

Bidirectional Interactive Multi-Scale Aggregation Network for Vehicle Detection in Urban Traffic

Chaojun Dong¹, Wenkang Qiu¹, Ye Li¹, Yikui Zhai¹, *Senior Member, IEEE*, Xiankun Liu², *Member, IEEE*, Chaoyun Mai¹, Hufei Zhu, Jianhong Zhou, *Member, IEEE*, Pasquale Coscia³, *Senior Member, IEEE*, Angelo Genovese³, *Senior Member, IEEE*, and C. L. Philip Chen⁴, *Life Fellow, IEEE*

Abstract—Existing UAV vehicle-detection datasets, typically captured under static and uniform illumination, fail to adequately represent the variable lighting conditions, dense traffic, and frequent occlusions observed in real-world transportation hubs. To bridge this gap, a new dataset, UAV-HubSurveillance, is introduced to capture complex vehicle interactions across urban transportation nodes under diverse environmental scenarios. Although UAV-HubSurveillance provides rich and multidimensional interaction data, it still suffers from severe occlusions and adverse weather conditions that hinder detection and identification accuracy. To address these limitations, a novel vehicle detection framework, termed bidirectional interactive multi-scale aggregation-yolo (BIMSA-YOLO), is proposed, which integrates bidirectional feature interaction with adaptive multi-scale aggregation to enhance detection robustness. First, the bidirectional shallow fusion module (BSFM) facilitates cross-resolution information exchange through a lightweight gating strategy, preserving fine-grained details of small objects. Second, the interactive deep fusion module (IDFM) reinforces contextual coherence via attention-guided cross-level semantic fusion. Third, the multi-scale adaptive aggregation module (MSAAM) dynamically aligns and integrates multi-scale features to improve robustness against scale variation. Extensive experiments conducted on the UAV-HubSurveillance dataset demonstrate that BIMSA-YOLO significantly enhances detection

performance under dynamic, occluded, and adverse-weather conditions. Specifically, the proposed model achieves an mAP_{0.5} of 63.2%, surpassing the baseline by 5.3 percentage points. Furthermore, BIMSA-YOLO also exhibits strong generalization capabilities on VisDrone and CARPK datasets. Our code and dataset are available at <https://github.com/yikuizhai/BIMSA-YOLO>

Index Terms—Uncrewed aerial vehicle, bidirectional interactive fusion, vehicle detection, multi-scale feature aggregation.

I. INTRODUCTION

IN RECENT years, rapid advancements in uncrewed aerial vehicle (UAV) hardware and communication technologies have stimulated increasing interest in deploying camera-equipped UAVs for a wide range of intelligent sensing tasks [1], [2]. These UAVs, equipped with lightweight onboard processors, are capable of real-time data acquisition and analysis, supporting applications such as intelligent city governance [3], security surveillance [4], [5], and large-scale panoramic environmental monitoring [6], [7]. In practical deployments, UAVs have demonstrated considerable value in traffic congestion management by enabling fine-grained spatiotemporal analysis of traffic density and facilitating real-time decision-making for congestion mitigation [8]. Furthermore, UAVs play an essential role in emergency response by rapidly assessing disaster-affected areas, identifying damaged infrastructure, and supporting coordinated rescue operations [9]. The high-resolution aerial imagery provided by UAV platforms enables real-time monitoring of urban development, detection of unauthorized construction activities, and optimization of traffic flow in modern smart cities [10]. Recent studies have demonstrated the effectiveness of panoramic semantic segmentation for comprehensive urban feature extraction [11], while also addressing challenges in multi-scale object detection in dynamic environments with moving vehicles [12]. Among these applications, vehicle detection based on UAV imagery is particularly important for intelligent transportation systems, as it supports accurate estimation of vehicle counts, trajectories, and motion patterns, which are essential for real-time congestion analysis, traffic signal optimization, and road safety assessment. For example, Li and Wu [13] proposed a UAV-based detection and tracking framework combining YOLOv5-OB with DeepSORT-OB, achieving 98.5% accuracy in estimating vehicle trajectories and speeds at highway interchanges,

Received 31 May 2025; revised 25 September 2025; accepted 28 November 2025. This work was supported in part by Guangdong Higher Education Innovation and Strengthening School Project under Grant 2022ZDZX1032 and Grant 2023ZDZX1029, in part by Wuyi University Hong Kong and Macao Joint Research and Development Fund under Grant 2022WGLH19, and in part by Guangdong Jiangmen Science and Technology Research Project under Grant 2220002000246 and Grant 2023760300070008390. The Associate Editor for this article was W. Ma. (Chaojun Dong, Wenkang Qiu, Ye Li, Xiankun Liu, Chaoyun Mai, Hufei Zhu, and Jianhong Zhou contributed equally to this work.) (Corresponding author: Yikui Zhai.)

Chaojun Dong, Wenkang Qiu, Ye Li, Yikui Zhai, Chaoyun Mai, and Hufei Zhu are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, Guangdong 529020, China (e-mail: cjun-dong@163.com; 173350147@qq.com; leyelee@163.com; yikuizhai@163.com; maichaoyun@foxmail.com; zhuhufeie@aliyun.com).

Xiankun Liu is with the School of Mechanical and Automation Engineering, Wuyi University, Jiangmen, Guangdong 529020, China (e-mail: lxxlml@163.com).

Jianhong Zhou is with the School of Electronics and Information Engineering, South China University of Technology, Guangzhou, Guangdong 510641, China (e-mail: jackychoulab@126.com).

Pasquale Coscia and Angelo Genovese are with the Department of Computer Science and the Dipartimento di Informatica, Università degli Studi di Milano, 20133 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, Guangdong 510006, China (e-mail: philipchen@scut.edu.cn).

Digital Object Identifier 10.1109/TITS.2025.3639483

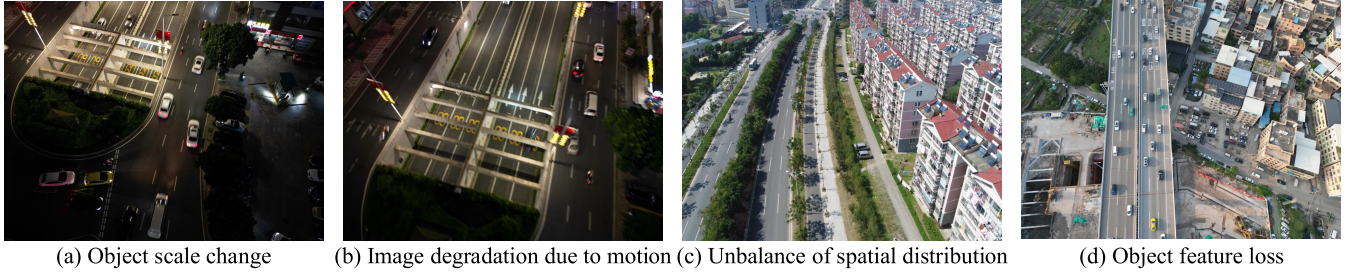


Fig. 1. The figure highlights key challenges in UAV-based vehicle detection: (a) scale variation due to differing vehicle sizes and perspectives, (b) motion blur and low-light image degradation, (c) uneven spatial distribution with vehicle clustering in specific regions, and (d) multi-source feature degradation in dense occlusion.

thereby facilitating refined traffic flow management. Similarly, Ying et al. [14] introduced a pyramid dual-pooling attention mechanism to enhance the detection performance of small vehicles in complex urban environments, contributing to more accurate congestion forecasting and faster emergency response. Compared with ground-based sensors and satellite imaging systems, UAVs provide unmatched deployment flexibility, high-resolution low-altitude perspectives, and rapid maneuverability across key traffic infrastructures such as freeway ramps, urban intersections, and bottleneck corridors. These advantages make UAVs particularly suitable for dynamic traffic monitoring, anomaly detection, and the implementation of adaptive control strategies in real-world scenarios. Nevertheless, the benefits of higher spatial resolution also introduce new challenges. The high density of visual content and the presence of numerous mobile objects increase background complexity and inter-class interference, which can significantly undermine detection robustness. Addressing these challenges requires advanced preprocessing pipelines that emphasize salient foreground targets, integrate anomaly-aware perception models, and design high-performance real-time object detectors capable of maintaining accuracy in densely populated and visually cluttered environments. However, many current approaches still rely heavily on precise model calibration and often require extensive manual validation, limiting their scalability and applicability in mission-critical scenarios such as large-scale traffic congestion monitoring and rapid-response disaster management.

Vehicle detection has been extensively studied in both aerial imagery and intelligent transportation research [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], leading to the development of several representative UAV datasets, such as UAVDT [27] and UAV123 [28]. Although these datasets consider challenging conditions, including low illumination and motion blur, they are usually collected in relatively constrained environments, which limits their ability to capture the complexity of real urban traffic. For high-precision traffic surveillance, UAVs are required to monitor transition zones between urban and rural areas over extended periods to capture dynamic changes in land use and traffic flow. However, existing datasets frequently emphasize static scenes, such as parking lots and bus terminals, whereas practical deployments demand robust perception in more complex environments,

including multi-level transportation hubs, waterfront logistics areas, and congested ramp interchanges. These scenarios are characterized by dense vehicle distributions, severe occlusions, and significant background interference, all of which reduce the robustness of detection algorithms. In addition, common preprocessing procedures may modify image statistics, thereby decreasing the generalizability of models trained on standard datasets. In real UAV applications, vehicle detection plays a vital role in traffic congestion management, emergency incident response, and road surveillance under varying illumination. For example, congestion monitoring is hindered by dense occlusions and frequent scale variations; emergency response, particularly during traffic accidents, is affected by motion blur and strong background clutter [29]; and road surveillance is challenged by sudden illumination changes or low-visibility conditions that cause missed detections [30]. UAV-based vehicle detection has also shown promise in real-time traffic management and public safety. For instance, Ke et al. [31] demonstrated the effectiveness of UAV monitoring for highway traffic and accident detection using both RGB and thermal video streams, underscoring the importance of diverse scene data for improving model robustness. Nevertheless, limitations in existing datasets, such as insufficient adaptation to poor lighting conditions and the lack of realistic occlusion scenarios, frequently result in missed detections and delayed responses, including ambulance dispatch and traffic rerouting. To overcome these limitations, the UAV-HubSurveillance dataset was constructed to capture a broad spectrum of complex urban scenes with varying illumination and dense occlusion patterns, thereby supporting more robust and practical UAV-based vehicle detection in real-world applications.

As illustrated in Fig. 1, UAV-HubSurveillance captures multiple detection challenges, including scale variation caused by differing vehicle sizes and aerial perspectives, motion blur combined with low-light degradation, spatial imbalance resulting from clustered object distributions, and feature degradation arising from dense occlusion and environmental clutter. These factors make UAV-based vehicle detection particularly challenging, especially for small targets whose visual features are easily lost or suppressed by surrounding contextual noise. Addressing these challenges requires detection models with strong multi-scale representation, semantic consistency

under occlusion, and resilience to visual degradation. The feature pyramid network (FPN) [32] aggregates hierarchical semantic information through lateral connections and top-down fusion, thereby improving object perception across varying scales. Building on this foundation, the path aggregation network (PAN) [33] introduces bidirectional paths to enhance spatial and semantic integration, while the high-level semantic filtering pyramid network (HSFPN) [34] further refines semantic selection for occluded objects. Similarly, the global fusion pyramid network (GFPN) [35] dynamically adjusts receptive fields to maintain detection performance across heterogeneous scale distributions. Inspired by these advances and tailored to UAV-specific requirements, the bidirectional interaction and multi-scale adaptation path aggregation network (BIMSA-PAN) is proposed, incorporating dynamic receptive-field modulation and inter-layer feature interaction to enhance robustness in small-object detection under visually complex conditions. Although previous fusion architectures such as FPN and PAN have contributed significantly to multi-scale detection, they are often integrated as modular components within broader pipelines. In contrast, recent frameworks such as YOLOv8 [36] and RT-DETR [37] emphasize real-time inference by simplifying detection heads and employing transformer-based or decoupled designs. YOLOv8 introduces a reparameterized decoupled detection head designed for anchor-free detection, whereas RT-DETR eliminates non-maximum suppression through set prediction enabled by transformer decoders. Nevertheless, these approaches primarily emphasize speed and accuracy trade-offs in general-purpose detection and lack specific adaptations for UAV-based scenarios, where extreme scale variation, heavy occlusion, and high spatiotemporal density prevail. To address these challenges, the proposed BIMSA-YOLO framework integrates BIMSA-PAN into a YOLO-based detection pipeline, combining bidirectional semantic aggregation with real-time anchor-free inference. This design improves localization sensitivity and precision for small objects while maintaining deployment efficiency on resource-constrained UAV platforms. Furthermore, BIMSA-YOLO explicitly mitigates detection degradation under occluded and lowlight conditions by enforcing interscale feature consistency and constructing illumination-resilient representations, thereby achieving superior performance on UAV-specific benchmarks such as UAV-HubSurveillance and VisDrone2021.

The key contributions of this paper can be summarized as follows:

- 1) The UAV-HubSurveillance dataset is proposed, comprising 8,790 images and 252,475 annotated vehicle instances captured across diverse real-world scenarios. Multi-level stereoscopic traffic hub environments are featured in the dataset, offering significantly higher scene complexity than existing UAV-based vehicle-detection datasets.
- 2) BIMSA-YOLO is proposed, combining bidirectional cross-layer interaction and multi-scale feature aggregation, enhanced by a cross-stage multi-receptive field pyramid pooling module to improve vehicle detection.

- 3) Three specialized modules are proposed for UAV-based vehicle detection: BSFM, which addresses occluded small targets through cross-resolution feature interaction; IDFM, which preserves semantic consistency via attention-guided alignment; and MSAAM, which enhances feature adaptability through deformable convolutions and dynamic multi-scale fusion.
- 4) We extensively evaluated our method on UAV-HubSurveillance, VisDrone, and CarPK datasets, demonstrating superior robustness and effectiveness over existing approaches.

II. RELATED WORK

This section reviews publicly available UAV-based vehicle detection datasets and state-of-the-art detection algorithms, emphasizing their relevance to aerial imagery. In addition, representative multi-scale feature fusion techniques are examined to improve detection performance in complex urban traffic environments.

A. Vehicle Detection Dataset From the UAV Perspective

In recent years, numerous UAV-based vehicle detection datasets have been developed to address a range of real-world challenges. For example, the AU-AIR dataset [38] contains 29,647 annotated instances collected under eight weather conditions, integrating 1920×1080 RGB imagery with GPS and IMU data to support multimodal dynamic monitoring. DroneSURF [39] focuses on urban traffic hubs and includes 12,470 high-resolution (4000×3000) dual-view images annotated for abnormal vehicle trajectories. Okutama-Action [40] comprises 43 videos recorded at 30 fps, capturing an average of 7.2 moving vehicles per frame with additional motion blur annotations. UAV-DeOT [41] consists of 47 video sequences collected in structurally complex environments such as overpasses, annotated with continuous vehicle identities and an average occlusion rate of 62%. UAVDT [42] provides 80,000 frames labeled with five levels of occlusion information for traffic surveillance tasks. VisDrone [43] covers ten vehicle categories with object sizes ranging from 5 to 400 pixels across diverse urban scenes. DroneVehicle [44] offers 28,439 aligned RGB-infrared image pairs for vehicle detection under low-light conditions. LH-UAV-Vehicle [45] contributes 18,663 images captured at altitudes between 250 and 400 meters, annotated with 180,358 vehicle instances to support small-object detection. CARPK [45] focuses on fixed-angle UAV surveillance over parking lots, consisting of 1,448 images densely annotated with cars.

Despite these contributions, existing datasets still exhibit important limitations in modeling realistic urban traffic scenarios [14]. Most benchmarks focus on planar road environments and fail to represent the structural complexity of multi-level traffic systems such as elevated highways and underpasses. For instance, AU-AIR and DroneVehicle emphasize environmental variability but do not account for vertical scene diversity. DroneSURF facilitates trajectory-based studies but lacks annotations for occlusion levels and localized microclimatic effects. Okutama-Action provides high frame rates yet

TABLE I
COMPARISON OF AERIAL VEHICLE DETECTION DATASETS

Dataset Categories	Balanced Scenes	Night Platform	UAV Scenes	Complex Scenes	Images (10^3)	Instances (10^3)	Resolution (pixels)	Scenes Types
AU-AIR [38]	✓		✓		29.6	8.0	1920×1080	3
DroneSURF [39]			✓	✓	12.5	5.0	4000×3000	2
Okutama [40]	✓			✓	4.5	32.4	1920×1080	1
UAV-DeOT [41]			✓	✓	14.1	94.3	1280×720	3
UAVDT [42]			✓	✓	80.0	840.0	1080×540	2
VisDrone [43]		✓	✓	✓	10.2	179.0	2000×1500	4
DroneVehicle [44]	✓		✓	✓	28.4	452.5	840×712	2
LH-UAV-Vehicle [14]			✓	✓	18.6	180.3	640×640	4
CARPK [45]				✓	1.4	90.0	720×360	1
UAV-HubSurveillance (ours)	✓	✓	✓	✓	8.8	252.5	1280×1280	4

TABLE II
OCCLUSION, CLASS IMBALANCE AND ALTITUDE OF AERIAL DATASETS

Dataset	Occlusion	Class Imbalance	UAV Altitude (m)
AU-AIR [38]	0.23	1.09	25 ± 5
DroneSURF [39]	0.31	1.25	30 ± 10
Okutama [40]	0.18	1.02	15 ± 3
UAV-DeOT [41]	0.40	1.35	40 ± 12
UAVDT [42]	0.35	1.42	60 ± 15
VisDrone [43]	0.28	1.18	45 ± 20
DroneVehicle [44]	0.22	1.05	50 ± 10
LH-UAV-Vehicle [14]	0.26	1.20	35 ± 8
CARPK [45]	0.15	1.00	40 ± 5
UAV-HS(ours)	0.37	1.09	45 ± 20

offers limited viewpoints, making it insufficient for capturing multi-angle traffic behavior. UAV-DeOT incorporates LiDAR sensing but suffers from illumination instability ranging from 200 to 2000 lux. UAVDT and VisDrone consider object scale variations but rely exclusively on bounding box annotations without modeling dynamic motion or semantic behavior. LH-UAV-Vehicle operates at a fixed resolution and lacks occlusion and temporal information, which restricts its applicability in cluttered or low-light environments. CARPK, although densely annotated, includes only a single vehicle category and lacks spatial diversity. These deficiencies highlight the need for a more comprehensive dataset that incorporates structural diversity, altitude variation, occlusion information, and class imbalance. To address this gap, UAV-HubSurveillance is proposed as a benchmark specifically designed to capture layered urban overpasses, varied vehicle densities, and motion-induced degradation, while providing category-level balance and detailed occlusion statistics. As summarized in Table I and Table II, this dataset offers a more realistic foundation for evaluating UAV-based vehicle detection algorithms.

B. Vehicle Detection

The objective of vehicle detection is to accurately localize vehicles within an image and classify them into predefined categories. In recent years, the rapid advancement of deep learning based object detection techniques has significantly promoted aerial vehicle detection, establishing it as a fundamental task in remote sensing image analysis. Representative methods, such as YOLOv3 [46] and Faster R-CNN [47], have demonstrated strong performance in both natural and aerial domains. As a one stage anchor based detector, YOLOv3 performs multi scale predictions through the integration of a feature pyramid network (FPN), effectively addressing scale variation. In contrast, Faster R-CNN employs a two stage architecture, where a region proposal network (RPN) first generates candidate bounding boxes, followed by region of interest pooling and subsequent refinement through classification and regression heads. To improve detection in dynamic urban traffic conditions, PP-YOLO [48] extends YOLOv3 by incorporating a multi scale feature fusion layer together with FPN and PAN structures, thereby enhancing semantic alignment and enabling more effective detection of small and densely distributed objects. YOLOX [49] further improves coverage of small objects by directly predicting bounding boxes from backbone features while integrating an adaptive spatial feature fusion mechanism that leverages channel attention and spatial pyramid encoding to strengthen robustness across scales. For oriented object detection in aerial imagery, RoITransformer [50] aligns proposals with object orientation, whereas Oriented R-CNN [51] refines rotated bounding box prediction within Faster R-CNN, achieving higher accuracy with reduced parameter complexity. Recent studies have also investigated challenges including motion blur, background clutter, and low visibility in UAV imagery. For instance, the edge guided perceptual network for infrared small target detection (EGPNet) [52] enhances feature representation through contour aware attention and multi scale spatial encoding, which is particularly advantageous for detecting small and low resolution vehicle targets in UAV imagery. Similarly, the

multibranch mutual guiding network (MMGN) [53] improves discriminability under cluttered backgrounds by applying inter branch supervision. From the perspective of robustness, SNIS [54] and FAMM [55], originally designed for forgery detection, employ noise resilient feature separation and channel based muscle motion modeling respectively, thereby providing inspiration for resilience against compression artifacts and noise in UAV visual data. Although these frameworks have achieved significant progress, complex environmental conditions such as dense occlusion and dynamic scale transitions continue to present challenges for traditional anchor based methods.

C. Multi-Scale Feature Aggregation

Multi scale feature aggregation plays a crucial role in vehicle detection by integrating image features across multiple resolutions. This strategy enables models to capture objects of varying sizes while maintaining spatial and semantic consistency, particularly in complex urban environments. Classical designs such as PANet introduced bottom up pathways to enhance the retention of low level details. GFPN further advanced this concept by incorporating a gating mechanism to adaptively recalibrate feature weights across scales, whereas AFPN improved the recall of occluded vehicles through direct interaction between non adjacent levels [56]. These methods established the foundation for robust hierarchical fusion in deep learning architectures. Subsequent research has focused on structural refinements to enhance detection accuracy. TridentNet [57] employed multi branch dilated convolutions within a unified backbone to learn scale adaptive receptive fields. NAS FPN [58] utilized neural architecture search to identify optimal cross scale connectivity, resulting in asymmetric skip pathways and improved performance. M2Det [59] adopted a U shaped encoder decoder architecture to bridge low level and high level features, while BiFPN [60] introduced a weighted bidirectional fusion mechanism with efficient normalization. HRFPN [61] preserved high resolution representations through parallel subnetworks, enabling full scale interaction and minimizing feature loss. Attention based modules such as CBAM [62], SE [63], and ECA [64] have also been applied to enhance channel sensitivity; however, they often lack adaptability to dynamic traffic scenarios and demonstrate limited robustness under occlusion or variable illumination. While state of the art methods such as BiFPN and HRFPN emphasize efficient bidirectional fusion or hierarchical interaction, the proposed bidirectional shallow fusion module (BSFM), interactive deep fusion module (IDFM), and multi scale adaptive aggregation module (MSAAM) are designed to address fine grained spatial inconsistencies and semantic misalignments across feature levels. Unlike BiFPN, which primarily balances contribution weights through learned scalars, BSFM introduces spatially aware gating with directional refinement. In comparison with the staged pyramid interaction in HRFPN, IDFM models inter level dependency using content aware dynamic modulation, resulting in improved information preservation across depth. MSAAM further captures fine scale contextual cues through joint spatial channel reasoning, outperforming generic fusion blocks in environments with heavy

occlusion. Furthermore, RS SE enhances conventional channel attention by incorporating residual spatial encoding, which is absent in SE, CBAM, and ECA. Essentially, it can be regarded as a squeeze and excitation mechanism augmented with residual connections, allowing identity mapping to be preserved while amplifying informative channels. Empirical evaluations confirm that RS SE achieves superior robustness under illumination variation and dense object occlusion, thereby offering stronger generalization in UAV based traffic scenarios. Despite these advances, critical challenges remain in UAV based traffic monitoring [65], including computational redundancy, weak spatial interaction modeling across hierarchical layers, and vulnerability to motion blur and visual noise.

III. UAV-HUBSURVEILLANCE DATASET

In this section, a comprehensive overview of the UAV-HubSurveillance dataset is provided, including image pre-processing, annotation methods, and dataset statistics and properties.

A. Dataset Description

The UAV HubSurveillance dataset was specifically constructed to address the inherent limitations of existing aerial traffic benchmarks, including VisDrone. As shown in the left portion of Fig. 2, VisDrone presents several constraints in terms of environmental diversity and observational flexibility. These include a predominantly top down surveillance perspective confined to planar urban environments, simplified low light scenes with limited visual complexity, homogeneous traffic flows restricted to flat road surfaces, and a narrow field of view resulting from monocular camera setups. In contrast, the right panel of Fig. 2 illustrates the richer and more varied scene compositions captured by UAV HubSurveillance. The dataset includes multi view aerial footage from elevated highway networks, logistic zones influenced by complex shadow occlusions, industrial freight corridors characterized by heavy duty vehicular motion, and multi lane intersections integrated with exclusive bus lanes. By systematically modeling perspective variations introduced by flight altitude, scene specific illumination changes, and heterogeneous traffic compositions, UAV HubSurveillance more accurately reflects the structural and semantic variability of real world urban mobility systems. In addition to enhanced scene diversity, UAV HubSurveillance incorporates detailed quantitative annotations that address key challenges in aerial vehicle detection. As summarized in Table II, the dataset exhibits a high occlusion rate of 0.37, reflecting the complexity of urban overpass environments with frequent partial and heavy vehicle occlusions. The class imbalance metric of 1.09 indicates a relatively balanced distribution across four major vehicle categories: cars, vans, trucks, and buses, which supports fairer evaluation and robust model training. Furthermore, UAV flight altitudes range from 10 to 120 meters, with an average altitude of approximately 45 meters and a standard deviation of 20 meters, thereby capturing multi scale perspectives that simulate realistic operational scenarios. This design philosophy is consistent with recent dataset driven optimization strategies in other domains,

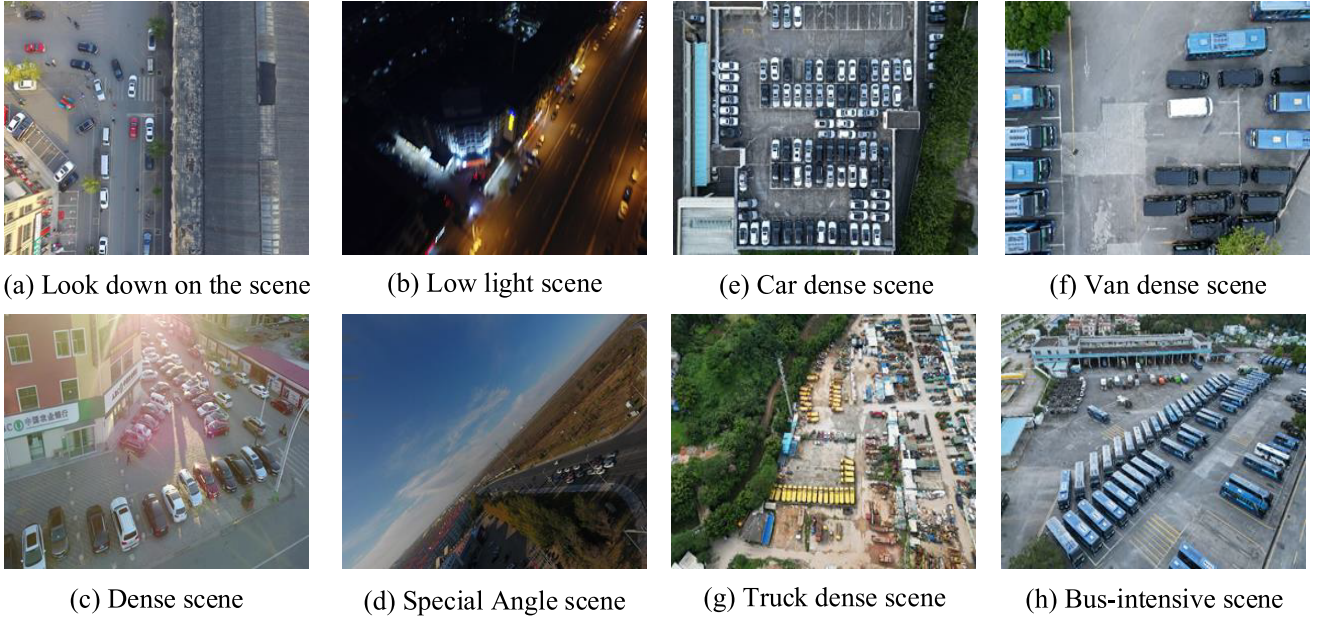


Fig. 2. The left panel illustrates four key environmental challenges in urban aerial observation: altitude-dependent perspective shifts (a), low-light image degradation (b), high-density vehicle clustering (c), and non-canonical viewpoint distortions (d). The right panel presents scenario-specific vehicle concentration patterns, including passenger car predominance in parking areas (a), logistics van aggregation (b), industrial truck formations (c), and public transit bus prioritization (d). This dual-panel visualization demonstrates the UAV-HubSurveillance dataset's ability to capture real-world traffic dynamics through multi-layer scene acquisition, integrating altitude-adaptive monitoring with class-balanced vehicle distributions to support robust urban mobility analysis.

such as the large scale benchmark and model improvements introduced for optimized YOLOv8 in fallen person detection [66], where rich scene diversity and carefully curated annotations directly enhance detector robustness. Collectively, these characteristics establish UAV HubSurveillance as a comprehensive and structurally diverse benchmark, enabling thorough assessment of detection algorithms under varied and challenging urban traffic conditions.

B. Images Preprocessing

In this study, three essential preprocessing steps were applied to the UAV HubSurveillance dataset to optimize it for vehicle detection. These steps included radiometric correction, geometric correction, and image pruning. Radiometric correction was performed to standardize lighting conditions across all images, thereby reducing environmental inconsistencies such as variations in sunlight, shadow patterns, and glare, and ensuring that the model focused on vehicle features rather than illumination artifacts. Geometric correction was implemented using ground control points and camera calibration parameters to eliminate perspective distortions and preserve accurate spatial relationships across all frames. Image pruning involved the removal of irrelevant regions, including distant background and sky areas, which reduced computational load and directed model attention toward vehicle relevant content. In combination with a fixed resolution of 1280 by 1280 pixels, these steps improved image quality, reduced geometric and radiometric distortion, and enhanced the signal to noise ratio. Consequently, the training process converged more rapidly, and detection accuracy and robustness were significantly

improved, making the dataset highly suitable for uncrewed aerial vehicle based vehicle detection tasks.

C. Annotation Method

In this study, the UAV HubSurveillance dataset was meticulously annotated to enhance vehicle detection in uncrewed aerial vehicle imagery. Building on the VisDrone dataset, targeted improvements were implemented to address challenges encountered in dense urban scenes and low light conditions. The PaddleLabel tool, an interactive annotation framework with custom adaptations for aerial imagery, was employed to enable efficient and precise labeling of complex traffic scenarios characterized by occlusions and variable object distances. Real time adjustments for low light conditions further ensured annotation accuracy. A rigorous quality control process resolved discrepancies through peer review, thereby maintaining consistency and precision. Annotators followed comprehensive guidelines to address challenging cases, including partially visible or overlapping vehicles. These procedures produced a high quality and reliable dataset for training robust detection models capable of operating effectively in dynamic urban environments.

D. Statistics and Attributes

To enable effective model training and rigorous performance evaluation, the UAV HubSurveillance dataset was partitioned into training, validation, and testing subsets with a ratio of 75:5:20. In total, the dataset contains 252,475 annotated vehicle instances distributed across four categories, namely

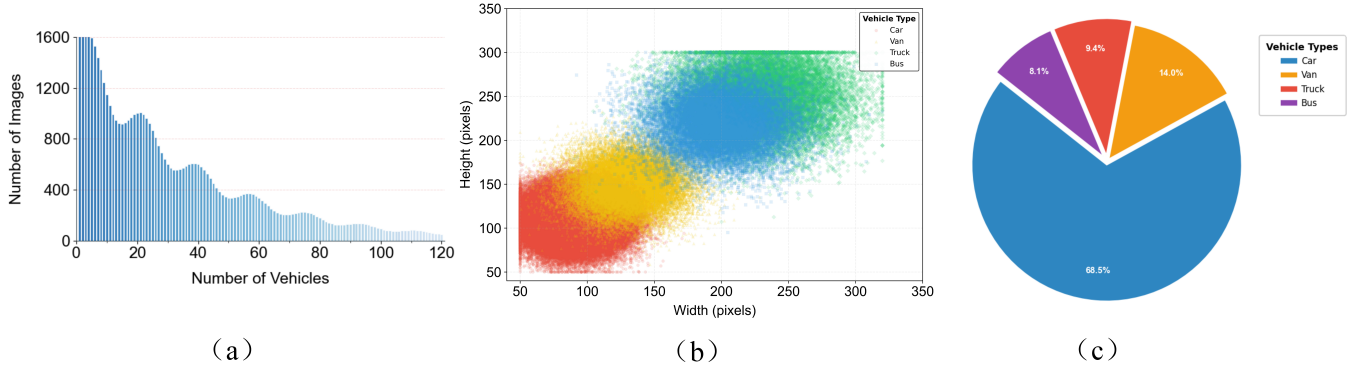


Fig. 3. (a) The distribution of all vehicle instances. (b) The scatter plot of the width and height of all vehicle instances. (c) The proportion of pixels occupied by different vehicle instances of an image in the foreground.

TABLE III
VEHICLE CATEGORY DISTRIBUTION ACROSS UAV-HUBSURVEILLANCE, VISDRONE, AND CARPK

Subset	UAV-HubSurveillance					VisDrone					CARPK	
	Total	Car	Van	Truck	Bus	Total	Car	Van	Truck	Bus	Total Images	Car Instances
Training	176,682	121,149	24,676	16,567	14,290	104,785	76,401	13,428	8,720	6,236	1,013	31,842
Validation	50,481	30,201	6,113	4,102	3,372	28,131	19,842	3,568	2,287	2,434	217	6,867
Testing	25,312	21,110	4,502	2,980	2,720	16,884	12,012	2,110	1,350	1,412	218	6,961

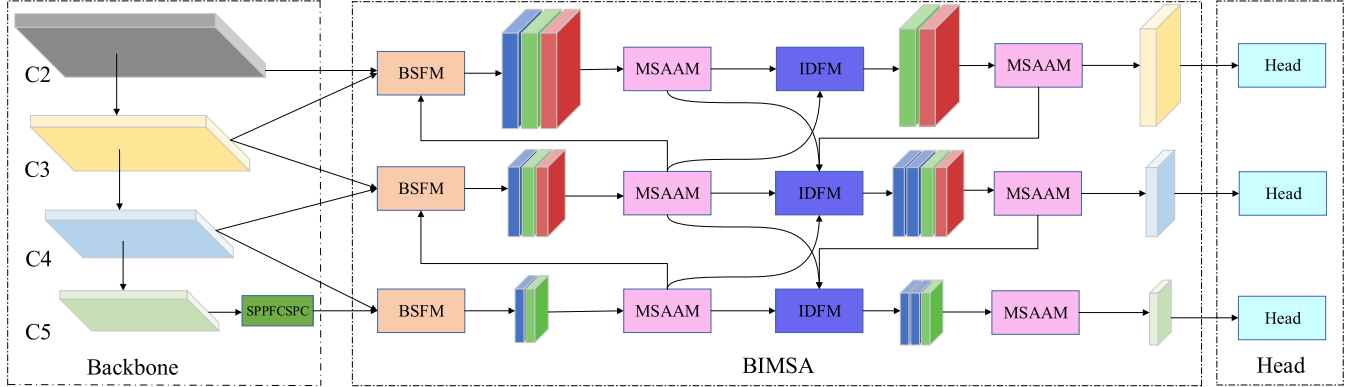


Fig. 4. Overall architecture of the proposed BIMSA-YOLO framework. The backbone extracts hierarchical feature maps from stages C_2 to C_5 , where $C_2 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$ preserves fine-grained spatial details, $C_3 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 2C}$ encodes mid-level structures, $C_4 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 4C}$ captures high-level semantics, and $C_5 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 8C}$ provides the most abstract representations. The C_4 feature is first processed by SPPFCSPC to expand the receptive field and integrate multi-scale context, after which BSFM aggregates information from adjacent layers through resolution alignment and efficient fusion. The fused features are then refined via IDFM, which enhances cross-scale interactions and semantic consistency, and further processed by MSAAM to adaptively align spatial and semantic representations. The final outputs C'_n are fed into three detection heads at $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$, enabling accurate object localization and classification across multiple scales.

cars (173,259), vans (35,291), trucks (23,650), and buses (20,275). Specifically, the training subset consists of 6,583 images with 176,682 instances, the validation subset includes 598 images with 50,481 instances, and the testing subset comprises 1,526 images while preserving the overall category distribution. The training subset covers diverse urban traffic scenarios such as congested arterial roads, high density parking zones, and mixed category traffic flows, thereby improving the

generalization ability of the model. The validation subset is used for hyperparameter tuning, whereas the testing subset is reserved for final benchmarking, both containing complex scenes that reflect real world deployment conditions. As summarized in Table III, this proportional partitioning not only ensures a balanced representation of all categories across subsets but also provides category level statistics of the VisDrone and CARPK datasets. In addition, Fig. 3 illustrates

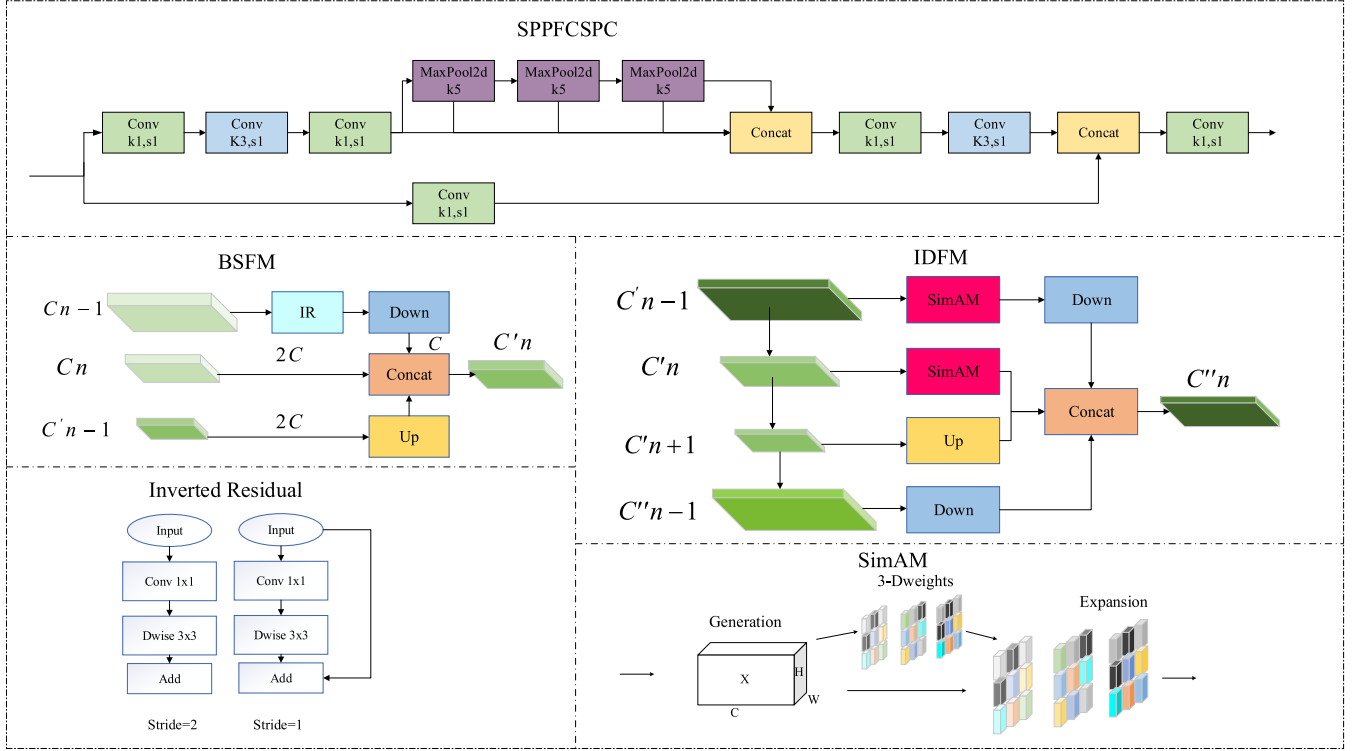


Fig. 5. Overall architecture and functional details of the BSFM, IDFM, and SPPFCSPC components within the proposed BIMSA-YOLO framework. The BSFM receives multi-resolution feature maps from C_{n-1} , C_n , and C'_{n+1} , which correspond to spatial resolutions of $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, and $\frac{H}{16} \times \frac{W}{16}$, respectively. These inputs are first brought into spatial correspondence via resolution alignment, implemented through precise downsampling or upsampling operations, ensuring that structural integrity and semantic consistency are preserved during transformation. Following alignment, depthwise separable convolutions are employed to reduce computational overhead while maintaining feature diversity, and 1×1 convolutions are applied to adapt channel dimensions and facilitate efficient information blending. The fused representation, denoted as C'_n , is subsequently delivered to the IDFM, which integrates it with C'_{n-1} , C'_{n+1} , and C'_n . All incoming features in the IDFM are uniformly rescaled to a common resolution of $\frac{H}{8} \times \frac{W}{8}$, enabling direct element-wise correspondence across scales. An attention weighting mechanism is applied to selectively emphasize regions with high task-relevance, thereby enhancing discriminative spatial focus. The aggregated features are concatenated along the channel dimension and further refined via convolutional fusion, resulting in C''_n with significantly enriched semantic representation and spatial granularity. The SPPFCSPC operates on high-level features by embedding cross-stage partial connections within a spatial pyramid pooling structure, allowing the simultaneous capture of fine-to-coarse contextual cues across multiple receptive field sizes. This design improves multi-scale feature aggregation without incurring excessive computational cost, ensuring that the final feature maps retain both strong global context and localized detail, thereby providing robust inputs for the subsequent detection heads.

instance level characteristics of UAV HubSurveillance: (a) distribution of vehicle categories, (b) scatter plot of object width versus height, and (c) proportion of foreground pixels per class.

IV. METHODS

A novel framework, termed bidirectional interactive multi-scale aggregation-yolo (BIMSA-YOLO), is proposed, which integrates the BIMSA module with the spatial pyramid pooling-fast module combined with cross stage partial convolution (SPPFCSPC). The BIMSA module consists of three components: BSFM, IDFM, and MSAAM. Information is fused across all layers to strengthen the network's ability to focus on vehicle objects. The overall architecture of the proposed method is illustrated in Fig. 4.

A. BIMSA Frame Structure

Fig. 4 presents the overall BIMSA YOLO architecture. The backbone extracts hierarchical features by progressively reducing spatial resolution and increasing channel depth, beginning

with a convolutional layer at C_1 that halves the resolution and expands the channels to 64. Stages C_2 to C_5 follow this pattern, increasing the channel depth from 128 to 512. To preserve spatial detail while enriching semantic information, intermediate C_{2f} modules refine the feature maps. The deepest stage C_5 incorporates the SPPFCSPC module to capture multi scale representations through pooling and cross stage connections. These hierarchical features are then forwarded to BSFM (Fig. 5), which performs bottom up multi scale fusion across $1/4$, $1/8$, and $1/16$ resolutions. Each fusion block consists of an inverted residual unit composed of a 1×1 convolution for channel expansion, a depthwise separable convolution to capture spatial patterns with low computational cost, and a final 1×1 convolution to project back to the desired dimension, with shortcut paths facilitating information preservation [67]. Formally, given input features $C_{n-1}, C_n, C_{n+1} \in \mathbb{R}^{H \times W \times C}$ from the backbone, BSFM computes the enhanced representation C'_n through a sequence of up sampling $U(\cdot)$, down sampling via 3×3 strided convolutions (denoted as Down), channel adjustment with 1×1 convolution (denoted as C), and non linear activation using the SiLU function δ . This design ensures

that semantic and spatial cues are effectively integrated before being transmitted to subsequent modules. The output after the BSFM operation is expressed as follows:

$$C'_n = \text{Concat} \left(\begin{array}{c} \delta(C(\text{Down}(C_{n-1}))), \\ \text{DepthwiseConv}(\text{PointConv}(C_n)), \\ U(C'_{n+1}) \end{array} \right) \quad (1)$$

These operations generate feature maps optimized for detecting objects of varying sizes and orientations, which is essential for multi scale detection in UAV based scenarios. After refinement by BSFM and MSAAM, the features are forwarded to IDFM (Fig. 5), which applies a cross scale fusion strategy to aggregate multi level information and enhance the detection of small and overlapping objects. Specifically, IDFM fuses the shallow high resolution feature map C'_{n-1} (1/4 resolution), the shallow low resolution feature C'_{n+1} (1/16 resolution), the current layer feature C'_n (1/8 resolution), and the previous low resolution feature C''_{n-1} (1/16 resolution), thereby enriching the output C_4 with both spatial precision and semantic consistency. Dense inter layer connections preserve critical structural information, while spatial attention suppresses irrelevant background interference. To further enhance feature discriminability, IDFM integrates the simple attention module (SimAM) [68], a parameter free attention mechanism that jointly models spatial and channel wise saliency by estimating neuron importance through a three dimensional energy function. This function emphasizes neurons exhibiting strong local contrast and semantic relevance without introducing additional parameters, thereby improving semantic density and reducing noise in the feature maps. Related approaches integrating dual attention for robust feature representation have also been reported in recent studies on consumer imaging devices [69]. The resulting representations maintain consistent resolutions (1/8 or 1/16) and provide robust inputs for downstream detection modules. Here, X denotes the input feature map and $\sigma(y)$ denotes the sigmoid activation applied to the attention response y . The next section introduces MSAAM, which further refines multi scale semantic aggregation. The output after applying IDFM is expressed as follows:

$$\text{SimAM}(x) = x \times \sigma(y) \quad (2)$$

$$\begin{cases} \hat{C}_{n-1} = C(\text{Down}(\text{SimAM}(C'_{n-1}))) \\ \hat{C}_n = \text{SimAM}(C'_n) \\ \hat{C}_{n-1}'' = \delta(C(\text{Down}(C''_{n-1}))) \\ \hat{C}_{n+1} = C(\text{Up}(C'_{n+1})) \end{cases} \quad (3)$$

$$C''_n = \text{Concat}(\hat{C}_{n-1}, \hat{C}_n, \hat{C}_{n-1}'', \hat{C}_{n+1}) \quad (4)$$

B. Multi-Scale Adaptive Aggregation Module

In the BIMS architecture, the MSAAM addresses the challenges of multi scale feature aggregation and vehicle detection across varying object sizes, which are common in UAV based traffic scenarios. As illustrated in Fig. 6, MSAAM processes the input feature I through two parallel pathways. The first pathway applies a 1×1 convolution, followed by feature concatenation and refinement through the RS-SE module.

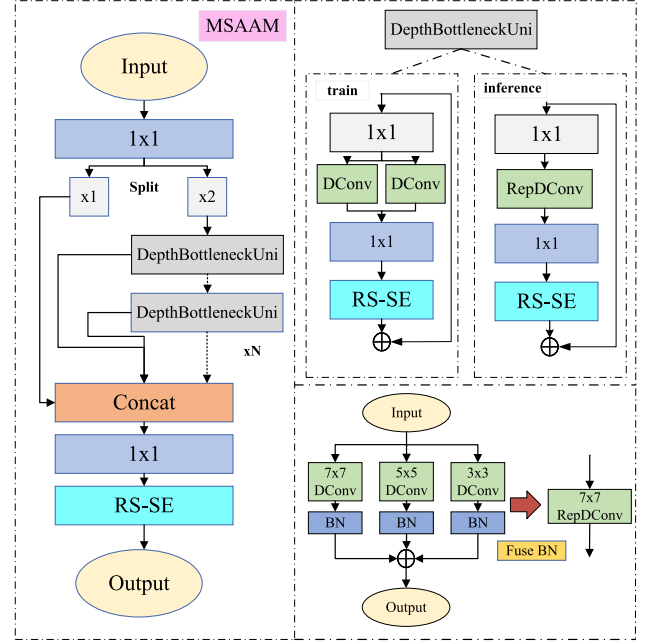


Fig. 6. MSAAM architecture and component level processing. The left panel shows the per path deep bottleneck unit that merges depthwise separable convolution and standard 3×3 convolution streams with an RS-SE attention block. Input features are first projected by a 1×1 convolution and split into parallel branches. Each branch applies its designated convolutional operators and residual connections then the branch outputs are concatenated and projected to restore the desired channel dimensionality. This bottleneck arrangement preserves fine spatial detail while constraining parameter growth and computational cost. The central panel illustrates the dynamic receptive field adaptation mechanism used for cross path feature interaction. Intermediate feature streams are adaptively rescaled and aggregated so that information flows across scales with element wise correspondence, which improves robustness to object scale variation and reduces interference from cluttered backgrounds. The right panel depicts the RepDConv based multi scale fusion unit. During training the module explicitly includes large kernel branches such as 5×5 and 7×7 convolutions together with batch normalization and channel concatenation to enrich multi scale representations. At inference the large kernel branches are reparameterized into a single RepDConv kernel to reduce runtime overhead while retaining the representational benefits of multi scale kernels. Together these components perform attention guided, multi scale alignment and fusion that enhances localization and discrimination of small and occluded vehicle instances in aerial imagery.

RS-SE integrates residual learning into the squeeze and excitation mechanism by embedding global average pooling and channel wise recalibration within a residual structure, thereby preserving identity mappings while amplifying semantically relevant channels. The second pathway enhances semantic richness by passing I through N inverted bottleneck units that expand feature dimensionality. This is followed by the RepDConv module, which adaptively adjusts receptive field sizes through reparameterizable convolutional branches. RepDConv can be regarded as a training and inference decoupled convolutional design. During training, multiple convolution branches with different kernel sizes jointly learn scale adaptive features, while at inference these branches are merged into a single equivalent kernel, ensuring efficiency without sacrificing representational power. Specifically, RepDConv applies parallel convolutions with kernel sizes 3×3 , 5×5 , and 7×7 during training, which are later fused into a single equivalent kernel at inference, thereby enabling dynamic scale adaptation without

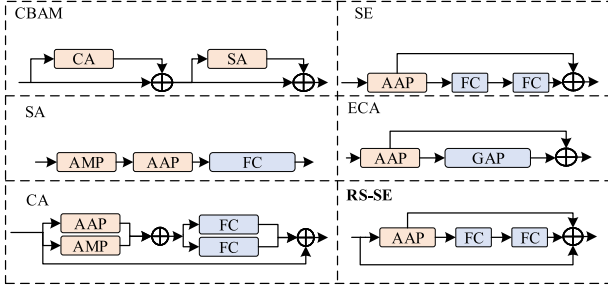


Fig. 7. This figure illustrates several channel and spatial attention modules, including RS-SE, CBAM, SE, ECA, SA, and CA. Among these modules, RS-SE is specifically designed to integrate receptive field adaptability with spatial context modeling. It enhances feature representation by dynamically recalibrating channel-wise responses while preserving critical spatial information. The figure highlights the core operations and architectural components of each attention mechanism, demonstrating that RS-SE differs from traditional methods such as SE and ECA by embedding spatial cues and receptive field awareness.

introducing additional computational cost. A subsequent 1×1 convolution compresses the feature map, preserving critical spatial information. Swish activation is consistently applied to enhance non linear representation, particularly improving the detection of small or occluded vehicles in cluttered scenes. The final output retains the input resolution while integrating semantically rich and spatially robust multi scale features, ensuring compatibility with downstream detection heads. Fig. 7 provides a comparative illustration of RS-SE and other attention mechanisms, including CBAM, SE, ECA, SA, and CA, highlighting their core architectures and operational differences. Let K_{k_i} and B_{k_i} denote the convolution weights and biases corresponding to kernel size $k_i \times k_i$, where $k_i \in \{3, 5, 7\}$, and let m represent the number of convolution branches. A similar dual attention fusion strategy was also explored in [70] for efficient deepfake detection on constrained devices, demonstrating the effectiveness of attention based multi branch re parameterization. The resulting output feature map O is:

$$F_i = DConv_{k_i}(I) + B_{k_i}, \quad i = 1, 2, \dots, m \quad (5)$$

$$F = C^{1 \times 1}(\text{Concat}(F_1, F_2, \dots, F_m)) \quad (6)$$

$$RS - SE = F + SE(F) \quad (7)$$

C. BIMS Integrated Into Yolo-BIMS-Yolo

In this chapter, BIMS-YOLO is introduced, which integrates the BIMS and SPPFCSPC modules to enhance multi-scale feature fusion and spatial representation for UAV-based vehicle detection. The proposed framework improves detection performance in scenarios characterized by varying object sizes and complex environments. By leveraging real-time efficiency and advanced fusion strategies, BIMS-YOLO achieves superior accuracy, particularly in detecting small, occluded, and overlapping vehicles in aerial imagery. In this framework, the conventional FPN is replaced with BIMS, which consists of BSFM, IDFM, and MSAAM. Specifically, BSFM preserves finegrained spatial details that are critical for

small object detection, IDFM aggregates cross-scale features to improve recognition of overlapping or occluded vehicles, and MSAAM dynamically adjusts receptive fields and applies channel wise attention to enhance robustness under motion blur and lighting variation. This hierarchical fusion mechanism enables the model to effectively capture both large and small objects in dynamic UAV scenarios. Furthermore, the original SPPF module is replaced by SPPFCSPC, which incorporates Cross-Stage Partial Connections to better process multi-scale features while retaining detailed spatial information. These design choices are supported by recent advances in lightweight multi-scale attention networks [71], and the structure of SPPFCSPC used in BIMS-YOLO is illustrated in Figure 4.

V. EXPERIMENTS

This section presents a comprehensive evaluation of the proposed BIMS and BIMS-YOLO across the UAV-HS, VisDrone, and CARPK datasets. Extensive comparisons with state-of-the-art detectors and feature fusion networks are conducted to verify the superiority of the proposed architecture in terms of accuracy and efficiency. In addition, ablation studies are performed to analyze the contributions of individual modules within BIMS and BIMS-YOLO, thereby validating the effectiveness of the overall design.

A. Experimental Setup

All experiments were conducted on a Windows 10 workstation with an Intel Core i9-11900K CPU and an NVIDIA RTX 3090 GPU (24 GB VRAM). Models were trained for 300 epochs using stochastic gradient descent (SGD) with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 5×10^{-4} . The learning rate was linearly decayed to zero over the training period. A batch size of 16 was used throughout. Mosaic augmentation was applied during the first 280 epochs and disabled in the final 20 to ensure training stability. To ensure transparency, the computational profile of the full BIMS-YOLO model is reported: 6.7M parameters and 27.4 GFLOPs. Under the specified environment, each epoch requires approximately 3 minutes, resulting in a total training time of around 15 hours. Peak GPU memory usage reaches 16.0 GB. During inference, the model achieves 67 FPS with a batch size of 16, corresponding to a latency of approximately 14.9 ms per image. These results reflect the model's efficiency and deployment practicality.

B. Model Evaluation Metrics

The performance of deep learning models in object detection is evaluated using four key metrics: number of parameters (Param), floating-point operations (FLOPs), mean Average Precision at IoU 0.5 (mAP₅₀), and frame rate (FPS).

C. Ablation Experiment

1) *BIMS Ablation Experiment*: To comprehensively assess the effectiveness of the proposed BIMS framework, which integrates BSFM, IDFM, and MSAAM, extensive ablation studies were conducted on both the VisDrone and UAV-HS

TABLE IV
MODULE-WISE ABLATION RESULTS OF BIMSA-YOLO ON VisDRONE AND UAV-HS

Method	BSFM	MSAAM	IDFM	SPPFCSPC	mAP _{0.5} (%)	mAP _{0.75} (%)	mAP _{0.5:95} (%)	Params (M)	Latency (ms)
VisDrone									
CSPDarknet					57.8 ± 0.5	45.0 ± 0.5	38.8 ± 0.5	3.1	9.5
CSPDarknet	✓				58.4 ± 0.5	46.0 ± 0.5	39.4 ± 0.4	3.8	10.1
CSPDarknet		✓			58.2 ± 0.4	45.8 ± 0.4	39.2 ± 0.4	3.7	10.0
CSPDarknet			✓		58.3 ± 0.4	45.9 ± 0.4	39.3 ± 0.4	3.8	10.2
CSPDarknet	✓	✓			59.6 ± 0.4	47.5 ± 0.3	40.9 ± 0.3	4.5	10.7
CSPDarknet	✓		✓		59.4 ± 0.3	47.3 ± 0.3	40.7 ± 0.3	4.4	10.6
CSPDarknet		✓	✓		59.5 ± 0.3	47.4 ± 0.3	40.8 ± 0.3	4.5	10.7
CSPDarknet	✓	✓	✓		61.2 ± 0.3	48.8 ± 0.3	41.5 ± 0.2	5.3	11.3
CSPDarknet	✓	✓	✓	✓	62.5 ± 0.3	50.2 ± 0.2	42.7 ± 0.2	6.7	12.1
UAV-HS (ours)									
CSPDarknet					57.9 ± 0.4	45.3 ± 0.4	39.3 ± 0.4	3.1	9.7
CSPDarknet	✓				58.7 ± 0.3	46.2 ± 0.3	39.7 ± 0.3	3.8	10.3
CSPDarknet		✓			58.6 ± 0.3	46.1 ± 0.3	39.6 ± 0.3	3.7	10.2
CSPDarknet			✓		58.6 ± 0.3	46.1 ± 0.3	39.6 ± 0.3	3.8	10.3
CSPDarknet	✓	✓			61.5 ± 0.3	49.0 ± 0.3	41.8 ± 0.2	4.5	10.9
CSPDarknet	✓		✓		61.4 ± 0.3	48.9 ± 0.3	41.7 ± 0.2	4.5	10.8
CSPDarknet		✓	✓		61.4 ± 0.3	48.9 ± 0.3	41.7 ± 0.2	4.5	10.8
CSPDarknet	✓	✓	✓		62.3 ± 0.2	49.6 ± 0.2	42.5 ± 0.2	5.3	11.4
CSPDarknet	✓	✓	✓	✓	63.2 ± 0.2	50.8 ± 0.2	43.4 ± 0.2	6.7	12.6

TABLE V
ABLATION STUDY OF MSAAM ON VisDRONE AND UAV-HS

Method	DConv	RepConv	RS-SE	mAP _{0.5} (%)
VisDrone				
CSPDarknet				56.9 ± 0.4
CSPDarknet	✓			57.6 ± 0.3
CSPDarknet	✓	✓		58.2 ± 0.3
CSPDarknet	✓	✓	✓	58.3 ± 0.2
UAV-HS (ours)				
CSPDarknet				57.2 ± 0.4
CSPDarknet	✓			57.9 ± 0.3
CSPDarknet	✓	✓		58.6 ± 0.3
CSPDarknet	✓	✓	✓	58.7 ± 0.2

datasets. The module-wise results are reported in Table IV, with CSPDarknet serving as the backbone. Each individual component consistently improved the baseline detector, yielding mAP_{0.5} gains ranging from approximately 0.4% to 1.2%. When two modules were combined, the improvements were further amplified, highlighting their complementary characteristics. Notably, the combination of BSFM and MSAAM achieved mAP_{0.5} scores of 59.6% on VisDrone and 61.5% on UAV-HS, which represent significant improvements over the baseline. The full BIMSA configuration achieved the best overall performance, increasing mAP_{0.5} from 57.8% to 62.5% on VisDrone and from 57.9% to 63.2% on UAV-HS, while also showing consistent improvements on stricter metrics such as mAP_{0.75} and mAP_{0.5:95}. These results clearly demonstrate that the proposed modules substantially enhance detection accuracy without introducing excessive computational cost. To further isolate the contributions of individual submodules, controlled ablation experiments were performed on MSAAM

TABLE VI
ABLATION STUDY OF IDFM ON VisDRONE AND UAV-HS

Method	Sigmoid	SimAM	mAP _{0.5} (%)
VisDrone			
CSPDarknet			57.8 ± 0.4
CSPDarknet	✓		58.6 ± 0.3
CSPDarknet	✓	✓	59.1 ± 0.2
UAV-HS (ours)			
CSPDarknet			57.9 ± 0.4
CSPDarknet	✓		58.8 ± 0.3
CSPDarknet	✓	✓	59.4 ± 0.2

and IDFM. The results in Table V illustrate that progressively integrating deformable convolution, reparameterized convolution, and residual squeeze-excitation within MSAAM steadily enhanced detection accuracy. On VisDrone, the complete MSAAM improved mAP_{0.5} by 1.4% compared to the baseline, while on UAV-HS the improvement was 1.5%. Similarly, Table VI shows the ablation results of IDFM. For instance, on UAV-HS, IDFM increased mAP_{0.5} from 57.9% to 59.4%, while on VisDrone, the improvement was from 57.8% to 59.1%. These consistent improvements validate the effectiveness of both MSAAM and IDFM in strengthening spatial and semantic representation. Taken together, the findings across Tables IV, V, and VI confirm that the proposed modules contribute complementary enhancements, and their integration within BIMSA leads to substantial performance gains. By improving both feature fusion and spatial-semantic modeling, BIMSA provides a robust and efficient solution for UAV-based vehicle detection in complex real-world scenarios.

2) *BIMSA-Yolo Ablation Experiment:* The experimental results summarized in Tables IV, V, and VI comprehensively

validate the effectiveness and robustness of the proposed BIMSA-YOLO framework across multiple UAV-based vehicle detection benchmarks. As presented in Table IV, the CSPDarknet baseline achieves 57.9% mAP_{0.5}, 45.3% mAP_{0.75}, and 39.3% mAP_{0.5:95} on the UAV-HS dataset. The progressive integration of the BSFM, MSAAM, IDFM, and SPPFCSPC modules consistently leads to improvements in detection accuracy. Each component contributes complementary benefits. The BSFM is primarily responsible for preserving fine-grained spatial details, which are crucial for detecting small-scale or partially occluded vehicles that frequently appear in UAV surveillance imagery. The MSAAM enhances multi-scale feature alignment and effectively suppresses spurious activations, which is particularly beneficial in complex traffic environments with cluttered backgrounds. The IDFM facilitates cross-level semantic fusion, enabling the network to integrate contextual information across different feature hierarchies and thereby improving recognition of vehicles under scale variation. The SPPFCSPC module strengthens high-level semantic representation by aggregating spatially diverse features, which enhances the discrimination ability of the network in dense traffic scenarios. With the complete BIMSA-YOLO configuration, the UAV-HS mAP_{0.5}, mAP_{0.75}, and mAP_{0.5:95} are elevated to 63.2%, 50.8%, and 43.4%, respectively. These correspond to absolute gains of 5.3%, 5.5%, and 4.1%, thereby confirming the complementary effect of the proposed modules. A similar trend is observed on the VisDrone dataset, where mAP_{0.5} increases from 57.8% to 62.5%, accompanied by corresponding improvements at higher IoU thresholds. This consistency across datasets highlights the robustness of the framework when applied to aerial imagery with diverse object scales, background clutter, and occlusion patterns. The component-wise contributions are further examined in Tables V and VI. In the MSAAM ablation study (Table V), replacing standard convolution with deformable convolution (DConv) and re-parameterized convolution (RepConv) alone raises the VisDrone mAP_{0.5} from 56.9% to 58.2%, and the UAV-HS mAP_{0.5} from 57.2% to 58.6%. This demonstrates that multi-scale adaptive alignment substantially enhances feature expressiveness for both small and medium-sized vehicles. The further inclusion of the RS-SE attention mechanism improves VisDrone and UAV-HS performance to 58.3% and 58.7%, respectively, confirming the importance of selectively emphasizing informative channels and suppressing background noise. Table VI evaluates the IDFM and shows that sequentially applying ReLU and SimAM activation functions significantly benefits cross-level feature fusion. The VisDrone mAP_{0.5} increases from 57.8% to 59.1%, while UAV-HS improves from 57.9% to 59.4%. These results suggest that IDFM plays a critical role in consolidating contextual information across scales, which is particularly advantageous in densely populated vehicle clusters and urban environments with complex layouts. Taken together, the ablation studies provide clear evidence of the complementary nature of BIMSA-YOLO's components. BSFM is effective in preserving fine-grained details that are often lost in down-sampling operations. MSAAM improves multi-scale spatial alignment, ensuring better handling of objects with size variation. IDFM strengthens semantic aggregation across

TABLE VII
PERFORMANCE COMPARISON OF KERNELS AND ATTENTION MODULES IN MSAAM. REPDCONV COMBINES 3×3 , 5×5 AND 7×7 KERNELS

Method	Kernels	Attention	mAP _{0.5} (%)
VisDrone			
CSPDarknet	–	–	56.9 ± 0.5
CSPDarknet	3×3	–	57.6 ± 0.4
CSPDarknet	5×5	–	57.9 ± 0.4
CSPDarknet	7×7	–	58.1 ± 0.4
CSPDarknet	RepDConv	–	58.2 ± 0.3
CSPDarknet	RepDConv	CBAM [62]	58.1 ± 0.3
CSPDarknet	RepDConv	ECA [64]	58.0 ± 0.3
CSPDarknet	RepDConv	CA [72]	58.1 ± 0.3
CSPDarknet	RepDConv	SE [63]	58.2 ± 0.2
CSPDarknet	RepDConv	RS-SE	58.3 ± 0.2
UAV-HS (ours)			
CSPDarknet	–	–	57.2 ± 0.4
CSPDarknet	3×3	–	57.9 ± 0.3
CSPDarknet	5×5	–	58.2 ± 0.3
CSPDarknet	7×7	–	58.4 ± 0.3
CSPDarknet	RepDConv	–	58.6 ± 0.2
CSPDarknet	RepDConv	CBAM [62]	58.5 ± 0.2
CSPDarknet	RepDConv	ECA [64]	58.4 ± 0.2
CSPDarknet	RepDConv	CA [72]	58.5 ± 0.2
CSPDarknet	RepDConv	SE [63]	58.6 ± 0.2
CSPDarknet	RepDConv	RS-SE	58.7 ± 0.2

different feature levels, while SPPFCSPC ensures that high-level features are better consolidated for robust recognition. The cumulative effect of these modules produces substantial performance improvements, while the computational overhead remains moderate. The parameter count increases from 3.1M in the baseline to 6.7M in the full BIMSA-YOLO model, and the computational cost grows from 18.7 GFLOPs to 27.4 GFLOPs. Importantly, inference latency remains below 13 ms per image, thereby preserving real-time processing capability. In summary, the results demonstrate that BIMSA-YOLO successfully advances UAV-based vehicle detection by striking a balance between accuracy, robustness, and efficiency. The framework provides significant improvements in detection accuracy across multiple datasets, while maintaining computational feasibility for deployment in real-world aerial surveillance applications. These advantages make BIMSA-YOLO a practical and effective solution for challenging operational conditions, including dense traffic, occlusion, scale variation, and complex urban environments.

D. Comparison Experiment

1) *BIMSA Comparison Experiment*: Table VII presents a comprehensive ablation study of the proposed MSAAM feature extractor, systematically evaluating the impact of convolutional kernel configurations and attention mechanisms on detection performance. In this table, a dash indicates the absence of a specific convolutional or attention module. On the VisDrone dataset, the baseline CSPDarknet achieves 56.9% mAP_{0.5}. When standard convolutions are replaced with re-parameterized deformable convolutions (RepDConv) of progressively larger kernel sizes, specifically 3×3 , 5×5 , and 7×7 , the detection performance improves incrementally, ultimately reaching 58.2%. This result demonstrates that

TABLE VIII
COMPARATIVE EXPERIMENTS OF DIFFERENT FEATURE
EXTRACTION AND NECK MODULES

Module	Neck	mAP _{0.5} (%)	C mAP _{0.5} (%)
VisDrone			
C3 [73]	PAFPN	57.2 ± 0.4	96.8 ± 0.5
C2f [74]	PAFPN	57.8 ± 0.3	97.2 ± 0.3
ELAN4 [75]	PAFPN	58.1 ± 0.5	97.4 ± 0.4
MSAAM (ours)	PAFPN	58.3 ± 0.3	97.6 ± 0.3
MSAAM (ours)	HSFPN	56.6 ± 0.4	97.3 ± 0.4
MSAAM (ours)	GFPN	58.0 ± 0.3	97.7 ± 0.2
MSAAM (ours)	BiFPN	59.2 ± 0.2	97.9 ± 0.2
MSAAM (ours)	BIMSA (ours)	61.0 ± 0.2	98.2 ± 0.2
UAV-HS (ours)			
C3 [73]	PAFPN	57.3 ± 0.5	96.9 ± 0.4
C2f [74]	PAFPN	57.9 ± 0.3	97.3 ± 0.3
ELAN4 [75]	PAFPN	58.4 ± 0.4	97.5 ± 0.4
MSAAM (ours)	PAFPN	58.7 ± 0.3	97.8 ± 0.3
MSAAM (ours)	HSFPN	57.0 ± 0.5	97.5 ± 0.4
MSAAM (ours)	GFPN	58.6 ± 0.3	97.9 ± 0.3
MSAAM (ours)	BiFPN	60.8 ± 0.2	98.3 ± 0.2
MSAAM (ours)	BIMSA (ours)	61.7 ± 0.2	98.6 ± 0.2

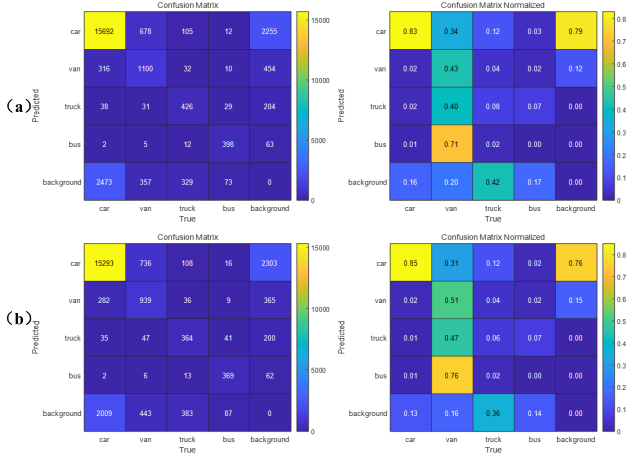


Fig. 8. Confusion matrices on the UAV-Hub surveillance test set. (a) Baseline detector: absolute counts (left) and row-wise normalized percentages (right). (b) Proposed BIMSA-YOLO: absolute counts (left) and row-wise normalized percentages (right).

RepDConv kernels are highly effective in capturing richer spatial features across different receptive fields, enhancing the network’s sensitivity to small and occluded vehicles. The addition of the RS-SE attention module further improves performance, increasing mAP_{0.5} to 58.3%, indicating that adaptive channel-wise recalibration provides a meaningful refinement by emphasizing informative feature channels while suppressing irrelevant background responses. A similar pattern is observed on the UAV-HS dataset, where the baseline performance of 57.2% mAP_{0.5} is increased to 58.6% through the use of RepDConv and further enhanced to 58.7% with the incorporation of RS-SE attention. These findings collectively suggest that the combination of RepDConv and RS-SE attention substantially improves feature expressiveness and contributes to more precise vehicle localization under challenging aerial imaging conditions. Beyond feature extraction,

Table VIII investigates the effect of integrating MSAAM with different neck architectures. When coupled with the widely adopted PAFPN, MSAAM achieves 58.3% and 58.7% mAP_{0.5} on VisDrone and UAV-HS, respectively, outperforming baseline feature extractors such as C3, C2f, and ELAN4 [75]. Despite these incremental improvements, the limited capacity of these necks to aggregate multi-scale features and model cross-scale interactions constrains their overall effectiveness. Alternative designs, including GFPN and BiFPN, enhance multi-scale fusion by providing additional weighted connections or repeated summation of feature maps; however, this approach can introduce redundancy and restrict further gains, with VisDrone and UAV-HS mAP_{0.5} capped at 59.2% and 60.8%, respectively. In contrast, integrating MSAAM with the proposed BIMSA neck achieves a substantial improvement, yielding 61.0% mAP_{0.5} on VisDrone and 61.7% on UAV-HS. The framework also maintains strong performance on the dense CARPK dataset, achieving 98.2% and 98.6% mAP_{0.5}. The superior accuracy of BIMSA arises from its capacity to effectively aggregate hierarchical features while leveraging the attention-enhanced representations generated by MSAAM. By mitigating the limitations of existing neck designs, such as shallow feature reuse, redundant feature fusion, and inadequate context modeling, BIMSA provides a more balanced trade-off between detection accuracy and computational efficiency. Taken together, the results in Tables VII and VIII provide strong empirical evidence supporting the proposed design. The combination of RepDConv kernels and RS-SE attention significantly enhances the feature representation capability of MSAAM, while the BIMSA neck strengthens cross-scale feature integration, resulting in substantial and consistent performance gains across VisDrone, UAV-HS, and CARPK datasets. These findings highlight the robustness and generalization ability of the proposed framework and underscore its practical applicability in real-world UAV-based vehicle detection scenarios, particularly under conditions characterized by dense traffic, small object sizes, and complex environmental factors. The integration of these components ensures that BIMSA-YOLO achieves both high detection accuracy and operational efficiency, making it a compelling solution for autonomous aerial surveillance and intelligent traffic monitoring applications.

2) *BIMSA-Yolo Comparison Experiment*: As demonstrated in Table IX, Table XI, and Table X, the proposed BIMSA-YOLO consistently delivers superior detection performance across multiple UAV datasets and backbone architectures, underscoring both task-specific optimization and generalization capability. On the UAV-HS dataset, BIMSA-YOLO achieves 63.2% mAP_{0.5} and 43.4% mAP_{0.5:0.95}, surpassing YOLOv5-n and YOLOX-n by 7.1% and 3.2%, respectively. Importantly, these gains are obtained with only 6.7M parameters and 27.4 GFLOPs, which is substantially more efficient than traditional two-stage detectors such as Faster R-CNN and Cascade R-CNN. These baselines not only yield lower accuracy (55.4% and 56.9% mAP_{0.5}) but also incur computational costs exceeding 190 GFLOPs, rendering them impractical for real-time UAV deployment. Table XI further demonstrates the per-class advantages of BIMSA-YOLO. The

TABLE IX
COMPARATIVE EXPERIMENTS OF DIFFERENT MODELS ON VISDRONE AND UAV-HS

Method	mAP _{0.5} (VisDrone)	mAP _{0.5:0.95} (VisDrone)	mAP _{0.5} (UAV-HS)	mAP _{0.5:0.95} (UAV-HS)	Params (M)	GFLOPs
Faster-RCNN [47]	55.3 ± 0.5	34.2 ± 0.5	55.4 ± 0.4	34.4 ± 0.4	41.3	191
Cascade-RCNN [76]	56.7 ± 0.5	39.2 ± 0.5	56.9 ± 0.4	39.4 ± 0.4	69.3	219
CenterNet [77]	49.4 ± 0.5	29.5 ± 0.4	49.7 ± 0.4	29.8 ± 0.4	32.3	184
FCOS [78]	50.5 ± 0.4	30.1 ± 0.4	50.7 ± 0.4	30.4 ± 0.3	32.2	184
EfficientDet-b0 [60]	50.1 ± 0.4	27.3 ± 0.4	50.2 ± 0.4	27.6 ± 0.3	3.8	13.8
ATSS [79]	51.6 ± 0.4	30.5 ± 0.4	51.9 ± 0.4	30.8 ± 0.3	32.3	189
YOLOv5-n [73]	56.6 ± 0.4	37.8 ± 0.3	58.1 ± 0.3	38.8 ± 0.3	7.0	31.7
YOLOX-n [49]	60.8 ± 0.3	39.7 ± 0.3	62.1 ± 0.3	40.9 ± 0.3	8.9	53.2
ObjectBox [80]	57.3 ± 0.4	36.7 ± 0.4	57.4 ± 0.4	36.9 ± 0.3	86.1	203.7
YOLOv6-n [81]	58.1 ± 0.3	39.3 ± 0.3	58.6 ± 0.3	39.8 ± 0.3	4.2	11.9
DINO [82]	59.4 ± 0.4	39.3 ± 0.4	60.0 ± 0.4	39.9 ± 0.3	47.5	252
TOOD [83]	56.1 ± 0.4	38.8 ± 0.4	56.5 ± 0.3	39.3 ± 0.3	34.1	185
RT-DETR [84]	58.7 ± 0.3	39.2 ± 0.3	59.5 ± 0.3	39.8 ± 0.3	21.0	233.2
YOLOv9-n [85]	59.8 ± 0.3	40.7 ± 0.3	60.2 ± 0.3	41.1 ± 0.3	7.6	7.6
YOLOv10-n [86]	57.5 ± 0.5	38.7 ± 0.4	58.2 ± 0.4	39.3 ± 0.4	2.7	8.2
YOLO11-n [87]	57.0 ± 0.5	38.4 ± 0.4	57.8 ± 0.4	39.0 ± 0.4	2.6	6.3
Hyper-YOLO-n [88]	60.4 ± 0.3	40.7 ± 0.3	60.8 ± 0.3	41.1 ± 0.3	4.1	10.8
YOLO12-n [89]	58.4 ± 0.4	39.6 ± 0.4	58.7 ± 0.3	39.9 ± 0.3	4.8	12.1
Baseline	57.8 ± 0.3	39.1 ± 0.3	57.9 ± 0.3	39.3 ± 0.3	3.1	8.9
BIMSA-YOLO (ours)	62.5 ± 0.3	50.2 ± 0.2	63.2 ± 0.2	43.4 ± 0.2	6.7	27.4

TABLE X
mAP_{0.5} COMPARISON OF BASELINE AND BIMSA-YOLO MODELS ACROSS DIFFERENT BACKBONES ON VISDRONE AND UAV-HS

Backbone	VisDrone (mAP _{0.5})		UAV-HS (mAP _{0.5})	
	Baseline	BIMSA-YOLO	Baseline	BIMSA-YOLO
CSPDarknet	57.8 ± 0.5	61.2 ± 0.2	57.9 ± 0.5	62.3 ± 0.2
ResNet-50 [90]	55.6 ± 0.5	59.4 ± 0.2	56.3 ± 0.5	60.1 ± 0.2
ResNet-18 [90]	54.0 ± 0.5	57.5 ± 0.3	54.8 ± 0.5	58.2 ± 0.3
MobileNetV2 [91]	52.5 ± 0.6	56.0 ± 0.3	53.2 ± 0.5	57.0 ± 0.3
Swin-T [92]	56.8 ± 0.4	60.4 ± 0.2	57.5 ± 0.4	61.1 ± 0.2
MetaFormer [93]	57.0 ± 0.4	60.6 ± 0.2	57.2 ± 0.4	61.3 ± 0.2

TABLE XI
PER-CLASS RESULTS ON UAV-HS

Method	Car AP	Van AP	Truck AP	Bus AP
Faster-RCNN [47]	63.5 ± 0.5	54.1 ± 0.7	52.8 ± 0.8	49.5 ± 0.9
YOLOv5-n [73]	66.3 ± 0.4	57.4 ± 0.6	58.0 ± 0.7	54.6 ± 0.8
TOOD [83]	66.0 ± 0.4	56.2 ± 0.6	55.1 ± 0.6	51.3 ± 0.7
Hyper-YOLO-n [88]	69.2 ± 0.3	60.5 ± 0.5	61.0 ± 0.5	58.3 ± 0.5
RT-DETR [84]	68.5 ± 0.4	59.8 ± 0.5	60.5 ± 0.6	57.2 ± 0.6
Baseline [74]	67.8 ± 0.4	58.9 ± 0.6	59.5 ± 0.7	55.9 ± 0.7
BIMSA-YOLO (ours)	72.3 ± 0.3	63.7 ± 0.4	66.4 ± 0.5	63.5 ± 0.4

model achieves 72.3%, 63.7%, 66.4%, and 63.5% AP for cars, vans, trucks, and buses, respectively, outperforming all competing methods. Compared with YOLOv8-n, BIMSA-YOLO provides absolute improvements of 4.5%, 4.8%, 6.9%, and 7.6% across these classes, while also increasing confidence accuracy to 89.3% and F1-score to 85.0%. These results indicate that BIMSA-YOLO not only boosts overall detection accuracy but also improves robustness against false positives and inter-class confusion, which are prevalent issues in UAV imagery characterized by small, densely packed, or occluded vehicles. Figure 8 presents the confusion matrices on the UAV-HubSurveillance test set, further highlighting the ability of BIMSA-YOLO to mitigate inter-class misclassification compared with the baseline detector. The

backbone comparison in Table X highlights the versatility of BIMSA-YOLO. Regardless of the underlying feature extractor, ranging from lightweight networks such as MobileNetV2 and ResNet-18 to transformer-based models such as Swin-T and MetaFormer, BIMSA-YOLO consistently achieves 3 to 5% absolute improvements in mAP_{0.5} on both VisDrone and UAV-HS. This indicates that the proposed bidirectional multi-scale aggregation and attention-enhanced modules are highly effective in enhancing feature representation and cross-scale information fusion, independent of the backbone architecture. Moreover, the consistent performance gains across both CNN and transformer backbones confirm that BIMSA-YOLO can robustly handle varying receptive fields, feature hierarchies, and spatial resolutions inherent in different architectures. By contrast, existing alternatives exhibit clear limitations. Anchor-free methods such as FCOS and CenterNet fail to achieve competitive accuracy due to inadequate handling of small-scale or occluded objects. Transformer-based approaches such as DINO and RT-DETR improve accuracy moderately but demand excessive computational resources, limiting real-time applicability. Lightweight models including EfficientDet-b0 and YOLOv6-n maintain efficiency but sacrifice precision, resulting in lower F1-scores and unstable class-wise performance. TOOD and ObjectBox, although achieving acceptable overall mAP, are less capable of discriminating closely spaced vehicles or small targets, which is reflected in their

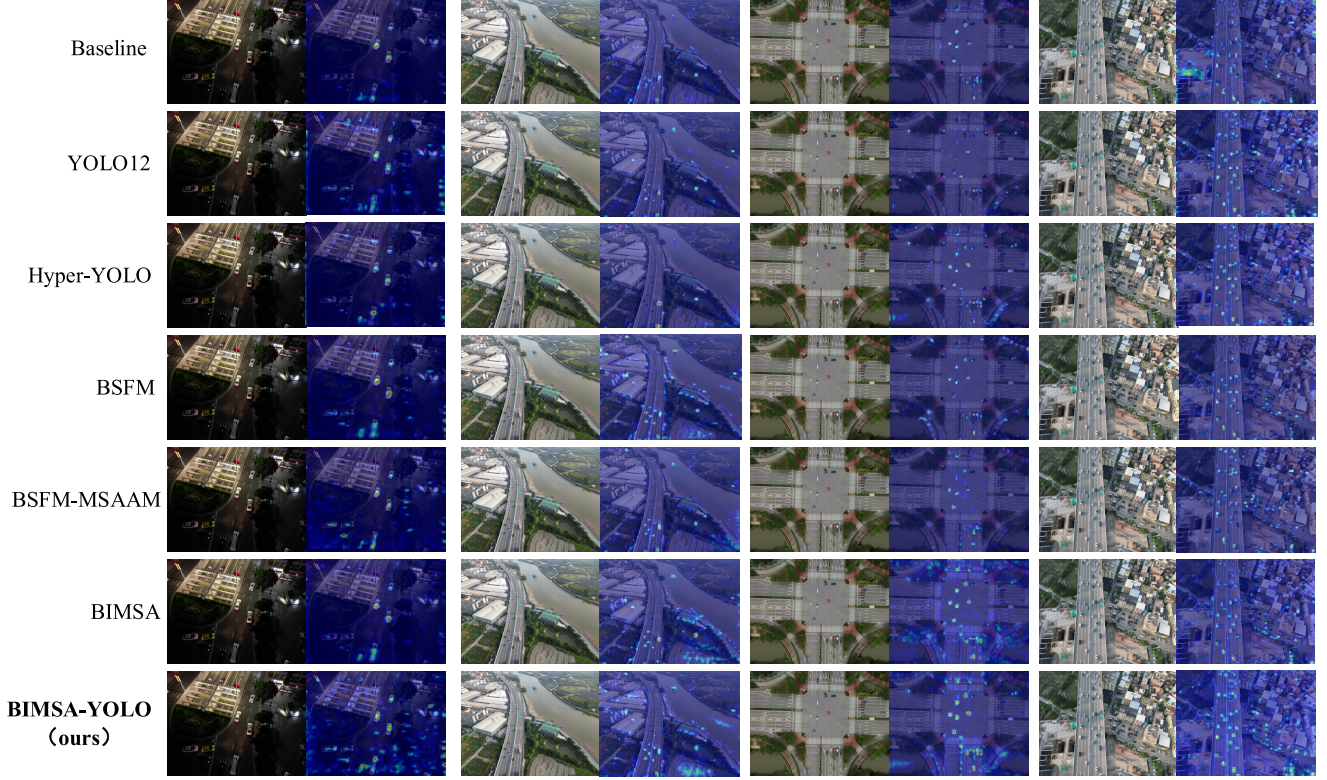


Fig. 9. Grad-CAM visualizations compare BIMSAYOLO with incremental module integration against recent YOLO-based detectors. Column 1 presents samples from the VisDrone dataset, while columns 2 to 4 show examples from the UAV-HubSurveillance dataset. To quantitatively assess attention concentration, the Average Focus Score (AFS) is calculated within the ground-truth bounding boxes. BIMSAYOLO achieves a superior AFS of 0.72, outperforming YOLO12 (0.53), Hyper-YOLO (0.60), and BSFM-augmented variants (0.65). The visualization reveals clear distinctions among the modules: the BSFM enhances multi-scale feature extraction leading to more concentrated activation areas, whereas the IDFM improves feature discrimination by selectively highlighting relevant spatial regions. The integrated SimAM attention within MSAAM further refines activation maps by suppressing noise and enhancing informative features. Consequently, BIMSAYOLO exhibits stable and focused activation even under challenging conditions such as occlusion and blur, effectively reducing over-activation, under-activation, and limiting spatial drift seen in models relying solely on BSFM. This comprehensive comparison demonstrates how the complementary roles of the modules contribute to improved detection performance and robustness.

per-class AP values in Table XI. BIMSAYOLO mitigates these limitations through its novel multi-scale aggregation, RepDConv-based feature extraction, and RS-SE attention mechanisms, which jointly enhance feature representation, contextual awareness, and small-object localization. The comprehensive experimental results in Tables IX, XI, and X collectively validate that BIMSAYOLO not only outperforms existing lightweight and heavyweight detectors in both accuracy and efficiency but also maintains robustness across backbone variations and multiple UAV datasets. These findings demonstrate that BIMSAYOLO constitutes a highly effective and generalizable solution for real-time UAV-based vehicle detection.

E. Discussion

The experimental results summarized in Table X demonstrate that the proposed BIMSAYOLO framework consistently improves detection performance across diverse backbone architectures on both the VisDrone and UAV-HS datasets. Quantitative comparisons reveal that integrating BIMSAYOLO into standard backbone networks yields substantial gains relative to their corresponding baseline

TABLE XII
PSEUDOCODE FOR COMPUTING AFS AND BSR METRICS

Algorithm: AFS and BSR Computation

Input: Grad-CAM heatmap $H(x)$, ground-truth binary mask $M(x)$
Output: Activation Focus Score (AFS), Background Suppression Rate (BSR)

```

1: For each image  $i = 1 \dots N$  do
2:    $H^+(x) \leftarrow \max(0, H(x))$  // apply ReLU
3:    $\tilde{H}(x) \leftarrow H^+(x) / \sum_x H^+(x)$  // normalize heatmap
4:    $AFS_i \leftarrow \sum_x \tilde{H}(x) \cdot M(x)$  // focus on object regions
5:    $BSR_i \leftarrow 1 - AFS_i$  // activation outside object regions
6: End For
7:  $AFS \leftarrow \frac{1}{N} \sum_{i=1}^N AFS_i$ 
8:  $BSR \leftarrow \frac{1}{N} \sum_{i=1}^N BSR_i$ 

```

configurations. On the VisDrone dataset, improvements in mAP_{0.5} range from approximately 3.5 to 4.7 percentage points, whereas on the UAV-HS dataset, gains span from 3.3 to 4.9 percentage points. Notably, the CSPDarknet backbone shows an increase from 57.8% to 61.2% on VisDrone and from 57.9% to 62.3% on UAV-HS, underscoring the efficacy of BIMSAYOLO when applied to conventional convolutional architectures. Transformer based backbones, including

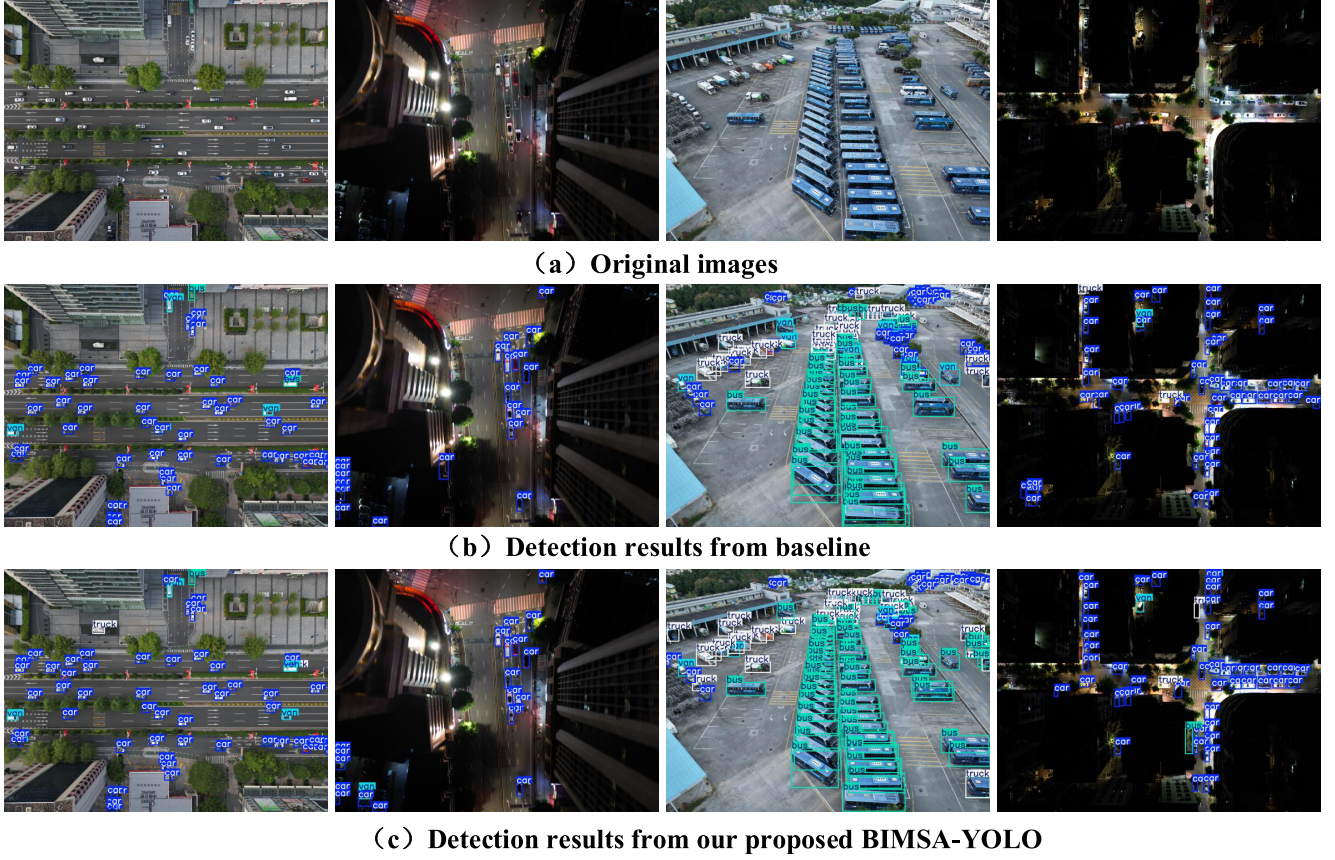


Fig. 10. The qualitative analysis of UAV-HubSurveillance was performed on BIMSA-YOLO. (a) Original images. (b) Detection results from baseline. (c) Detection results from our proposed BIMSA-YOLO.

TABLE XIII
INTERPRETABILITY RESULTS ON THE VISDRONE DATASET
(AFS INDICATES FOCUS, BSR INDICATES SUPPRESSION)

Method	AFS $\uparrow$	BSR $\downarrow$
Faster-RCNN [47]	0.48	0.52
RT-DETR [84]	0.55	0.45
TOOD [83]	0.58	0.42
Hyper-YOLO-n [88]	0.60	0.40
Baseline [74]	0.57	0.43
BIMSA-YOLO (ours)	0.72	0.28

Swin-T and MetaFormer, also achieve substantial improvements, confirming the compatibility of the framework with attention driven feature extractors. These consistent gains across both CNN and transformer models indicate that BIMSA-YOLO is highly versatile, capable of enhancing feature representation and detection robustness under varied architectural designs and computational constraints. Beyond these quantitative improvements, qualitative visualizations provide further insights into the effectiveness of BIMSA-YOLO. Figure 9 presents Grad CAM visualizations that compare BIMSA-YOLO with incremental module integrations against recent YOLO based detectors. In these visualizations, the baseline model and YOLO12 exhibit weak and spatially inconsistent activations, particularly for small, partially

occluded, or distant vehicles. Hyper-YOLO [88] produces stronger responses; however, activations are often dispersed and misaligned, with attention leaking to irrelevant background elements such as road markings or vegetation. The incorporation of the BSFM enhances multi scale feature extraction and preserves fine grained details, leading to more concentrated activation around vehicle regions, although minor interference persists. The IDFM then improves cross scale feature discrimination by selectively highlighting relevant spatial regions, which helps mitigate errors caused by overlapping or occluded vehicles. The integration of the MSAAM further refines activation maps through channel aware attention and dynamic receptive field adjustment, suppressing residual noise and improving robustness under motion blur and lighting variation. When the full BIMSA architecture is applied, activation maps become cohesive and continuous across complex urban intersections and multilane roads, maintaining strong focus on relevant vehicle regions while effectively suppressing background distractions. These results confirm that the cascaded design of BIMSA-YOLO progressively enhances spatial attention, leading to substantial improvements over baseline and recent YOLO based variants. Figure 10 further provides a qualitative comparison between the baseline detector and BIMSA-YOLO on the UAV-HubSurveillance dataset. Panel (a) shows the original input images, panel

(b) presents detections from the baseline model, and panel (c) displays results from BIMSA-YOLO. In these examples, the baseline model frequently fails to detect vehicles under severe occlusion or motion blur, leading to missed detections and incorrect classifications. By contrast, BIMSA-YOLO demonstrates markedly improved localization and classification, accurately detecting small, partially obscured, and fast moving vehicles under challenging urban traffic conditions. To rigorously evaluate these qualitative findings, interpretability metrics were computed following the pseudocode in Table XII. The Activation Focus Score (AFS) measures the proportion of activation energy concentrated within the ground truth bounding boxes, whereas the Background Suppression Rate (BSR) captures the proportion suppressed outside object regions. As summarized in Table XIII, BIMSA-YOLO achieves an AFS of 0.72 and a BSR of 0.28, outperforming Faster R CNN [47], RT DETR [84], TOOD [83], and Hyper-YOLO n [88]. These results confirm that BIMSA-YOLO produces spatially concentrated activations with strong background suppression, thereby enhancing interpretability and robustness in complex aerial detection scenarios. Furthermore, BIMSA-YOLO maintains stable attention even under conditions of partial occlusion, strong glare, or rapid vehicle motion, reducing both under activation and over activation observed in BSFM only variants. Such stability is essential in UAV based traffic monitoring, where vehicles appear at varying scales, orientations, and speeds. Collectively, the consistent improvements across quantitative metrics, qualitative detection outcomes, and interpretability analyses highlight BIMSA-YOLO as a robust and practically applicable solution for real time aerial vehicle detection in dense and dynamic urban environments [94].

VI. CONCLUSION AND FUTURE WORK

To address the lack of representative benchmarks for dynamic urban hub scenarios in UAV-based vehicle detection, we introduce the UAV-HubSurveillance dataset, which contains high-resolution real-world aerial imagery captured under diverse illumination conditions, varying traffic densities, and frequent occlusions. This dataset was specifically designed to reflect operational challenges encountered in complex urban environments, including multilane intersections, highway merging zones, and crowded parking areas. In addition to providing a realistic evaluation platform, the dataset exposes several persistent challenges for current detection frameworks, such as the presence of small-scale vehicles under extreme lighting conditions, intricate interactions among multiple overlapping objects, and frequent partial or full occlusions caused by surrounding infrastructure or other vehicles. These factors collectively increase the difficulty of accurate vehicle localization and classification, highlighting the need for advanced feature extraction and fusion strategies that can robustly handle both spatial and semantic complexity. To address these challenges, we propose BIMSA-YOLO, a bidirectional interaction-enhanced multi-scale aggregation framework designed to improve both detection accuracy and robustness in UAV scenarios. BIMSA-YOLO integrates three core modules: (1) the Bidirectional Spatial Feature Module

(BSFM), which captures and preserves fine-grained details of small or distant vehicles, mitigating the loss of information typically encountered in deep feature hierarchies; (2) the Inter-level Dynamic Fusion Module (IDFM), which facilitates semantic information exchange across multiple feature levels, enabling coherent representation of vehicles at different scales; and (3) the Multi-Scale Adaptive Attention Module (MSAAM), which performs adaptive alignment and weighting of multi-scale features, emphasizing critical regions while suppressing irrelevant background information. These modules collectively enhance feature representation and improve localization precision, particularly in scenarios with dense traffic, overlapping vehicles, or rapid illumination changes. Empirical results demonstrate that BIMSA-YOLO significantly outperforms baseline detectors. On the UAV-HubSurveillance dataset, $mAP_{0.5}$ increased from 57.9% to 63.2%, while on Vis-Drone2021, the metric improved from 57.8% to 62.5%, reflecting the framework's strong generalization capability across distinct UAV datasets. In addition, the model achieves real-time inference at 64 FPS on embedded hardware, confirming its practical applicability for time-sensitive aerial monitoring tasks. Notably, qualitative assessments reveal that BIMSA-YOLO maintains high spatial focus on small and occluded vehicles, reducing false positives in complex urban scenes and effectively handling partial overlaps and layered structures that confound standard detectors. Despite these advances, BIMSA-YOLO retains certain limitations that provide directions for future research. First, the bidirectional multi-scale processing pipeline introduces non-negligible computational overhead, which can hinder deployment on resource-constrained micro-UAV platforms or scenarios demanding ultra-low latency. Second, detection accuracy still decreases under extreme occlusion conditions, particularly when multiple vehicles overlap or when structural elements introduce severe cross-scale feature confusion. Addressing these limitations may require lightweight attention mechanisms, optimized feature fusion strategies, or dynamic inference scheduling to balance accuracy, robustness, and computational efficiency. Overall, BIMSA-YOLO represents a substantial step forward in UAV-based urban vehicle detection, combining multi-scale aggregation, attention-guided refinement, and bidirectional feature interaction to deliver state-of-the-art performance in highly challenging aerial surveillance scenarios.

Future work will primarily concentrate on further reducing inference latency and memory consumption to facilitate deployment on resource-constrained UAV platforms. To achieve this, we intend to investigate a combination of model compression techniques, including structured pruning and knowledge distillation, as well as low-bit quantization strategies that minimize computational overhead while preserving detection accuracy. Hardware-aware optimization, tailored to specific embedded processors and edge devices, will also be explored to fully leverage parallelism and memory hierarchies, thereby enhancing real-time performance without compromising robustness. In parallel, we plan to advance occlusion-aware learning methodologies to improve detection reliability in highly crowded urban scenes. This includes the development of adaptive feature weighting mechanisms,

attention-guided occlusion handling, and cross-frame temporal modeling for video sequences, enabling the network to infer partially occluded or overlapping vehicles more accurately. To support these enhancements, the UAV-HubSurveillance dataset will be systematically expanded to incorporate low-light and nighttime scenarios, additional fine-grained semantic labels, and temporally continuous sequences that provide rich contextual information across frames. This augmentation will facilitate the design and evaluation of temporal and context-aware detection models, further bridging the gap between controlled experimental settings and real-world operational environments. Moreover, we aim to conduct extensive validation and benchmarking on various embedded and edge computing devices, assessing not only detection accuracy but also energy efficiency, thermal stability, and end-to-end inference latency. These evaluations will provide practical insights into the deployability of BIMSA-YOLO for real-world UAV-based traffic monitoring applications, including urban intersection management, highway surveillance, and automated incident detection. Collectively, these ongoing and future efforts are expected to promote the development and practical integration of highly efficient, robust, and adaptive UAV-based vehicle detection systems, ultimately contributing to intelligent urban traffic management frameworks and the advancement of autonomous aerial monitoring technologies.

REFERENCES

- [1] K. Guan, H. Yi, D. He, B. Ai, and Z. Zhong, "Towards 6G: Paradigm of realistic terahertz channel modeling," *China Commun.*, vol. 18, no. 5, pp. 1–18, May 2021.
- [2] H. Jiang, M. Mukherjee, J. Zhou, and J. Lloret, "Channel modeling and characteristics for 6G wireless communications," *IEEE Netw.*, vol. 35, no. 1, pp. 296–303, Jan. 2021.
- [3] T. Ye, W. Qin, Z. Zhao, X. Gao, X. Deng, and Y. Ouyang, "Real-time object detection network in UAV-vision based on CNN and transformer," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.
- [4] S. Wang, F. Jiang, B. Zhang, R. Ma, and Q. Hao, "Development of UAV-based target tracking and recognition systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3409–3422, Aug. 2020.
- [5] P. Zhu, J. Zheng, D. Du, L. Wen, Y. Sun, and Q. Hu, "Multi-drone-based single object tracking with agent sharing network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 4058–4070, Oct. 2021.
- [6] L. Sun et al., "Aerial-PASS: Panoramic annular scene segmentation in drone videos," in *Proc. Eur. Conf. Mobile Robots (ECMR)*, Aug. 2021, pp. 1–6.
- [7] R. Makrigiorgis, C. Kyrkou, and P. Kolios, "How high can you detect? Improved accuracy and efficiency at varying altitudes for aerial vehicle detection," in *Proc. Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, Jun. 2023, pp. 167–174.
- [8] H. Zhou, X. Li, W. Wang, and M. Chen, "UAV-enabled traffic monitoring and congestion management based on spatiotemporal data analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 23210–23222, Dec. 2022.
- [9] Y. Li, W. Zhang, Q. Sun, and Z. Liu, "Rapid post-disaster assessment using UAV imagery and deep learning: A case study of earthquake-damaged infrastructure," *ISPRS J. Photogramm. Remote Sens.*, vol. 198, pp. 156–170, Jan. 2023.
- [10] P. Chen, Y. Dang, R. Liang, W. Zhu, and X. He, "Real-time object tracking on a drone with multi-inertial sensing data," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 131–139, Jan. 2018.
- [11] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.
- [12] H. Fang, S. Guo, X. Wang, S. Liu, C. Lin, and P. Du, "Automatic urban scene-level binary change detection based on a novel sample selection approach and advanced triplet neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601518.
- [13] X. Li and J. Wu, "Developing a more reliable framework for extracting traffic data from a UAV video," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 12272–12283, Nov. 2023.
- [14] Z. Ying et al., "Large-scale high-altitude UAV-based vehicle detection via pyramid dual pooling attention path aggregation network," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 14426–14444, Oct. 2024.
- [15] J. Yu, H. Gao, J. Sun, D. Zhou, and Z. Ju, "Spatial cognition-driven deep learning for car detection in unmanned aerial vehicle imagery," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 4, pp. 1574–1583, Dec. 2022.
- [16] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian, "Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1758–1770, Jun. 2020.
- [17] X. Li, X. Li, Z. Li, X. Xiong, M. O. Khyam, and C. Sun, "Robust vehicle detection in high-resolution aerial images with imbalanced data," *IEEE Trans. Artif. Intell.*, vol. 2, no. 3, pp. 238–250, Jun. 2021.
- [18] Z. Zhang, Y. Liu, T. Liu, Z. Lin, and S. Wang, "DAGN: A real-time UAV remote sensing image vehicle detection framework," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1884–1888, Nov. 2020.
- [19] A. Bouguettaya, H. Zarzour, A. Kechida, and A. M. Taberkit, "Vehicle detection from UAV imagery with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6047–6067, Nov. 2022.
- [20] Z. Wang, J. Zhan, C. Duan, X. Guan, P. Lu, and K. Yang, "A review of vehicle detection techniques for intelligent vehicles," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 3811–3831, Aug. 2023.
- [21] J. Shen, W. Zhou, N. Liu, H. Sun, D. Li, and Y. Zhang, "An anchor-free lightweight deep convolutional network for vehicle detection in aerial images," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24330–24342, Dec. 2022.
- [22] H. Xu, M. Guo, N. Nedjah, J. Zhang, and P. Li, "Vehicle and pedestrian detection algorithm based on lightweight YOLOv3-promote and semi-precision acceleration," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19760–19771, Oct. 2022.
- [23] R. Liu, Z. Yuan, and T. Liu, "Learning TBox with a cascaded anchor-free network for vehicle detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 321–332, Jan. 2022.
- [24] M. Hassaballah, M. A. Kenk, K. Muhammad, and S. Minaee, "Vehicle detection and tracking in adverse weather using a deep learning framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4230–4242, Jul. 2021.
- [25] T. Ye, W. Qin, Y. Li, S. Wang, J. Zhang, and Z. Zhao, "Dense and small object detection in UAV-vision based on a global-local feature enhanced network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, Aug. 2022, Art. no. 2515513.
- [26] H. Wang, Y. Yu, Y. Cai, X. Chen, L. Chen, and Y. Li, "Soft-weighted-average ensemble vehicle detection method based on single-stage and two-stage deep learning models," *IEEE Trans. Intell. Vehicles*, vol. 6, no. 1, pp. 100–109, Mar. 2021.
- [27] D. Du et al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 375–391.
- [28] M. Mueller, N. G. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 445–461.
- [29] Q. Luan, H. Cui, X. Yu, L. Zhang, and B. Sikdar, "An asynchronous shuffled frog-leaping with feasible jaya algorithm for uncertain task rescheduling problem in UAV emergency networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 13106–13121, Oct. 2024.
- [30] H. Zhang, K.-F. Yang, Y.-J. Li, and L. L.-H. Chan, "Night-time vehicle detection based on hierarchical contextual information," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 14628–14641, Oct. 2024.
- [31] R. Ke, Z. Li, J. Tang, Z. Pan, and Y. Wang, "Real-time traffic flow parameter estimation from UAV video based on ensemble classifier and optical flow," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 54–64, Jan. 2019.
- [32] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
- [33] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
- [34] X. Zhang, P. Liu, X. Zhao, N. Xu, Y. Wu, and Z. Li, "Hsfpn: Hierarchical spatial feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2081–2090.

- [35] Z. Li, Y. Liu, X. Liu, Z. He, and L. Wu, "Gfnp: Graph feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4218–4227.
- [36] G. Jocher et al., "YOLO by ultralytics," Ultralytics Inc., Frederick, MD, USA, Tech. Rep. 8.0.0, 2023.
- [37] X. Li, L. Wang, J. Wang, P. Luo, and T. Lu, "RT-DETR: Fast one-stage detr with receptive field enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 21904–21914.
- [38] Y. Zhang et al., "Au-air: A large-scale aerial urban dataset for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, IEEE, 2021, pp. 4512–4520.
- [39] J. Liu et al., "Dronesurf: A UAV dataset for real-time object detection with deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1600–1609.
- [40] K. Hara, K. Matsui, Y. Taniguchi, T. Yamaguchi, and Y. Kuno, "Okutama-action: A benchmark for UAV-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 1980–1990.
- [41] J. Wang, P. Li, Z. Zhang, J. Liu, H. Wang, and J. Wang, "UAV-DEOT: UAV-based detection of objects and tracking in real-time," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2558–2567.
- [42] H. Yu et al., "The unmanned aerial vehicle benchmark: Object detection, tracking and baseline," *Int. J. Comput. Vis.*, vol. 128, no. 1, pp. 1141–1159, 2020.
- [43] Y. Zhu, X. Chen, Q. Han, W. Liu, M. Xu, and J. Lu, "Visdrone: A large-scale dataset for UAV-based object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4028–4037.
- [44] Z. Yuan, R. Liu, W. Zhang, Z. Zhang, X. He, and L. Yuan, "Dronevehicle: A dataset for vehicle detection from aerial UAV images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2540–2548.
- [45] M. Hsieh, Y.-L. Lin, C.-H. Chen, and C.-Y. Wang, "Counting cars in parking lots using deep learning and drone imagery," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.
- [46] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
- [47] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [48] X. Long et al., "PP-YOLO: An effective and efficient implementation of object detector," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Mar. 2020, pp. 824–831.
- [49] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 812–813.
- [50] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Roitransformer: Learning to predict oriented objects in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2849–2858.
- [51] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3500–3509.
- [52] Z. Liang and H. Shen, "Edge-guided feature pyramid networks: An edge-guided model for enhanced small target detection," *Sensors*, vol. 24, no. 23, p. 7767, Dec. 2024.
- [53] W. Zhao, L. He, Y. Liu, and C. Wang, "Multibranch mutual-guiding learning for infrared small vehicle detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 281–293, Jan. 2023.
- [54] J. Chen, X. Liao, W. Wang, and Y. Wang, "Signal-to-noise inference separation for forgery detection in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8477–8489, Dec. 2022.
- [55] H. Zhou, M. Zhan, J. Li, and Y. Liu, "FAMM: Facial action muscle modeling for deepfake detection in compressed videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4412–4426, Sep. 2023.
- [56] G. Yang, J. Lei, H. Tian, Z. Feng, and R. Liang, "Asymptotic feature pyramid network for labeling pixels and regions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 7820–7829, Sep. 2024.
- [57] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, "Scale-aware trident networks for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6054–6063.
- [58] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7036–7045.
- [59] Q. Zhao et al., "M2Det: A single-shot object detector based on multi-level feature pyramid network," in *Proc. AAAI Conf. Art. Intel. (AAAI)*, vol. 33, Jan. 2019, pp. 9259–9266.
- [60] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.
- [61] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.
- [62] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
- [63] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.
- [64] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
- [65] Z. Chen, H. Ji, Y. Zhang, Z. Zhu, and Y. Li, "High-resolution feature pyramid network for small object detection on drone view," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 475–489, Jan. 2024.
- [66] H. Khan et al., "Visionary vigilance: Optimized YOLOv8 for fallen person detection with large-scale benchmark dataset," *Image Vis. Comput.*, vol. 149, Sep. 2024, Art. no. 105195, doi: [10.1016/j.imavis.2024.105195](https://doi.org/10.1016/j.imavis.2024.105195).
- [67] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.
- [68] L. Yang, R. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 11863–11874. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/17186>
- [69] M. T. Usman, H. Khan, S. K. Singh, M. Y. Lee, and J. Koo, "Efficient deepfake detection via layer-frozen assisted dual attention network for consumer imaging devices," *IEEE Trans. Consum. Electron.*, vol. 71, no. 1, pp. 281–291, Feb. 2025.
- [70] S. K. Nalluri, X. Chen, Z. Pan, J. Zeng, and Q. Wu, "Efficient deepfake detection via layer-frozen assisted dual attention network for consumer imaging devices," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 2876–2888, Jul. 2024.
- [71] M. T. Usman, H. Khan, I. Rida, and J. Koo, "Lightweight transformer-driven multi-scale trapezoidal attention network for saliency detection," *Eng. Appl. Artif. Intell.*, vol. 155, Sep. 2025, Art. no. 110917.
- [72] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.
- [73] G. Jocher et al., (2020). YOLOv5. [Online]. Available: <https://github.com/ultralytics/yolov5>
- [74] (2023). YOLOv8-Ultralytics. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [75] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.
- [76] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
- [77] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.
- [78] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.
- [79] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.
- [80] M. Zand, A. Etemad, and M. Greenspan, "ObjectBox: From centers to boxes for anchor-free object detection," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2022, pp. 390–406.
- [81] C. Li et al., "YOLOv6: Towards efficient hardware deployment of object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 210–211.
- [82] H. Zhang et al., "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, 2023, pp. 1–19.

- [83] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3490–3499.
- [84] Y. Guo, J. Yu, J. Li, S. Lan, Y. Liu, and D. Lin, "RT-DETR: Real-time detection transformer," *IEEE Trans. Image Process.*, vol. 32, pp. 2774–2786, 2023.
- [85] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Agave crop segmentation and maturity classification with deep learning data-centric strategies using very high-resolution satellite imagery," 2023, *arXiv:2303.11564*.
- [86] A. Wang et al., "YOLOv10: Real-time end-to-end object detection," in *Proc. 38th Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, 2024, pp. 107984–108011.
- [87] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," 2024, *arXiv:2410.17725*.
- [88] Y. Feng et al., "Hyper-YOLO: When visual object detection meets hypergraph computation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2388–2401, Apr. 2025.
- [89] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric real-time object detectors," 2025, *arXiv:2502.12524*.
- [90] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [91] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
- [92] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
- [93] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10819–10829.
- [94] H. Khan, S. U. Khan, W. Ullah, and S. W. Baik, "Optimal features driven hybrid attention network for effective video summarization," *Eng. Appl. Artif. Intell.*, vol. 158, Oct. 2025, Art. no. 111211.

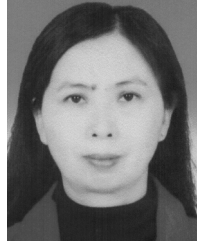


Ye Li is currently an Engineer with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, China. He has authored many core articles, with research direction in control engineering and automation.



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

Since October 2007, he has been with the Department of Intelligence Manufacturing, Wuyi University, Jiangmen, China, where he is currently 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.



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 articles. Her research directions are intelligent detection and intelligent information processing.



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.



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

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



Wenkang Qiu received the B.S. degree in transportation engineering from Wuyi University, Jiangmen, China, in 2023, where he is currently pursuing the M.S. degree in electronic information (Artificial Intelligence) with the School of Electronics and Information Engineering.

His research interests include computer vision, object detection, and their applications in intelligent transportation systems and artificial intelligence.



Hufei Zhu received the B.E. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1998, the M.E. degree in computing from the National University of Singapore, Singapore, in 2004, and the Ph.D. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014.

He is currently a Distinguished Professor with the School of Electronics and Information Engineering, Wuyi University, Guangdong, China, and a Research Fellow with the National Engineering Laboratory for

Big Data System Computing Technology, Shenzhen University, Shenzhen, China. His research interests include neural networks, signal processing, and wireless communications.



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.

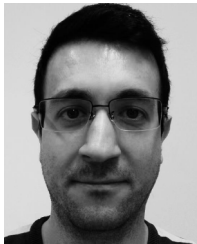
His research interests include computer vision, representational contrastive learning, image processing, and object detection.



Angelo Genovese (Senior Member, IEEE) received the Ph.D. degree in computer science from the Università degli Studi di Milano, Crema, Italy, in 2014.

Since 2014, he has been a Post-Doctoral Research Fellow of computer science with the Università degli Studi di Milano. 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. He is currently an Associate Editor of *Journal of Ambient Intelligence and Humanized Computing* (Springer).



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 with the Università degli Studi di Padova. 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 multi-agents systems. He is currently a member of the

program committees for numerous the IEEE international conferences and workshops. He has received the 2021 Best Paper Award of the Elsevier's Image and Vision Computing Journal for his work on human motion prediction. Since 2023, he has held the position of Co-Chair for the Intelligent Measurement Systems Technical Committee (TC-22) with the IEEE Instrumentation and Measurement Society.



C. L. Philip Chen (Life 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, cybernetics, and computational intelligence.

He is a fellow of American Association for the

Advancement of Science (AAAS), the International Association for Pattern Recognition (IAPR), Chinese Association of Automation (CAA), and 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 IEEE TRANSACTION ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS since 2014 and an Associate Editor of several IEEE TRANSACTIONS.