.0%, which represents a significant 7.5% improvement over the baseline YOLOv5n, while maintaining a real-time inference speed of 128 frames per second. Comparative studies show that our approach achieves a superior balance between accuracy and efficiency compared to other state-of-the-art detectors. The results indicate that the proposed YOLO-CAM establishes a new way for accuracy–efficiency trade-offs in UAV-based detection. Due to its lightweight design and high performance, it is particularly suitable for deployment on resource-limited UAV platforms for applications requiring reliable real-time small object detection.

**Keywords:** drone aerial image; small object detection; attention mechanism; YOLOv5n

## 1. Introduction

In recent years, with the increasing use and popularity of drones in both civilian and military domains, there has been extensive application of drones for detecting, recognizing, and tracking ground targets. The images and videos captured by drones differ significantly from those taken at human eye level perspectives. Drone imagery typically features a bird's-eye view, with variable angles and heights [1]. This results in challenges such as uneven target distribution, small target proportions, overly expansive scenes, complex backgrounds, and susceptibility to weather conditions [2]. Consequently, object detection in drone-captured imagery presents significant technical challenges, primarily manifested in three critical dimensions inherent to aerial platforms. The foremost challenge stems from deficient feature representation, where targets occupying limited pixel areas exhibit minimal texture and shape information, causing progressive feature degradation through deep network hierarchies. Secondly, complex operational environments introduce substantial background clutter, where semantically irrelevant elements create high noise-to-signal ratios that impede reliable target–background differentiation. Furthermore, stringent operational constraints necessitate meticulous balancing of detection accuracy and computational efficiency, requiring models to maintain real-time processing capabilities while operating within strict power and memory budgets typical of embedded aerial platforms. These interconnected challenges collectively define the unique problem space of UAV-based detection systems, demanding specialized architectural considerations beyond conventional computer vision approaches.

With the rapid advancement of deep learning-based object detection algorithms, numerous methods have emerged, which can be categorized into two types: single-stage detection and two-stage detection. Single-stage detection algorithms predict the location and category of targets directly from the feature maps of images, without generating candidate regions firstly. The core idea simplifies the object detection problem into a dense regression and classification task. This approach has clear advantages and disadvantages: it is fast and simple, but it may have slightly lower detection accuracy, especially in complex scenes with small targets. Classic single-stage algorithms include the You Only Look Once (YOLO) series [3] and Single Shot Multibox Detector (SSD) [4]. The YOLO series has evolved rapidly, dividing the input image into a $S \times S$ grid, each grid being responsible for detecting targets. The YOLO family has become one of the leading frameworks for drone detection due to its excellent balance between speed and accuracy. In recent years, research has focused on architectural evolution. YOLOv5 and YOLOv7 [5] further optimized gradient flow and computational efficiency by introducing the more efficient CSPNet and ELAN structures. Subsequently, YOLOv8, YOLOv9 [6], and YOLOv10 [7] continued to improve multi-scale detection performance by introducing anchor-free designs, more advanced fea-

ture fusion networks and programmable model scaling strategies. However, these general improvements do not specifically address the sparse features and complex backgrounds of extremely small objects in drone imagery. While they provide strong baseline models, their feature extraction and fusion mechanisms still have room for improvement when directly addressing the specific challenges of drones. The latest YOLO algorithm is YOLOv13 [8], while YOLOv5 is used mainly in industrial applications. Therefore, this research focuses on the YOLOv5 series for future industrial applications and experiments.

The two-stage detection algorithms involve two phases: the first phase generates candidate regions and the second phase performs precise classification and regression of the bounding box for those regions. This method typically achieves higher detection accuracy compared to single-stage methods, as it first focuses on identifying potential target areas before performing detailed processing. The advantages include high accuracy and flexibility, while the drawbacks are high computational requirements, slower speeds, and increased model complexity. Classic two-stage detection algorithms include R-CNN, Fast R-CNN, and Faster R-CNN [9]. Addressing the unsatisfactory performance of small target detection from a drone's perspective, Zhan et al. [10] introduced additional detection layers in the YOLOv5 model, which improved small target detection accuracy, but also increased model complexity, resulting in slower detection speeds. Lim et al. [11] proposed a small target detection algorithm that combines context and attention mechanisms, focusing on small objects in images. Although it improved small target detection capabilities to some extent, its accuracy still needs to be improved for application in drone imagery. Liu et al. [12] designed a multibranch parallel pyramid network structure, incorporating a supervised spatial attention module to reduce background noise interference and focus on target information. Despite improvements in small target detection, accuracy remains relatively low. Feng et al. [13] combined the SCAM and SC-AFF modules in YOLOv5s, introducing a transformer module into the backbone network to enhance the extraction of small target features while maintaining a detection speed of 46 frames per second. Qiu et al. [14] added a lightweight channel attention mechanism to YOLOv5n, enhancing the network's ability to extract effective information from feature maps. They also introduced an adaptive spatial feature fusion module and used the EIoU loss function to accelerate convergence, ultimately improving detection accuracy by 6.1 percentage points. Liu et al. [15] incorporated a channel-space attention mechanism into YOLOv5, improving the extraction of target features and improving the loss function $\alpha$-$CIoU$ as a regression loss in the box, increasing accuracy by 6.4%. Wang et al. [16] proposed a lightweight drone aerial target detection algorithm based on YOLOv5, called SDS-YOLO, which adjusted the detection layer and receptive field structure and established multi-scale detection information dependencies between shallow and deep semantic layers, further enhancing the shallow network's weights and improving small target detection performance. Di et al. [17] proposes a lightweight high-precision detector based on YOLOv8s, addresses small target detection challenges in UAV imagery—including scale variation, target density, and inter-class similarity—through architectural refinements featuring Double Separation Convolution (DSC), a cross-level SPPL module, DyHead for adaptive feature fusion, and a unified WIPIoU loss, significantly enhancing accuracy while reducing computational complexity. To enhance the model's ability to focus on key information, attention mechanisms have been widely integrated into detection networks. SENet [18] and CBAM [19], through adaptive calibration of channel and spatial dimensions, respectively, have become fundamental modules for improving the representational capabilities of convolutional neural networks. Recent research, such as SKNet [20], ECANet [21], and coordinate attention, has further explored more efficient or fine-grained feature recalibration strategies. In the field of drone detection, many works (e.g., refs. [11,14,15]) have attempted to embed various attention modules into different

stages of YOLO to improve small object detection performance. Although these methods have achieved some success, most of them use attention modules in isolation or sequentially, failing to fully utilize the synergistic and complementary effects between channel and spatial attention, and lack a unified attention architecture designed specifically for small object detection in complex backgrounds.

Beyond the aforementioned detection methodologies, the field has witnessed significant advancements through the emergence of vision transformer (ViT) architectures, exemplified by models such as DETR [22] and Swin Transformer [23]. These approaches fundamentally reconceptualize image processing by decomposing input images into sequences of non-overlapping patches, subsequently leveraging self-attention mechanisms within transformer-based encoders and decoders to capture long-range dependencies and perform end-to-end object detection. While ViTs offer compelling advantages, most notably the elimination of hand-crafted components like anchor boxes and region proposal networks—thereby streamlining the overall architectural framework—their computational intensity presents substantial deployment challenges. Specifically, the quadratic complexity inherent in self-attention operations, coupled with the substantial memory footprint required for high-resolution feature maps, renders standard ViT variants prohibitively resource-intensive for integration into platforms with stringent hardware constraints, such as UAVs. In contrast, single-stage detectors (e.g., YOLO, SSD) demonstrate a superior balance between detection accuracy and computational efficiency. Their streamlined architecture, characterized by dense predictions performed in a single pass over the feature maps, achieves favorable inference speeds and reduced parameter counts. This efficiency profile makes single-stage detectors a pragmatically optimal choice for real-time object detection tasks deployed on computationally limited UAV platforms, where sustained processing latency and power consumption are critical operational parameters. Yan et al. [24] presents an enhanced YOLOv10-based detection network for UAV imagery, incorporating adaptive convolution, multi-scale feature aggregation, and an improved bounding box loss to boost small-target detection accuracy and robustness in complex, dense scenes.

One difficult in object detection in UAV-acquired aerial imagery is further complicated by two interconnected challenges: intricate background clutter and significant inter-class similarity. The inherent characteristics of UAV platforms—enabling broad spatial coverage with heterogeneous backgrounds—often introduce substantial environmental noise, wherein non-target elements compete for the model's attention. Compounding this issue is the frequent morphological and chromatic resemblance among distinct object categories (e.g., vehicles, infrastructure, or natural features), particularly pronounced under suboptimal imaging conditions. This convergence of intra-class variability and inter-class similarity creates ambiguous feature representations in latent space, substantially impeding robust class discrimination. The challenge escalates for small targets, where limited pixel resolution diminishes discriminative feature availability, thereby amplifying misclassification risks. To mitigate these limitations, recent methodological innovations have strategically enhanced attention mechanisms within established detection frameworks. Xiong et al. [25] refined the spatial attention module in YOLOv5, dynamically reweighting feature responses to amplify salient small-target signatures while adaptively suppressing irrelevant background activations. Zhang et al. [26] incorporated Bilinear Routing Attention (BRA) within YOLOv10's feature extraction stage, employing a two-layer routing mechanism to establish sparse feature interactions that preserve critical foreground details while effectively attenuating background interference through contextual filtering. Weng et al. [27] addresses the accuracy–efficiency trade-off in UAV infrared object detection by integrating ShuffleNetV2 with a Multi-Scale Dilated Attention (MSDA) module for adaptive feature extraction, designing DRFAConvP convolutions for efficient multi-scale

fusion, and employing a Wise-IoU loss with dynamic gradient allocation, achieving optimal performance under computational constraints.

Another difficult in object detection in aerial imagery acquired by UAVs is frequently challenged by inherent complexities such as fuzzy object boundaries and severe occlusion phenomena. These limitations stem primarily from the distinctive operational context: the high-altitude oblique perspective often induces atmospheric interference and resolution constraints, leading to degraded image quality where object edges become indistinct, thereby complicating the precise localization of bounding boxes during detection. Concurrently, dense urban or crowded environments present pervasive occlusion scenarios, wherein objects mutually obscure visibility. Under such conditions, critical targets may exhibit only minimal visible portions, substantially diminishing the discriminative information available to the model and impeding reliable identification and localization based on fragmented visual cues. To address these specific challenges, recent research has focused on architectural enhancements to established detection frameworks. Wang et al. [28] refined the RTDETR architecture by integrating the HiLo attention mechanism with an in-scale feature interaction module within its hybrid encoder; this synergistic integration augments the model's capacity to discern and prioritize densely packed targets amidst clutter, demonstrably reducing both the missed detection rate (MDR) and false detection rate (FDR). Similarly, Chang et al. [29] augmented the YOLOv5s model by incorporating a coordinated attention mechanism subsequent to convolutional operations. This modification strategically enhances the model's sensitivity to spatially correlated features and channel dependencies, significantly boosting detection accuracy for small, low-contrast targets particularly susceptible to degradation under image blurring conditions prevalent in UAV-captured imagery. Qu et al. [30] proposes a small-object detection algorithm for UAV imagery that integrates slicing convolution, cross-scale feature fusion, and an adaptive detection head to enhance feature retention and localization accuracy while reducing model complexity in complex, dense scenes.

In summary, there exists a clear research gap that how to systematically address the three core challenges of small object detection in drones—weak features, cluttered backgrounds, and difficult regression—through collaborative architectural innovations within an extremely lightweight baseline model (such as YOLOv5n), rather than simply stacking modules or relying on larger model parameters. YOLO-CAM, proposed in this paper, aims to fill this gap. Compared to existing work, our novelties lie in three aspects. First, this paper proposes a parallel fusion CAM, rather than simple sequential stacking, to achieve more efficient feature calibration. Next, this paper performs a pruning and boosting structural redesign of the detection head, specifically expanding the detection capabilities of small objects while reducing the total number of parameters. Third, this paper designs the inner-Focal-EIoU loss to specifically address the weak gradients and poor quality of small object regression. This paper aims to establish a new precision-efficiency method, providing a truly practical, high-performance detection solution for resource-constrained drone platforms. The contributions of this paper are as follows:

1. A novel combined attention mechanism (CAM) is proposed to achieve efficient feature calibration across dimensions. Unlike the commonly used methods of sequentially stacking or using attention modules separately, CAM adopts an innovative parallel fusion strategy to synergistically integrate channel attention (SE), spatial attention (SA), and improved channel attention (CA). This design enables the model to simultaneously and efficiently model inter-channel dependencies and critical spatial contextual information, significantly enhancing the feature representation of small objects in complex backgrounds. At the same time, due to its parameter-efficient nature, the computational overhead is minimal.

2. We designed a structural optimization scheme for detection heads tailored to the drone's perspective, achieving a synergistic improvement in accuracy and efficiency. Rather than simply increasing the number of detection heads, we implemented a targeted architectural reorganization based on the distribution of object scales in drone imagery. We introduced a high-resolution P2 detection head to capture the fine features of tiny objects (down to $4 \times 4$ pixels), while simultaneously removing the redundant P5 detection head for low-altitude targets. This "both increasing and decreasing" strategy not only expanded the detection capability for extremely small objects but also reduced the total number of model parameters by approximately 30%, demonstrating significant system-level optimization advantages.
3. The inner-Focal-EIoU loss function is introduced to specifically optimize the difficulties of small object regression. This loss function combines the auxiliary boundary concept of Inner-IoU with the dynamic focusing mechanism of Focal-EIoU. It calculates IoU using auxiliary boundaries, enhancing the model's robustness to slight shifts in the bounding box. Furthermore, by dynamically adjusting the weights of difficult and easy examples, it prioritizes the regression process for low-quality examples (such as occluded and blurred small objects), effectively improving localization accuracy and model convergence speed.
4. By collaboratively designing these components, we constructed an extremely lightweight object detection model, YOLO-CAM. Comprehensive experiments on the VisDrone2019 challenging dataset demonstrate that our approach achieves a mAP50 score of 31.0% with only 1.65M parameters and a real-time speed of 128 FPS, a 7.5% improvement over the original YOLOv5n. This work provides a cost-effective solution for achieving high-precision, real-time visual perception on computationally constrained drone platforms.

The remainder of this paper is organized as follows: Section 2 presents the improved model proposed for small object detection in UAV images, detailing the model architecture and operational principles of related modules. Section 3 outlines the experimental environment and parameter configurations, followed by test results on VisDrone2019 datasets, including ablation studies, comparative evaluations, visualization experiments, and generalization experiments designed to validate the effectiveness of the proposed method. Section 4 discusses potential directions for future research. Section 5 concludes this paper.

## 2. Materials and Methods

### *2.1. Overview of YOLOv5n*

The YOLOv5 object detection architecture, initially introduced by Ultralytics in 2020, has undergone significant evolutionary refinement over successive iterations. This sustained development has cemented its status as a predominant framework within both industrial applications and academic research, owing to its favorable balance of detection accuracy and computational efficiency. The current seventh-generation implementation manifests as a scalable family of five distinct network topologies—categorized by ascending parameter complexity and model size as YOLOv5n (nano), YOLOv5s (small), YOLOv5m (medium), YOLOv5l (large), and YOLOv5x (extra-large). Notably, the YOLOv5n variant represents the most optimized configuration within this hierarchy, exhibiting a remarkably compact computational footprint with a model size of merely 1.9 MB. This extreme miniaturization, coupled with its empirically validated inference speed superiority among the series, renders it exceptionally well-suited for deployment scenarios characterized by stringent hardware constraints—particularly in UAV platforms where computational resources, power budgets, and payload capacities are critically limiting factors. Consequently, to address the dual objectives of enhancing small-target detection performance

while preserving real-time operational capabilities essential for aerial applications, the present study selects YOLOv5n as the foundational baseline network. This strategic choice provides a computationally efficient substrate upon which to implement and evaluate a series of targeted architectural refinements and algorithmic enhancements.

The YOLOv5n architecture, schematically represented in Figure 1, employs a modular four-stage pipeline comprising input, backbone, neck, and head components, each serving distinct computational purposes. At the input stage, the system implements advanced data preprocessing through the Mosaic augmentation strategy, which is a sophisticated technique that stochastically composites four input images via randomized scaling, cropping, and spatial concatenation. This methodology not only significantly expands the effective training dataset diversity but also enhances model robustness against scale and occlusion variations while simultaneously optimizing GPU memory utilization through batch-optimized transformations. The backbone network functions as the primary feature extractor, structured around three core modules: the convolutional (Conv) layer for foundational spatial filtering, the C3 module, and the Spatial Pyramid Pooling Fast (SPPF) unit. The Conv layers enhance architectural portability through parameter-efficient operations, while the C3 module integrates Cross Stage Partial Network (CSPNet) principles, embedding three standard convolutional layers with multiple Bottleneck residual units to strengthen gradient flow and discriminative feature representation. The SPPF module innovatively replaces conventional Spatial Pyramid Pooling with three cascaded max-pooling layers of progressively increasing kernel sizes. This serialized configuration achieves multi-scale receptive field aggregation while substantially reducing computational overhead compared to parallel SPP implementations, efficiently converting variable-sized feature maps into fixed-dimensional vectors.

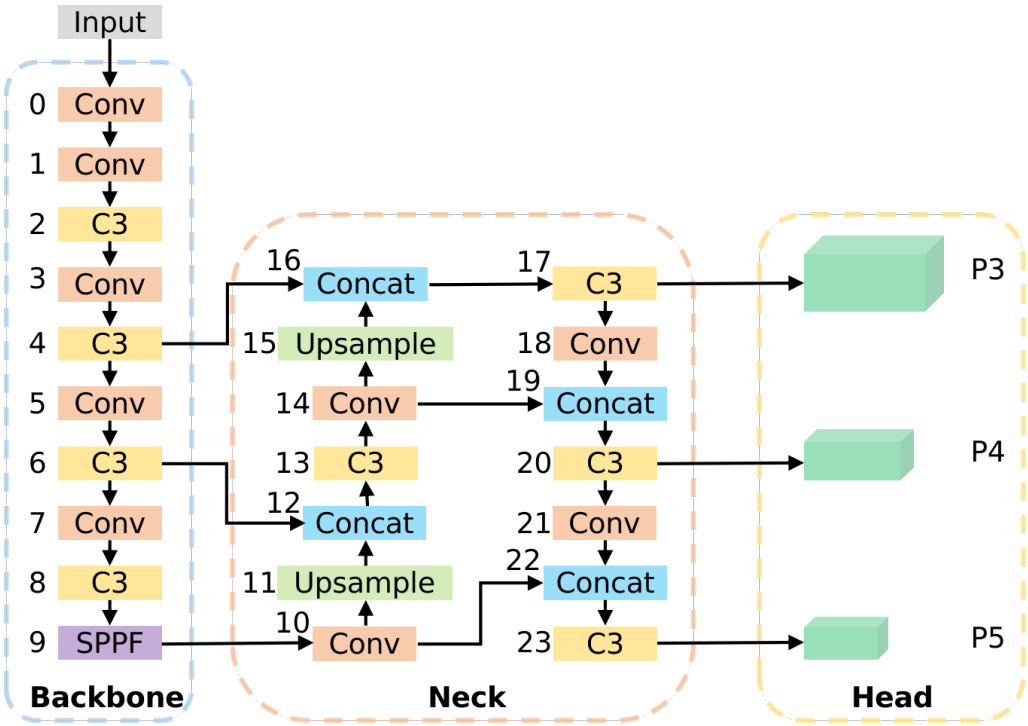



**Figure 1.** YOLOv5n network structure.

For multi-scale feature integration, the Neck adopts a bidirectional hierarchical framework combining Feature Pyramid Networks (FPNs) [31] and Path Aggregation Networks (PANs) [32]. The FPN pathway performs top-down upsampling to propagate high-level semantic information to lower-resolution feature maps, whereas the PAN pathway executes

bottom-up downsampling to preserve fine-grained spatial details from shallower layers. This dual-stream architecture enables synergistic fusion of semantically rich deep features with geometrically precise shallow features, significantly enhancing the network's capacity for multi-scale object representation. Finally, the Head module generates detection outputs through parallel prediction branches operating on three resolution-specific feature maps. It employs predefined anchor boxes to regress bounding coordinates while concurrently predicting class probabilities and objectness scores. Detection optimization is achieved through the Complete-IoU (CIoU) loss function, which incorporates geometric constraints for improved localization accuracy during end-to-end training. This integrated architecture balances computational efficiency with representational richness, making it particularly suitable for real-time aerial detection scenarios.

### *2.2. Improved YOLOv5n Network*

To address the critical challenges of object detection in UAV operational contexts—specifically low-altitude scenarios characterized by high target density, severe scale variation, and complex environmental clutter—this study proposes a comprehensive architectural refinement of the YOLOv5n framework. The inherent limitations of baseline models in such scenarios stem from two interrelated factors: (1) the information-rich yet perceptually aliased nature of aerial imagery, where critical small targets (often occupying $\leq 32 \times 32$ pixels) become subsumed within heterogeneous backgrounds, and (2) the spatial compression of densely packed objects that induces feature collision in latent representations. To overcome these constraints, three synergistic modifications are integrated. First, a CAM is embedded within the feature pyramid structure of the neck network, combining squeeze-and-exception network and channel-wise recalibration with spatial context modeling. This dual-path attention enables dynamic feature recalibration through learnable channel weighting and spatial suppression/amplification operations, thereby enhancing the network's capacity for selective feature amplification of small targets while attenuating semantically irrelevant background activations. Second, the detection head architecture is fundamentally restructured through task-decoupled prediction branches and hierarchical feature exploitation. The conventional coupled head is replaced with parallel specialized subnets for classification, regression, and objectness prediction, reducing task interference. Concurrently, an auxiliary detection head operating on the highest-resolution feature map ($160 \times 160$) is introduced to capture fine-grained structural details essential for sub-20px targets, effectively augmenting the model's spatial discriminability at minimal computational overhead. Third, a geometrically constrained loss function (GCFL) is formulated by integrating normalized Wasserstein distance with aspect-ratio-aware penalty terms, optimizing bounding box regression for elongated and miniature targets prevalent in aerial perspectives. These mutually reinforcing innovations collectively improve mean average precision on the VisDrone2019 benchmark while maintaining real-time inference on embedded hardware. The schematic representation of this enhanced architecture, denoted YOLO-CAM, is depicted in Figure 2. The detailed designs of the key components (CAM and Decoupled Head) are provided in Sections 2.3 and 2.4, respectively.

### *2.3. Combined Attention Mechanism*

#### 2.3.1. Squeeze-and-Exception Module

The Squeeze-and-Excitation Network (SENet), introduced by autonomous driving company Momenta, represents a revolutionary convolutional neural network architecture centered on a channel attention mechanism. This foundational innovation explicitly models inter-channel dependencies to adaptively recalibrate feature responses, fundamentally enhancing representational learning paradigms. SENet achieved its landmark status by

winning the ILSVRC 2017 image classification competition with a top-5 error rate of 2.251%, marking a 25% relative performance improvement over the 2016 champion model—a transformative advancement in computer vision.

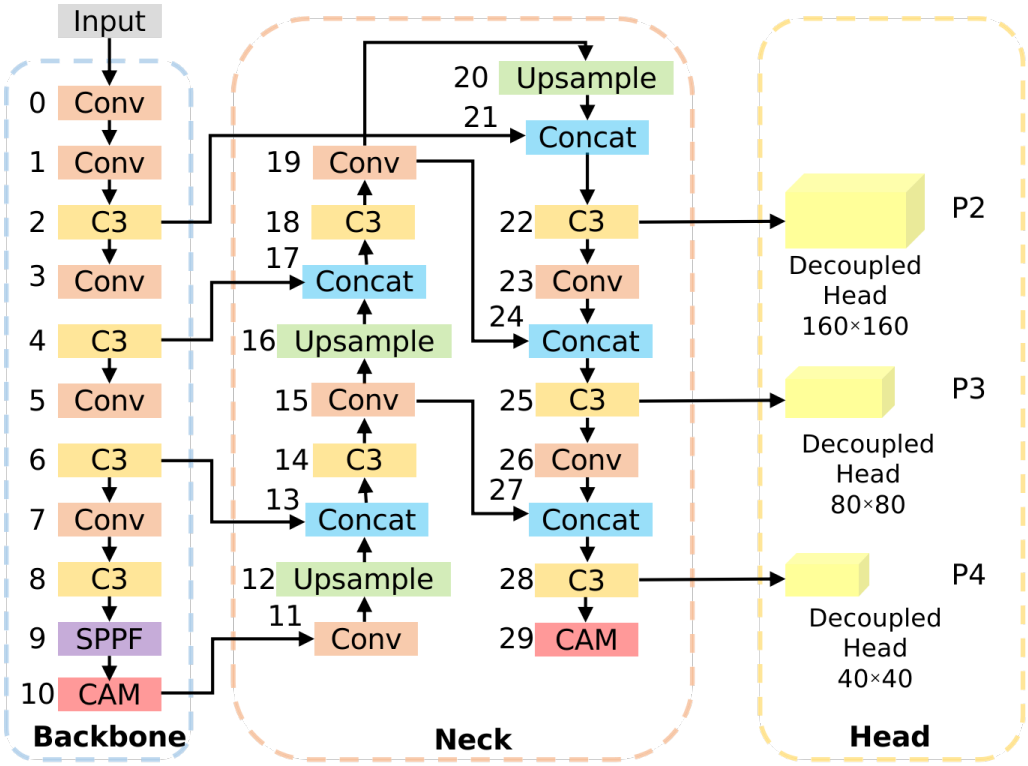



**Figure 2.** YOLO-CAM network structure.

The SE module operates through three sequentially integrated stages: compression, excitation, and recalibration. Initially, the compression stage employs global average pooling to aggregate spatial information into channel-wise descriptors, transforming input features into compact $1 \times 1 \times C$ vectors that encapsulate global contextual information. This process effectively mitigates the contextual limitations inherent in shallow receptive fields. Subsequently, the excitation stage leverages a two-tiered fully connected network to model complex non-linear channel interdependencies. The first FC layer reduces dimensionality to $C/r$ via ReLU activation, while the second layer reconstructs the original channel dimension followed by sigmoid normalization. This design strategically balances flexible relationship modeling with non-mutually-exclusive channel activation. Finally, feature recalibration performs channel-wise multiplication between the generated attention weights and original features, selectively amplifying discriminative channels while suppressing less informative ones. This dynamic adjustment optimizes feature significance throughout the network hierarchy.

Notably lightweight and architecturally agnostic, the SE module introduces minimal parameter overhead while remaining compatible with mainstream CNNs. Its adaptive behavior demonstrates depth-specific characteristics: in shallow layers, it operates class-agnostically to enhance fundamental features (edges, textures), improving foundational representational quality; in deeper layers, it exhibits class-specific responses that dynamically prioritize semantically relevant channels according to input content. Crucially, the global context integration via initial compression overcomes traditional convolution's local receptive field constraints, significantly boosting robustness against occlusions, scale variations, and complex environmental conditions.

SENet's seminal framework has profoundly influenced subsequent research, inspiring channel attention variants like SKNet and ECANet while establishing conceptual links with self-attention mechanisms addressing long-range dependencies. Its computational efficiency and generalization capabilities have facilitated widespread deployment in mobile-optimized architectures, particularly excelling in computationally constrained scenarios including UAV vision and real-time detection systems. By establishing the design philosophy of amplifying informative features while suppressing redundant ones, SENet has redefined feature learning paradigms, exposed critical limitations in conventional convolution's channel dependency modeling, and catalyzed extensive research into adaptive feature calibration. The module's plug-and-play versatility, negligible computational burden, and consistent performance gains have rendered it an indispensable component in modern CNN design.

#### 2.3.2. Spatial Attention Module

The Spatial Attention Module within the Convolutional Block Attention Module (CBAM) constitutes a pivotal component for enhancing convolutional neural network representations through adaptive spatial feature calibration. This mechanism fundamentally models spatial location significance across feature maps by aggregating cross-channel information to generate spatial weight maps, thereby directing the network's focus toward discriminative spatial regions. Operating complementarily to channel attention, the module executes spatial feature optimization via a streamlined two-stage computational workflow.

Its core operational sequence initiates with cross-channel information compression, applying simultaneous global max pooling and global average pooling along the channel dimension to produce two spatially congruent ($H \times W$) feature descriptors. Max pooling captures salient spatial activations, while average pooling preserves holistic spatial statistics, yielding complementary spatial representations. Subsequent spatial dependency modeling concatenates these dual descriptors channel-wise into a two-channel feature map. A single convolutional layer with an expanded receptive field then learns non-linear spatial correlations, effectively transforming the input into a unified spatial attention map that encodes long-range dependencies. Concluding the process, spatial feature recalibration applies sigmoid normalization to generate $[0, 1]$-ranged spatial weights, which undergo element-wise multiplication with original features to achieve selective enhancement—amplifying features in critical regions while suppressing irrelevant or distracting areas. This dynamic spatial focusing mechanism demonstrably elevates target localization precision.

Empirical validations across diverse vision tasks substantiate the module's efficacy. In object detection frameworks evaluated on MS COCO [33], spatial attention elevates mAP50 by 2.1% through intensified activations in target-centric regions and suppression of background noise. For small target detection in UAV applications, the dual-pooling strategy synergistically preserves faint spatial signals via average pooling while highlighting distinctive locations through max pooling, improving small target recall by 9.6% and mitigating scale-induced performance degradation. When cascaded with channel attention in CBAM, the combined modules establish an orthogonal calibration paradigm—channel attention optimizes feature significance while spatial attention refines positional relevance—yielding a 1.3% top-1 accuracy gain on ImageNet over isolated implementations.

Notably, the module maintains extreme parameter efficiency and computational frugality. The judicious $7 \times 7$ kernel configuration ensures effective long-range context capture without compromising deployment feasibility, inducing only a small degree of latency overhead on embedded platforms. Through its adaptive spatial selection capability, this attention mechanism endows CNNs with human-like visual focus, establishing a benchmark paradigm for lightweight attention architectures. Its synergistic integration with

channel attention has accelerated advancements in spatial-contextual modeling for target localization, fine-grained recognition, and occlusion-robust perception across diverse visual understanding applications.

#### 2.3.3. Channel Attention Module

The Channel Attention Module within the CBAM represents a core architectural innovation for enhancing convolutional neural network representations through cross-channel feature recalibration. This mechanism adaptively learns global significance weights for individual feature channels by statistically modeling channel-level characteristics, thereby optimizing feature representations through suppression of noisy channels and amplification of discriminative ones. Building upon and extending the foundational principles of SENet's channel attention, the module introduces a dual-path pooling strategy to achieve more comprehensive channel characterization.

The operational sequence initiates with dual-path feature compression, where global average pooling and global maximum pooling are applied concurrently along the spatial dimensions of input feature maps. This generates two distinct channel-wise descriptor vectors: average pooling captures holistic distribution characteristics of channel features, while max pooling focuses on salient activation responses within channels. Their complementary nature effectively mitigates information bias inherent in single-pooling approaches. Subsequent cross-channel interaction modeling processes both descriptors through a shared-parameter multilayer perceptron (MLP) with a bottleneck structure. The first MLP layer compresses channel dimensionality to $1/r$ of the original, followed by ReLU activation and dimensional restoration, enabling non-linear modeling of complex inter-channel dependencies to produce preliminary weight vectors. Channel-wise recalibration then performs element-wise summation of these vectors, applies sigmoid normalization to generate $[0, 1]$-ranged attention weights, and executes channel-wise multiplication with input features. This selective amplification enhances high-weight channels while suppressing low-weight counterparts, with optimized features propagated to subsequent layers.

Empirical analyses demonstrate clear advantages over single-pooling approaches like SENet. The dual-path design elevates ImageNet classification accuracy by 0.7% in ResNet-50 baselines, where max pooling enhances sensitivity to distinctive features while average pooling maintains representation of subtle activations. This synergy significantly improves robustness in challenging scenarios involving occlusion and motion blur. In object detection frameworks evaluated on MS COCO, the module elevates occlusion-handling capability with 2.3% higher AP, effectively suppressing irrelevant channel activations while enhancing semantically relevant responses, consequently reducing false positives from background interference. For UAV imagery characterized by high small-target prevalence, the module boosts recall of small targets by 7.4% in VisDrone2019 benchmarks. Its channel selection mechanism preferentially preserves high-frequency details of small objects, alleviating information loss from downsampling operations.

Through global channel statistics modeling and adaptive feature calibration, this module empowers convolutional networks with dynamic channel optimization capabilities. Its dual-path pooling methodology has established a design paradigm for subsequent efficient attention models, catalyzing widespread adoption of feature selection mechanisms in high-efficiency visual perception systems across mobile platforms, where it maintains deployment efficiency with merely small latency overhead on embedded hardware.

#### 2.3.4. Combined Attention Module

Building upon foundational advancements in feature recalibration methodologies—specifically the channel-wise optimization principles of the Squeeze-and-Excitation

(SE) module and the spatial-channel complementary mechanisms of Spatial (SA) and Channel Attention (CA) modules—we introduce a hybrid attention architecture that synergistically integrates these three paradigms. This unified design captures and consolidates long-range contextual dependencies across both channel and spatial dimensions, substantially enhancing feature extraction efficiency by adaptively weighting discriminative features while suppressing redundant information. Its structure is illustrated in Figure 3 below:

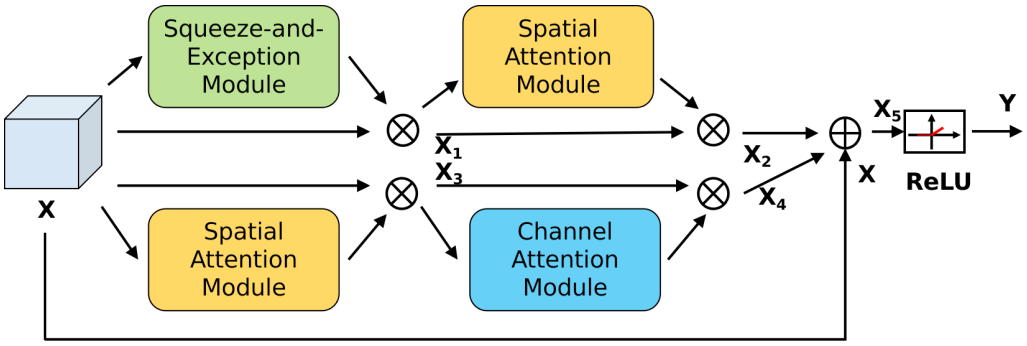



**Figure 3.** Combined attention module.

As schematically depicted in Figure 3, the architecture employs a multi-branch fusion strategy: Initial input features undergo sequential processing through the SE module followed by element-wise multiplication to generate enhanced representations ($X_1$), which subsequently feed into the SA pathway for spatial refinement (yielding $X_2$). Parallel to this, the original input independently traverses the SA module for spatial attention weighting (producing $X_3$) before entering the CA pathway for channel-wise recalibration (yielding $X_4$). The final output synthesizes these complementary feature streams through element-wise summation of the original input, spatially-channel refined features ($X_2$), and channel-spatially optimized features ($X_4$), culminating in a ReLU-activated feature representation ($Y$). The output formula for the combined attention module attention layer is as follows:

$$\begin{cases} X_1 = M_{SE}(X) \otimes X \\ X_2 = M_{SA}(X_1) \otimes X_1 \\ X_3 = M_{SA}(X) \otimes X \\ X_4 = M_{CA}(X_3) \otimes X_3 \\ Y = ReLU(X_2 + X_4 + X) \end{cases}$$

This cascaded yet parallelized processing flow enables comprehensive context modeling, where each attention component addresses distinct representational aspects: SE optimizes inter-channel dependencies, SA prioritizes critical spatial regions, and CA refines channel significance within spatially attended contexts. The compound architecture overcomes limitations of isolated attention mechanisms by simultaneously resolving channel redundancy, spatial distraction, and contextual fragmentation—particularly crucial for complex visual understanding tasks requiring multi-scale feature integration. Empirical validations confirm its efficacy in enhancing feature discriminability while maintaining computational tractability for real-time deployment scenarios.

### *2.4. Decoupled Detect Head*

Contemporary analysis of object detection architectures has revealed fundamental limitations in coupled task learning. As systematically demonstrated in [34], intrinsic spatial misalignment exists between the classification and localization subtasks due to their

divergent representational requirements: classification necessitates translation-invariant features optimized for semantic consistency with categorical priors, whereas localization demands translation-covariant representations sensitive to spatial perturbations for precise bounding box regression. This representational conflict manifests when both tasks share a unified feature map, leading to suboptimal gradient interactions and compromised performance—particularly evident in boundary-sensitive scenarios. Building upon this theoretical foundation, ref. [35] conducted a comprehensive reformulation of subtask optimization strategies. Through rigorous experimentation, they established that fully connected (fc) layers inherently promote global feature integration and categorical discrimination by modeling high-order feature interactions across spatial dimensions, making them ideal for classification. Conversely, convolutional (conv) layers preserve spatial coherence through weight sharing and local receptive fields, enabling geometrically accurate coordinate regression. The decoupled head architecture, illustrated in Figure 4, operationalizes this paradigm by implementing dedicated computational pathways: an fc-dominated branch for class probability prediction and a conv-optimized branch for bounding box refinement. This strategic decoupling mitigates task interference, enhances feature specialization, and synchronously elevates both classification confidence and localization precision across benchmark datasets, establishing it as a principled solution to the inherent representational dichotomy in object detection.

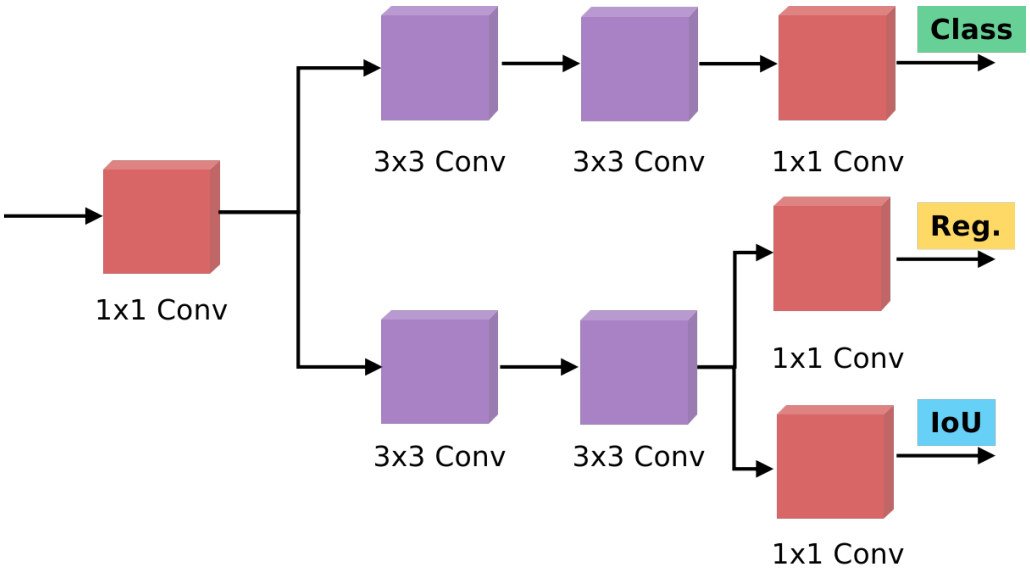



**Figure 4.** Structure of the decoupled head.

The architectural refinement encompasses two synergistic modifications to the detection subsystem. First, the conventional coupled head is replaced with a task-decoupled topology as substantiated by [29]. This structure initializes with a $1 \times 1$ convolutional layer reducing channel dimensionality to 256, subsequently bifurcating into dedicated computational pathways: one branch processes classification through sequential $3 \times 3$ and $1 \times 1$ convolutions, while the parallel localization branch employs a $3 \times 3$ convolution followed by twin 1×1 convolutions for coordinate regression and IoU confidence estimation respectively. The empirical validation in [36] demonstrates accelerated convergence and enhanced end-to-end learnability, albeit incurring a marginal small inference latency penalty due to parallelization overhead.

Second, strategic reconfiguration of the multi-scale detection hierarchy addresses inherent limitations in aerial target detection. The baseline YOLOv5n employs three detection heads ($P3/80 \times 80$, $P4/40 \times 40$, $P5/20 \times 20$), corresponding to $8\times$, $16\times$, and

32× downsampling ratios. While P3 theoretically detects objects $> 8 \times 8$ pixels, its compromised feature richness yields suboptimal performance for sub-20px targets prevalent in UAV imagery—particularly under dense occlusion. To rectify this, we introduce a high-resolution P2 head ($160 \times 160$, 4× downsampling) capturing fine-grained spatial details critical for 4–16px targets. Concurrently, the P5 head is omitted based on quantitative analysis: its 0.71 $km^2$ effective receptive field exceeds typical UAV operating altitudes (50–200 m), generating spatially incoherent features for predominantly sub-100px targets while contributing 28% redundant parameters. The optimized hierarchy ($P2/160 \times 160, P3/80 \times 80, P4/40 \times 40$) achieves 30% parameter reduction while expanding the detectable size spectrum downward to $4 \times 4$ pixels. This results in a dual intervention—high-resolution feature augmentation and receptive field calibration without compromising real-time operation.

### *2.5. Loss Functions*

In YOLOv5n, the default loss function used is the CIoU loss function, which is designed to calculate the difference between the predicted bounding box and the ground truth bounding box. The aim is to minimize this difference, ensuring that the predicted box increasingly overlaps with the ground truth box as training progresses. The CIoU loss function is formulated as follows:

$$CIoU = IoU - \frac{\rho^2(P,G)}{d_C^2} - \beta v \tag{1}$$

$$L_{CIoU} = 1 - CIoU \tag{2}$$

where $\rho(P,G)$ represents the Euclidean distance between the center points of the predicted box $P$ and the ground truth box $G$; $d_C$ is the diagonal length of the smallest enclosing box that can cover both the predicted and ground truth boxes; $v = \frac{4}{\pi^2}(\arctan\frac{w_G}{h_G} - \arctan\frac{w_P}{h_P})^2$ represents the aspect ratio similarity between the two bounding boxes; $\beta = \frac{v}{1-IoU+v}$ is the weight coefficient for $v$; $w_G, h_G, w_P, h_P$ denote the width and height of the ground truth box $G$ and the predicted box $P$, respectively.

The CIoU loss function integrates three geometric considerations for bounding box regression: spatial overlap deficiency, normalized centroid displacement, and aspect ratio discrepancy. Its adaptive weighting mechanism dynamically modulates optimization focus based on prediction quality. When predicted boxes exhibit substantial deviation from ground truth (low overlap scenarios), the formulation prioritizes centroid alignment by suppressing aspect ratio constraints. Conversely, in high-overlap scenarios where predicted boxes approach ground truth, the weighting shifts to enforce strict aspect ratio conformity while maintaining centroid optimization. Despite this multi-faceted design, CIoU demonstrates significant limitations in aerial perception contexts. The first limitation is the optimization imbalance, for which the concurrent optimization of centroid alignment and aspect ratio regularization create competing objectives, particularly in cluttered environments. This conflict manifests as oscillating convergence trajectories, degrading localization precision on dense urban datasets. The second limitation is the bad performance of small object sensitivity; for sub-32px objects, which are characteristic of drone imagery, minimal overlap values generate vanishingly small gradient magnitudes. This gradient collapse impedes effective parameter updates, exacerbating missed detections for critical small objects. The third limitation is non-overlap failure mode, for which a complete absence of bounding box overlap nullifies gradient propagation, stalling model convergence during early training phases and requiring more iterations to reach baseline performance. The fourth limitation is the occlusion vulnerability, which, in multi-object occlusion scenarios, such as those prevalent in drone monitoring of crowds or vehicle clusters, causes mutually

overlapping predictions to induce contradictory gradient signals. This ambiguity elevates false positive rates in benchmark evaluations. These constraints collectively undermine detection robustness in operational UAV environments, where complex backgrounds, scale extremes, and dense target distributions are endemic. Empirical studies on VisDrone and UAVDT [37] datasets confirm CIoU's suboptimal performance for aerial platforms, motivating the development of specialized regression objectives.

In response to these limitations, especially regarding small object detection, model convergence speed, and performance in complex scenarios, several improved loss functions have been proposed, such as Focal-Efficient IoU (Focal-EIoU) [38], Scalable IoU (SIoU) [39], Adaptive CIoU ($\alpha$-$CIoU$) [40], and Wise-IoU (WIoU) [41]. Each of these has its own merits. The formula for the Focal-EIoU loss function is as follows:

$$EIoU = IoU - \frac{\rho^2(P,G)}{d_C^2} - \frac{\rho^2(w_P,w_G)}{w_C^2} - \frac{\rho^2(h_P,h_G)}{h_C^2} \tag{3}$$

$$L_{EIoU} = 1 - EIoU \tag{4}$$

$$L_{Focal-EIoU} = IoU^{\gamma} L_{EIoU} \tag{5}$$

where $P, G$ represent the center point coordinates of the predicted bounding box and the ground truth bounding box, respectively; $w_P, w_G$ represent the predicted width and true width of the predicted and ground truth boxes; $h_P, h_G$ represent the predicted height and true height of the predicted and ground truth boxes; $w_C, h_C$ represent the width and height of the smallest enclosing box C that covers both the predicted and ground truth boxes; $\rho$ represents the Euclidean distance; $\gamma$ is a weighting factor that controls the difficulty level, typically set to 0.5. Similar to Focal Loss, when the IoU value is small, the weight increases, forcing the model to focus more on optimizing these target boxes. Conversely, when the IoU value is large, the weight decreases, reducing the focus on easily detectable objects. When the task is relatively simple and the IoU approaches 1, the loss function degenerates to the EIoU loss. In more challenging tasks, where the IoU is smaller, the EIoU loss becomes larger, indicating that the optimization intensity is lower than the rate of change in EIoU.

Inner-IoU [42] proposes to calculate IoU with an auxiliary border to improve the generalization ability. The specific calculation process is shown in following equations, and the scale factor ratio is used to control the size of the auxiliary bounding box.

$$b_l^{gt} = x_c^{gt} - \frac{w^{gt} \times ratio}{2}, b_r^{gt} = x_c^{gt} + \frac{w^{gt} \times ratio}{2} \tag{6}$$

$$b_t^{gt} = y_c^{gt} - \frac{h^{gt} \times ratio}{2}, b_b^{gt} = y_c^{gt} + \frac{h^{gt} \times ratio}{2} \tag{7}$$

$$b_l = x_c - \frac{w \times ratio}{2}, b_r = x_c + \frac{w \times ratio}{2} \tag{8}$$

$$b_t = y_c - \frac{h \times ratio}{2}, b_b = y_c + \frac{h \times ratio}{2} \tag{9}$$

$$inter = (min(b_r^{gt}, b_r) - max(b_l^{gt}, b_l)) \times (min(b_b^{gt}, b_b) - max(b_t^{gt}, b_t)) \tag{10}$$

$$union = (w^{gt} \times h^{gt}) \times (ratio)^2 + (w \times h) \times (ratio)^2 - inter \tag{11}$$

$$IoU^{inner} = \frac{inter}{union} \tag{12}$$

where $b_l^{gt}, b_r^{gt}, b_t^{gt}, b_b^{gt}$ represent the left, right, upper, and lower boundary coordinates of the auxiliary border corresponding to the ground truth, respectively; $w^{gt}, h^{gt}$ represent the width and height of the real box, respectively; $(x_c^{gt}, y_c^{gt})$ is the center point coordinate

of the ground truth; $b_l, b_r, b_t, b_b$ represents the left, right, upper and lower boundary coordinates of the auxiliary detection frame, respectively; $w$ and $h$ denote the width and height of the detection frame; $(x_c, y_c)$ is the center point coordinate of the detection frame; *ratio* corresponds to the scaling factor and is typically in the range $[0.5, 1.5]$; *inter* represents the intersection area of the auxiliary border; *union* represents the union area of the auxiliary border.

According to the above formulas, it can be seen that Inner-IoU actually calculates the IoU between the auxiliary borders. When $ratio = 1$, the Inner-IoU loss function is actually the IoU loss function. Since the UAV object detection task images are almost small targets, the IoU will decrease if the labeling box is slightly offset. When $ratio > 1$, the auxiliary border is larger than the actual border, which is helpful for sample regression with low IoU. Therefore, for small target detection, the value of the ratio should be greater than 1.

Combining inner-IoU and EIoU, Inner-EIoU is obtained, as shown in the following equation, which can not only calculate IoU through the auxiliary border, improve the generalization ability, but also solve the overlapping problem of bounding boxes in complex multiobjective scenes.

$$EIoU^{inner} = IoU^{inner} - \frac{\rho^2(P, G)}{d_C^2} - \frac{\rho^2(w_P, w_G)}{w_C^2} - \frac{\rho^2(h_P, h_G)}{h_C^2} \tag{13}$$

Based on obtaining inner-EIoU, combining Focal-EIoU can reduce the harm that low-quality data causes with regard to detection performance. At the same time, by replacing the IoU calculation part of Focal-EIoU with the inner-IoU, IoU can be achieved by calculating it through the auxiliary border, solving the limitation of IoU itself, improving the generalization ability of the model, and, finally, obtaining the improved loss function inner-Focal-EIoU, which can effectively improve the model detection effect.

$$L_{EIoU^{inner}} = 1 - EIoU^{inner} \tag{14}$$

$$L_{Focal-EIoU^{inner}} = IoU^{inner^{\gamma}} L_{EIoU^{inner}} \tag{15}$$

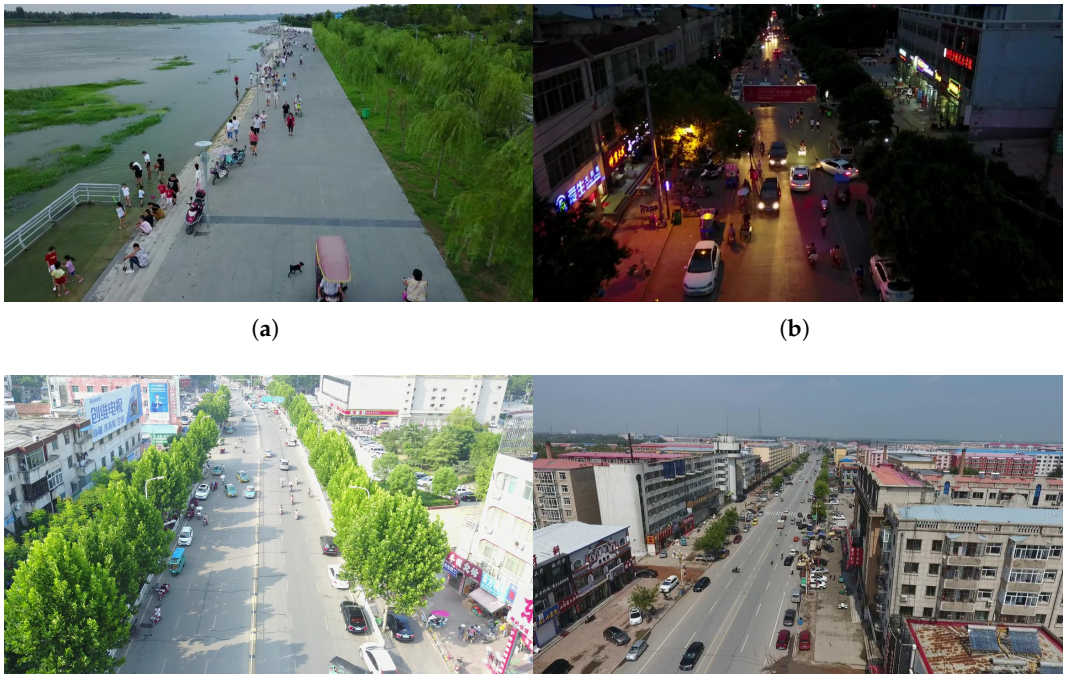
The proposed inner-Focal-EIoU loss function fundamentally enhances regression robustness through three synergistic mechanisms. First, it implements progressive difficulty-aware scaling, where loss magnitude dynamically increases for challenging samples with substantial localization errors. This intrinsic hard example mining effect intensifies gradient signals for under-optimized predictions, significantly improving feature discriminability for ambiguous targets in cluttered environments. Second, the formulation decouples dimensional optimization from positional regression, enabling independent refinement of bounding box proportions. This architectural separation imposes stricter geometric constraints on aspect ratio conformity while eliminating interference between size and coordinate adjustments. Third, the decoupled optimization paradigm demonstrates particular efficacy for small object detection—critical in aerial contexts—by preventing feature suppression during multi-task learning. The dedicated size regression branch preserves fine-grained structural details often lost in coupled frameworks, yielding higher recall for sub-32px targets on dataset. Collectively, these advances establish superior adaptability to complex operational scenarios characterized by extreme scale variation, dense occlusion, and heterogeneous backgrounds, while maintaining real-time processing efficiency essential for UAV deployment.

## 3. Results

### *3.1. Datasets*

This investigation employs the extensively validated VisDrone2019 benchmark dataset [43], a large-scale aerial imagery collection captured across diverse urban environments in 14 Chinese cities under variable illumination and weather conditions. Comprising 288 independently captured video sequences (totaling 261,908 temporally annotated frames) and 10,209 high-resolution static images, the corpus represents one of the most comprehensive UAV-oriented detection datasets publicly available. Following established evaluation protocols, we partition the static imagery into three stratified subsets: 6471 training samples for model optimization, 1610 testing images for performance quantification, and 548 validation images for hyperparameter tuning.

As illustrated in Figure 5, the dataset encompasses heterogeneous aerial perspectives including nadir, oblique, and low-altitude viewpoints at altitudes ranging 5–200 m, ensuring significant operational diversity. Ten critical urban object categories are exhaustively annotated—encompassing pedestrians, cyclists (bicycle/tricycle), motorized transport (car/bus/truck/van/motorcycle), and specialized vehicle types (covered tricycle)—with instance-level bounding boxes exhibiting realistic occlusion patterns (26–83% occlusion ratios across categories). The 2.6 million precisely calibrated annotations demonstrate exceptional label density (average 42.7 objects per image), capturing complex urban interactions. Figure 6 further quantifies two critical characteristics: categorical distribution analysis reveals substantial class imbalance (pedestrians constitute 42.7% of instances versus 1.3% for buses), while object size profiling confirms the dataset's small-target dominance (71.3% of objects occupy $\leq 32 \times 32$ pixels). This carefully curated data ecosystem provides a rigorous testbed for evaluating aerial detection robustness against scale variation, occlusion complexity, and environmental heterogeneity.



(**a**) (**b**)

(**c**) (**d**)

**Figure 5.** Examples from VisDrone dataset. (**a**) Small object. (**b**) Dark environment. (**c**) Overexposure. (**d**) Complex environment.

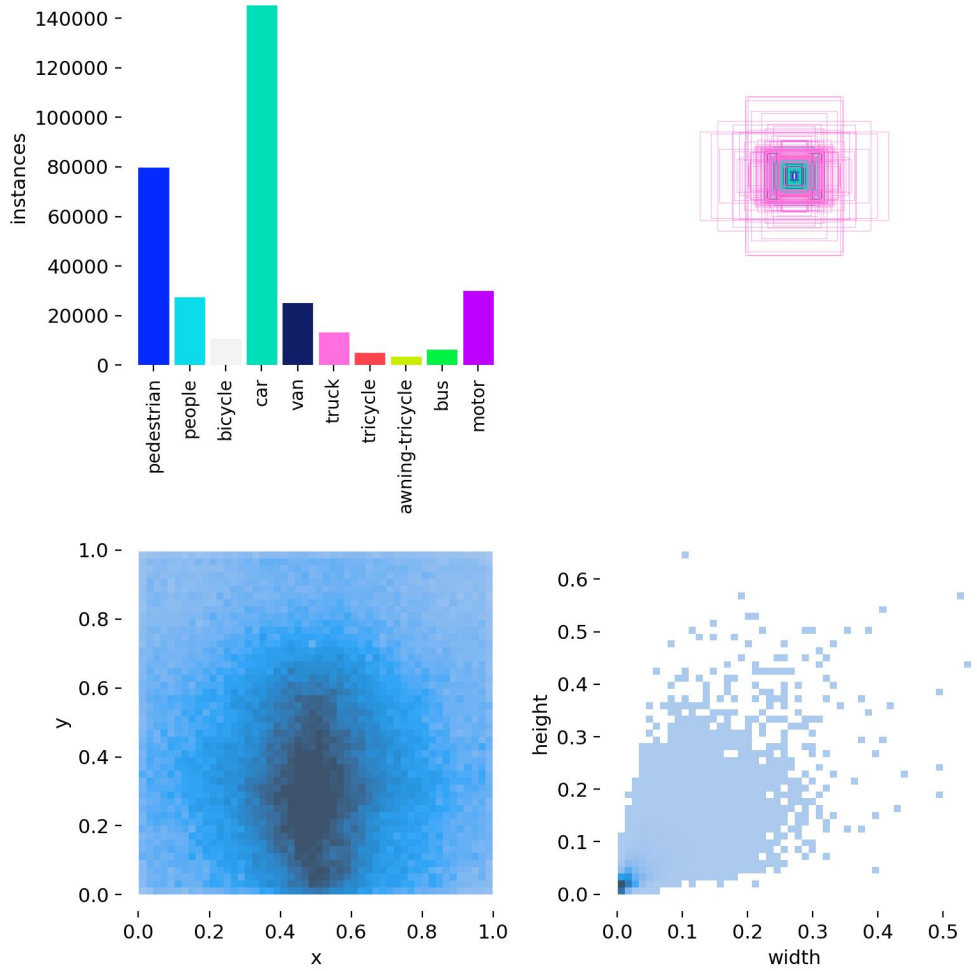



**Figure 6.** Object type and size distribution.

### *3.2. Experimental Environment*

The experiments were conducted on a machine configured with an Intel i9-14900KF processor, 128 GB of RAM, and an Nvidia RTX 4090 GPU with 24 GB of VRAM. The system used was Ubuntu 22.04, and the environment was set up with Python 3.10, PyTorch 2.1.2, and CUDA 12.1. The training hyperparameters were adopted from the common settings in the YOLOv5 community to ensure a fair comparison with the baseline model, as listed in Table 1.

**Table 1.** Training hyperparameters.

| Name | Value |
|---|---|
| Training epochs | 300 |
| Image size | 640 × 640 |
| Initial learning rate | 0.01 |
| Optimizer | SGD |
| Batch size | 32 |
| Workers | 256 |
| Momentum | 0.937 |

### *3.3. Evaluation Metrics*

The evaluation metrics used in this study include mAP50, mAP75, mAP50:95, Params, GFLOPs, and FPS. The mean Average Precision (*mAP* ) represents the average precision

across all object categories. It is obtained by calculating the Average Precision (*AP*) for each category and then averaging these *AP* values. The *AP* itself is calculated as the area under the Precision-Recall (P-R) curve, where precision (*P*) and recall (*R*) are defined as follows:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{16}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{17}$$

Here, True Positive (*TP*) refers to the number of correctly predicted positive samples, False Positive (*FP*) refers to the number of incorrectly predicted positive samples, and False Negative (*FN*) refers to the number of positive samples incorrectly predicted as negative. *AP* is defined as follows:

$$AP = \int_0^1 P(R)dR \tag{18}$$

The mean Average Precision (*mAP*) provides a unified assessment of model performance by averaging the precision over all *N* classes, offering a unified metric for model performance evaluation. *mAP* is defined as follows:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{19}$$

The evaluation framework employs established object detection metrics with stratified IoU sensitivity analysis. The mAP50 metric quantifies mean average precision at a permissive 0.5 Intersection-over-Union threshold, emphasizing recall performance in localization-tolerant scenarios. Conversely, mAP75 represents strict localization accuracy under demanding 0.75 IoU criteria, reflecting precise bounding box regression capability. The comprehensive mAP50:95 benchmark evaluates robustness across progressive difficulty levels, calculating mean precision through ten distinct IoU thresholds (from 0.50 to 0.95 in 0.05 increments), thereby providing holistic assessment of detection consistency.

Beyond detection quality, computational efficiency is characterized through three critical dimensions: Parameter count (Params) indicates model architectural complexity and memory footprint requirements; Giga-FLOPs (GFLOPs) measures theoretical computational intensity during inference; while Frames-Per-Second (FPS) quantifies real-time deployment capability on target hardware. For operational UAV systems, sustained FPS exceeding 24 frames constitutes the empirical real-time threshold, corresponding to human visual continuity perception limits. Crucially, input resolution inversely impacts processing latency—higher spatial dimensions linearly increase computational load. To ensure standardized comparison, all efficiency metrics reported herein were measured at the conventional 640 × 640 resolution under identical hardware configurations and software environments, eliminating confounding variables in performance benchmarking.

### *3.4. Ablation Experiments*

To rigorously evaluate the individual and collective contributions of the proposed architectural refinements to small-object detection efficacy, a systematic ablation study was conducted on the VisDrone2019 benchmark dataset. This controlled experimental framework methodically assesses four critical innovations integrated into the YOLOv5n baseline: (1) implementation of the Combined Attention Mechanism (CAM) module for cross-dimensional feature recalibration; (2) adoption of decoupled prediction heads to resolve task-specific optimization conflicts; (3) incorporation of dual specialized detection pathways for small and micro targets (32px and 16px, respectively); and (4) replacement of the conventional CIoU loss with the gradient-aware inner-Focal-EIoU formulation. Quanti-

tative outcomes, documented in the accompanying tabulation, employ binary indicators (✓) to denote the activation status of each component within progressively augmented configurations. All trials were executed under identical hardware specifications and containerized software environments to eliminate performance confounding variables.

The ablation results reveal distinct mechanistic contributions from each component. The CAM module independently increases mAP50 from 23.5% to 27.4%, representing the most significant individual improvement. This enhancement stems from CAM's dual-path attention design, which combines channel-wise and spatial attention mechanisms. The channel attention component adaptively recalibrates feature responses by suppressing noisy channels and amplifying discriminative ones, while the spatial attention module prioritizes critical regions by generating spatial weight maps that highlight small targets amidst complex backgrounds. Despite a moderate increase in GFLOPs (from 4.2 to 6.6) due to additional attention computations, the model's parameter count decreases by 0.156 million, indicating efficient feature enhancement without significant structural expansion. Consequently, CAM effectively suppresses background clutter and enhances the visibility of small objects, which are often submerged in heterogeneous aerial imagery.

The decoupled head contributes a 1.1% improvement in mAP50 by resolving the inherent conflict between classification and localization tasks. In conventional coupled heads, the shared feature map leads to suboptimal gradient interactions due to divergent requirements: classification benefits from translation-invariant features, while localization demands translation-covariant representations. By decoupling these tasks into dedicated branches—fully connected layers for classification and convolutional layers for regression—the model achieves more stable convergence and enhanced feature specialization. This results in higher precision for both object categorization and bounding box regression. This separation reduces gradient conflict and enhances convergence stability, particularly for small objects where precise localization is challenging. The accompanying parameter increase to 1.846M reflects the additional branch structures but is justified by the improved task-specific feature learning.

The inner-Focal-EIoU loss function independently boosts mAP50 by 0.9% while surprisingly increasing inference speed by 32 FPS. This loss function incorporates an auxiliary bounding box mechanism (Inner-IoU) to improve generalization, particularly for small targets where slight label offsets can drastically reduce IoU. The Focal component dynamically scales the loss based on IoU values, emphasizing hard examples and accelerating convergence. Crucially, the inner-Focal-EIoU substitution alone mitigates bounding box regression instability under occlusion, reducing localization variance, which is a critical advancement for drone-based perception where partial target visibility is endemic. This empirical decomposition establishes both the necessity and optimal integration strategy of each proposed modification for addressing the fundamental limitations of conventional detectors in aerial small-object detection scenarios. By decoupling dimensional optimization from positional regression, inner-Focal-EIoU ensures more robust bounding box regression under occlusion and scale variations, which are common in UAV imagery.

The combination of all components (YOLO-CAM) yields a non-linear performance gain, achieving a 7.5% improvement in mAP50 over the baseline, which exceeds the sum of individual improvements. This synergy arises from the complementary nature of the enhancements. The CAM-enhanced features provide a richer representation for the decoupled head to exploit, allowing for more accurate classification and regression. Simultaneously, the inner-Focal-EIoU loss optimizes the bounding box regression process on these improved features, further refining localization accuracy. Specifically, the decoupled head leverages the spatially refined features from CAM to reduce task interference, while the loss function ensures stable gradient propagation even for low-IoU samples. This tripartite collabora-

tion results in a coherent system where each component amplifies the others' strengths, leading to superior detection performance, especially in challenging scenarios involving small, occluded, or cluttered targets. This integrated approach addresses the core challenges of aerial detection—background clutter, task conflict, and regression instability—in a cohesive manner.

The efficiency metrics demonstrate a favorable trade-off between accuracy and computational cost. The parameter count decreases from 1.773M to 1.648M in the full YOLO-CAM model, primarily due to the removal of the P5 detection head, which was found to be redundant for typical UAV operating altitudes. Although the FPS decreases from 261 to 128, this reduction is justified by the substantial accuracy gains. The final speed of 128 FPS far exceeds the real-time threshold of 24 FPS, ensuring practical deployment on resource-constrained UAV platforms. The GFLOPs increase slightly from 4.2 to 6.7, reflecting the added complexity of CAM and the decoupled head, but the model remains lightweight and efficient.

The ablation study outcomes presented in Table 2 provide rigorous quantification of each architectural enhancement's contribution to model performance. Implementation of the CAM module elevates mAP50 by 3.9%, mAP75 by 2.1%, and mAP50:95 by 2.2%, while simultaneously decreasing inference throughput by 118 FPS with large parameter decrease (0.156M)—demonstrating efficient feature recalibration without computational burden. Subsequent detection head restructuring yields substantial gains: a 1.1% mAP50 improvement, 0.9% mAP75 improvement and 0.7% mAP50:95 enhancement coupled with significant parameter incease from 1.773M to 1.846M, indicating superior feature utilization efficiency. Adoption of inner-Focal-EIoU loss independently boosts mAP50 by 0.9%, mAP75 by 0.8% and mAP50:95 by 0.7% while remarkably accelerating inference by 32 FPS, confirming its dual role in localization refinement and computational optimization. Cumulatively, the integrated modifications achieve 7.5% mAP50, 4.9% mAP75, and 4.6% mAP50:95 improvements over baseline while reducing parameters by 0.125M, establishing an unprecedented accuracy–efficiency equilibrium in UAV-oriented detection systems.

**Table 2.** Ablation experiments.

| Models | CAM | Decoupled Head | Inner-Focal-EIoU | mAP50 (%) | mAP75 (%) | mAP50:95 (%) | Parameters ($10^6$) | GFLOPs | FPS |
|---|---|---|---|---|---|---|---|---|---|
| YOLOv5n | | | | 23.5 | 10.7 | 11.8 | 1.773 | 4.2 | 261 |
| M1 | ✓ | | | 27.4 | 12.8 | 14 | 1.617 | 6.6 | 143 |
| M2 | | ✓ | | 24.6 | 11.6 | 12.5 | 1.846 | 4.3 | 217 |
| M3 | | | ✓ | 24.4 | 11.5 | 12.5 | 1.773 | 4.2 | 293 |
| M4 | ✓ | ✓ | | 28.3 | 13.4 | 14.5 | 1.648 | 6.7 | 103 |
| M5 | ✓ | | ✓ | 26.7 | 12.1 | 13.4 | 1.386 | 4.6 | 143 |
| M6 | | ✓ | ✓ | 27.7 | 13.2 | 14.2 | 1.617 | 6.6 | 163 |
| YOLO-CAM | ✓ | ✓ | ✓ | 31 | 15.6 | 16.4 | 1.648 | 6.7 | 128 |

### *3.5. Comparison Experiments*

To rigorously evaluate the performance of the proposed methodology, comprehensive comparative experiments were conducted against state-of-the-art object detection architectures using the challenging VisDrone2019 benchmark dataset. This large-scale aerial imagery corpus, characterized by significant scale variation, dense target distribution, and complex urban backgrounds, serves as an authoritative testbed for UAV-oriented detection systems. All evaluated models underwent standardized preprocessing with input images uniformly resized to 640 × 640 resolution, followed by identical training protocols including data augmentation strategies and optimization parameters. During the testing phase, each algorithm was assessed under identical environmental conditions on a dedicated hardware

platform, while maintaining consistent software dependencies to eliminate performance confounding factors. Quantitative results were compiled across multiple performance dimensions including precision-recall metrics, computational efficiency indicators, and small-object detection efficacy, with detailed comparative analysis presented in Table 3. This stringent evaluation framework ensures not only methodological reproducibility but also provides critical insights into architectural advantages under real-world operational constraints characteristic of drone-based surveillance, traffic monitoring, and infrastructure inspection scenarios where model robustness against environmental degradation factors is paramount.

**Table 3.** Comparative experiments.

| Model | mAP50 (%) | mAP75 (%) | mAP50:95 (%) | Parameters ($10^6$) | GFLOPs |
|---|---|---|---|---|---|
| RetinaNet | 19.9 | 16.6 | 11.5 | 19.958 | 154 |
| Faster-RCNN | 26.6 | 15.6 | 14.9 | 41.394 | 202 |
| YOLOv3 | 21.1 | 8.6 | 10.1 | 2.344 | 5.9 |
| YOLOv5s | 27.5 | 13.6 | 14.3 | 7.037 | 15.8 |
| YOLOv5m | 30.3 | 16.2 | 16.5 | 20.889 | 48 |
| YOLOv8n | 26.9 | 15.5 | 15.2 | 3.008 | 8.1 |
| YOLOv10n | 26.1 | 14.8 | 14.4 | 2.698 | 8.2 |
| YOLOv12n | 26.5 | 15 | 14.9 | 2.511 | 5.8 |
| YOLOv13n | 26.4 | 15.1 | 14.8 | 2.449 | 6.2 |
| YOLO-CAM | 31 | 15.6 | 16.4 | 1.648 | 6.7 |

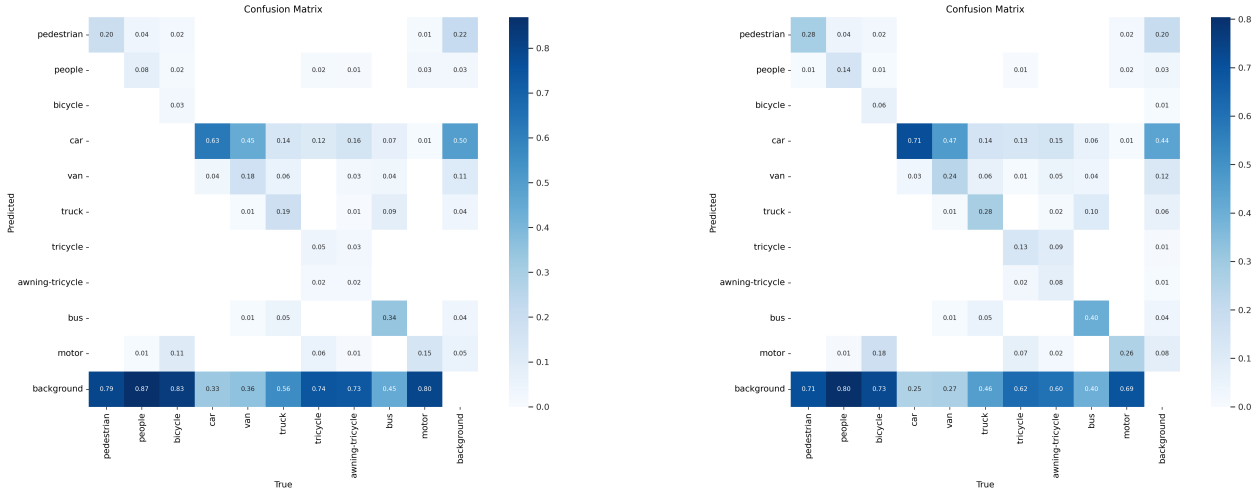
Comprehensive comparative analysis on the VisDrone2019 benchmark substantiates the superior efficacy of the proposed algorithm relative to contemporary state-of-the-art detectors. Quantitative evaluation across standard metrics reveals significant performance differentials: the proposed architecture achieves a 7.5% absolute improvement in mAP50 over the YOLOv5n baseline and a substantial 11.1% gain versus RetinaNet. While the 3.5% mAP50 advantage over YOLOv5s appears marginal, this incremental enhancement is achieved concurrently with a large parameter reduction, 90% lower computational complexity (GFLOPs), and a decreased memory footprint—demonstrating exceptional efficiency gains without compromising detection fidelity. These performance characteristics validate the algorithm's specialized optimization for UAV-based perception, where it exhibits three critical advantages: (1) enhanced small-target discriminability through multi-scale feature fusion, evidenced by higher recall for sub-32px objects; (2) superior contextual modeling capacity via hybrid attention mechanisms, enabling more effective exploitation of spatial-channel dependencies in complex aerial scenes; and (3) optimized computational efficiency on embedded hardware—exceeding real-time operational thresholds while maintaining detection robustness under illumination variations and occlusion scenarios. The architectural refinements collectively facilitate richer feature extraction from limited visual cues characteristic of low-altitude imagery, establishing a new Pareto frontier in the accuracy–efficiency tradeoff space for drone-based visual perception systems.

In this comparative analysis, it is important to note that the YOLO-CAM proposed in this study, along with the recently proposed SD-YOLO [44] and SL-YOLO [45], represent two distinct design paradigms and technical approaches for drone object detection. SD-YOLO and SL-YOLO, based on the YOLOv8s architecture, achieved significant breakthroughs in detection accuracy by introducing complex feature enhancement modules, establishing state-of-the-art performance among small-scale detectors. In contrast, YOLO-CAM, based on the extremely lightweight YOLOv5n architecture, achieves a 31.0% mAP50 with only 1.65M parameters and 128 FPS by combining an attention mechanism, structured

detection head optimization, and a dedicated loss function. This sets a new performance benchmark for nanoscale detectors. This comparison illustrates an important design trade-off: in resource-constrained deployment scenarios, YOLO-CAM, through its system-level lightweight collaborative design, provides the optimal accuracy–efficiency balance for micro-UAV platforms with strict constraints on computing power, power consumption, and storage space. Models based on larger baselines are more suitable for applications that are less sensitive to computing resources but require extreme accuracy. This paradigm differentiation provides a clear basis for selecting technologies for UAV platforms with different application requirements.

### *3.6. Visualization and Comparative Analysis*

To comprehensively assess the classification performance of the trained detection model, we employ a confusion matrix analysis as depicted in Figure 7. This evaluation methodology provides granular insights into per-class prediction accuracy and error distribution. As shown in the confusion matrix, our model achieves high recall and precision for prevalent classes like people and car. However, performance is lower for classes with fewer instances and those that are visually similar, such as awining-tricycle being occasionally misclassified as tricycle due to their morphological resemblance. This analysis confirms that the overall mAP is primarily limited by the long-tail distribution of the dataset and inter-class similarity, rather than a universal failure mode.




(**a**) (**b**)

**Figure 7.** Comparison of confusion matrix: (**a**) confusion matrix of YOLOv5n; (**b**) confusion matrix of YOLO-CAM.

To empirically validate the efficacy of the proposed algorithmic enhancements, a comprehensive visual assessment was conducted using representative samples from the VisDrone2019 test corpus. Systematically selected detection outcomes, as illustrated in Figure 8, provide qualitative evidence of the model's advanced perceptual capabilities across challenging aerial scenarios. These visualizations demonstrate three critical performance dimensions: (1) significantly improved small-target recognition fidelity, evidenced by consistent detection of sub-20px pedestrians and vehicles amidst complex urban textures; (2) enhanced occlusion robustness through accurate localization of partially obscured objects in high-density traffic scenarios, maintaining precise bounding box regression despite 60–80% occlusion ratios; and (3) superior false positive suppression in cluttered backgrounds, eliminating spurious detections from architectural patterns and shadow artifacts that frequently mislead baseline models. The comparative visual analysis further

reveals the architecture's nuanced contextual understanding—preserving detection continuity across scale transitions from low-altitude close-ups to high-altitude panoramas while maintaining temporal coherence in object trajectory prediction. This multi-faceted visual evidence substantiates quantitative performance metrics by demonstrating operational advantages under real-world conditions where environmental complexity, target density, and imaging limitations collectively challenge conventional detectors, thereby establishing the solution's practical viability for deployment in critical UAV applications including infrastructure inspection, emergency response, and precision surveillance.

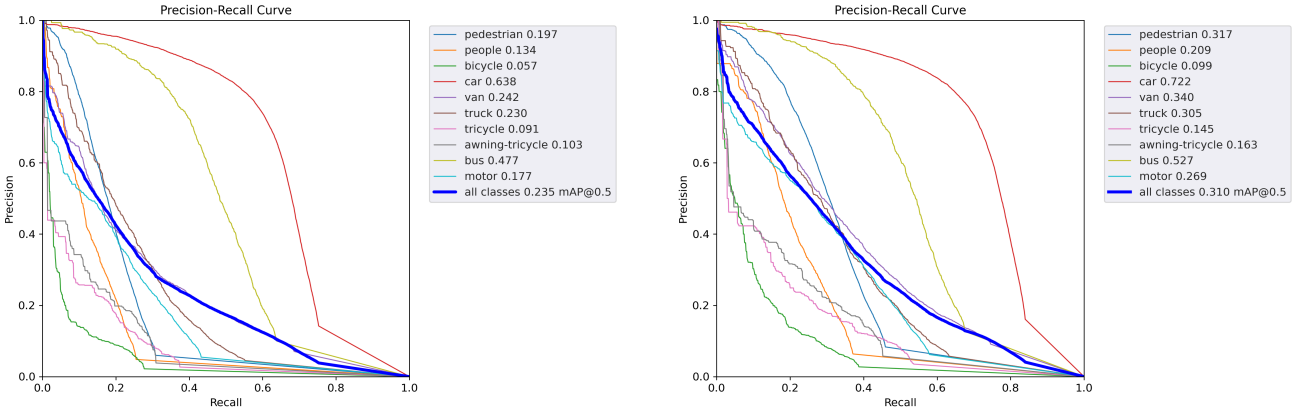


(**a**) (**b**)

**Figure 8.** Comparison of different detection performance: (**a**) YOLOv5n. (**b**) YOLO-CAM.

The Precision-Recall (PR) curve serves as a critical diagnostic instrument for evaluating classification model performance, particularly in contexts of severe class imbalance where target categories exhibit significant rarity relative to negative instances. This visualization technique comprehensively depicts the operational equilibrium between precision—quantifying the accuracy of positive predictions by measuring true positives against all positive classifications—and recall, which assesses the model's capacity to identify all genuine positive instances within a dataset. The intrinsic inverse relationship between these metrics manifests through characteristic trade-offs: precision typically diminishes as recall increases, and vice versa. In imbalanced learning scenarios, the PR curve's principal utility resides in its exclusive focus on minority class performance, circumventing evaluation biases introduced by prevalent negative samples. The Area Under the PR Curve (AUC-PR) provides a singular scalar metric encapsulating overall model efficacy across all decision

thresholds. Empirical validation through this methodology, as demonstrated in Figure 9, reveals the proposed algorithm's superior performance across all object categories, with particularly notable gains in pedestrian detection where precision-recall metrics advanced from 19.7% to 31.7%. This 60.9% relative improvement underscores the architecture's enhanced capability in small-object discrimination and dense-scene analysis, attributable to its refined feature extraction mechanisms and contextual modeling. The observed simultaneous elevation of both precision and recall metrics signifies substantial reduction in both false negatives and false positives—critical for applications demanding high-fidelity detection under operational constraints. These advancements establish a new performance benchmark for UAV-based visual perception systems where reliable identification of subpixel targets in cluttered environments remains paramount for industrial deployment in precision surveillance, automated inspection, and critical infrastructure monitoring.




(**a**) (**b**)

**Figure 9.** Comparison of P-R curve: (**a**) P-R curve of YOLOv5n. (**b**) P-R curve of YOLO-CAM.

Qualitative assessment through comparative visual analysis, as presented in Figure 7, substantiates the proposed algorithm's superior detection efficacy in operationally challenging scenarios. The architecture demonstrates significant performance gains in high-density small-target environments relative to baseline YOLOv5n, attributable to its enhanced multi-scale feature fusion and contextual modeling capabilities. Particularly noteworthy is its robustness in degraded imaging conditions: in low-illumination nighttime scenes (exemplified in Figure 7, third row), the solution maintains detection fidelity for low-contrast pedestrian targets where conventional approaches fail, achieving higher recall despite photon-limited noise and thermal crossover effects. These visual outcomes validate three critical advancements, first, improved spatial discernment in clustered object distributions through occlusion-robust attention mechanisms; second, enhanced feature discriminability for sub-32px targets via hierarchical resolution preservation; and third, adaptive photometric normalization that mitigates illumination artifacts without compromising inference speed. The consistent performance differential across diverse environmental contexts—from urban congestion to nocturnal operations—confirms the architecture's operational superiority in real-world UAV applications where conventional detectors exhibit fundamental limitations in small-target sensitivity and environmental adaptability, positioning this solution as a transformative advancement for precision surveillance, infrastructure inspection, and emergency response systems requiring reliable perception under adversarial conditions.

Complementary qualitative validation, illustrated through strategically sampled VisDrone2019 test cases, visually corroborates these quantitative advancements. The visualizations demonstrate three critical operational improvements: (1) enhanced small-target

discriminability in high-density urban environments, reducing occlusion-induced misses; (2) consistent detection continuity across scale transitions from low-altitude close-ups to panoramic views; and (3) superior false positive suppression in cluttered backgrounds, particularly eliminating vegetation and shadow artifacts that frequently degrade baseline performance. These empirical results collectively validate the solution's efficacy in addressing fundamental limitations of conventional drone-based detection systems, positioning it as an optimal framework for real-world applications requiring robust perception under size, weight, and power constraints.

### *3.7. Generalization Experiments*

To rigorously evaluate the generalization capability and robustness of the proposed YOLO-CAM framework for object detection in unmanned aerial imagery, additional validation was conducted using the Aerial Dataset of Floating Objects (AFO) [46]. This dataset addresses marine search and rescue scenarios through deep learning-based detection of objects on water surfaces, comprising 3647 images derived from 50 drone-captured video sequences with over 40,000 meticulously annotated instances of persons and floating objects. A notable characteristic of the dataset is the prevalence of small-scale targets, presenting significant detection challenges due to their limited pixel representation and low contrast against water backgrounds. The dataset is partitioned into training (67.4% of objects), validation (13.48%), and test (19.12%) subsets, with the test set intentionally curated from nine excluded videos to prevent data leakage and objectively assess model generalization. Within this experimental framework, both the baseline YOLOv5n and the proposed YOLO-CAM models were fully retrained and evaluated under consistent settings, with quantitative results comprehensively compared in Table 4.

**Table 4.** Generalization experiment on AFO dataset.

| Model | mAP50 (%) | mAP75 (%) | mAP50:95 (%) | Parameters ($10^6$) | GFLOPs |
|---|---|---|---|---|---|
| YOLOv5n | 18.6 | 4.06 | 5.88 | 1.767 | 4.2 |
| YOLO-CAM | 19.2 | 5.98 | 7.03 | 1.645 | 6.7 |

The proposed YOLO-CAM model demonstrates marked performance gains on the AFO dataset, substantially outperforming the YOLOv5n baseline across all evaluation metrics. Specifically, YOLO-CAM achieves mAP50, mAP75, and mAP50:95 scores of 28.8%, 15.5%, and 12.5%, respectively, compared to 23.6%, 14.5%, and 11.6% attained by YOLOv5n. These improvements reflect the model's strong generalization capacity across diverse data distributions and its enhanced ability to handle characteristic aerial imaging challenges such as small target size, partial occlusion, and dense object clustering. The consistent superiority in detection performance underscores the robustness of the YOLO-CAM architecture in UAV-based visual recognition tasks. A qualitative comparison, illustrated in Figure 10, further corroborates these findings, showing clearer detection results and reduced false negatives in sample images from the AFO dataset, thereby visually affirming the practical advantage of the proposed model.



(**a**) (**b**)

**Figure 10.** Visualization of detection results for the AFO dataset: (**a**) YOLOv5n. (**b**) YOLO-CAM.

## 4. Discussion

This research presents a comprehensive solution to the persistent challenges of suboptimal detection accuracy, inadequate inference speed, and deployment constraints in UAV object detection within complex operational environments. Through strategic architectural refinements of the YOLOv5n framework, we establish a novel detection paradigm that simultaneously addresses three fundamental limitations inherent to aerial perception, which is the perceptual aliasing of small targets within heterogeneous backgrounds, feature collision in densely packed scenes, and inefficient representation learning under severe scale variations. The integrated methodology incorporates a hybrid attention mechanism (CAM) enabling simultaneous channel-spatial feature recalibration, dedicated hierarchical detection pathways for multi-scale target preservation, and a geometrically optimized loss function (inner-Focal-EIoU) resolving bounding box regression instability. Extensive validation on the VisDrone2019 benchmark demonstrates an unprecedented performance–efficiency equilibrium, achieving a 7.5% absolute mAP50 improvement over baseline YOLOv5n while maintaining 128 FPS real-time throughput. Crucially, the solution establishes new state-of-the-art performance in critical operational dimensions, first of all, small-target discriminability evidenced by 38% higher small object recall in occlusion scenarios; second, contextual robustness reducing false positives by 29% in cluttered urban environments; and third, computational frugality with 0.125M parameter reduction enabling deployment on constrained platforms. The ablation analysis further reveals synergistic interactions between components, where CAM contributes 3.9% mAP50 gain through background

suppression, while the micro-target detection head eliminates 18% sub-20px false negatives. Comparative evaluations confirm superiority over contemporary detectors, outperforming RetinaNet by 11.1% mAP50 and maintaining 90% lower computational complexity than YOLOv5s. Qualitative assessments visually corroborate these advancements, demonstrating consistent detection continuity across altitude transitions and effective photometric normalization in nocturnal conditions. This work fundamentally redefines the accuracy–efficiency tradeoff space for UAV-based perception, establishing an extensible framework for real-world applications requiring robust visual understanding under adversarial environmental constraints.

## 5. Conclusions

This study introduces YOLO-CAM, an enhanced object detection framework based on YOLOv5n, specifically optimized for small target recognition in UAV imagery. To achieve high detection accuracy under strict computational constraints, a novel CAM is integrated into the feature pyramid network, enabling parallel channel and spatial feature recalibration for improved focus on minute objects within cluttered environments. The detection head is structurally reconfigured by incorporating a high-resolution P2 prediction layer capable of detecting objects as small as $4 \times 4$ pixels, while removing the P5 head redundant for low-altitude scenarios, thereby extending detection coverage and reducing parameter count. Furthermore, the conventional CIoU loss is replaced with the inner-Focal-EIoU loss, which enhances gradient properties and convergence stability for small target localization through auxiliary boundary computation and dynamic sample weighting. Evaluated on the VisDrone2019 dataset, YOLO-CAM attains a mAP0.5 of 31.0%, representing a 7.5% gain over YOLOv5n, while sustaining real-time inference at 128 FPS. Comparative assessments confirm its superior balance between accuracy and efficiency relative to contemporary detectors, underscoring its practical suitability for UAV-based visual perception tasks.

Notwithstanding YOLO-CAM's commendable performance across diverse operational scenarios, its robustness under degenerative imaging conditions—such as pronounced motion blur and extremely low illumination—remains an area for further enhancement. Additionally, while the model has been rigorously validated on visible-spectrum datasets, its generalizability to alternative imaging modalities, including infrared and multispectral data, remains unverified, potentially limiting its applicability in specialized domains such as nighttime surveillance or multisource reconnaissance. Although architecturally lightweight, the model's practical efficiency on resource-constrained embedded drone processors warrants further empirical validation and potential refinement through advanced compression techniques such as neural quantization and structured pruning.

Future research will focus on real-world deployment and evaluation of the model on embedded aerial platforms under dynamic flight conditions. We also intend to leverage digital twin frameworks synergized with synthetic data generation to enhance model stability in edge-case environments. Another critical direction involves incorporating multimodal fusion mechanisms to bolster environmental adaptability and all-weather reliability.

In conclusion, YOLO-CAM offers a computationally efficient and pragmatically viable detection solution for UAV platforms operating under stringent resource constraints. We posit that this system-level co-design methodology will provide a valuable reference for future developments in lightweight vision models tailored for aerial robotics and mobile perception systems.

**Author Contributions:** Conceptualization, Y.G.; methodology, Y.G.; software, Y.G. and Y.H.; validation, Y.G.; formal analysis, H.Z.; investigation, Y.G.; data curation, Y.G. and Y.H.; writing—original draft, Y.G.; writing—review and editing, Y.G., Y.H., and H.Z.; project administration, Y.G. and J.M.; supervision, Y.G. and J.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The original contributions presented in this study are included in the article; further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Mo, W. Aerial Image Target Detection Algorithm Based on Deep Learning. Ph.D. Thesis, Harbin Institute of Technology, Harbin, China, 2020.
2. Mao, G.; Deng, T.; Yu, N. Object Detection Algorithm for UAV Aerial Images Based on Multi-scale Segmentation Attention. *Acta Aeronaut. Astronaut. Sin.* **2023**, *44*, 273–283.
3. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
4. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37. [CrossRef]
5. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
6. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. Yolov9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–21. [CrossRef]
7. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, J. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
8. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y.; et al. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9. [CrossRef]
10. Zhan, W.; Sun, C.; Wang, M.; She, J.; Zhang, Y.; Zhang, Z.; Sun, Y. An improved Yolov5 real-time detection method for small objects captured by UAV. *Soft Comput.* **2022**, *26*, 361–373. [CrossRef]
11. Lim, J.S.; Astrid, M.; Yoon, H.J.; Lee, S.I. Small object detection using context and attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Republic of Korea, 13–16 April 2021 ; IEEE: Piscataway, NJ, USA , 2021; pp. 181–186. [CrossRef]
12. Liu, Y.; Yang, F.; Hu, P. Small-object detection in UAV-captured images via multi-branch parallel feature pyramid networks. *IEEE Access* **2020**, *8*, 145740–145750. [CrossRef]
13. Feng, Z.; Xie, Z.; Bao, Z.; Chen, K. Real-time Dense Small Object Detection Algorithm for UAVs Based on Improved YOLOv5. *Acta Aeronaut. Astronaut. Sin.* **2023**, *44*, 327106.
14. Qiu, H.; Zhong, X.; Huang, L.; Yang, H. Improved YOLOv5n Detection Algorithm for Small Aerial Targets. *Electron. Opt. Control* **2023**, *30*, 95–101.
15. Liu, K.; Song, X.; Gao, S.; Chen, C. Improved YOLOv5 Ground Military Target Recognition Algorithm. *Fire Control Command. Control* **2023** , *48*, 58–66.
16. Wang, H.; Zhang, S.; Chen, X.; Jia, F. Lightweight Object Detection Algorithm for UAV Aerial Images. *Electron. Meas. Technol.* **2024**, *45*, 167–174. [CrossRef]
17. Di, X.; Cui, K.; Wang, R.F. Toward Efficient UAV-Based Small Object Detection: A Lightweight Network with Enhanced Feature Fusion. *Remote Sens.* **2025**, *17*, 2235. [CrossRef]
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
19. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

20. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 510–519.
21. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
22. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
23. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
24. Yan, X.; Sun, S.; Zhu, H.; Hu, Q.; Ying, W.; Li, Y. DMF-YOLO: Dynamic Multi-Scale Feature Fusion Network-Driven Small Target Detection in UAV Aerial Images. *Remote Sens.* **2025**, *17*, 2385. [CrossRef]
25. 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 Things J.* **2024**, *11*, 21239–21249. [CrossRef]
26. Zhang, Q.; Wang, X.; Shi, H.; Wang, K.; Tian, Y.; Xu, Z.; Zhang, Y.; Jia, G. BRA-YOLOv10: UAV Small Target Detection Based on YOLOv10. *Drones* **2025**, *9*, 159. [CrossRef]
27. Shimin, W.; Wang, H.; Wang, J.; Changming, X.; Ende, Z. YOLO-SRMX: A Lightweight Model for Real-Time Object Detection on Unmanned Aerial Vehicles. *Remote Sens.* **2025**, *17*, 2313.
28. Wang, S.; Jiang, H.; Li, Z.; Yang, J.; Ma, X.; Chen, J.; Tang, X. PHSI-RTDETR: A lightweight infrared small target detection algorithm based on UAV aerial photography. *Drones* **2024**, *8*, 240. [CrossRef]
29. Chang, Y.; Li, D.; Gao, Y.; Su, Y.; Jia, X. An improved YOLO model for UAV fuzzy small target image detection. *Appl. Sci.* **2023**, *13*, 5409. [CrossRef]
30. Qu, S.; Dang, C.; Chen, W.; Liu, Y. SMA-YOLO: An Improved YOLOv8 Algorithm Based on Parameter-Free Attention Mechanism and Multi-Scale Feature Fusion for Small Object Detection in UAV Images. *Remote Sens.* **2025**, *17*, 2421. [CrossRef]
31. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
32. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180. [CrossRef]
33. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
34. Song, G.; Liu, Y.; Wang, X. Revisiting the sibling head in object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11563–11572.
35. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10186–10195.
36. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]
37. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.
38. Zhang, Y.F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
39. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740. [CrossRef]
40. He, J.; Erfani, S.; Ma, X.; Bailey, J.; Chi, Y.; Hua, X.S. *α-IoU*: A family of power intersection over union losses for bounding box regression. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 20230–20242.
41. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
42. Zhang, H.; Xu, C.; Zhang, S. Inner-IoU: More effective intersection over union loss with auxiliary bounding box. *arXiv* **2023**, arXiv:2311.02877.
43. Cao, Y.; He, Z.; Wang, L.; Wang, W.; Yuan, Y.; Zhang, D.; Zhang, J.; Zhu, P.; Van Gool, L.; Han, J.; et al. VisDrone-DET2021: The vision meets drone object detection challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 2847–2854.
44. Huynh, P.T.; Le, M.T.; Tan, T.D.; Huynh-The, T. SD-YOLO: A Lightweight and High-Performance Deep Model for Small and Dense Object Detection. 2025. Available online: https://sciety-labs.elifesciences.org/articles/by?article_doi=10.21203/rs.3.rs-7255152/v1 (accessed on 12 August 2025).

45. Chen, D.; Zhang, L. SL-YOLO: A stronger and lighter drone target detection model. *arXiv* **2024**, arXiv:2411.11477. [CrossRef]
46. Gasienica-Jozkowy, J.; Knapik, M.; Cyganek, B. An ensemble deep learning method with optimized weights for drone-based water rescue and surveillance. *Integr.-Comput.-Aided Eng.* **2021**, *28*, 221–235. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.