DCA-YOLO: A Dynamic Cross-Attention Network for Small Vehicle Detection in UAV Aerial Imagery

Chaojun Dong
Wuyi University, Jiangmen, Guangdong, China

Yizhi Wang
Wuyi University, Jiangmen, Guangdong, China

Yikui Zhai, Senior Member, IEEE
Wuyi University, Jiangmen, Guangdong, China

Ye Li
Wuyi University, Jiangmen, Guangdong, China

Xiankun Liu, Member, IEEE
Wuyi University, Jiangmen, Guangdong, China

Chaoyun Mai
Wuyi University, Jiangmen, Guangdong, China

Jianhong Zhou, Member, IEEE
South China University of Technology, Guangzhou, Guangdong, China
School of Computer Science and Engineering, Guangzhou, Guangdong, China

Pasquale Coscia, Senior Member, IEEE
Universita Degli Studi di Milano, Milan, Italy
Department of Computer Science and the Dipartimento di Informatica

Angelo Genovese, Senior Member, IEEE
Universita Degli Studi di Milano, Milan, Italy
Department of Computer Science and the Dipartimento di Informatica

C. L. Philip Chen, Fellow, IEEE
South China University of Technology, Guangzhou, Guangdong, China
School of Computer Science and Engineering, Guangzhou, Guangdong, China

***Abstract*— To address challenges in UAV aerial imagery, such as complex backgrounds, drastic scale variations, and partial occlusions. We propose DCA-YOLO, a high-performance vehicle detection framework featuring a newly designed neck architecture named the dynamic cross-attention path aggregation network (DCA-PAN). Our approach introduces three synergistic modules in the neck to enhance feature expressiveness for small vehicle detection: dynamic weighted adaptive feature fusion, cross-layer dynamic interaction, and coordinate-aware dynamic attention. These modules synergistically improve multi-scale representation, seamlessly integrate shallow details with deep semantics, and dynamically emphasize key object features. To this end, we leverage the efficient and powerful FasterNet as the backbone to achieve an effective balance between computational overhead and detection accuracy. Experiments on the VisDrone and UAV-HS datasets demonstrate that DCA-YOLO achieves an optimal accuracy-efficiency tradeoff among compact detectors. On the challenging VisDrone dataset, DCA-YOLO significantly outperforms the baseline YOLOv8n, boosting mAP_{50} from 57.2% to 60.9%, $\text{mAP}_{50:95}$ from 38.5% to 41.6%, and mAP_s from 14.3% to 16.1%. Our code and dataset are available at https://github.com/yikuizhai/DCA-YOLO.**

***Index Terms*— real-time vehicle detection, UAV, adaptive fusion, multi-scale feature fusion.**

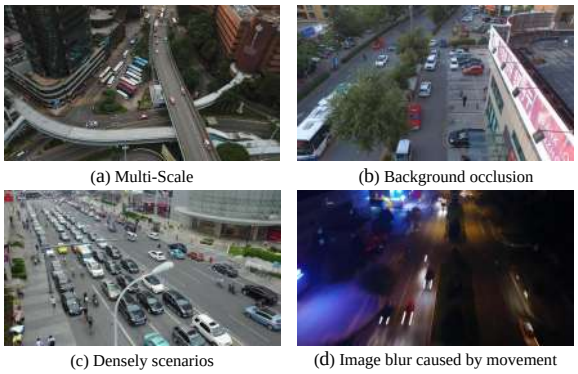
I. INTRODUCTION

WITH the increasing demand for fine-grained visual perception in low-altitude scenarios, unmanned aerial vehicles have become indispensable tools in remote sensing due to their flexible deployment, low cost, and capacity for capturing high resolution images from diverse aerial perspectives [1], [2]. These features have enabled UAVs to play a crucial role in a wide range of civilian applications, such as land use mapping, environmental surveillance, urban infrastructure inspection, and wildlife monitoring [3], [4]. Among these applications, UAV-based vehicle detection has garnered extensive attention as it supports efficient monitoring over large areas without reliance on fixed ground infrastructure. However, UAV imagery poses significant challenges for vehicle detection, such as extreme scale variation, dense object distribution, complex backgrounds, and motion blur, as shown in Fig. 1. The inherent drastic scale variation in aerial perspectives, in particular, requires more adaptive and robust detection frameworks [5].

To address the aforementioned challenges, a wide range of deep learning based object detection frameworks have been developed. Two-stage detectors such as Faster R-CNN [6] offer strong localization accuracy but often suffer from high computational complexity, making

Manuscript received XXXXX 00, 0000; revised XXXXX 00, 0000; accepted XXXXX 00, 0000.
This study was funded by the National Natural Science Foundation of China (No. 62273109), Guangdong Province Key Disciplines Scientific Research Capacity Enhancement Project (No. 2024ZDJS034), Guangdong Higher Education Innovation and Strengthening School Project (No. 2022ZDZX1032, No. 2023ZDZX1029), and Wuyi University Hong Kong and Macao Joint Research and Development Fund (No. 2022WGALH19). This work was supported by the High Performance Computing Center of Wuyi University. *(Corresponding author: Chaoyun Mai.)*(Chaojun Dong, Yizhi Wang, Yikui Zhai, Ye Li, Xiankun Liu , Chaoyun Mai and Jianhong Zhou contributed equally to this work.)

Chaojun Dong, Yizhi Wang, Yikui Zhai, Ye Li and Chaoyun Mai are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, China, E-mail: (cjun-dong@163.com; wangyizhi1998@gmail.com; yikuizhai@163.com; leyelee@163.com; maichaoyun@foxmail.com). Xiankun Liu is with the School of Mechanical and Automation Engineering, Wuyi University, Jiangmen 529020, China, E-mail: (lxklml@163.com). C. L. Philip Chen and Jianhong Zhou are with the School of Electronics and Information Engineering, South China University of Technology, Guangzhou 510641, China, E-mail: (jackychou_lab@126.com). Pasquale Coscia and Angelo Genovese are with the Department of Computer Science, Universita Degli Studi di Milano, Milan, Italy, E-mail: (pasquale.coscia@unimi.it; angelo.genovese@unimi.it). C. L. Philip Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China, E-mail: (philipchen@scut.edu.cn).



(a) Multi-Scale (b) Background occlusion

(c) Densely scenarios (d) Image blur caused by movement

Fig. 1. Illustrative examples of challenging scenarios for UAV-based vehicle detection.

them unsuitable for real-time UAV applications. One-stage models like YOLO and RetinaNet [7] achieve better efficiency and are more suitable for onboard deployment [8]. In addition, recent literature has reviewed state-of-the-art detection approaches that provide valuable insights into the evolution of object detection frameworks [9]. However, these general-purpose detectors are typically designed for natural scene datasets and struggle to maintain performance in UAV imagery, where objects often appear small, densely distributed, and vary drastically in scale. Moreover, conventional feature fusion strategies in these models lack the adaptability needed to capture subtle spatial details in aerial views, limiting their effectiveness in detecting small and occluded vehicles. To enhance detection of small-scale objects, many methods employ multi-scale feature fusion [10], anchor-free architectures [11], and attention mechanisms [12], typically integrated into the backbone or neck of the network. Among these efforts, lightweight backbones such as MobileNet [13], ShuffleNet [14], and GhostNet [15] have gained popularity for their reduced computational complexity. Nevertheless, such architectures may sacrifice semantic richness, which is crucial for accurately detecting small and densely packed aerial objects. To compensate for this, feature fusion strategies like feature pyramid network (FPN) [16], path aggregation network (PAN) [10], and BiFPN [17] have been introduced to enhance hierarchical feature learning. Yet, when applied to UAV data, these methods can encounter issues such as redundant features and scale misalignment due to complex aerial backgrounds. Additionally, attention mechanisms including channel-based modules like SE [12] and ECA [18], spatial mechanisms like LSKnet [19], and transformer-based approaches [20] have shown effectiveness in remote sensing detection tasks. However, many of these designs are computationally expensive or insufficiently adaptive to UAV-specific challenges such as diverse object scales and low object-to-background contrast. These limitations collectively highlight the need for a unified framework that integrates a powerful yet efficient architecture with robust multi-scale feature fusion and UAV-oriented attention refinement to achieve high-performance vehicle detection in aerial scenarios.

To address the challenges of UAV-based vehicle detection, such as scale variation, occlusion, motion blur, and complex backgrounds, we propose dynamic cross-attention YOLO (DCA-YOLO), an enhanced adaptive detection framework based on YOLOv8. This framework improves detection performance in aerial scenarios by achieving an optimal balance of accuracy, inference speed, and model complexity. Our primary innovation is the dynamic cross-attention path aggregation network (DCA-PAN), a specialized neck structure designed to enhance multi-scale feature aggregation. At its core, DCA-PAN integrates three complementary modules: the dynamic weighted adaptive feature fusion (DWAF) module, which adaptively reweights multi-level features to effectively address object scale variations; the cross-layer dynamic interaction (CDI) module, which promotes semantic consistency to improve context awareness and mitigate occlusion; and the coordinate-aware dynamic attention (CoDA) module, which incorporates position-sensitive priors to suppress background interference. These modules are jointly optimized to improve semantic-spatial alignment and strengthen the representation of small objects. Furthermore, an early-stage detection head is introduced to specifically enhance the detection of extremely small vehicles. To complement this advanced neck design, we adopt the lightweight. FasterNet [21] as the backbone, replacing the original from YOLOv8. This choice reduces the overall parameter count and computational cost, as FasterNet offers a superior balance of efficiency and representation capacity ideal for resource-constrained UAV-ground collaborative environments.

The main contributions of this article can be summarized as follows.

1) A detection framework, DCA-YOLO, is proposed, in which the DCA-PAN is integrated to enhance multi-scale feature fusion and spatial-semantic interaction. This design is complemented by the replacement of the original YOLOv8 backbone with FasterNet, which enables a more efficient and powerful feature extraction pipeline and significantly boosts vehicle detection performance in UAV imagery.
2) The DCA-PAN is designed to enhance multi-scale feature fusion and spatial-semantic interaction. Through the support of the DWAF module for adaptive multi-level feature re-weighting and the CDI module for improved semantic consistency, the representation of small, dense, and partially occluded vehicles in complex UAV scenes is effectively strengthened.
3) The CoDA module is proposed as an advanced fusion unit, designed to mitigate feature conflicts and suppress noise when integrating fine-grained details with deep semantic cues. This significantly

improves the model's focus and robustness in detecting dense or occluded targets.

4) We conducted extensive experiments on the VisDrone [22]and UAV-HS datasets, achieving up to 60.9% mAP_{50} on VisDrone and similar gains on the others, demonstrating that our method consistently outperforms representative compact and lightweight detectors.

The remainder of this paper is organized as follows: Section II reviews related work. Section III details our proposed DCA-YOLO framework and its core modules. Section IV presents the experimental evaluation of our method, Section VI provides a discussion of failure case categorization, with particular emphasis on frequently observed missed detections. Finally, Section VII concludes the paper and outlines directions for future work.

II. RELATED WORK

In this section, we first discuss UAV-based vehicle detection algorithms under a top-down perspective, followed by a brief overview of multi-scale feature fusion methods. Finally review various attention mechanisms applicable to vehicle detection.

A. UAV Vehicle Detection Method

Over the past decade, vehicle detection has transitioned from handcrafted feature-based methods, including Viola-Jones [23], [24], HOG+SVM [25], and DPM [26], to advanced deep learning paradigms. While foundational, early approaches lacked the semantic depth to handle scale variations and occlusions. Convolutional neural networks significantly advanced performance; however, two-stage detectors in the R-CNN series [6], [27], [28] remain computationally expensive and suboptimal for small targets in UAV scenarios. Conversely, one-stage detectors, such as SSD [29] and YOLO [30], offer real-time speeds but struggle with small or occluded vehicles due to insufficient context modeling. Although anchor-free architectures, including FCOS [31], CenterNet [11], and YOLOX [32], improve generalization by eliminating predefined anchors, they still demand robust context reasoning for dense UAV imagery.

More recently, transformer-based frameworks like DETR [33] have leveraged global attention to remove hand-crafted components, yet their high computational overhead and reduced sensitivity to small objects hinder real-time applications. Addressing these persistent challenges, we propose DCA-YOLO. By integrating a lightweight FasterNet [21] backbone with modular enhancements tailored for small and occluded objects, our method achieves an optimal balance among speed, accuracy, and model complexity for UAV-based vehicle detection.

B. Multi-scale Features Fusion For Vehicle Detection

In practical UAV scenarios, vehicle detection is severely challenged by extreme scale variations, demanding the effective integration of high-resolution spatial details from shallow layers and rich semantics from deep layers. To bridge this feature gap, various multi-scale fusion frameworks have been developed. Early architectures like FPN [16] and PANet [10] established foundational top-down and bidirectional pathways, while subsequent models such as NAS-FPN [34] and BiFPN [17] optimized fusion topologies through neural architecture search and adaptive learnable weights to balance efficiency and accuracy.

Beyond these general frameworks, task-specific designs have further refined multi-scale representations. For instance, AFPN [35] mitigates semantic gaps via progressive fusion, EMIFF [36] enhances autonomous driving perception using cross-attention, and MFO-Net [37] improves small vehicle detection in aerial imagery through optimized local context modeling. Similarly, ASF-YOLO [38] utilizes adaptive spatial fusion, though its original medical imaging focus limits its robustness against the extreme scale variations and complex backgrounds typical of UAV imagery. To explicitly address these challenges of semantic inconsistency and spatial degradation, we propose the DCA-PAN neck architecture. By integrating Dynamic Weighted Adaptive Fusion and Cross-Layer Dynamic Interaction modules, our approach enables robust bidirectional adaptive aggregation and strengthens cross-level feature interactions, specifically optimizing multi-scale representation for UAV-based detection.

C. Attention Mechanism

Attention mechanisms have revolutionized vehicle detection by enhancing task-relevant features while suppressing background noise [39]. Early approaches primarily focused on inter-channel dependencies; for instance, SENet [12] explicitly models channel-wise correlations to recalibrate feature responses, though its global pooling often sacrifices spatial context—a limitation later addressed by SPANet [40]. To capture multi-scale information, mechanisms like SKNet [41] and LSKNet [19] introduced selective kernel convolutions. By adaptively fusing features from varying receptive fields, these methods extend attention to the spatial dimension, enabling the network to dynamically adjust its focus based on object scale and spatial dependencies.

To achieve a more comprehensive feature representation, hybrid attention models such as scSE [42], BAM [43], and CBAM [44] synergistically combine channel and spatial attention. Whether through parallel integration or sequential refinement, these frameworks effectively highlight salient regions while suppressing irrelevant channel responses. Building upon the empirical success of these hybrid paradigms and targeting the specific challenges of complex imagery, we design

the Coordinate-Aware Dynamic Attention module. By seamlessly integrating coordinate-aware spatial attention with dynamic channel weighting, CoDA empowers the network to precisely isolate and emphasize critical object features, demonstrating exceptional robustness in highly cluttered or occluded environments.

III. PROPOSED METHODS

To address the challenges of UAV-based aerial imagery, we propose a detection framework, DCA-YOLO, built upon the YOLOv8 baseline. At the core of DCA-YOLO lies the DCA-PAN neck, which integrates three tailored modules: DWAF for multi-scale feature interaction, CDI for strengthening semantic consistency, and CoDA for highlighting task-relevant information. The feature flow proceeds by first aggregating multi-scale features with DWAF, then refining cross-scale semantics through CDI, and finally emphasizing critical cues with CoDA to achieve precise vehicle detection. Furthermore, to alleviate computational overhead, we replace the original backbone with the lightweight FasterNet [21]. The overall framework is depicted in Fig. 2, while the details of DWAF, CDI, and CoDA are elaborated in the subsequent sections.

A. Dynamic Weighted Adaptive Feature Fusion Module

To enhance the fusion of localization details in the bottom-up pathway, we introduce the CDC module. Unlike conventional FPN or PAN that rely on straightforward additive or concatenative fusion, the CDC module explicitly structuralizes the aggregation of fine-grained spatial details and robust semantic cues. This structuralization is critical to addressing the drastic multi-scale variations inherent in UAV perspectives. Specifically, the CDC operates as a composite block comprising initial convolutional layers, our proposed DWAF module, whose architecture is illustrated in Fig. 3(a), and a terminal C2f block.

The core novelty and empirical effectiveness of the CDC module stem primarily from this DWAF block. While advanced architectures like BiFPN utilize scalar learnable weights for feature node scaling, they often fail to disentangle the distinct characteristics of shallow features. In contrast, DWAF extends beyond standard attention mechanisms by introducing a novel, learnable dual-pooling architecture tailored specifically for low-level features. By employing parallel adaptive max-pooling and average-pooling branches to independently extract sharp object boundaries and broader background context, and dynamically arbitrating their contributions through learnable weights, DWAF actively mitigates aliasing artifacts and semantic dilution. This targeted design streamlines the multi-level feature fusion process, yielding a highly discriminative, multi-scale representation that significantly boosts localization accuracy for complex UAV-based detection tasks.

B. Cross-layer Dynamic Interaction Fusion Module

Detection accuracy in UAV-based vehicle detection is commonly degraded by background occlusion and motion blur, which disrupt feature continuity and obscure critical object boundaries. To overcome these persistent challenges, we propose the CDI module, whose architecture is detailed in Fig. 3(b). Unlike conventional multi-scale fusion strategies that rely on rigid element-wise addition or concatenation, the CDI module introduces a fundamentally novel paradigm by explicitly modeling the cross-scale interplay of features. It bridges the semantic gap between shallow spatial details and deep semantics through dynamic channel alignment and learnable attention-based aggregation, ensuring content-aware transformations that adapt dynamically to diverse input scales rather than applying uniform linear transformations.

The core novelty and empirical impact of the CDI module lie in its unique 3D convolutional enhancement stage, which fundamentally advances beyond traditional 2D feature aggregation. Rather than simply flattening or merging features, CDI constructs a unified pseudo-3D tensor across a newly introduced 'scale' dimension, seamlessly embedding the attention-weighted fused features between the aligned low-level spatial details and high-level semantic cues. By deploying 3D convolutions directly across this scale dimension, the module actively captures inter-scale correlations and deeply integrates multi-level information without the need for redundant low-level operators. This structuralized interaction inherently reconstructs disrupted feature continuities and sharpens blurred boundaries, yielding a highly robust, unified feature representation that significantly improves the detection of heavily occluded and fast-moving vehicles in complex aerial environments.

C. Coordinate-aware Dynamic Attention Module

To address multi-scale vehicle detection in high-resolution aerial scenes, particularly for small targets embedded in complex backgrounds, we propose the CoDA module, whose architecture is illustrated in Fig. 3(c). Unlike conventional attention mechanisms like SENet that suffer from information loss due to dimensionality reduction and incur high computational overhead from fully connected layers, CoDA explicitly structuralizes feature refinement into two efficient, parallel pathways. For the channel dimension, we introduce an adaptive local cross-channel interaction mechanism. Instead of relying on global dense connections, it employs an adaptive one-dimensional convolution where the kernel size K is dynamically determined by the channel dimension C:

$$K = \max\left(\min\left(\left\lfloor \frac{|\log_2(C) + b|}{\gamma} \right\rfloor_{\text{odd}}, C\right), 3\right) \quad (1)$$

where γ and b are configurable hyperparameters. This targeted adaptation allows the network to capture local channel dependencies without the burden of global com-

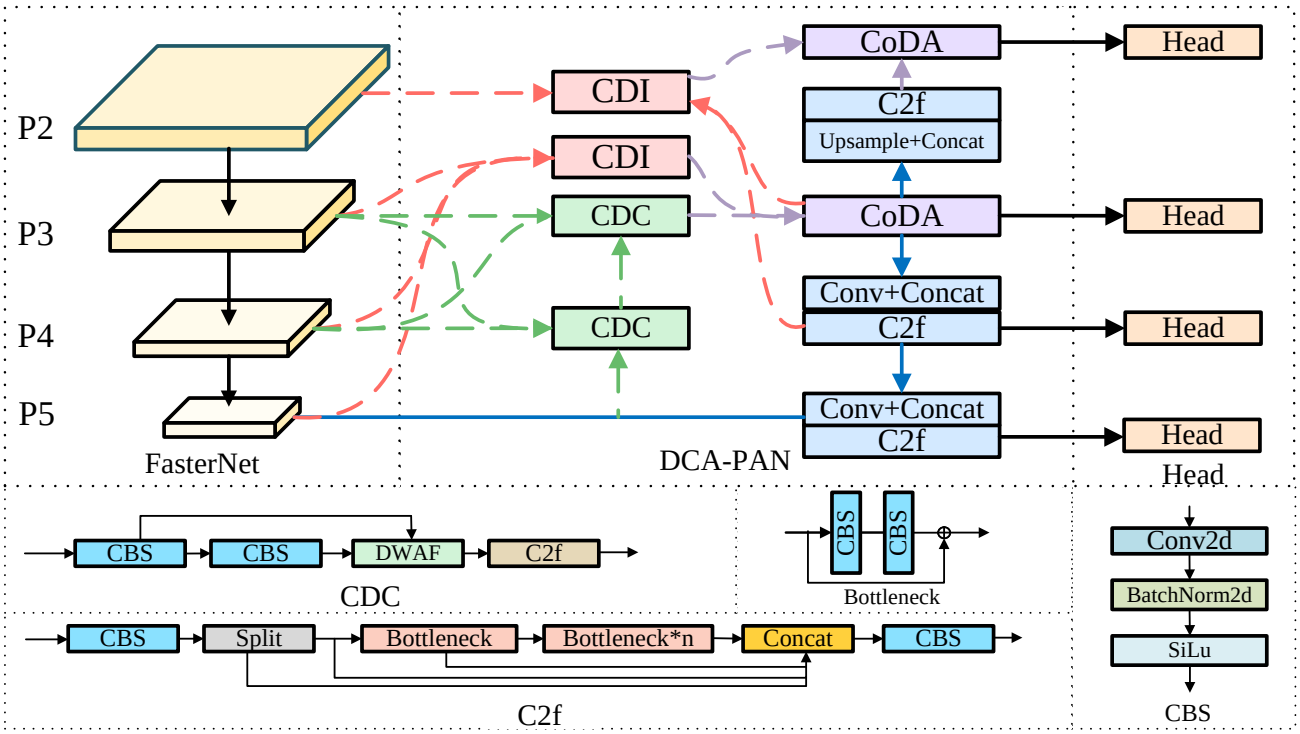


Fig. 2. Overall architecture of the proposed DCA-YOLO. The model integrates a lightweight backbone with the DCA-PAN neck, which incorporates DWAF for adaptive multi-scale fusion, CDI for semantic interaction, and CoDA for spatial attention. These fused features are used by the detection head to generate multi-scale predictions.

putation, and is augmented by a learnable scaling factor to autonomously calibrate attention intensity.

Concurrently, to overcome the limitations of traditional spatial attention—which either blurs fine details through global pooling or incurs massive parameter costs with fixed-size, non-adaptive large kernels—CoDA adapts a coordinate attention mechanism specifically tailored for aerial perspectives. It decouples the spatial dimension into orthogonal height and width directions, effectively capturing anisotropic positional context. The directional contexts are aggregated via adaptive 1D pooling along each axis:

$$p_w(i) = \frac{1}{H}\sum_{0\le j\le H} E(i,j) \tag{2}$$

$$p_h(j) = \frac{1}{W}\sum_{0\le i\le H} E(i,j) \tag{3}$$

where $E(i,j)$ denotes the feature value at position (i,j). To maintain computational efficiency, the aggregated coordinate features are compressed using depthwise separable convolutions prior to reconstructing the spatial attention masks. Ultimately, the features from both pathways are synergistically unified:

$$f_{\text{CoDA}}(X_1, X_2) = \text{Conv}_{\text{fusion}}\Big(\text{Concat}\big(f_{\text{Cha}}(X_1), f_{\text{Spt}}(X_2)\big)\Big) \tag{4}$$

where X_1 and X_2 represent the inputs to the channel and spatial pathways, respectively. This streamlined fusion yields a highly discriminative feature map that powerfully suppresses background noise while accentuating fine-grained target localization for UAV-based detection.

D. Dynamic Cross-Attention Path Aggregation Network

To construct our framework's core, we introduce DCA-PAN, a bidirectional feature pyramid architecture that strategically integrates the DWAF, CDI, and CoDA modules. While built upon the topology of FPN [16] and PAN [10], DCA-PAN abandons their rigid, element-wise fusion. Furthermore, it explicitly overcomes the limitations of BiFPN. Although BiFPN employs learnable scalar weights, it still relies on simple weighted addition, treating features as uniform entities and failing to decouple intricate spatial details from semantic context. In contrast, DCA-PAN structurally revolutionizes this paradigm through a content-aware integration: leveraging CDI for inter-scale 3D correlations, CoDA for asymmetric cross-attention, and DWAF for adaptive boundary preservation. This synergistic design transforms the standard feature pyramid into a highly responsive network tailored for the drastic multi-scale variations of aerial scenes.

The functional superiority of DCA-PAN stems from its meticulously designed pathways. In the top-down semantic pathway, the CDI module replaces scalar-weighted merging by constructing a unified pseudo-3D representation across scales to establish a robust semantic baseline. The CoDA module then performs asymmetric cross-channel and spatial attention, selectively fusing these semantic features with backbone representations to suppress background noise. In the complementary bottom-up localization pathway, CDC blocks anchored by DWAF

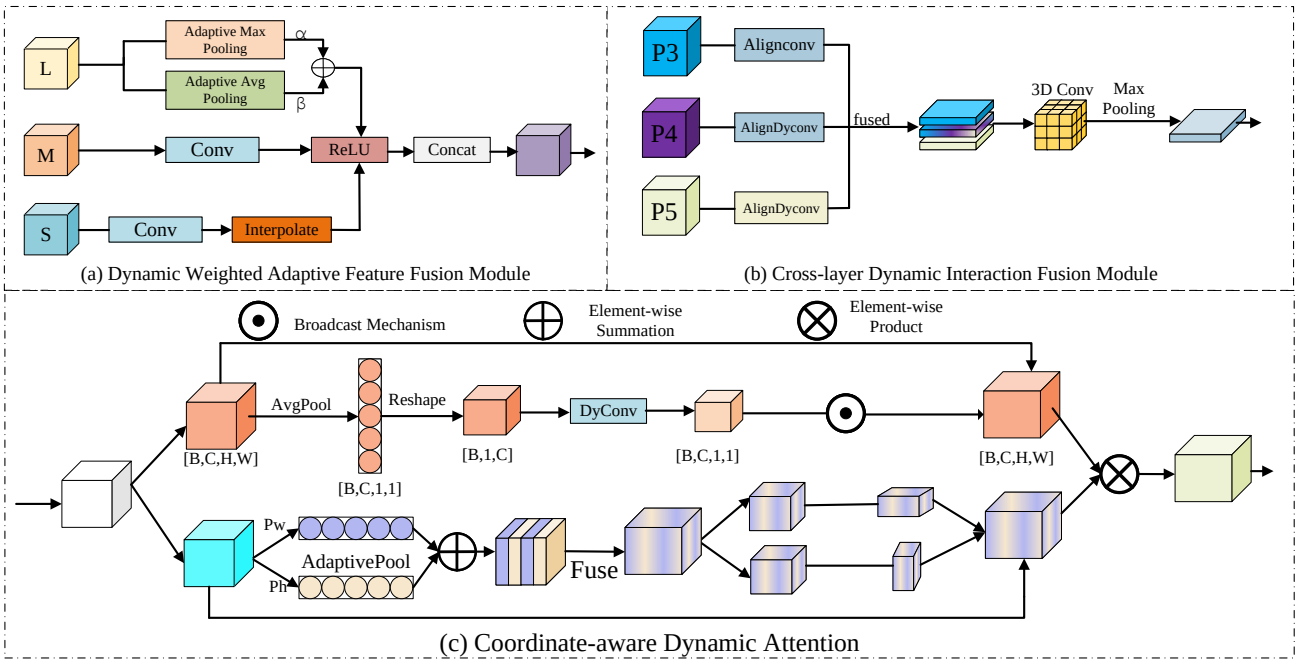


Fig. 3. onceptual illustration of the proposed modules: (a) The DWAF module adaptively fuses three feature levels using weighted dual-pooling on low-level features before concatenation. (b) The CDI module aligns and sums three input scales, then applies a 3D convolution to a pseudo-3D tensor for deep cross-scale fusion. (c) The CoDA module performs asymmetric parallel fusion by refining separate pathways with channel and spatial attention before a final convolutional merge.

execute an adaptive dual-pooling fusion. This ensures sharp boundaries and contextual cues propagate upward without the aliasing artifacts of standard downsampling. Through this cohesive interplay, DCA-PAN transcends simple weighted additions, achieving an optimal equilibrium between semantic richness and spatial precision for robust UAV-based vehicle detection.

IV. Experiments and Results

In this section, we first introduce the dataset used in this article, and then preprocess the dataset to improve the accuracy of vehicle detection. Secondly, we introduced various evaluation indicators and comprehensively evaluated the superiority of our model through a large number of experiments. Finally, the advantages of the DCA-YOLO model are demonstrated intuitively through the visualizing detection results.

A. Experimental Databases

To evaluate the performance of the proposed DCA-YOLO framework, we conduct experiments on the widely used public dataset VisDrone [22]. This dataset covers diverse UAV-based vehicle detection scenarios with varying levels of object density, scale variation, and background complexity.

1. *VisDrone*

The VisDrone is a large-scale dataset collected by Tianjin University and the AISKYEYE team. It consists of thousands of aerial images captured by drones from various altitudes, viewing angles, and illumination conditions across urban environments. The original annotations include ten object categories such as pedestrian, car, van, truck, and bus. As our study focuses on vehicle detection, we used only the four relevant classes: car, van, truck, and bus. After this filtering, the dataset contains 5,917 training images, 515 validation images, and 1,526 test images. The dataset is characterized by high object density, extreme scale variation, and cluttered backgrounds, making it a representative benchmark for evaluating small vehicle detection in UAV scenarios. Examples of the dataset captured at different times of day are illustrated in the first row of Fig. 4.

2. *UAV-HS*

The UAV-HS is a complex-scene dataset extended from VisDrone, designed to address the limitations of benchmark datasets captured in controlled environments. It focuses on real-world complexities by documenting intricate vehicle behaviors in challenging urban hubs, such as multi-level interchanges, logistics centers, and bus-priority intersections. The dataset provides multi-dimensional data encompassing significant scale variations, motion blur, and spatial imbalance. Furthermore, it offers dedicated coverage of nocturnal scenes, ensures a balanced class distribution, and incorporates advanced preprocessing techniques like radiometric correction. By capturing authentic lighting fluctuations and dynamic traffic ecosystems, UAV-HS establishes a more robust benchmark for developing models with enhanced resilience to occlusion and superior performance in small vehicle de-

Fig. 4. Example images from the primary datasets: VisDrone and UAV-HS

tection, thereby facilitating advancements in fine-grained urban traffic management. Examples are shown in the bottom side of Fig. 4.

B. Evaluation Indicators and Training Parameters

This article focuses on developing a high-performance model for real-time vehicle detection in challenging UAV aerial scenes. Therefore, evaluating the model's detection accuracy, inference speed, and overall computational efficiency is particularly important. We will now introduce the performance evaluation criteria, experimental platform, and relevant parameters used during training.

1. Evaluation Indicators

To holistically evaluate the model, we assess both its detection accuracy and computational efficiency. The model's detection accuracy is quantified using several standard mAP metrics, each offering a different perspective on performance. The mAP_{50} metric, calculated at an IoU threshold of 0.5, primarily gauges the model's recognition and classification capability. The stricter mAP_{75} metric, requiring an IoU greater than 0.75, places a greater emphasis on localization precision. For the most comprehensive assessment, we utilize $\text{mAP}_{50:95}$, which averages the mAP across ten IoU thresholds from 0.50 to 0.95 and represents the overall detection quality. Given the prevalence of small targets in aerial imagery, we also report mAP_s, a specialized metric representing the average precision for objects with an area smaller than 32x32 pixels. To complement these accuracy metrics, the model's computational cost is measured by its total number of parameters and GFLOPs.

2. Training Parameters

The hardware platform and training environment used in our experiments consisted of an Intel I9-14900KF CPU, an NVIDIA GeForce RTX 4090 GPU with 24GB memory, and a software environment based on Python 3.9.19 with PyTorch 2.1.2 running on Windows 10. To ensure stable and efficient model convergence, the training process was configured with 300 epochs, an input image size of 1280, and a batch size of 8. The SGD optimizer was employed with a learning rate of 0.01, momentum of 0.8, weight decay of 0.0005, and 3 warmup epochs. Furthermore, we explicitly clarify that all experimental results reported in this study for both the proposed DCA-YOLO and the various baseline models were strictly obtained using this identical training configuration and computational budget. This standardized approach constitutes a controlled, uniform-training study designed purely to isolate architectural effectiveness, ensuring that performance gains are attributed solely to structural innovations rather than disparities in training hyperparameters. Consequently, our comparative analysis is not intended to serve as a definitive ranking of all state-of-the-art detectors under their respective best-optimized recipes. Instead, under this controlled setup, the proposed framework demonstrates highly competitive performance specifically among compact architectures, scientifically validating its optimal accuracy-efficiency tradeoff for UAV-based detection.

C. Comparison Experiment

In this section, we present the experimental results of DCA-YOLO on two distinct datasets: VisDrone and UAV-HS. For both datasets, our model is benchmarked against a series of contemporary advanced methods, including high-capacity YOLO variants and transformer-based detectors such as EfficientVIT and Deformable-DETR, to demonstrate its competitiveness in complex scenarios. As shown in Table I, DCA-YOLO achieves superior performance among compact architectures, reaching 61.1% mAP_{50} and 16.3% mAP_s on VisDrone. While medium-scale models like YOLOv11-m and YOLOv12-s exhibit higher absolute accuracy, DCA-YOLO significantly outperforms other nano-level detectors and transformer-enhanced models like EfficientVIT and Deformable-DETR. On the more challenging UAV-HS dataset, DCA-YOLO further achieves 61.5% mAP_{50}, highlighting its robustness.

Beyond detection accuracy, we analyze the structural efficiency of DCA-YOLO by comparing its parameter count and computational complexity with various contemporary advanced models, as summarized in Table II. Although high-capacity models such as YOLOv11-m, YOLOv11-s, and YOLOv12-m achieve superior precision, their significantly higher parameter volumes and GFLOPs impose substantial computational demands that often exceed the operational limits of performance-constrained ground-based edge devices. The multi-fold increase in GFLOPs in these medium-scale variants typically leads to prohibitive latency and hardware requirements, making them less practical for real-time deployment on localized hardware where computing resources are restricted. Therefore, we consider Nano-level architectures to be a more suitable choice for such environments, as they effectively address the critical balance between architectural compactness and detection performance. By maintaining a low parameter footprint and optimized GFLOPs, DCA-YOLO ensures efficient real-

TABLE I

Comparison with state of the art models on VisDrone and UAV-HS datasets.

Method	**mAP$_{50}$-VisDrone**	**mAP$_{50:95}$-VisDrone**	**mAPs-VisDrone**	**mAP$_{50}$-UAV-HS**	**mAP$_{50:95}$-UAV-HS**	**mAPs-UAV-HS**
Faster R-CNN [6]	55.3	34.2	12.8	55.4	34.4	12.8
Cascade R-CNN [45]	56.7	39.2	15.7	56.9	39.4	15.8
CenterNet [46]	49.4	29.5	11.0	49.7	29.8	11.1
EfficientDet [17]	50.1	27.3	11.5	50.2	27.6	11.5
EfficeintVIT [47]	55.6	38.5	17.4	56.2	38.9	17.6
Deformable-Detr [48]	54.9	34.1	13.5	55.6	34.4	13.7
RT-DETR [49]	50.1	32.4	11.7	50.3	32.5	11.8
DINO [50]	53.0	33.7	14.5	53.7	34.1	14.7
TOOD [51]	56.1	38.8	13.9	56.5	39.3	14.0
ATSS [52]	57.0	36.6	14.7	57.6	38.9	14.9
YOLOv5-n [53]	56.6	37.8	13.4	58.1	38.8	13.8
YOLOv6-n [54]	58.1	39.3	13.6	58.6	39.8	13.6
YOLOv9-n [55]	59.8	40.7	14.3	60.2	41.1	14.5
YOLOv10-n [56]	57.5	38.7	13.6	58.2	39.3	13.7
YOLOv11-n [57]	57.0	38.4	13.4	57.8	39.0	13.7
YOLOv11-s	62.3	42.7	16.4	62.8	43.0	16.7
YOLOv11-m	**65.9**	**45.9**	**19.0**	**66.6**	**46.1**	**19.3**
Hyper-YOLO-n [58]	60.4	40.7	15.3	60.8	41.1	15.2
YOLOv12-n [59]	58.4	39.6	13.6	58.7	39.9	13.6
YOLOv12-s	63.1	43.4	16.6	63.6	43.6	16.7
DCA-YOLO (ours)	60.9	41.6	16.1	61.5	42.2	16.3

† The mAP values are reported in percentage (%).

(a) Detection result of baseline (b) Detection result of DCA-YOLO

Fig. 5. Detection results of (a) the baseline YOLOv8n versus (b) our DCA-YOLO on challenging scenes. Red boxes indicate instances where our model corrects the baseline's failures, such as false detections and poor localization.

time processing while achieving competitive accuracy, thus providing a more viable solution for ground-based application scenarios under hardware constraints.

To provide a more intuitive validation of our model's superiority, we present a qualitative comparison of detection results between DCA-YOLO and the baseline YOLOv8n in Fig. 5. The figure illustrates challenging scenarios where the baseline model produces false negatives or localization errors (red boxes). In contrast, DCA-YOLO yields more accurate detections with fewer false positives. These visual results are consistent with the quantitative evaluation, confirming the robustness of DCA-YOLO in real-world UAV vehicle detection.

TABLE II

Parameters and FLOPs of representative vehicle detection architectures

Module	Params (M)	FLOPs (G)
Faster R-CNN [6]	41.3	191
Cascade R-CNN [45]	69.3	219
CenterNet [46]	32.3	184
EfficientDet [17]	3.8	13.8
EfficeintVIT [47]	10.7	27.2
Deformable-Detr [48]	40.8	312
RT-DETR [49]	41.9	125
DINO [50]	47.5	252
TOOD [51]	34.1	185
ATSS [52]	32.3	189
YOLOv5-n [53]	7.0	13.8
YOLOv6-n [54]	4.2	31.7
YOLOv9-n [55]	7.6	11.9
YOLOv10-n [56]	2.7	8.2
YOLOv11-n [57]	**2.6**	**6.3**
YOLOv11-s	9.4	21.3
YOLOv11-m	20.0	67.7
Hyper-YOLO-n [58]	4.1	10.8
YOLOv12-n [59]	4.8	12.1
YOLOv12-s	9.2	21.2
DCA-YOLO (ours)	3.6	14.8

TABLE III

Ablation study of multi-scale fusion structures.

Configuration	P2	mAP_{50}	$\text{mAP}_{50:95}$	mAP_s
BIFPN-YOLO	×	58.1	39.2	14.4
ASF-YOLO	×	59.0	39.9	14.5
ASF-YOLO	✓	57.7	38.7	14.8
DCA-YOLO (ours)	×	59.9	40.6	15.3
DCA-YOLO (ours)	✓	**60.9**	**41.6**	**16.1**

† The mAP values are reported in percentage (%).

D. Ablation Experiment

1. *Multi-Scale Feature Fusion Ablation Experiment*

To evaluate the effectiveness of the proposed multi-scale feature fusion strategy, we conducted a comprehensive ablation experiment comparing several architectural configurations: the baseline ASF-YOLO, an ASF-YOLO variant with an integrated P2 layer, BiFPN-YOLO (where the neck is replaced with a standard BiFPN structure), and our complete DCA-YOLO. The experimental results, summarized in Table III, reveal that the baseline ASF-YOLO achieved an mAP_{50} of 59.0%, an $mAP_{50:95}$ of 39.9%, and an mAP_s of 14.5%. While adding a P2 layer slightly improved small target detection to 14.8% mAP_s, the overall performance declined to 57.7% mAP_{50}, suggesting that simple shallow feature integration is insufficient for complex UAV environments. Furthermore, although the BiFPN-YOLO configuration leverages learnable weights for bidirectional fusion, it often fails to adequately resolve the extreme scale variations and background clutter inherent in aerial imagery. In contrast, DCA-YOLO achieves the highest performance across all metrics, reaching 60.9% mAP_{50}, 41.6% $mAP_{50:95}$, and 16.1% mAP_s. This significant improvement demonstrates that our innovative fusion framework effectively enhances semantic–spatial interaction, providing superior robustness and accuracy for detecting small, densely distributed vehicles in UAV aerial imagery

2. *Ablation Study on Backbone Networks*

To determine a more effective backbone for UAV-based vehicle detection than the original CSPDarknet53 in YOLOv8, we conducted a systematic comparison by substituting the backbone with ConvNeXtV2 [60], GhostNetV2 [15], ResNet34 [61], FasterNet [21], ShuffleNetV2 [62], MobileNetV4 [63], and StarNet [64]. All variants were evaluated under identical training settings on the VisDrone [22] test set, and the quantitative results are summarized in Table IV.

The results demonstrate that FasterNet consistently delivers the strongest detection performance, achieving an mAP_{50} of 58.3%, surpassing both CSPDarknet53 and other competitive lightweight architectures. Importantly, this accuracy gain is obtained without incurring additional computational burden; FasterNet exhibits lower FLOPs with only a marginal parameter increase. This superior accuracy–efficiency balance makes it particularly well-suited for UAV small-object detection scenarios, where fine-grained spatial representation and computational practicality are both critical. Therefore, FasterNet was adopted as the backbone of the proposed framework.

3. *Ablation Study on Learnable Parameters*

To ascertain the efficacy of the learnable parameters within our proposed DWAF module, we conducted a targeted ablation study. In this experiment, we compared our full model against a variant where the learnable parameters in the DWAF module were removed, while all other components of the network were held constant. The results demonstrate a notable improvement when learnable parameters are utilized. Specifically, the model equipped with learnable parameters achieved an mAP_{50} of 60.9%, an $\text{mAP}_{50:95}$ of 41.6%, and an mAP_s of 16.1%, while the variant without these parameters yielded 58.2%, 39.3%, and 15.4%, respectively. As shown in Table V, incorporating learnable parameters in the DWAF module improves the mAP50 by 2.7% and the mAP50:95 by 2.3%, clearly demonstrating the benefit of adaptive weight adjustment. This comparison indicates that the learnable parameters are crucial, enabling the DWAF module to adaptively adjust its weights and thereby enhance the model's overall detection accuracy, particularly for small and densely distributed vehicles in UAV imagery.

In summary, the DCA-YOLO model demonstrates significant improvements over the baseline in all aspects, confirming that each proposed module consistently enhances the performance of DCA-YOLO without introducing conflicts between them.

TABLE IV
Model Performance of UAV Vehicle Detection Using YOLOv8 Different Backbones in VisDrone.

Backbone	**mAP$_{50}$**	**mAP$_{75}$**	**mAP$_{50:95}$**	**mAP$_s$**	**Params (M)**	**FPS**	**FLOPs (G)**
CSPDarkNet	57.2	44.4	38.5	14.3	2.3	219.0	8.9
ConvNeXt	55.0 (-2.2)	41.9 (-2.5)	37.0 (-1.5)	13.7 (-0.6)	5.66 (+3.36)	138.8 (-80.2)	14.1 (+5.2)
ResNet34	52.8 (-4.4)	39.9 (-4.5)	35.4 (-3.1)	12.1 (-2.2)	**2.18 (-0.12)**	**285.0 (+66.0)**	**6.4 (-2.5)**
GhostNetV2	56.9 (-0.3)	44.2 (-0.2)	38.6 (+0.1)	13.4 (-0.9)	2.3 (+0.0)	192.3 (-26.7)	6.8 (-2.1)
StarNet	53.4 (-3.8)	40.2 (-4.2)	35.7 (-2.8)	12.5 (-1.8)	2.21 (-0.09)	204.0 (-15.0)	6.5 (-2.4)
ShuffleNetV2	55.6 (-1.6)	42.3 (-2.1)	37.6 (-0.9)	12.9 (-1.4)	2.79 (+0.49)	169.8 (-49.2)	7.4 (-1.5)
MobileNetV4	58.0 (+0.8)	44.7 (+0.3)	39.3 (+0.8)	13.9 (-0.4)	5.7 (+3.4)	198.1 (-20.9)	22.5 (+13.6)
FasterNet	**58.3 (+1.1)**	**44.9 (+0.5)**	**39.7 (+1.2)**	**14.3 (+0.0)**	4.17 (+1.87)	185.1 (-33.9)	10.7 (+1.8)

† The values in parentheses indicate the performance change compared with CSPDarkNet.

TABLE V
Ablation study of DWAF parameters.

Setting	**mAP$_{50}$**	**mAP$_{50:95}$**	**mAP$_s$**
w/o params	58.2	39.3	15.4
w/ params	**60.9**	**41.6**	**16.1**

† The mAP values are reported in percentage (%).

TABLE VI
Comparison of different attention mechanisms on VisDrone vehicle detection.

Attention Module	**mAP$_{50}$**	**mAP$_{50:95}$**	**mAPs**
SE	56.7	38.3	14.7
CA	57.4	38.7	15.1
ECA	57.6	38.8	14.8
CBAM	57.4	38.6	14.9
SimAM	57.2	38.4	14.9
CoDA	**60.9**	**41.6**	**16.1**

† The mAP values are reported in percentage (%).

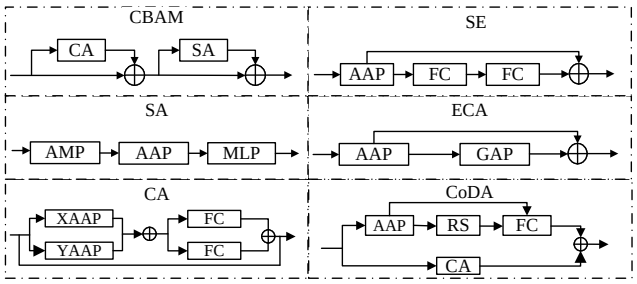


Fig. 6. Illustrates the structural layouts of representative attentionmechanisms, namely CBAM, SE,ECA, SA, and CA, with the essentialoperations and components of each module clearly depicte.

4. *Ablation Study on Attention Mechanisms*

To demonstrate the effectiveness of our proposed CoDA module, we conducted a targeted ablation study comparing it with several classic and recent attention mechanisms. As shown in Table VI and Fig. 6, this experiment involved replacing the CoDA module in our final architecture with other mature attention modules including SE [12], CA [65], ECA [18], CBAM [44], and SimAM [66] while keeping all other components of the model unchanged. The results confirmed the superiority of our approach, with the CoDA equipped model achieving the highest scores across all metrics: 60.9% in mAP$_{50}$, 41.6% in mAP$_{50:95}$, and 16.1% in mAP$_s$. This performance represents a significant improvement over the best-performing alternatives, surpassing the second-best mechanism, ECA, by 3.3% in mAP$_{50}$ and the next best on small targets, CA, by 1.0% in mAP$_s$. This confirms that CoDA's design, which employs an asymmetric parallel approach to fuse channel and spatial information, is an effective attention mechanism.

5. *Ablation Study on the DCA-YOLO Architecture*

Following the validation of the CoDA module, we conducted a comprehensive ablation study on the VisDrone dataset to systematically evaluate the contribution of each component within the DCA-YOLO framework. As detailed in Table VII, the baseline YOLOv8n model achieves an mAP_{50} of 57.2%, which increases to 58.3% when the original backbone is replaced with the lightweight FasterNet. While the individual integration of DWAF, CDI, and CoDA yields mAP_{50} scores of 59.2%, 58.8%, and 58.7% respectively, a critical observation arises when specific modules are combined. Notably, the pairwise combinations of DWAF+CDI and DWAF+CoDA result in a performance decline, with mAP_{50} dropping to 57.6% and 57.7% respectively. This phenomenon underscores a fundamental feature construction conflict between spatial precision and semantic abstraction, where the high-frequency spatial information preserved by DWAF to maintain sharp object boundaries may fundamentally interfere with the abstract cross-scale semantic consistency targeted by CDI. In the absence of a specialized coordination mechanism, these disparate feature representations lead to feature misalignment or

TABLE VII
EXPERIMENTAL RESULTS OF INTRODUCING DIFFERENT IMPROVEMENT STRATEGIES IN VisDrone

FasterNet	DWAF	CDI	CoDA	P2	mAP$_{50}$	mAP$_{75}$	mAP$_{50:95}$	mAPs	Params(M)	FPS	FLOPs (G)
					57.2	44.4	38.5	14.3	2.3	219.0	8.9
✓					58.3 (+1.1)	44.9 (+0.5)	39.7 (+1.2)	14.3 (+0.0)	4.17 (+1.9)	185.1 (-33.9)	10.7 (+1.8)
✓	✓				59.2 (+2.0)	44.8 (+0.4)	40.6 (+2.1)	14.9 (+0.6)	3.0 (+0.7)	217.0 (-2.0)	8.3 (-0.6)
✓		✓			58.8 (+1.6)	45.1 (+0.7)	39.9 (+1.4)	14.8 (+0.5)	4.21 (+1.9)	178.5 (-40.5)	11.0 (+2.1)
✓			✓		58.7 (+1.5)	45.0 (+0.6)	39.6 (+1.1)	14.9 (+0.6)	4.18 (+1.9)	182.8 (-36.2)	10.8 (+1.9)
✓				✓	58.8 (+1.6)	45.7 (+1.3)	40.0 (+1.5)	14.8 (+0.5)	2.9 (+0.6)	204.1 (-14.9)	12.2 (+3.3)
✓	✓	✓			57.6 (+0.4)	44.1 (-0.3)	38.9 (+0.4)	14.9 (+0.6)	4.19 (+1.9)	176.1 (-42.9)	11.1 (+2.2)
✓	✓		✓		57.7 (+0.5)	44.4 (+0.0)	39.1 (+0.6)	14.9 (+0.6)	4.22 (+1.9)	197.1 (-21.9)	11.2 (+2.3)
✓		✓	✓		59.0 (+1.8)	45.2 (+0.8)	39.9 (+1.4)	15.0 (+0.7)	4.47 (+2.2)	174.9 (-44.1)	11.6 (+2.7)
✓	✓	✓	✓		59.9 (+2.7)	46.2 (+1.8)	40.6 (+2.1)	15.3 (+1.0)	4.34 (+2.0)	142.8 (-76.2)	12.2 (+3.3)
✓	✓	✓	✓	✓	**60.9 (+3.7)**	**47.6 (+3.2)**	**41.6 (+3.1)**	**16.1 (+1.8)**	3.6 (+1.3)	200.0 (-19.0)	14.8 (+5.9)

† The values in parentheses indicate the performance change compared with the baseline.

TABLE VIII
Generalization experiment.

Method	mAP_{50}	$mAP_{50:95}$	mAP_s
Faster R-CNN	55.3	34.4	12.8
Cascade R-CNN	56.8	39.4	15.8
CenterNet	49.6	29.6	11.0
EfficientDet	50.0	27.3	11.4
EfficientVIT	55.9	38.6	17.6
Deformable-DETR	55.2	34.1	13.5
RT-DETR	50.2	32.4	11.7
DINO	53.3	37.7	14.4
TOOD	56.4	39.0	13.9
ATSS	57.3	37.2	13.9
YOLOv5-n	57.1	38.1	13.6
YOLOv6-n	58.4	39.5	13.6
YOLOv9-n	60.0	38.9	13.7
YOLOv10-n	57.8	38.9	13.6
YOLOv11-n	57.3	38.6	13.5
YOLOv11-s	62.6	42.9	16.7
YOLOv11-m	**66.1**	**46.0**	**19.1**
YOLOv12-n	58.5	39.6	13.6
YOLOv12-s	63.3	43.4	16.6
HyperYOLO-n	60.6	40.8	15.2
DCA-YOLO (ours)	61.1	41.9	16.3

TABLE IX
Robustness evaluation under diverse operating scenarios

Scenario	mAP_{50}	mAP_{75}	$mAP_{50:95}$	mAP_s
Daytime (Baseline)	65.8	52.0	45.2	18.8
Nighttime (Baseline)	50.2	36.5	32.6	10.0
High-altitude (Baseline)	55.9	43.6	37.7	16.4
Low-altitude (Baseline)	57.9	43.8	38.9	11.8
Daytime (Ours)	**69.4**	**55.7**	**48.3**	**21.6**
Nighttime (Ours)	**51.8**	**38.7**	**34.4**	**11.0**
High-altitude (Ours)	**59.3**	**47.0**	**40.4**	**18.3**
Low-altitude (Ours)	**59.7**	**46.1**	**41.0**	**13.2**

mutual interference, which confuses the detector in complex UAV backgrounds.

These quantitative observations are further corroborated by the Grad-CAM [67] visualizations presented in Fig. 7, where the first column highlights representative samples from VisDrone and the subsequent columns illustrate diverse challenging scenarios from the UAV-HS benchmark. The heatmaps for the pairwise combinations reveal a distinct spatial drift, manifested as scattered activation maps that inappropriately highlight irrelevant background noise, thereby providing a direct visual explanation for the precision degradation observed in our quantitative results. In contrast, the complete DCA-YOLO framework effectively mitigates these conflicts through the strategic coordination provided by the CoDA module. By utilizing coordinate-aware spatial attention and dynamic channel weighting, the model suppresses background interference and refocuses activation precisely on critical vehicle targets, as evidenced by the tri-module integration which restores the mAP_{50} to 59.9%. Finally, with the inclusion of the P2 detection head tailored for small objects, the complete DCA-YOLO reaches a peak mAP_{50} of 60.9%, representing a total improvement of 3.7% over the baseline while maintaining a high inference speed of 200.0 FPS. Furthermore, it is necessary to elaborate on why the final model, despite an increase in theoretical computational load (FLOPs), achieves both a reduction in parameters and an improvement in inference speed. Specifically, processing the high-resolution P2 layer inherently inflates FLOPs due to its massive spatial grid; however, its extremely narrow channels introduce minimal learnable parameters. Concurrently, the synergy among DWAF, CDI, and CoDA facilitates the structural pruning of redundant, parameter-dense deep layers. While intermediate variants suffered from structural fragmentation and unoptimized routing, the final DCA-YOLO framework achieves an efficient network design. By unifying the feature flow and eliminating sequential bottlenecks, this architectural refinement ensures highly efficient forward propagation, delivering superior real-time throughput despite the higher theoretical computational volume.

E. Generalization Experiments

To evaluate the generalization capability and practical applicability of the proposed DCA-YOLO under diverse UAV scenarios, comprehensive experiments are

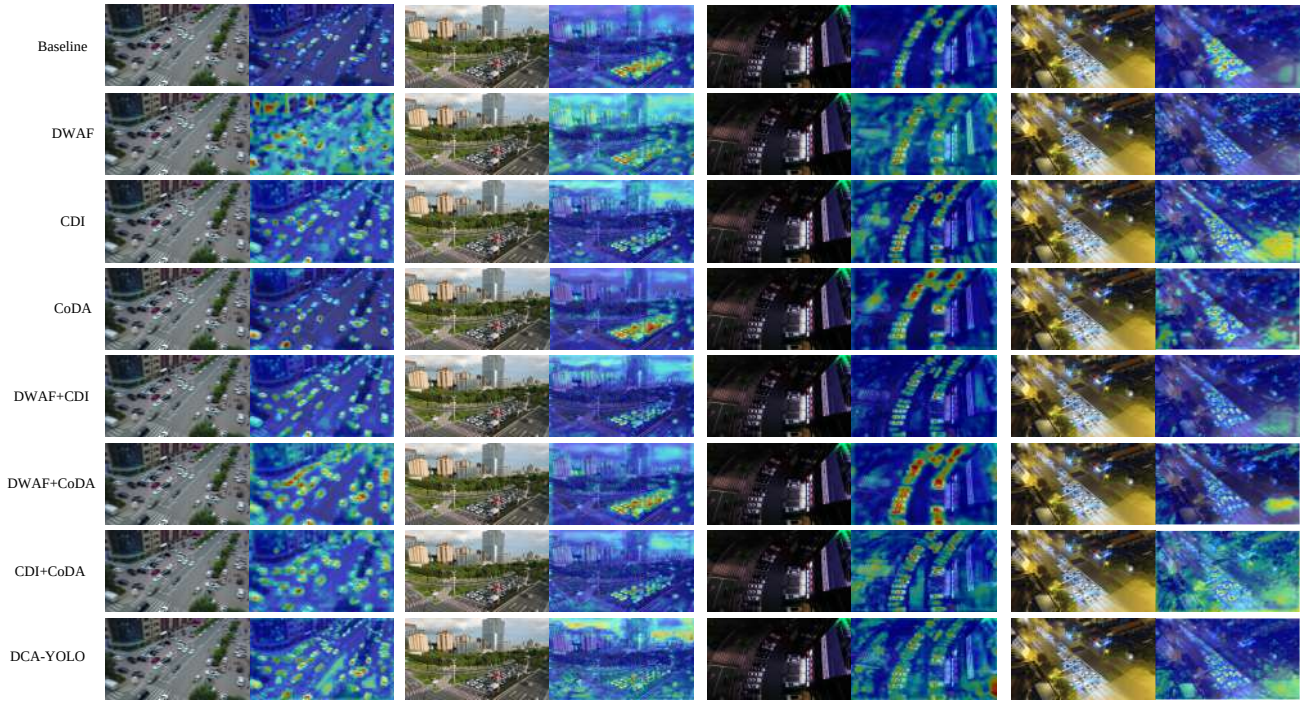


Fig. 7. Grad-CAM-based feature visualizations extracted from the final layer of the DCA-PAN neck. The first column presents representative samples from the VisDrone dataset , while Columns 2–4 illustrate challenging scenarios from the UAV-HS benchmark. The heatmaps display the activation intensity where the color transition from cool blue to deep red signifies increasing focus and confidence of the detection network. Visualizations of individual modules reveal specialized characteristics: the DWAF module primarily emphasizes fine-grained spatial details , whereas the CDI module promotes cross-scale semantic consistency. In contrast, certain pairwise module combinations exhibit more dispersed feature responses and significant background noise, highlighting the potential conflict between high-frequency spatial precision and abstract cross-layer semantics. The complete DCA-YOLO framework, strategically coordinated by the CoDA module, effectively suppresses environmental interference and produces the most concentrated activations on key vehicle targets. This integrated design maintains stable feature representations even under target distortion and various occlusion scenarios.

TABLE X
Inference performance on NVIDIA RTX 3060

Method	FPS	GPU Memory
Faster R-CNN	12.8	572 MB
Cascade R-CNN	13.1	681 MB
CenterNet	22.8	232 MB
EfficientDet	33.3	488 MB
EfficientVIT	18.5	406 MB
Deformable-DETR	13.4	495 MB
RT-DETR	11.4	807 MB
DINO	17.2	679 MB
TOOD	21.4	495 MB
ATSS	28.6	387 MB
YOLOv5-n	54.6	199 MB
YOLOv6-n	62.4	220 MB
YOLOv9-n	34.3	205 MB
YOLOv10-n	51.2	254 MB
YOLOv11-n	61.3	274 MB
YOLOv11-s	31.8	300 MB
YOLOv11-m	25.3	523 MB
YOLOv12-n	48.0	322 MB
YOLOv12-s	22.4	448 MB
HyperYOLO-n	45.4	453 MB
DCA-YOLO (ours)	**52.3**	**330 MB**

conducted considering cross-dataset generalization, performance under different flight conditions, failure mode analysis, and inference efficiency.

First, cross-dataset generalization experiments are performed to assess robustness under domain shifts. Specifically, the model is trained on the VisDrone dataset and directly evaluated on UAV-HS, without any additional fine-tuning. As summarized in Table VIII, the results show that, compared with baseline detectors, DCA-YOLO exhibits a slower performance degradation under cross-dataset settings, particularly on small-object detection metrics. Notably, our proposed method achieves an mAP_{50} of 61.1% and an mAP_s of 16.3% on the target dataset. This indicates that the proposed DCA-PAN effectively mitigates feature distribution shifts caused by variations in imaging altitude, viewpoints, and scene layouts , thereby enhancing generalization to unseen UAV environments.

Furthermore, to analyze adaptability to realistic UAV operating conditions, the test samples are divided according to illumination conditions and flight altitude, forming daytime versus nighttime subsets and low-altitude versus high-altitude subsets. It should be noted that changes in flight altitude not only significantly affect object scale but also alter spatial distribution density, with high-altitude scenarios naturally involving more sparse small-object distributions. As detailed in Table IX, experimental results demonstrate that DCA-YOLO maintains excellent performance across all environmental conditions. In day-

TABLE XI
Generalization comparison results on the CARPK dataset

Method	mAP_{50}	mAP_{75}	$mAP_{50:95}$
Faster R-CNN	89.5	76.3	56.3
Cascade R-CNN	91.0	79.1	58.7
CenterNet	87.8	73.9	53.0
EfficientDet	83.2	67.9	46.5
EfficientVIT	96.6	86.1	74.4
RT-DETR	94.8	83.4	61.6
DINO	94.6	83.1	61.4
TOOD	91.3	79.2	58.5
ATSS	89.0	75.6	55.5
YOLOv5-n	92.3	80.3	59.6
YOLOv6-n	93.8	82.2	60.8
YOLOv9-n	95.1	84.0	62.0
YOLOv10-n	93.3	81.5	59.5
YOLOv11-n	93.0	81.1	59.1
YOLOv11-s	94.7	83.2	60.8
YOLOv11-m	96.0	85.0	62.4
YOLOv12-n	95.7	85.0	62.7
YOLOv12-s	96.6	86.2	63.8
HyperYOLO-n	95.4	84.4	62.3
DCA-YOLO (ours)	**98.8**	**89.8**	**67.3**

time scenarios, the model achieves a high accuracy with an mAP_{50} of 69.4%. While nighttime conditions pose substantially greater challenges due to reduced illumination and contrast, DCA-YOLO remains highly effective, reaching an mAP_{50} of 51.8%, which represents a significant lead over the baseline's 50.2%. This suggests that the proposed dynamic cross-layer interaction mechanism more effectively enhances weak object features under adverse lighting. Similarly, under varying flight altitudes, the model demonstrates superior adaptability, achieving mAP_{50} scores of 59.3% and 59.7% in high-altitude and low-altitude scenes, respectively, consistently outperforming the baseline in both metrics. Notably, the small-object detection capability reached 18.3% at high altitudes, an improvement over the baseline's 16.4%, validating the complementary effects of the DWAF and CDI modules in complex multi-scale environments.

Finally, to validate the cross-domain robustness of the proposed architecture, generalization experiments are conducted on the CARPK dataset. As demonstrated in Table XI, DCA-YOLO maintains high mean Average Precision when detecting vehicles in unseen environments, underscoring its strong adaptability to varying target densities and complex backgrounds. Finally, to evaluate deployment efficiency, inference benchmarks are performed on an NVIDIA RTX 3060 GPU using a batch size of one, reflecting real-time UAV operational constraints. As detailed in Table X, DCA-YOLO achieves a stable inference speed of 52.3 FPS with a peak GPU memory consumption of only 330 MB. Compared to advanced detectors like YOLOv11-m and YOLOv12-s, DCA-YOLO delivers a superior frame rate alongside a significantly lower memory footprint. These results confirm that our modular enhancements incur minimal computational overhead, achieving an optimal balance among detection accuracy, generalization capability, and deployment efficiency for resource-constrained UAV applications.

TABLE XII
Impact of image resolution on performance.

Resolution	mAP_{50}	mAP_{75}	$mAP_{50:95}$
320	28.5	14.0	14.9
640	47.1	32.2	29.6
1280	**60.9**	**47.6**	**41.6**

V. Key findings and guidelines

Extensive evaluations validate the efficacy of the proposed DCA-YOLO framework in advancing UAV-based vehicle detection, demonstrating substantial performance gains across challenging benchmarks. These architectural improvements are realized while strictly adhering to a resource-efficient, edge-oriented design philosophy. By integrating the lightweight FasterNet backbone with the highly efficient DCA-PAN neck, the model minimizes parameter count and computational complexity while maintaining low memory consumption and high-throughput real-time inference. This structurally optimized design ensures that the framework fully satisfies the stringent computational constraints typical of ground-based edge devices in collaborative UAV operations.

Consequently, the DCA-YOLO framework is highly recommended as a superior alternative to standard lightweight object detectors in operational scenarios characterized by drastic multi-scale variations, dense target distributions, and severe background occlusions. Although higher-capacity models may yield marginal gains in absolute precision, their prohibitive parameter volumes and elevated inference latency render them impractical for real-time deployment on localized hardware. DCA-YOLO systematically transcends this limitation by establishing an optimal equilibrium between architectural compactness and detection accuracy. It ultimately serves as a robust architectural paradigm for resource-constrained aerial perception tasks, particularly for missions necessitating the reliable detection of minuscule targets under adverse illumination or severe motion blur.

VI. Discussion

As illustrated in Fig. 8, a side-by-side comparison of the original image (a), the baseline model inference (b), and the DCA-YOLO inference (c) clearly delineates the detection disparities under identical conditions. Consistent with our ablation studies, the complete DCA-YOLO framework generally produces more concentrated and stable activation responses in complex aerial environments, significantly reducing the overall occurrence of missed detections. Nevertheless, despite these substantial performance improvements, highly challenging visual scenarios still expose the inherent limitations of

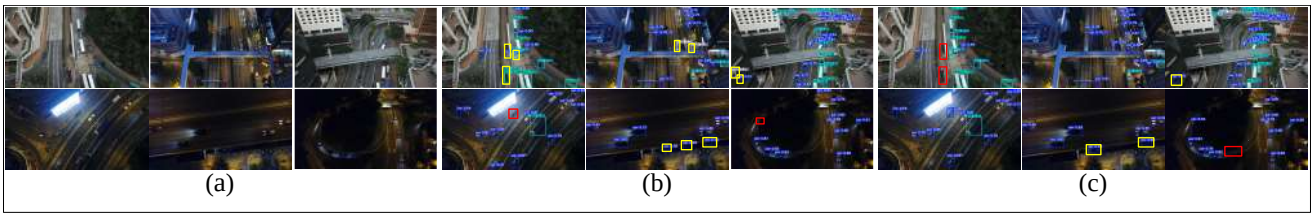


Fig. 8. Visualization of representative failure cases in diverse scenarios. (a) Original remote sensing images; (b) baseline inference results; (c) DCA-YOLO inference results. Red dashed boxes indicate False Negatives (FN), mainly missed black vehicles in low-contrast scenes, and red solid boxes denote False Positives (FP). A comparative view shows improved detection stability of DCA-YOLO, although certain challenging cases remain.

the proposed model, necessitating a deeper analysis of the underlying causes of persistent detection errors.

Frequent false negatives predominantly occur with black vehicles and heavily occluded targets operating under severely low-contrast conditions. In daytime vertical-view scenes, black buses and cars exhibit negligible pixel variance against dark asphalt surfaces. This severe visual homogeneity deprives the convolutional network of the distinct gradient signatures and edge features necessary for accurate boundary delineation, compelling the model to incorrectly suppress these targets as background noise. Similarly, in low-illumination nighttime environments, the profound degradation of structural details and the loss of color fidelity exacerbate this issue. Although DCA-YOLO effectively leverages its dynamic attention mechanisms to capture more lighting-related cues and recovers many difficult targets compared to the baseline, it fundamentally struggles when the visual contrast falls below the intrinsic feature-extraction threshold required by the network.

Conversely, specular reflection constitutes a persistent source of false positives for both models. In densely populated urban landscapes, reflections off glass building facades, metallic structures, or wet surfaces frequently project symmetrical boundaries and artificial textures that closely mimic the geometric properties of real vehicles. Because DCA-YOLO relies exclusively on appearance-based RGB features, it lacks the multi-modal context, such as depth cues or thermal signatures, required to reliably decouple these optical artifacts from actual physical objects. Consequently, while DCA-YOLO excels at mitigating scale variations and partial occlusions, our analysis indicates that its most suitable deployment scenarios are moderately to well-lit aerial environments, such as daytime traffic monitoring and highway surveillance, where fundamental object-to-background contrast is preserved. For operational environments dominated by extreme low-contrast conditions or severe reflective interference, augmenting the vision-based framework with supplementary sensor modalities, such as LiDAR or infrared imaging, remains a highly recommended practical strategy.

VII. Conclusion and Future Work

Building upon the comprehensive evaluations, including cross-scenario generalization analysis, resolution scaling experiments, and representative failure case studies, we summarize the main findings of this work.

In this paper, we introduced DCA-YOLO, a detection framework designed to address the inherent challenges of UAV-based aerial vehicle detection. By integrating the efficient FasterNet backbone with the proposed DCA-PAN neck incorporating DWAF, CDI, and CoDA modules, the model enhances multi-scale representation while maintaining computational efficiency. Quantitative evaluations on the VisDrone and UAV-HS datasets demonstrate consistent improvements over baseline models in mAP_{50} and $mAP_{50:95}$, supporting the effectiveness of the proposed architecture for real-time monitoring in complex aerial scenarios.

The extended resolution analysis further shows that input image resolution plays a critical role in small-object detection performance. As reported in Table XII, increasing the resolution from 320 to 1280 substantially improves detection accuracy, with mAP_{50} rising from 28.5% to 60.9% and $mAP_{50:95}$ increasing from 14.9% to 41.6%. High-resolution inputs preserve sufficient pixel-level details for small and distant targets, which reduces information loss during successive downsampling operations. Although lower resolutions provide higher throughput, the 1280×1280 configuration achieves a more favorable balance between detection quality and the intrinsic characteristics of UAV imagery.

The failure case analysis reveals the operational boundaries of the proposed framework. While DCA-YOLO improves robustness under challenging illumination and partial occlusion, extremely low-contrast vehicles and reflection-dominated scenes remain difficult for vision-based detection systems. These findings indicate that certain limitations are inherent to appearance-based modeling in the absence of additional contextual or multimodal information.

Finally, considering the intended deployment scenario in which UAVs perform image acquisition and detection is conducted on computationally constrained ground-based platforms, the framework prioritizes a practical balance between parameter efficiency and real-time performance rather than absolute peak accuracy. Future research will focus on incorporating temporal-aware mechanisms to leverage video continuity and exploring multimodal fusion strategies to further enhance robustness and scalability for comprehensive aerial perception.

REFERENCES

[1] S. Srivastava, S. Narayan, and S. Mittal, "A survey of deep learning techniques for vehicle detection from uav images," *Journal of Systems Architecture*, vol. 117, p. 102152, 2021.

[2] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128 837–128 868, 2019.

[3] L. Sun, J. Wang, K. Yang, K. Wu, X. Zhou, K. Wang, and J. Bai, "Aerial-pass: Panoramic annular scene segmentation in drone videos," in *2021 European Conference on Mobile Robots (ECMR)*, 2021, pp. 1–6.

[4] N. Samir Labib, G. Danoy, J. Musial, M. R. Brust, and P. Bouvry, "Internet of unmanned aerial vehicles—a multilayer low-altitude airspace model for distributed uav traffic management," *Sensors*, vol. 19, no. 21, p. 4779, 2019.

[5] H. Li and H. Qu, "Dassf: Dynamic-attention scale-sequence fusion for aerial object detection," in *Computational Visual Media*, P. Didyk and J. Hou, Eds. Singapore: Springer Nature Singapore, 2025, pp. 212–227.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[8] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[9] Z. Ying, J. Zhou, Y. Zhai, H. Quan, W. Li, A. Genovese, V. Piuri, and F. Scotti, "Large-scale high-altitude uav-based vehicle detection via pyramid dual pooling attention path aggregation network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 10, pp. 14 426–14 444, 2024.

[10] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768.

[11] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," in *arXiv preprint arXiv:1904.07850*, 2019.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

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

[14] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.

[15] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1577–1586.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[17] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 778–10 787.

[18] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 531–11 539.

[19] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 16 748–16 759.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: Chasing higher flops for faster neural networks," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 12 021–12 031.

[22] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang *et al.*, "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 213–226.

[23] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–I.

[24] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, pp. 137–154, 2004.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893 vol. 1.

[26] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[28] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.

[30] A. Nazir and M. A. Wani, "You only look once - object detection models: A review," in *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2023, pp. 1088–1095.

[31] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9626–9635.

[32] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 213–229.

[34] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7029–7038.

[35] G. Yang, J. Lei, H. Tian, Z. Feng, and R. Liang, "Asymptotic feature pyramid network for labeling pixels and regions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 7820–7829, 2024.

[36] Z. Wang, S. Fan, X. Huo, T. Xu, Y. Wang, J. Liu, Y. Chen, and Y.-Q. Zhang, "Emiff: Enhanced multi-scale image feature fusion for vehicle-infrastructure cooperative 3d object detection," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 16 388–16 394.

[37] Z. Lan, F. Zhuang, Z. Lin, R. Chen, L. Wei, T. Lai, and C. Yang, "Mfo-net: A multiscale feature optimization network for uav image object detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[38] M. Kang, C.-M. Ting, F. F. Ting, and R. C.-W. Phan, "Asf-yolo: A novel yolo model with attentional scale sequence fusion for cell instance segmentation," *Image and Vision Computing*, vol. 147, p. 105057, 2024.

[39] J. Wang, Z. Luan, Z. Yu, J. Ren, J. Gao, K. Yuan, and H. Xu, "Superpixel segmentation with squeeze-and-excitation networks," *Signal, Image and Video Processing*, vol. 16, no. 5, pp. 1161–1168, 2022.

[40] J. Guo, X. Ma, A. Sansom, M. McGuire, A. Kalaani, Q. Chen, S. Tang, Q. Yang, and S. Fu, "Spanet: Spatial pyramid attention network for enhanced image recognition," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[41] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 510–519.

[42] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 540–549, 2019.

[43] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[44] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 3–19.

[45] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

[46] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6568–6577.

[47] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "Efficientvit: Memory efficient vision transformer with cascaded group attention," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14 420–14 430.

[48] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[49] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16 965–16 974.

[50] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.

[51] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2021, pp. 3490–3499.

[52] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9756–9765.

[53] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, J. Fang, C. Wong, Z. Yifu, D. Montes *et al.*, "ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations," *Zenodo*, 2022.

[54] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[55] C.-Y. Wang, I.-H. Yeh, and H.-Y. Y. Mark Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in *European conference on computer vision*. Springer, 2024, pp. 1–21.

[56] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han *et al.*, "Yolov10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107 984–108 011, 2024.

[57] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[58] Y. Feng, J. Huang, S. Du, S. Ying, J.-H. Yong, Y. Li, G. Ding, R. Ji, and Y. Gao, "Hyper-yolo: When visual object detection meets hypergraph computation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 47, no. 4, pp. 2388–2401, 2024.

[59] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.

[60] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16 133–16 142.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[62] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 122–138.

[63] D. Qin, C. Leichner, M. Delakis, M. Fornoni, S. Luo, F. Yang, W. Wang, C. Banbury, C. Ye, B. Akin, V. Aggarwal, T. Zhu, D. Moro, and A. Howard, "Mobilenetv4: Universal models for the mobile ecosystem," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 78–96.

[64] X. Ma, X. Dai, Y. Bai, Y. Wang, and Y. Fu, "Rewrite the stars," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 5694–5703.

[65] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 708–13 717.

[66] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 11 863–11 874.

[67] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

Chaojun Dong received the B.S. and M.S. degrees in thermal engineering from the Central South University of Technology, Changsha, Hunan, China, in 1990 and 1993, respectively, and the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, Shanxi, China, in 2006.

He is currently a Professor with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, China. His research interests include intelligent information processing, AI, and intelligent traffic detection and control.

Yizhi Wang received the associate degree from Shanghai University of Engineering Science, Shanghai, China, in 2020. He is currently pursuing the M.S. degree in Pattern Recognition and Intelligent Systems with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, China. His research interests include computer vision, object detection, and their applications in intelligent transportation systems and artificial intelligence.

Yikui Zhai (Senior Member, IEEE) received the bachelor's degree in optical electronics information and communication engineering from Shantou University, Shantou, China, in 2004, the master's degree in signal and information processing from Shantou University, in 2007, and the Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in June 2013.

Since October 2007, he has been working with the Department of Intelligence Manufacturing, Wuyi University, Jiangmen, China, where he is a Professor. He has been a Visiting Scholar with the Department of Computer Science, University of Milan, Milan, Italy, since 2016. His research interests include image processing, deep learning, and pattern recognition.

Ye Li is currently an Engineer with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, China. He has authored many core papers, with research direction in control engineering and automation.

Xiankun Liu (Member, IEEE) is currently a Senior Experimentalist with the School of Mechanical and Automation Engineering, Wuyi University, Jiangmen, China. She has authored many core papers. Her research directions are intelligent detection and intelligent information processing.

Chaoyun Mai received the Ph.D. degree in information and communication engineering from Beihang University, Beijing, China, in 2017.

He is currently Associate Professor with the School of Electronics and Information Engineering, Wuyi University, Guangdong, China. His current research interests include image processing and signal processing.

Jianhong Zhou (Member, IEEE) received the B.S. degree in communication engineering and the M.S. degree in information and communication engineering from Wuyi University, Jiangmen, China, in 2020 and 2024, respectively. He is currently pursuing the Ph.D. degree with the School of Electronics and Information Engineering, South China University of Technology, Guangzhou, China.

His research interests include computer vision, representational contrastive learning, image processing, and object detection.

Pasquale Coscia (Senior Member, IEEE) received the Ph.D. degree in Industrial and Information Engineering from the Università degli Studi di Milano, Italy, in 2019. From 2019 to 2022, he was a post-doctoral researcher at the Università degli Studi di Padova. Dr. Coscia has received the 2021 Best Paper Award of the Elsevier's Image and Vision Computing Journal for his work on human motion prediction. Additionally, he is a member of the program committees for numerous IEEE international conferences and workshops. Since 2023, he has held the position of Co-chair for the Intelligent Measurement Systems Technical Committee (TC-22) within the IEEE Instrumentation and Measurement Society.

His main research interests include the areas of applied aspects of computational intelligence for signal and image processing, computer vision, machine learning and probabilistic models applied to multiagents systems.).

Angelo Genovese (Senior Member, IEEE) received the Ph.D. degree in computer science from the Università degli Studi di Milano, Crema, Italy, in 2014. He has been a postdoctoral Research Fellow in computer science with the Università degli Studi di Milano since 2014. He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. Original results have been published in over 30 papers in international journals, proceedings of international conferences, books, and book chapters.

His current research interests include signal and image processing, three-dimensional reconstruction, computational intelligence technologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems. Dr. Genovese is an Associate Editor of the Journal of Ambient Intelligence and Humanized Computing (Springer).

C. L. Philip Chen (Fellow, IEEE) received the M.S. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 1985, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1988.

He is currently the Head of the School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, China. His current research interests include systems, cyber netics, and computational intelligence. Dr. Chen is a fellow of the American Association for the Advancement of Science (AAAS), the International Association for Pattern Recognition (IAPR), the Chinese Association of Automation (CAA), and the Hong Kong Institution of Engineers (HKIE). He was the Chair of TC 9.1 Economic and Business Systems of the International Federation of

Automatic Control from 2015 to 2017 and a Program Evaluator of the Accreditation Board of Engineering and Technology Education (ABET), USA, for computer engineering, electrical engineering, and software engineering programs. He has been the Editor-in-Chief of the IEEE TRANSACTION ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS since 2014 and an associate editor of several IEEE TRANSACTIONS.