SICNet: Structured Integrated Cascade Network for UAV-Based Vehicle Detection

Wenkang Qiu, Chaojun Dong, Yikui Zhai, *Senior Member, IEEE*, Cuilin Yu, Ye Li, Xiankun Liu, *Member, IEEE*, Chaoyun Mai, Jianhong Zhou, *Member, IEEE*, Pasquale Coscia, *Senior Member, IEEE*, Angelo Genovese, *Senior Member, IEEE*, and Chun-Lung Philip Chen, *Life Fellow, IEEE*

***Abstract*—Accurate vehicle detection in drone imagery remains a critical challenge due to substantial object scale variation and severe feature degradation arising from small target sizes and complex urban environments. To address these challenges, a novel detection framework, termed the structured integrated cascade network (SICNet), is proposed to improve detection performance in aerial scenes, particularly under dense traffic conditions and small-object scenarios. The core methodological innovation of SICNet lies in a structured attention-guided feature interaction paradigm, which explicitly models the coordinated enhancement of global semantic representations and local spatial details across hierarchical features. Specifically, the proposed structured integrated balanced attention path aggregation network (SIBA-PAN) reformulates conventional feature fusion by introducing an attention-guided mechanism for efficient semantic–spatial information exchange. Within this architecture, a dual-branch attention module, referred to as the structured integrated balanced attention module (SIBAM), is constructed, where the structured grouped attention module and the integrated spatial attention module operate in parallel to capture complementary channel-wise dependencies and spatial responses, respectively, thereby enhancing discriminative feature representation while suppressing background interference. In addition, a spatial cross-stage module (SCM) is integrated into the backbone to strengthen multi-scale context modeling through the coordinated use of cross-stage connections, ELAN-style aggregation, and depth-wise separable convolutions, achieving efficient feature fusion with minimal computational overhead. Extensive experiments conducted on the VisDrone dataset demonstrate that SICNet achieves a 3.6% improvement in $mAP_{0.5}$ over the baseline, with notable gains in detection accuracy for small and densely distributed vehicles, thereby validating the effectiveness of the proposed structured attention mechanism and multi-scale feature interaction strategy. The code and dataset are available at https://github.com/yikuizhai/SICNet**

***Index Terms*—Drone, vehicle detection, feature fusion, attention mechanism.**

Received 8 December 2025; revised 25 April 2026 and 11 June 2026; accepted 18 July 2026. This work was supported in part by the China Postdoctoral Science Foundation under Grant 2026M790616, in part by the National Natural Science Foundation of China under Grant 62273109, in part by the Guangdong Province Key Disciplines Scientific Research Capacity Enhancement Project under Grant 2024ZDJS034, in part by the Guangdong Higher Education Innovation and Strengthening School Project under Grant 2023ZDZX1029, in part by the Wuyi University (WYU) Hong Kong and Macao Joint Research and Development Fund under Grant 2022WGALH19, in part by the Jiangmen Science and Technology Commissioner Research Project under Grant 2024760000580010502, and in part by the WYU High Performance Computing Center and WYU-BUAA Joint Laboratory of Aerospace Information Technology under Grant 3717037. The Associate Editor for this article was S. Mumtaz. *(Wenkang Qiu, Chaojun Dong, Yikui Zhai, Cuilin Yu, Ye Li, Xiankun Liu, Chaoyun Mai, Hufei Zhu, and Jianhong Zhou contributed equally to this work.) (Corresponding author: Cuilin Yu.)*

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

Cuilin Yu is with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, Guangdong 518107, China (e-mail: yuclin3@mail.sysu.edu.cn).

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

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

Pasquale Coscia and Angelo Genovese are with the Department of Computer Science and the Dipartimento di Informatica, Universitá Degli Studi di Milano, 20133 Milan, Italy (e-mail: pasquale.coscia@unimi.it; angelo.genovese@unimi.it).

Chun-Lung 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.2026.3717450

I. Introduction

With the rapid development of unmanned aerial vehicle (UAV) technologies, camera-equipped UAVs have been widely adopted in urban monitoring due to their high mobility, flexible deployment, and real-time data acquisition capabilities [1], [2]. When integrated with onboard computing, UAVs enable efficient edge-level processing for applications such as traffic surveillance, infrastructure inspection, crowd monitoring, and public safety management [3]. Among these tasks, vehicle detection in aerial imagery is essential for intelligent transportation, urban planning, and emergency response [4]. However, reliable detection remains challenging because UAV imagery exhibits pronounced variations in object scale, orientation, and spatial distribution [5]. Although modern detectors, including two-stage frameworks represented by Faster R-CNN [6], anchor-based one-stage models exemplified by the YOLO series [7], anchor-free methods such as FCOS [8], and Transformer-based detectors such as DETR [9] and its variants [10], have achieved strong performance on natural scene benchmarks [11], their effectiveness is often degraded when applied to UAV scenarios. This degradation is primarily caused by drastic scale variation, dense object layouts, and complex urban backgrounds. As illustrated in Fig. 1, UAV-based vehicle detection typically faces two fundamental

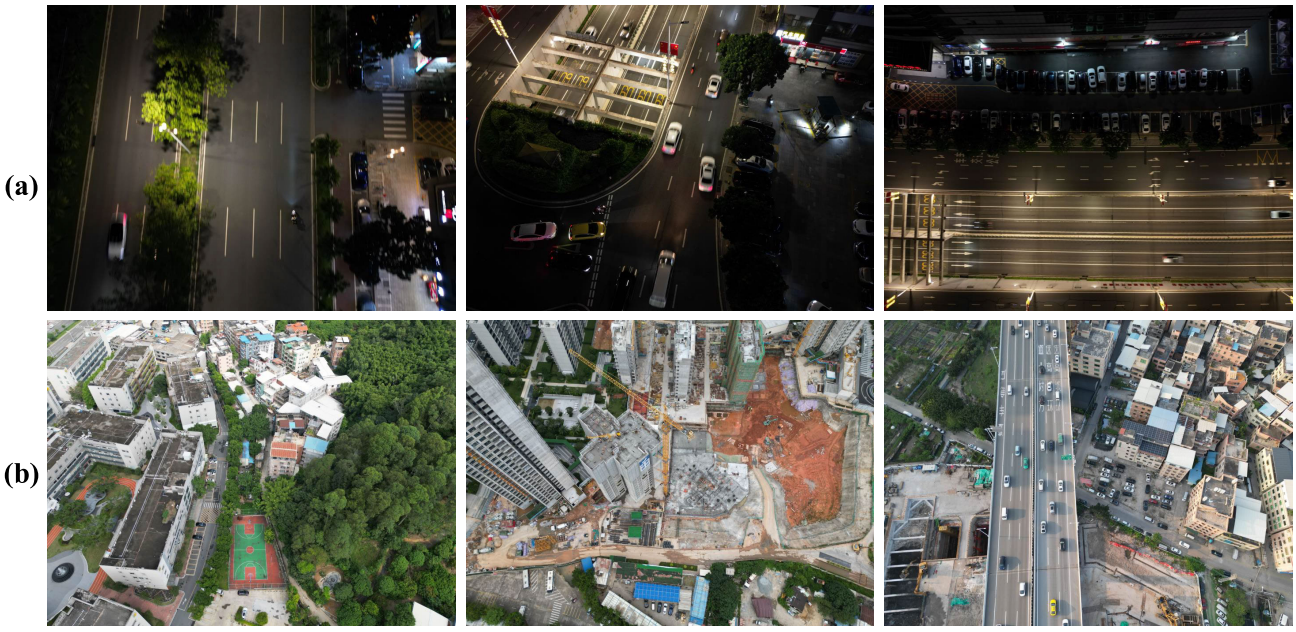

Fig. 1. Challenges are encountered in urban aerial scenes by UAV-based vehicle detection. (a) Object scale change. (b) Object feature loss. Images are from VisDrone dataset.

challenges. The first is severe scale variation, which produces inconsistent feature representations and hinders the detection of small vehicles, including car, van, bus, and truck. The second is feature degradation induced by inter-class overlap and background clutter, leading to inaccurate localization, misclassification, and missed detections, which ultimately affect downstream tasks such as vehicle tracking and traffic flow analysis.

To address the critical challenges illustrated in Fig. 1, namely drastic object scale variation and feature degradation, recent studies have focused on enhancing feature fusion and contextual modeling. Among these efforts, various path aggregation network (PAN) variants [12] have been proposed to improve multi-scale feature fusion by optimizing the flow and alignment of semantic and spatial information across hierarchical layers. Representative designs such as the bidirectional feature pyramid network [13] and neural architecture search-based feature pyramid network [14] incorporate bidirectional fusion and adaptive weighting strategies to strengthen cross-scale connectivity. Although these methods have demonstrated promising results, they still suffer from increased architectural complexity, redundant feature propagation, and limited sensitivity to small and densely distributed vehicles in UAV imagery. In addition, while PAN-based models partially alleviate the scale variation problem, feature degradation for small or occluded vehicles remains a persistent challenge, particularly in aerial perspectives where weak textures and low contrast result in poor feature representations. Spatial pyramid-based modules, such as Spatial Pyramid Pooling Fast Cross Stage Partial Connections (SPPFCSPC) and Atrous Spatial Pyramid Pooling (ASPP) [15], have been introduced to enrich contextual information through multi-scale pooling or dilated convolutions. However, these approaches remain constrained by fixed receptive field configurations, pooling redundancy, and limited adaptability to detection backbones. More importantly, existing methods generally rely on homogeneous feature aggregation or loosely coupled attention mechanisms, lacking explicit modeling of semantic–spatial interactions, which restricts their ability to jointly address scale variation and feature degradation in a unified manner.

To overcome these limitations, a vehicle detection framework tailored for UAV imagery, termed the structured integrated cascade network (SICNet), is proposed to enhance detection accuracy in scenarios involving dense traffic, small objects, and complex urban backgrounds. As illustrated in Fig. 1, UAV imagery is characterized by severe scale variation and feature degradation caused by cluttered backgrounds and weak object responses. To address these issues, SICNet introduces the structured integrated balanced attention path aggregation network (SIBA-PAN) to improve the flow of semantic and spatial information across hierarchical feature layers through an attention-guided fusion strategy. At its core, a structured integrated balanced attention module (SIBAM) is embedded to jointly address scale variation and background-induced feature degradation via coordinated channel–spatial recalibration. Specifically, the structured grouped attention module (SGAM) enhances multi-scale channel dependencies by capturing discriminative responses across different spatial resolutions, thereby improving robustness to scale variation, while the integrated spatial attention module (ISAM) refines spatially salient regions and suppresses background interference, effectively mitigating feature degradation in complex urban scenes. These two components operate in parallel, forming a complementary attention mechanism that strengthens vehicle-relevant features while preserving structural

consistency. Meanwhile, SICNet leverages efficient design principles, including grouped channel operations, depthwise separable convolutions, and parameter-sharing mechanisms within attention branches, which enhance computational organization and feature processing despite the introduction of parallel structures. In addition, a multi-scale enhancement module, referred to as the spatial cross-stage module (SCM), is integrated into the backbone. SCM incorporates cross-stage connections, ELAN-style feature aggregation, and depthwise separable convolutions to extend the conventional spatial pyramid pooling mechanism, thereby enhancing multi-scale feature fusion and contextual representation capability. Extensive experiments conducted on three public UAV-based vehicle detection benchmarks, namely VisDrone [16], PUCPR [17], and CARPK, demonstrate the effectiveness of SICNet. The proposed framework achieves substantial improvements in detection performance, particularly for small and densely distributed vehicles, validating the capability of the structured attention mechanisms and multi-scale feature fusion strategy to address the inherent challenges in UAV-based vehicle detection. The key contributions of this paper can be summarized as follows:

1) A structured integrated balanced attention path aggregation network (SIBA-PAN) has been developed, in which semantic and spatial features are integrated across hierarchical levels to improve vehicle localization accuracy. In addition to architectural integration, the design is grounded in an explicit feature interaction mechanism that models the complementary roles of global semantic abstraction and local spatial detail preservation, providing a principled basis for enhanced multi-scale representation.
2) To enhance contextual perception and spatial discrimination, a structured grouped attention module (SGAM) and an integrated spatial attention module (ISAM) have been incorporated. Long-range dependencies are captured through grouped channel-wise encoding, while vehicle-relevant spatial regions are highlighted via spatial processing. These modules are applied in parallel to improve detection robustness without increasing model complexity. This parallel formulation establishes a structured dual-branch attention paradigm, enabling adaptive coordination between channel and spatial responses rather than relying on conventional sequential attention schemes.
3) A spatial cross-stage module (SCM) based on spatial pyramid pooling was integrated into the backbone to aggregate multi-scale context, improving detection of small objects in UAV imagery. The module further introduces a reformulated combination of cross-stage connections, hierarchical aggregation, and progressive context modeling, which enhances feature diversity and stability under severe scale variation.
4) Experiments on VisDrone show a 3.6% $mAP_{0.5}$ gain over the baseline and competitive results with state-of-the-art methods. Beyond empirical improvements, the proposed framework demonstrates consistent gains across multiple evaluation metrics, reflecting enhanced feature discriminability and robustness in complex UAV scenarios.

II. Related Work

This section reviews UAV-based vehicle detection from three perspectives, namely vehicle detection frameworks, feature fusion strategies, and attention mechanisms. In particular, it focuses on the challenges of object scale variation and object feature loss in aerial imagery, and analyzes how existing methods address these issues by improving multi-scale perception, enhancing feature aggregation, and emphasizing task-relevant representations while suppressing background interference.

A. Vehicle Detection

In recent years, vehicle detection has become a crucial component of intelligent traffic surveillance systems, particularly within UAV-based remote sensing. The primary objective is to accurately localize and classify vehicles in aerial imagery captured from diverse altitudes and viewpoints. With continuous progress in deep learning, mainstream detectors such as RetinaNet, YOLO, and Faster R-CNN have been effectively adapted to remote sensing tasks. RetinaNet, a single-stage detector, utilizes the feature pyramid network to address object scale variation and incorporates Focal Loss to mitigate foreground–background imbalance, enabling robust performance in dense traffic scenes [18]. In contrast, Faster R-CNN adopts a two-stage design: it first generates candidate regions using a region proposal network and subsequently performs region-wise classification and regression for accurate object localization. These architectures have demonstrated competitive performance on generic object detection benchmarks and are widely employed as baselines in UAV-based vehicle detection studies.

Despite these advances, vehicle detection in UAV imagery remains highly challenging due to the small physical size of vehicles, dense spatial layouts, and complex background textures inherent to aerial perspectives. Two fundamental challenges, namely object scale variation and feature degradation, have been identified as the primary factors limiting detection accuracy. To address these issues, numerous improvements over traditional detectors have been explored. For instance, VFNet extends RetinaNet by introducing an IoU-aware classification strategy and varifocal loss to improve performance in crowded scenes [19]. Within the YOLO family, YOLOv4-CSP enhances feature reuse and gradient propagation through cross stage partial connections, while YOLOX [20] introduces decoupled detection heads and dynamic label assignment to better accommodate variable object distributions [20]. More recent advances, including YOLOv10 [21], YOLOv11 [22], and YOLOv12 [23], further improve detection efficiency and accuracy through optimized architectures and training strategies. In parallel, RTMDet [24] introduces an efficient real-time detection framework with improved label assignment and optimized training pipelines, achieving strong performance in dense and small-object detection scenarios. In the two-stage domain, Libra R-CNN and Cascade R-CNN enhance accuracy

through balanced feature sampling and progressive refinement strategies [25]. In addition, spatial pyramid pooling modules have been widely adopted to aggregate multi-scale contextual information, and attention mechanisms have been introduced to strengthen scale-invariant feature encoding. However, most existing methods address scale variation and feature degradation separately and often compromise either semantic richness or computational efficiency. To jointly resolve both challenges, this paper proposes SICNet, a structured vehicle detection framework tailored for UAV-based remote sensing. By integrating spatial-semantic attention with efficient multi-scale fusion within a unified design, SICNet enhances feature discriminability while maintaining real-time inference capability. Consequently, more accurate detection of small and densely distributed vehicles in complex aerial environments is achieved.

B. Feature Fusion

Multi-scale feature aggregation has been widely recognized as a key strategy in UAV-based vehicle detection for addressing feature degradation, which is primarily caused by small object sizes, densely packed layouts, and complex urban backgrounds. Traditional convolutional detectors often struggle to maintain discriminative representations for small-scale objects due to their limited receptive fields and inadequate scale adaptability [26]. To alleviate these issues, Various detection frameworks have introduced spatial pyramid pooling mechanisms to enhance multi-scale contextual encoding. Classical SPP modules partition feature maps into fixed spatial bins to capture hierarchical cues across scales, improving robustness under scale variation [27]. Variants such as Spatial Pyramid Pooling Fast (SPPF) and Simplified Spatial Pyramid Pooling Fast (SimSPPF) simplify the pooling structure while preserving essential multi-resolution information. Atrous spatial pyramid pooling expands the receptive field via dilated convolutions, enabling richer contextual awareness without increasing kernel size [28]. Meanwhile, Spatial Pyramid Pooling Cross Stage Partial Connections (SPPCSPC) and its improved variant SPPFCSPC incorporate cross-stage partial connections to enhance feature reuse and fusion efficiency across scales.

Although these SPP-based modules have proven effective for detecting small and overlapping vehicles in UAV imagery by mitigating feature degradation, several limitations persist. Fixed spatial binning often fails to accommodate irregular object shapes and varied distributions in aerial views, reducing spatial precision. While reducing computational burden, semantic richness is often compromised. Modules like ASPP may also introduce excessive parameters and struggle to retain fine-grained details during scale transitions [29]. To address such constraints, recent studies have explored adaptive or compact pyramid pooling strategies—such as multi-kernel parallel pooling and dynamic configuration schemes—that aim to balance representational capacity and efficiency. For example, designs like Lite-SPP and adaptive pyramid modules attempt to modulate pooling behavior based on semantic features within each layer. However, these methods still face trade-offs between semantic abstraction and spatial accuracy, limiting their ability to recover fine object cues under complex UAV scenarios. To overcome these challenges, this paper proposes a multi-scale enhancement module named SCM, which integrates an expressive fusion mechanism. By combining cross-stage connections, Efficient Layer Aggregation Network (ELAN) style aggregation, and depthwise convolutions, SCM extends traditional pyramid pooling designs to improve semantic detail preservation while maintaining efficiency. Integrated into the SICNet backbone, SCM enhances the network's capacity to recover fine-grained features across scales, enabling more accurate vehicle detection in dense and cluttered aerial scenes.

C. Attention Mechanisms

Attention mechanisms have been widely adopted in UAV-based vehicle detection to mitigate object scale variations and enhance discriminative feature representation across multiple spatial resolutions. Channel attention modules, such as squeeze-and-excitation (SE) [30] and efficient channel attention (ECA) [11], recalibrate feature responses by emphasizing informative channels; however, they generally lack explicit spatial modeling and thus fail to capture rich contextual dependencies. Spatial attention modules focus on salient regions within feature maps, yet their limited receptive fields and fixed kernel designs restrict their adaptability in complex aerial scenes. Hybrid approaches, such as the convolutional block attention module (CBAM) [31], sequentially integrate channel and spatial attention to model both channel-wise importance and spatial relevance. However, their rigid sequential formulation limits flexibility in capturing diverse object geometries and irregular spatial distributions. Coordinate attention incorporates directional encoding into channel attention to enhance orientation-aware representation, yet it remains less effective in modeling highly irregular layouts commonly observed in UAV imagery. Transformer-based self-attention mechanisms, as exemplified by the vision transformer [32], enable global dependency modeling through full token interaction and have demonstrated strong representational capacity in various vision tasks. Recent advances, such as the real-time detection transformer (RT-DETR) [33], further improve efficiency through optimized architecture design and training strategies, making transformer-based detectors more practical for real-time applications. However, in UAV-based vehicle detection scenarios, these methods still incur significantly higher computational overhead, posing challenges for deployment on resource-constrained edge devices. Moreover, their performance in small-object and densely distributed scenarios remains sensitive to feature degradation and scale variation, as further discussed in the experimental analysis.

Although these attention-based techniques enhance feature discrimination and scale awareness, significant challenges remain in balancing representational capacity, spatial adaptability, and computational efficiency under highly variable and cluttered UAV imagery. To address these limitations, this paper introduces SIBAM, a structured dual-branch attention module comprising SGAM and ISAM. SGAM refines semantic channel dependencies through structured grouping and multi-scale encoding, while ISAM enhances spatial

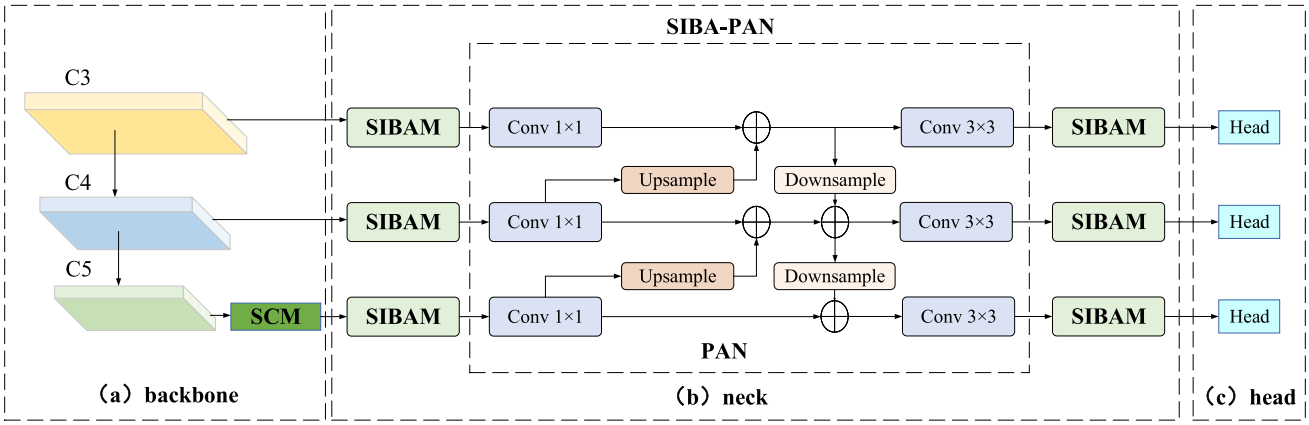

Fig. 2. The overall framework of our proposed vehicle detection method. The SIBA-PAN model comprises two components: PAN and SIBAM. To begin, the feature maps (C3-C5) produced by the backbone are processed by SIBAM to extract specific information from the corresponding feature layers. Subsequently, the hierarchical architecture of the PAN component, which incorporates both bottom-up and top-down pathways, is utilized to comprehensively fuse semantic and spatial information from distinct feature layers. Lastly, after undergoing extensive fusion, the feature maps are subjected to another round of processing by SIBAM in order to extract finer features. These refined feature maps are then directed to their respective detection heads.

responses via multi-statistical spatial reweighting and progressive refinement. Operating in parallel, these two branches capture complementary attention cues and establish a coordinated channel–spatial interaction mechanism. Embedded within the SIBA-PAN architecture, SIBAM enables more effective semantic–spatial fusion across hierarchical features, thereby improving robustness and discriminability in multi-scale vehicle detection for complex UAV scenarios.

III. Methods

In this section, the proposed SICNet for UAV-based vehicle detection is introduced. SICNet is constructed by integrating the SCM and the SIBA-PAN. Specifically, SIBA-PAN is composed of the SGAM and the ISAM, which collaboratively enhance feature representation by modeling both channel-wise and spatial dependencies. Through this hierarchical design, the proposed framework effectively alleviates the challenges of object scale variation and object feature loss, thereby improving detection performance in complex UAV scenes. The overall architecture of SICNet is illustrated in Fig. 2.

A. Structured Integrated Balanced Attention Path Aggregation Network

To address the critical challenges of object scale variation and feature degradation in UAV-based vehicle detection, an enhanced feature interaction framework is proposed to improve multi-scale representation and robustness. In vehicle detection, conventional feature pyramid network (FPN) and PAN frameworks perform multi-scale fusion through direct feature aggregation, implicitly treating semantic and spatial information as homogeneous components. Although effective, such designs lack an explicit mechanism to regulate the interaction between global semantic responses and local spatial details, particularly under severe object scale variations and feature degradation in UAV imagery. As illustrated in Fig. 5, existing attention mechanisms typically adopt simplified or sequential designs, which are insufficient for explicitly modeling the complementary relationship between channel and spatial representations.

To address this limitation, the proposed SIBA-PAN introduces an attention-guided feature interaction framework, in which feature aggregation is reformulated via structured attention modeling. Specifically, the Structured Integrated Balanced Attention Module (SIBAM) is constructed by jointly adapting channel and spatial attention mechanisms within a parallel architecture. Unlike conventional sequential designs, the channel branch incorporates structured multi-scale encoding to capture long-range semantic dependencies, while the spatial branch leverages multi-statistical aggregation to emphasize fine-grained local patterns. This parallel formulation transforms channel and spatial attention from isolated recalibration operations into mutually complementary components within a unified interaction framework, thereby enabling more effective coordination between semantic abstraction and spatial detail preservation. From a theoretical perspective, this design can be interpreted as a decomposition of feature representations into complementary subspaces, where channel attention captures global semantic distributions and spatial attention models localized structural variations, resulting in a more expressive joint representation for complex aerial scenes. Furthermore, the integration of the two attention branches is formulated as an adaptive balancing process rather than a fixed fusion operation. The fusion coefficient λ is explicitly defined as a learnable parameter and is jointly optimized with the network weights during training, enabling dynamic adjustment of the relative contributions of channel-enhanced and spatial-enhanced features. This design avoids the limitations of manually predefined fusion ratios and allows the model to adaptively emphasize global semantic responses or local spatial details according to input characteristics. From an optimization perspective, the introduction of λ can be interpreted as a data-driven weighting mechanism that approximates optimal feature selection under varying input distributions, thereby improving gradient flow stability and enhancing convergence

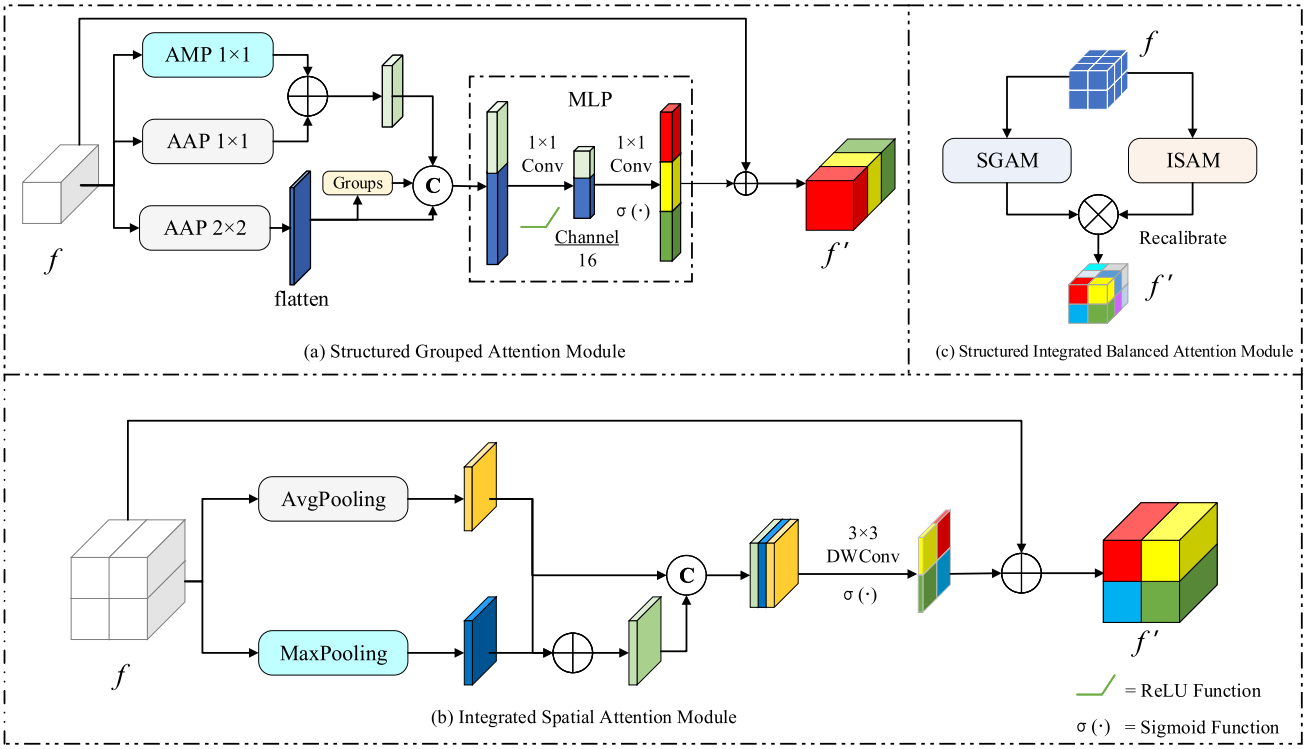

Fig. 3. Conceptual illustration of the proposed attention modules. (a) The SGAM extracts multi-scale channel-wise context using three adaptive pooling branches: AAP at 1×1 and 2×2 resolutions, and AMP at 1×1 resolution. Pooling outputs with the same spatial size are fused by element-wise addition, flattened, and grouped before being concatenated into a unified descriptor. This descriptor is passed through a MLP to compute the channel attention weights, which are then applied to the input feature via a residual connection. (b) The ISAM captures spatial dependencies by applying average and max pooling along the channel axis, followed by element-wise addition. The resulting three spatial maps are concatenated and processed by a 3×3 depthwise convolution and sigmoid activation to generate spatial attention weights, which are likewise applied with a residual connection. (c) The SIBAM integrates SGAM and ISAM in parallel. The outputs of both attention branches are fused to produce a refined feature representation, enabling balanced enhancement in both channel and spatial dimensions.

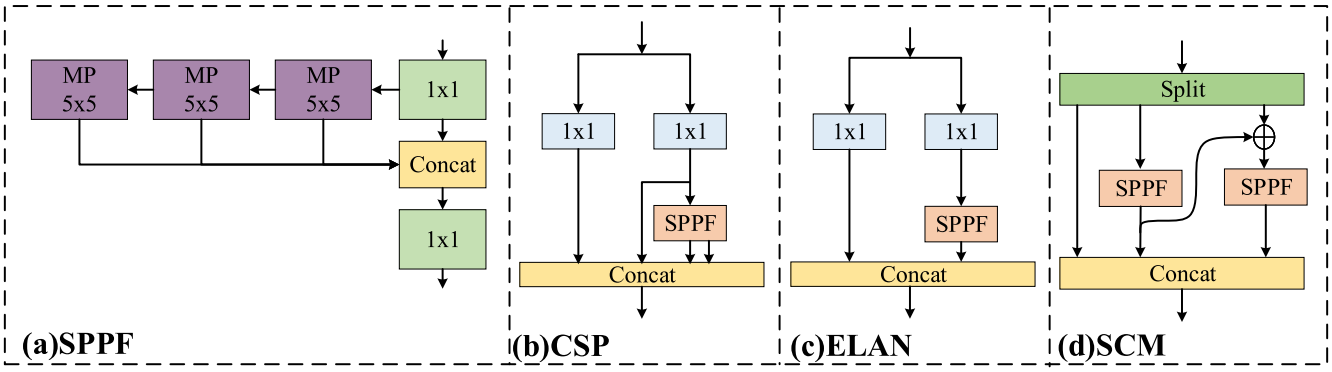

Fig. 4. Structural comparison of different spatial feature aggregation modules. (a) Spatial Pyramid Pooling Fast (SPPF) leverages sequential max pooling and concatenation. (b) Efficient Layer Aggregation Network (ELAN) splits features into parallel paths with partial fusion. (c) Cross Stage Partial (CSP) introduces a split-transform-merge strategy. (d) The proposed SCM integrates CSP and ELAN principles with cascaded depthwise convolutions to achieve efficient multi-scale feature extraction.

behavior. Given an input feature map $f \in \mathbb{R}^{C \times H \times W}$, two parallel attention transformations are first applied to generate complementary representations. The channel attention transformation is denoted as $\mathcal{A}_c(\cdot)$, corresponding to the SGAM branch, while the spatial attention transformation is denoted as $\mathcal{A}_s(\cdot)$, corresponding to the ISAM branch. These operations produce intermediate features f_{SGAM} and f_{ISAM}, both of which retain the same dimensionality as the input feature map. To ensure stable optimization and balanced feature contributions, both feature maps are normalized using the ℓ_2-norm, denoted as $\|\cdot\|_2$, which computes the Euclidean norm across channel and spatial dimensions. A small constant ϵ is introduced to prevent numerical instability. The normalized features are represented as $\hat{f}_{\text{SGAM}}$ and $\hat{f}_{\text{ISAM}}$. The fusion coefficient $\lambda \in [0, 1]$ is generated through a learnable projection mechanism, where GAP$(\cdot)$ denotes global average pooling that aggregates spatial information into a channel descriptor, $\mathbf{w} \in \mathbb{R}^C$ is a learnable weight vector, and $\sigma(\cdot)$ denotes the sigmoid activation function

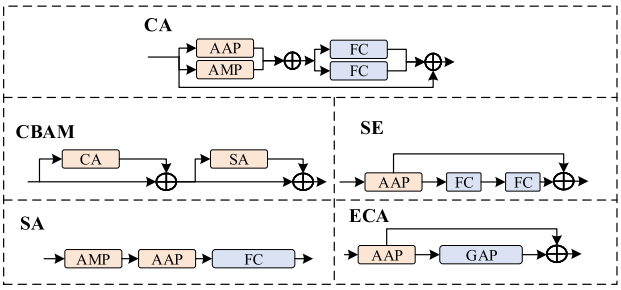

Fig. 5. This figure illustrates the structural layouts of representative attention mechanisms, namely CBAM, SE, ECA, SA, and CA, with the essential operations and components of each module clearly depicted.

that constrains the output within a bounded range. In terms of computational complexity, the proposed parallel attention design introduces only marginal overhead. Specifically, SGAM relies on multi-scale pooling and grouped one-dimensional convolution, while ISAM employs depthwise and pointwise convolutions with linear complexity with respect to spatial resolution. Consequently, the overall time and space complexity remains comparable to conventional attention modules while providing enhanced representational capacity.

$$
\begin{aligned}
f_{\text{SGAM}} &= \mathcal{A}_c(f), \\
f_{\text{ISAM}} &= \mathcal{A}_s(f), \\
\hat{f}_{\text{SGAM}} &= \frac{f_{\text{SGAM}}}{\|f_{\text{SGAM}}\|_2 + \epsilon}, \quad \hat{f}_{\text{ISAM}} = \frac{f_{\text{ISAM}}}{\|f_{\text{ISAM}}\|_2 + \epsilon}, \\
\lambda &= \sigma\left(\mathbf{w}^\top \cdot \text{GAP}(f)\right), \\
f' &= f + \lambda \cdot \hat{f}_{\text{SGAM}} + (1-\lambda) \cdot \hat{f}_{\text{ISAM}}
\end{aligned} \tag{1}
$$

B. Structured Grouped Attention Module

To address the critical challenge of object scale variation in UAV-based vehicle detection, a multi-scale channel attention mechanism is introduced to enhance feature representation across different spatial resolutions. In UAV-based vehicle detection, conventional single-scale channel attention mechanisms often fail to capture multi-resolution dependencies and fine-grained structural details, thereby limiting feature discriminability under significant scale variation and complex background conditions. To overcome these limitations, SGAM introduces a multi-scale adaptive pooling framework that generates channel descriptors at multiple spatial resolutions, enabling the network to simultaneously encode global semantic context and localized structural information. This design is particularly effective for handling scale variation, as features corresponding to small and large objects are jointly modeled across diverse receptive fields.

From a theoretical perspective, multi-scale adaptive pooling can be interpreted as a hierarchical feature decomposition process, where low-resolution pooling captures global semantic distributions, while higher-resolution pooling preserves localized structural responses. This complementary representation alleviates feature degradation for small objects by preventing excessive spatial information loss during global aggregation and retaining discriminative local cues that are critical for small-scale target recognition. Specifically, average pooling stabilizes feature responses and suppresses background interference, whereas max pooling emphasizes salient foreground activations. By hierarchically fusing these complementary descriptors through element-wise addition, SGAM constructs structured multi-scale channel representations that capture subtle variations in object appearance and enhance the discrimination of overlapping targets. Compared with conventional attention mechanisms that rely on single-path global descriptors or simple channel weighting, SGAM explicitly integrates multi-scale feature interactions and complementary pooling strategies, resulting in more discriminative, context-aware, and computationally efficient representations tailored for UAV imagery. Formally, given an input feature map $f \in \mathbb{R}^{C\times H\times W}$, SGAM constructs multi-scale channel descriptors by applying global average pooling and global max pooling, denoted as GAP$(\cdot)$ and GMP$(\cdot)$, respectively, where both operations aggregate spatial information into channel-wise statistics. The resulting descriptors $f_1^c \in \mathbb{R}^{C\times 1\times 1}$ and $f_2^c \in \mathbb{R}^{C\times 1\times 1}$ encode complementary statistical properties. These descriptors are reshaped into vectors $z_1, z_2 \in \mathbb{R}^C$ via the flatten$(\cdot)$ operation, which maps spatially aggregated tensors into one-dimensional channel representations. To introduce structured channel interaction, the max-pooled descriptor z_2 is processed by a grouped one-dimensional convolution operator Conv1D$_g(\cdot)$, where g denotes the number of channel groups controlling the granularity of local channel dependency modeling, producing $\tilde{z}_2 \in \mathbb{R}^C$. The vectors z_1 and $\tilde{z}_2$ are concatenated via concat$(\cdot)$ to form a unified representation $z \in \mathbb{R}^{2C}$, integrating multi-scale and multi-statistical information. A bottleneck multi-layer perceptron is then applied, where $W_1 \in \mathbb{R}^{\frac{C}{r}\times 2C}$ and $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ are learnable weight matrices, and r is the channel reduction ratio controlling feature compression. The nonlinear activation function $\delta(\cdot)$ denotes the ReLU function, while the sigmoid function $\sigma(\cdot)$ maps the output to the range $[0, 1]$, producing channel attention weights $s \in \mathbb{R}^C$. Finally, the refined feature map $f_c' \in \mathbb{R}^{C\times H\times W}$ is obtained by channel-wise modulation of the input feature f with attention weights s through element-wise multiplication $\otimes$, where s is broadcast along the spatial dimensions.

$$
\begin{aligned}
f_1^c &= \text{GAP}(f), \quad f_2^c = \text{GMP}(f), \\
z_1 &= \text{flatten}(f_1^c), \quad z_2 = \text{flatten}(f_2^c), \\
\tilde{z}_2 &= \text{Conv1D}_g(z_2), \\
z &= \text{concat}(z_1, \tilde{z}_2), \\
s &= \sigma\left(W_2 \cdot \delta\left(W_1 \cdot z\right)\right), \\
f_c' &= f \otimes s
\end{aligned} \tag{2}
$$

C. Integrated Spatial Attention Module

To explicitly address the problem of object feature loss in UAV-based vehicle detection, spatial attention plays a critical role in modeling positional dependencies within feature maps, particularly in UAV imagery where complex backgrounds, small targets, and densely distributed objects pose significant challenges. Conventional spatial attention mechanisms, such as CBAM, typically rely on aggregating channel-wise statistics through a single global descriptor to generate attention maps. This approach compresses rich spatial information into a limited representation and fails to adequately discriminate fine-grained structures, leading to insufficient sensitivity to

object boundaries and contextually important regions, especially under severe feature degradation caused by background clutter and low-contrast targets. To address these limitations, ISAM introduces a multi-statistical spatial encoding framework that fundamentally differs from traditional single-path designs. By jointly exploiting average-pooled and max-pooled spatial descriptors, ISAM simultaneously captures globally smooth contextual information and locally salient features, enabling more reliable recovery of degraded object cues. From a theoretical perspective, this design can be interpreted as a complementary statistical decomposition of spatial information, where average pooling approximates low-frequency background distributions while max pooling preserves high-frequency salient responses. Their joint modeling effectively mitigates feature degradation by enhancing signal-to-noise separation, allowing discriminative foreground structures to be preserved even under complex background interference. In addition, a mixed descriptor is constructed through element-wise interaction, enabling cross-statistical coupling that further enriches spatial representation and improves robustness to degraded feature conditions.

In implementation, ISAM constructs a structured multi-statistical spatial representation by jointly leveraging three complementary descriptors derived from the input feature map $f \in \mathbb{R}^{C\times H\times W}$. Specifically, $f^s_{\text{avg}} \in \mathbb{R}^{1\times H\times W}$ and $f^s_{\max} \in \mathbb{R}^{1\times H\times W}$ are obtained via channel-wise average pooling $\text{AvgPool}_c(\cdot)$ and max pooling $\text{MaxPool}_c(\cdot)$, respectively, where the pooling operations aggregate channel information while preserving spatial structure. The mixed descriptor $f^s_{\text{mix}} \in \mathbb{R}^{1\times H\times W}$ is defined through an element-wise interaction operator $\oplus$, which denotes adaptive fusion between f^s_{avg} and $f^s_{\max}$ to encode complementary statistical characteristics. These descriptors are concatenated along the channel dimension using $\text{concat}(\cdot)$ to form a unified representation $F_s \in \mathbb{R}^{3\times H\times W}$. To refine spatial representations while maintaining computational efficiency, a factorized convolutional structure is employed, where $\text{DWConv}(\cdot; W_d)$ denotes depthwise convolution with learnable parameters W_d, performing independent spatial filtering per channel, and $\text{PWConv}(\cdot; W_p)$ denotes pointwise convolution with parameters W_p, enabling cross-channel interaction and feature recombination. The refined feature $\tilde{F}_s$ is then mapped to a spatial attention map $M_s \in \mathbb{R}^{1\times H\times W}$ via the sigmoid activation function $\sigma(\cdot)$, which constrains values to the range $[0, 1]$. The attention map is broadcast along the channel dimension and applied to the input feature f through element-wise multiplication $\otimes$, enabling spatially adaptive feature recalibration. A residual connection is further introduced to preserve the original feature distribution and stabilize gradient propagation. From an optimization perspective, this formulation enhances gradient flow by maintaining identity mapping while selectively amplifying informative spatial regions, thereby improving convergence stability and robustness under feature degradation conditions.

$$
\begin{aligned}
f^s_{\text{avg}} &= \text{AvgPool}_c(f), \quad f^s_{\max} = \text{MaxPool}_c(f),\\
f^s_{\text{mix}} &= f^s_{\text{avg}} \oplus f^s_{\max},\\
F_s &= \text{concat}(f^s_{\text{avg}}, f^s_{\max}, f^s_{\text{mix}}),\\
\tilde{F}_s &= \text{PWConv}\big(\text{DWConv}(F_s; W_d); W_p\big),\\
M_s &= \sigma(\tilde{F}_s),\\
f'_s &= f \otimes M_s + f
\end{aligned}
\tag{3}
$$

D. Spatial Cross-Stage Module

To explicitly address the challenges of object scale variation and object feature loss in UAV-based vehicle detection, the representational capacity of the backbone network is enhanced through the proposed SCM. The SCM is constructed by reinterpreting and adapting the core principles of cross stage partial (CSP), efficient layer aggregation network (ELAN), and spatial pyramid pooling within a unified framework. Rather than directly inheriting their original formulations, each component is structurally redesigned to better facilitate multi-scale feature interaction in drone imagery. This integrated design is motivated by the need to simultaneously preserve fine-grained details and expand contextual perception, which are critical for accurately detecting small objects under complex aerial perspectives. From a theoretical perspective, the fusion of CSP, ELAN, and SPPF can be interpreted as a structured decomposition of feature propagation and aggregation, where CSP preserves gradient flow and low-frequency structural information through partial identity mapping, ELAN enhances feature diversity via multi-path transformations that approximate higher-order feature interactions, and SPPF captures multi-scale contextual dependencies through hierarchical receptive field expansion. This complementary design alleviates feature degradation for small objects by maintaining local detail fidelity while progressively enriching global contextual cues, forming a balanced representation between spatial precision and semantic abstraction.

In implementation, the input feature map $f \in \mathbb{R}^{C\times H\times W}$ is first compressed via a 1×1 convolution, denoted as $\text{Conv}_{1\times1}(\cdot)$, to reduce channel redundancy and facilitate efficient branch decomposition. The compressed feature $\tilde{f}$ is then partitioned into multiple branches, where one branch f_0 preserves low-level information through identity mapping, and the remaining branches $\{f_i\}_{i=1}^N$ perform depthwise separable convolutions $\text{DWConv}(\cdot)$ followed by hierarchical pooling operations $\{\mathcal{P}_k(\cdot)\}_{k=1}^K$ to capture multi-scale contextual features. Here, $\mathcal{P}_k(\cdot)$ denotes pooling with increasing receptive fields, enabling progressive context aggregation. The intermediate features are then aggregated through ELAN-style hierarchical concatenation, denoted as $\mathcal{H}(\cdot)$, which fuses multi-path representations while preserving intermediate diversity. Subsequently, SPPF operations, denoted as $\text{SPPF}(\cdot)$, are applied to selected branches to further enhance contextual encoding. Let $\text{Concat}(\cdot)$ denote channel-wise concatenation, and let $\mathcal{F}(\cdot)$ represent the final feature fusion via 1×1 convolution. From an optimization perspective, this formulation enables iterative refinement of spatial context while maintaining stable gradient propagation across branches, resulting in improved robustness for small and densely distributed object detection.

$$
\begin{aligned}
\tilde{f} &= \text{Conv}_{1\times1}(f),\\
f_0 &= \tilde{f},\\
f_i &= \mathcal{P}_k\big(\text{DWConv}(\tilde{f})\big), \quad i = 1, \ldots, N,
\end{aligned}
$$

$$
\begin{aligned}
F_{\mathrm{ELAN}} &= \mathcal{H}(f_0, f_1, \ldots, f_N), \\
F_{\mathrm{SPP}} &= \mathrm{Concat}\big(\mathrm{SPPF}(f_0),\ \mathrm{SPPF}(F_{\mathrm{ELAN}})\big), \\
f_{\mathrm{out}} &= \mathcal{F}(F_{\mathrm{SPP}})
\end{aligned} \tag{4}
$$

IV. Experimental and Results

In this section, the proposed vehicle detection network is systematically evaluated under a comprehensive experimental setting. Specifically, ablation experiments are first conducted to investigate the contribution of each component in addressing object scale variation and object feature loss. Subsequently, a detailed analysis is performed to examine the key operations that influence detection performance. Finally, comparative experiments are carried out against state-of-the-art methods using multiple evaluation metrics, and the corresponding visual results are analyzed to further validate the effectiveness and robustness of the proposed approach.

A. Experimental Setting

1) Implementation Details: All experiments were conducted on a Windows 10 platform equipped with an Intel Core i9-11900K CPU and an NVIDIA RTX 3090 GPU with 24 GB of memory. The models were trained for 300 epochs using stochastic gradient descent (SGD) with an initial learning rate of 0.01 and a batch size of 16. Mosaic data augmentation was applied during the first 280 epochs and disabled in the final 20 epochs to stabilize convergence. To ensure a fair comparison, all models were trained under a unified experimental protocol, including consistent input resolutions, training schedules, and data augmentation strategies. Specifically, identical input sizes were adopted across all experiments, and all compared methods were retrained under the same settings rather than relying on reported results. In addition, training configurations, including optimizer parameters, learning rate schedules, and augmentation pipelines, were strictly aligned to eliminate potential sources of bias. To further enhance experimental transparency, both training and validation loss curves were continuously monitored throughout the training process to assess convergence behavior and mitigate potential overfitting during extended training.

2) Experimental Data: Most experiments were conducted on the VisDrone dataset, focusing on four vehicle categories: car, van, bus, and truck. All images were resized to 1280×1280 pixels for training and evaluation. To further evaluate the generalization ability of the proposed method, additional comparative experiments were performed on the PUCPR and CARPK datasets, which represent diverse UAV-based scenarios. It should be noted that both CARPK and PUCPR contain only the car category. The dataset splits and experimental settings for VisDrone are summarized in Table I, while the partitioning statistics for the car category in CARPK and PUCPR are presented in Table II.

3) Evaluation Metric: Multiple evaluation metrics were employed, including mean Average Precision at an IoU threshold of 0.5, denoted as $mAP_{0.5}$, at an IoU threshold of 0.75, denoted as $mAP_{0.75}$, and the average over IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05, denoted as $mAP_{0.5:0.95}$. A prediction was considered correct for $mAP_{0.5}$ if the Intersection over Union (IoU) between the predicted and ground-truth bounding boxes exceeded 0.5, and for $mAP_{0.75}$ if it exceeded 0.75. The metric $mAP_{0.5:0.95}$ reflects the average precision across ten IoU thresholds, providing a comprehensive evaluation of detection performance. Among these metrics, $mAP_{0.5}$ primarily reflects classification capability, $mAP_{0.75}$ emphasizes localization accuracy, and $mAP_{0.5:0.95}$ represents the overall detection quality. All mAP values are reported in percentage %. In addition to accuracy metrics, computational efficiency and model complexity were evaluated using the number of parameters (Params (M)) and floating-point operations (GFLOPs), which respectively measure model size and computational cost. Furthermore, training and inference behaviors were analyzed through the loss value, recall rate, and confidence score, which reflect optimization convergence, detection completeness, and prediction reliability, respectively. The inference latency, measured in milliseconds (ms), was also reported to assess real-time performance.

TABLE I

Statistical Analysis of the Object Number Following the Partitioning of the VisDrone Dataset

Set of Type	Car	Bus	Truck	Van
Training Set	121,149	14,290	16,567	24,676
Validation Set	30,201	3,372	4,102	6,113
Test Set	21,110	2,720	2,980	4,502

TABLE II

Statistical Analysis of Car Objects in CARPK and PUCPR Datasets Following Partitioning

Dataset	Training Set	Validation Set	Test Set
CARPK	71,822	8,978	8,977
PUCPR	13,165	1,646	1,645

B. Comparative Experiment

In this section, we conduct a series of ablation studies to verify the effectiveness of each proposed component, including the SCM, dual-branch attention strategies, and the complete SIBA-PAN architecture. All experiments are performed on the VisDrone dataset using CSPNeXt as the backbone and CSPNeXt-PAN as the baseline neck unless otherwise specified. To ensure fair comparison, each module is evaluated under identical training settings, and the results are averaged over multiple runs to guarantee statistical robustness. The goal is to isolate the contribution of each component and validate its impact on both detection accuracy and computational efficiency in UAV-based vehicle detection scenarios.

1) Structured Integrated Cascade Network: In order to validate the effectiveness of SICNet for UAV-based vehicle detection, comprehensive comparative experiments were conducted on three challenging benchmarks, namely VisDrone, PUCPR, and CARPK. These datasets are characterized by densely distributed vehicles and complex urban environments, thereby providing a rigorous testbed for evaluating detection robustness. As reported in Table III, Table IV, and Table V,

TABLE III
COMPARATIVE RESULTS ON THE VISDRONE DATASET IN TERMS OF DETECTION ACCURACY AND SCALE-WISE PERFORMANCE

Method	$mAP_{0.5}$	$mAP_{0.75}$	$mAP_{0.5:0.95}$	mAP_S	mAP_M	mAP_L
Faster R-CNN [6]	55.3±0.3	42.0±0.4	34.2±0.3	18.5±0.3	35.2±0.3	48.0±0.3
Cascade R-CNN [25]	56.7±0.4	46.0±0.3	39.2±0.4	19.6±0.3	37.8±0.4	50.1±0.3
CenterNet [34]	49.4±0.3	36.3±0.2	29.5±0.3	14.2±0.2	30.5±0.3	42.6±0.3
FCOS [8]	50.5±0.2	37.1±0.3	30.1±0.2	15.0±0.3	31.2±0.2	43.5±0.2
EfficientDet-b0 [13]	50.1±0.3	34.9±0.4	27.3±0.3	13.8±0.3	28.6±0.3	40.2±0.3
ATSS [18]	51.6±0.4	37.9±0.3	30.5±0.4	15.6±0.3	32.1±0.3	44.3±0.4
YOLOv5-n [35]	56.6±0.3	44.0±0.4	37.8±0.3	21.9±0.3	38.7±0.3	52.4±0.3
YOLOv5-s [35]	61.8±0.3	48.2±0.4	40.9±0.3	25.6±0.3	42.3±0.3	56.5±0.3
YOLOv5-m [35]	65.0±0.3	51.0±0.4	42.8±0.3	27.8±0.3	43.8±0.3	58.2±0.3
YOLOX-n [20]	60.8±0.4	47.5±0.3	39.7±0.4	24.6±0.3	41.2±0.4	55.3±0.3
YOLOX-s [20]	61.9±0.3	48.6±0.3	41.0±0.3	25.8±0.3	42.5±0.3	56.7±0.3
YOLOX-m [20]	65.2±0.3	51.2±0.4	42.9±0.3	27.9±0.3	44.0±0.3	58.4±0.3
ObjectBox [36]	57.3±0.3	43.8±0.4	36.7±0.3	20.8±0.3	37.5±0.3	50.6±0.3
YOLOv6-n [37]	58.1±0.2	46.2±0.3	39.3±0.2	23.7±0.2	40.5±0.3	54.1±0.2
YOLOv6-s [37]	62.0±0.3	49.0±0.3	41.5±0.3	26.2±0.3	43.0±0.3	57.3±0.3
YOLOv6-m [37]	65.3±0.3	51.5±0.4	43.2±0.3	28.2±0.3	44.4±0.3	58.8±0.3
DINO [9]	59.4±0.4	46.1±0.3	39.3±0.4	22.9±0.4	40.8±0.3	54.7±0.4
TOOD [7]	56.1±0.3	45.0±0.4	38.8±0.3	22.5±0.3	39.6±0.3	52.9±0.3
YOLOv8-n [38]	57.8±0.2	46.0±0.3	39.1±0.2	23.1±0.2	40.2±0.2	53.8±0.2
YOLOv8-s [38]	62.2±0.3	49.5±0.3	41.8±0.3	26.5±0.3	43.2±0.3	57.6±0.3
YOLOv8-m [38]	65.5±0.3	51.8±0.4	43.5±0.3	28.6±0.3	43.8±0.3	59.2±0.3
RT-DETR [33]	58.7±0.3	45.0±0.3	39.2±0.3	24.0±0.3	40.6±0.3	54.5±0.3
YOLOv9-n [39]	59.8±0.3	48.0±0.4	40.7±0.3	25.3±0.3	42.1±0.3	56.2±0.3
YOLOv9-s [39]	62.4±0.3	50.0±0.4	42.2±0.3	27.0±0.3	43.7±0.3	58.2±0.3
YOLOv9-m [39]	65.8±0.3	52.0±0.4	43.7±0.3	28.8±0.3	44.2±0.3	59.8±0.3
YOLOv10-n [21]	57.5±0.3	45.1±0.2	38.7±0.3	22.8±0.3	39.5±0.3	53.0±0.2
YOLOv10-s [21]	61.7±0.3	48.8±0.3	41.2±0.3	26.0±0.3	42.8±0.3	57.0±0.3
YOLOv10-m [21]	65.2±0.3	51.5±0.4	43.1±0.3	28.2±0.3	44.6±0.3	58.9±0.3
YOLOv11-n [22]	57.0±0.2	44.6±0.3	38.4±0.2	22.2±0.2	39.0±0.2	52.6±0.3
YOLOv11-s [22]	61.5±0.3	48.5±0.3	41.0±0.3	25.7±0.3	42.5±0.3	56.8±0.3
YOLOv11-m [22]	65.0±0.3	51.2±0.4	42.8±0.3	27.9±0.3	44.2±0.3	58.5±0.3
Hyper-YOLO-n [40]	60.4±0.4	48.0±0.3	40.7±0.4	25.7±0.4	42.6±0.3	56.8±0.4
YOLOv12-n [23]	58.4±0.3	46.8±0.4	39.6±0.3	23.9±0.3	41.0±0.3	54.6±0.3
YOLOv12-s [23]	62.6±0.3	50.1±0.4	42.3±0.3	26.8±0.3	43.5±0.3	58.0±0.3
YOLOv12-m [23]	65.9±0.3	52.2±0.4	43.8±0.3	29.0±0.3	44.3±0.3	60.2±0.3
RTMDet [24]	64.6±0.3	50.2±0.4	41.2±0.3	28.4±0.3	44.1±0.3	59.8±0.3
SICNet (Ours)	**68.2±0.3**	**54.8±0.4**	**44.0±0.3**	**32.1±0.3**	**44.8±0.4**	**62.5±0.3**

SICNet consistently achieves superior performance across all datasets while maintaining competitive model efficiency. Specifically, it attains the highest $mAP_{0.5}$ scores of 68.2%, 96.5%, and 99.0% on VisDrone, PUCPR, and CARPK, respectively, while also achieving superior $mAP_{0.5:0.95}$ results of 44.0%, 64.1%, and 67.3%.

Compared with recent strong baselines, including RTMDet, YOLOv10, YOLOv11, and YOLOv12, SICNet consistently demonstrates enhanced detection performance, particularly in challenging scenarios involving small-scale and densely distributed vehicles. Notably, SICNet surpasses RTMDet by 3.6, 0.9, and 0.8 percentage points in $mAP_{0.5}$ on VisDrone, PUCPR, and CARPK, respectively. This consistent improvement can be attributed to the structured integration of semantic and spatial attention within SIBA-PAN, which enables more effective feature recalibration across hierarchical levels. By jointly modeling global channel dependencies and local spatial saliency, the proposed framework enhances the discriminability of small objects while suppressing background interference, thereby improving detection robustness in complex aerial scenes. Furthermore, the ablation results of attention combination strategies in Table VI and Table VII demonstrate that the weighted parallel design consistently achieves the best performance across all evaluation metrics. Compared with serial or naive parallel configurations, the weighted fusion mechanism enables dynamic balancing between channel-enhanced and spatial-enhanced features. From a theoretical perspective, this formulation can be interpreted as a soft selection mechanism over complementary feature subspaces, where the learnable coefficient adaptively modulates the contribution of global semantic information and local spatial details. Such a design avoids the information bottleneck introduced by sequential processing and mitigates feature redundancy in standard parallel fusion, resulting in a more stable and expressive feature representation. Consequently, the network is better equipped to capture complex contextual dependencies inherent in UAV imagery. Despite a moderate increase in computational complexity, SICNet maintains competitive inference efficiency, as evidenced in Table V, where it achieves a favorable trade-off between model size, computational cost, and inference

TABLE IV
COMPARATIVE RESULTS ON PUCPR AND CARPK DATASETS IN TERMS OF $mAP_{0.5}$, $mAP_{0.75}$, AND $mAP_{0.5:0.95}$ (%)

Method	$mAP_{0.5}$ - PUCPR	$mAP_{0.75}$ - PUCPR	$mAP_{0.5:0.95}$ - PUCPR	$mAP_{0.5}$ - CARPK	$mAP_{0.75}$ - CARPK	$mAP_{0.5:0.95}$ - CARPK
Faster R-CNN [6]	85.4±0.3	72.0±0.4	52.1±0.3	89.5±0.3	76.3±0.4	56.3±0.3
Cascade R-CNN [25]	87.2±0.4	75.5±0.3	55.3±0.4	91.0±0.4	79.1±0.3	58.7±0.4
CenterNet [34]	82.0±0.3	68.1±0.2	47.8±0.3	87.8±0.3	73.9±0.2	53.0±0.3
FCOS [8]	83.1±0.2	69.8±0.3	48.6±0.2	88.3±0.2	74.6±0.3	54.1±0.2
EfficientDet-b0 [13]	80.7±0.3	66.2±0.4	44.2±0.3	83.2±0.3	67.9±0.4	46.5±0.3
ATSS [18]	84.5±0.4	71.8±0.3	50.3±0.4	89.0±0.4	75.6±0.3	55.5±0.4
YOLOv5-n [35]	89.3±0.3	77.4±0.4	56.7±0.3	92.3±0.3	80.3±0.4	59.6±0.3
YOLOv5-s [35]	92.8±0.3	81.5±0.4	59.5±0.3	94.5±0.3	83.2±0.4	61.2±0.3
YOLOv5-m [35]	94.2±0.3	83.6±0.4	61.2±0.3	95.6±0.3	84.8±0.4	62.5±0.3
YOLOX-n [20]	91.2±0.4	79.9±0.3	58.1±0.4	93.5±0.4	81.7±0.3	60.4±0.4
YOLOX-s [20]	93.0±0.3	82.0±0.3	59.9±0.3	94.8±0.3	83.3±0.3	61.6±0.3
YOLOX-m [20]	94.5±0.3	84.0±0.4	61.5±0.3	95.9±0.3	85.2±0.4	62.8±0.3
ObjectBox [36]	89.0±0.3	76.8±0.4	55.9±0.3	91.8±0.3	79.6±0.4	58.0±0.3
YOLOv6-n [37]	91.8±0.3	80.6±0.2	58.4±0.3	93.8±0.3	82.2±0.2	60.8±0.3
YOLOv6-s [37]	93.6±0.3	82.5±0.3	60.1±0.3	95.2±0.3	84.0±0.3	61.9±0.3
YOLOv6-m [37]	94.8±0.3	84.3±0.4	61.8±0.3	96.1±0.3	85.5±0.4	63.2±0.3
DINO [9]	93.1±0.4	82.0±0.3	59.7±0.4	94.6±0.4	83.1±0.3	61.4±0.4
TOOD [7]	88.7±0.3	76.6±0.4	55.2±0.3	91.3±0.3	79.2±0.4	58.5±0.3
YOLOv8-n [38]	92.4±0.2	81.3±0.3	58.9±0.2	94.1±0.2	82.6±0.3	61.0±0.2
YOLOv8-s [38]	94.0±0.3	83.0±0.3	60.6±0.3	95.3±0.3	84.2±0.3	62.0±0.3
YOLOv8-m [38]	95.2±0.3	84.6±0.4	62.2±0.3	96.5±0.3	85.8±0.4	63.6±0.3
RT-DETR [33]	93.2±0.3	82.1±0.3	59.6±0.3	94.8±0.3	83.4±0.3	61.6±0.3
YOLOv9-n [39]	94.0±0.3	83.2±0.4	60.5±0.3	95.1±0.3	84.0±0.4	62.0±0.3
YOLOv9-s [39]	95.1±0.3	84.5±0.4	61.8±0.3	96.2±0.3	85.3±0.4	63.1±0.3
YOLOv9-m [39]	95.8±0.3	85.2±0.4	62.5±0.3	96.9±0.3	86.3±0.4	64.0±0.3
YOLOv10-n [21]	92.1±0.3	80.7±0.2	58.3±0.3	93.3±0.3	81.5±0.2	59.5±0.3
YOLOv10-s [21]	93.8±0.3	82.6±0.3	60.0±0.3	94.9±0.3	83.6±0.3	61.2±0.3
YOLOv10-m [21]	95.0±0.3	84.2±0.4	61.6±0.3	96.3±0.3	85.5±0.4	62.9±0.3
YOLOv11-n [22]	91.7±0.2	80.2±0.3	57.8±0.2	93.0±0.2	81.1±0.3	59.1±0.2
YOLOv11-s [22]	93.5±0.3	82.0±0.3	59.5±0.3	94.7±0.3	83.2±0.3	60.8±0.3
YOLOv11-m [22]	94.7±0.3	83.8±0.4	61.0±0.3	96.0±0.3	85.0±0.4	62.4±0.3
Hyper-YOLO-n [40]	94.5±0.4	83.8±0.3	61.0±0.4	95.4±0.4	84.4±0.3	62.3±0.4
YOLOv12-n [23]	95.0±0.3	84.5±0.4	61.4±0.3	95.7±0.3	85.0±0.4	62.7±0.3
YOLOv12-s [23]	95.8±0.3	85.5±0.4	62.6±0.3	96.6±0.3	86.2±0.4	63.8±0.3
YOLOv12-m [23]	96.1±0.3	86.2±0.4	63.2±0.3	97.2±0.3	87.2±0.4	64.5±0.3
RTMDet [24]	95.6±0.3	85.8±0.4	62.9±0.3	98.2±0.3	88.7±0.4	64.8±0.3
SICNet (Ours)	**96.5±0.3**	**87.4±0.4**	**64.1±0.3**	**99.0±0.3**	**89.8±0.4**	**67.3±0.3**

speed. This balance makes SICNet suitable for real-world UAV deployment scenarios.

2) Structured Integrated Balanced Attention Path Aggregation Network: To evaluate the effectiveness of the proposed SIBA-PAN, comprehensive comparisons were conducted against Various attention integration strategies and mainstream feature pyramid network designs. Table VI summarizes the results of four attention combination strategies implemented under the same CSPNeXt backbone [24]. Among these, the weighted parallel configuration employed by SIBA-PAN achieved the highest detection performance, reaching 67.7% $mAP_{0.5}$, 60.6% $mAP_{0.75}$, and 48.6% $mAP_{0.5:0.95}$. In contrast, conventional serial and basic parallel strategies exhibited lower accuracy, indicating that naïve stacking or merging of attention modules does not sufficiently capture the complex spatial and semantic dependencies present in UAV imagery. The effectiveness of SIBA-PAN is further supported by complementary feature response metrics, as reported in Table VII. The weighted parallel configuration demonstrates the highest Precision(B), P/R ratio, and lowest training loss among the four strategies, highlighting its ability to produce stronger feature activation in foreground regions while maintaining effective suppression of background responses. These metrics illustrate how the weighted parallel design enhances both discriminability and feature robustness across hierarchical features. Additional evaluations across different feature pyramid networks, shown in Table VIII, further validate the superiority of SIBA-PAN. Despite sharing the same CSPNeXt backbone [24], SIBA-PAN surpasses CSPNeXt-PAN and PDPA-PAN [44] by 3.1% and 0.9% $mAP_{0.5}$, respectively, and outperforms HSFPN [42] and GFPN [43]. These results confirm that SIBA-PAN effectively strengthens cross-scale feature interaction and achieves state-of-the-art detection of small, densely distributed vehicles in complex urban UAV imagery.

3) Structured Grouped Attention Module and Integrated Spatial Attention Module: To evaluate the effectiveness of different attention mechanisms for UAV-based vehicle detection, a comprehensive comparison was conducted involving both widely adopted modules and the proposed designs. As shown in Table IX, Table X, and Fig. 5, conventional mechanisms such as SE, ECA, CA, and CBAM [31] provide only marginal improvements over the baseline

TABLE V
COMPARISON OF DIFFERENT METHODS ON THE VISDRONE DATASET IN TERMS OF MODEL EFFICIENCY

Method	Params (M)	GFLOPs	FPS
Faster R-CNN [6]	41.3	191	21.0
Cascade R-CNN [25]	69.3	219	18.4
CenterNet [34]	32.3	184	31.4
FCOS [8]	32.2	184	30.1
EfficientDet-b0 [13]	3.8	13.8	88.5
ATSS [18]	32.3	189	29.3
YOLOv5-n [35]	7.0	31.7	74.6
YOLOv5-s [35]	8.2	32.5	68.3
YOLOv5-m [35]	21.4	102.3	52.6
YOLOX-n [20]	8.9	53.2	71.9
ObjectBox [36]	86.1	203.7	24.0
YOLOv6-n [37]	4.2	11.9	96.2
YOLOv6-s [37]	8.5	33.1	71.4
YOLOv6-m [37]	22.1	101.8	54.2
DINO [9]	47.5	252	17.4
TOOD [7]	34.1	185	29.9
YOLOv8-n [38]	3.1	8.9	105.3
YOLOv8-s [38]	8.2	32.7	76.5
YOLOv8-m [38]	21.8	101.2	56.8
RT-DETR [33]	21.0	233.2	19.2
YOLOv9-n [39]	7.6	7.6	82.6
YOLOv9-s [39]	8.4	31.9	72.8
YOLOv9-m [39]	22.5	103.6	55.3
YOLOv10-n [21]	2.7	8.2	103.1
YOLOv10-s [21]	8.1	32.4	79.6
YOLOv10-m [21]	21.9	102.7	58.1
YOLOv11-n [22]	**2.6**	6.3	**113.6**
YOLOv11-s [22]	7.9	31.5	82.4
YOLOv11-m [22]	22.3	101.9	60.2
Hyper-YOLO-n [40]	4.1	10.8	94.3
YOLOv12-n [23]	4.8	12.1	92.6
YOLOv12-s [23]	8.6	33.6	75.2
YOLOv12-m [23]	22.7	104.1	57.4
RTMDet [24]	4.1	**4.7**	107.5
SICNet (Ours)	6.7	16.1	84.0

CSPNeXt-PAN. Notably, SE and CBAM even lead to performance degradation, with $mAP_{0.5}$ decreasing to 64.2% and 64.5%, respectively. These results suggest that general-purpose attention mechanisms are not sufficiently tailored to the unique challenges of UAV imagery, such as small object scales and dense spatial distributions. In contrast, the proposed attention modules yield significant performance gains. SGAM enhances channel-wise feature discrimination, while ISAM improves spatial localization accuracy. As further evidenced in Table X, different convolutional structure designs within SGAM exhibit varying effectiveness, where the proposed combination of 1×1 and 2×2 convolutions achieves the best trade-off between representation capability and computational efficiency. When integrated into the unified SIBAM, the network achieves the best overall performance, with 67.7% $mAP_{0.5}$, 60.6% $mAP_{0.75}$, and 48.6% $mAP_{0.5:0.95}$. These improvements are obtained with only minimal additional computational overhead while maintaining real-time inference speed. Overall, the results validate both the effectiveness and practicality of the proposed attention designs for addressing dense object detection challenges in UAV-based scenarios.

4) Spatial Cross-Stage Module: To evaluate the effectiveness of the proposed SCM, a comprehensive comparison was conducted against several widely used SPP variants, including SPP, SPPF, ASPP, SPPCSPC, and SPPFCSPC. As shown in Table XI, all experiments utilized the CSPNeXt backbone with a consistent neck configuration to ensure a fair comparison. The proposed SCM achieves the highest $mAP_{0.5}$ of 65.2%, surpassing the conventional SPP and SPPF by 0.6% and 1.5%, respectively. Moreover, SCM outperforms advanced variants such as ASPP, SPPCSPC, and SPPFCSPC, demonstrating superior multi-scale feature aggregation capabilities. Furthermore, SCM attains the best $mAP_{0.75}$ and $mAP_{0.5:0.95}$ scores of 57.8% and 46.4%, respectively, indicating enhanced localization accuracy and overall detection performance. These results confirm SCM's practicality as a robust enhancement for spatial feature representation in dense UAV-based vehicle detection scenarios.

C. Ablation Experiments

In this section, a series of ablation studies are conducted to verify the effectiveness of each proposed component, including the SGAM, ISAM, weighted fusion, and SCMs. All experiments are performed on the VisDrone dataset using CSPNeXt as the backbone. To ensure a fair comparison, each module is evaluated incrementally under identical training settings, with performance measured by $mAP_{0.5}$ and $mAP_{0.5:0.95}$. The results demonstrate the individual and combined contributions of these components to detection accuracy while maintaining reasonable computational efficiency in UAV-based object detection scenarios.

1) Structured Integrated Cascade Network: Table XII presents the ablation study results of SICNet on the VisDrone, PUCPR, and CARPK datasets, highlighting the individual contributions of each component. By progressively integrating SGAM, ISAM, the weighted fusion strategy, and the SCM, consistent improvements in detection performance are observed across all datasets. Specifically, for the VisDrone dataset, $mAP_{0.5}$ increases from 64.6% to 68.2%, with $mAP_{0.5:0.95}$ rising from 41.2% to 49.1%. On the PUCPR dataset, $mAP_{0.5}$ improves from 95.6% to 96.5%, while $mAP_{0.5:0.95}$ increases from 62.9% to 64.1%. Similarly, on the CARPK dataset, $mAP_{0.5}$ rises from 98.2% to 99.0% and $mAP_{0.5:0.95}$ from 64.8% to 67.3%. The corresponding parameter counts also increase gradually, reaching 6.7M for the full SICNet, with a moderate GFLOPs growth from 4.7 to 16.1. These results validate the effectiveness of each proposed module in enhancing detection accuracy, while maintaining a computational cost that remains feasible for UAV-assisted edge processing. The ablation analysis demonstrates that the inclusion of SGAM and ISAM provides the largest marginal gains in feature representation, the weighted fusion further optimizes the integration of channel and spatial cues, and SCM contributes to multi-scale context aggregation, collectively enabling robust performance across diverse UAV imagery scenarios. This comprehensive evaluation confirms the design rationale and practical efficiency of SICNet in real-world aerial vehicle detection tasks.

TABLE VI
COMPARISON OF DIFFERENT ATTENTION COMBINATION STRATEGIES

Backbone	Combination Strategy				$mAP_{0.5}$	$mAP_{0.75}$	$mAP_{0.5:0.95}$
	Serial (S+I)	Serial (I+S)	Parallel	Weighted Parallel			
CSPNeXt	✓				66.5±0.3	59.3±0.4	46.9±0.3
CSPNeXt		✓			66.9±0.3	59.7±0.3	47.4±0.3
CSPNeXt			✓		67.3±0.2	60.2±0.3	48.0±0.2
CSPNeXt				✓	**67.7±0.3**	**60.6±0.4**	**48.6±0.3**

TABLE VII
EVALUATION OF ATTENTION COMBINATION STRATEGIES WITH COMPLEMENTARY FEATURE RESPONSE METRICS

Backbone	Combination Strategy				Prec(B)	P/R Ratio	Train Loss
	Serial (S+I)	Serial (I+S)	Parallel	Weighted Parallel			
CSPNeXt	✓				0.63	0.90	0.047
CSPNeXt		✓			0.65	0.86	0.044
CSPNeXt			✓		0.61	0.88	0.042
CSPNeXt				✓	**0.72**	**0.91**	**0.038**

TABLE VIII
COMPARISON OF DIFFERENT FEATURE PYRAMID NETWORKS ON UAV-BASED VEHICLE DETECTION

Feature Pyramid Network	$mAP_{0.5}$	$mAP_{0.75}$	$mAP_{0.5:0.95}$
FPN [41]	61.2±0.3	54.1±0.4	42.3±0.3
PAN [12]	62.8±0.3	55.3±0.3	43.9±0.3
CSPNeXt-PAN [24]	64.6±0.3	57.0±0.4	45.6±0.3
HSFPN [42]	65.2±0.2	57.5±0.3	46.2±0.2
GFPN [43]	65.7±0.3	57.8±0.3	46.5±0.3
PDPA-PAN [44]	66.8±0.4	59.4±0.4	47.6±0.3
SIBA-PAN (Ours)	**67.7±0.3**	**60.6±0.4**	**48.6±0.3**

TABLE IX
COMPARISON OF DIFFERENT ATTENTION MECHANISMS ON UAV-BASED VEHICLE DETECTION

Attention Module	$mAP_{0.5}$	$mAP_{0.75}$	$mAP_{0.5:0.95}$
None	64.6±0.3	57.0±0.4	45.6±0.3
SE [30]	64.2±0.3	56.6±0.3	45.3±0.3
ECA [11]	64.9±0.2	57.1±0.3	45.7±0.2
CA [45]	65.1±0.3	57.3±0.3	45.9±0.3
CBAM [31]	64.5±0.3	56.7±0.4	45.5±0.3
SGAM (Ours)	66.8±0.3	59.0±0.4	47.4±0.3
ISAM (Ours)	67.0±0.3	59.3±0.3	47.7±0.3
SIBAM (Ours)	**67.7±0.3**	**60.6±0.4**	**48.6±0.3**

TABLE X
COMPARISON OF DIFFERENT CONVOLUTIONAL STRUCTURES IN SGAM FOR UAV-BASED VEHICLE DETECTION

SGAM Structure	$mAP_{0.5}$	$mAP_{0.75}$	$mAP_{0.5:0.95}$
1×1	65.4±0.3	57.8±0.4	46.2±0.3
2×2	65.7±0.3	58.0±0.3	46.4±0.3
1×1 + 3×3	65.9±0.4	58.6±0.4	46.8±0.3
1×1 + 2×2 + 3×3	65.6±0.3	58.2±0.4	46.6±0.3
1×1 + 2×2 (Ours)	**66.8±0.3**	**59.0±0.4**	**47.4±0.3**

TABLE XI
COMPARISON OF DIFFERENT SPP MODULES ON UAV-BASED VEHICLE DETECTION

Backbone	Module	$mAP_{0.5}$	$mAP_{0.75}$	$mAP_{0.5:0.95}$
CSPNeXt	SPP [27]	64.6±0.3	57.0±0.3	45.6±0.3
CSPNeXt	SPPF [38]	63.7±0.3	56.0±0.3	44.7±0.3
CSPNeXt	ASPP [29]	63.9±0.2	56.2±0.3	45.0±0.2
CSPNeXt	SPPCSPC [35]	64.2±0.3	56.5±0.3	45.2±0.3
CSPNeXt	SPPFCSPC [37]	64.4±0.3	56.7±0.3	45.5±0.3
CSPNeXt	**SCM (Ours)**	**65.2±0.3**	**57.8±0.4**	**46.4±0.3**

2) Structured Grouped Attention Module and Integrated Spatial Attention Module: The effectiveness of the proposed SGAM module was thoroughly evaluated through an ablation study, as shown in Table XIII. By progressively incorporating global pooling, local context pooling, and group convolution, consistent improvements in detection performance were achieved. Specifically, $mAP_{0.5}$ increased from 64.6% to 66.8%, and $mAP_{0.5:0.95}$ rose from 45.6% to 47.4%, with the parameter count increasing moderately from 4.1M to 4.8M and GFLOPs from 15.4 to 16.8. These results demonstrate that SGAM effectively enhances multi-scale spatial feature aggregation while introducing minimal computational overhead. Similarly, the performance contribution of the ISAM module was assessed in Table XIV. Introducing the average pooling stream led to noticeable gains, and further adding the max pooling stream provided additional improvements. The full ISAM configuration, which integrates both streams through combined pooling fusion, achieved the highest detection scores, with $mAP_{0.5}$ reaching 66.1% and $mAP_{0.5:0.95}$ reaching 47.3%, while the parameter count increased from 4.1M to 4.5M and GFLOPs rose only marginally to 15.9. These findings

TABLE XII
ABLATION STUDY OF SICNET ON THE VISDRONE, PUCPR, AND CARPK DATASETS

Dataset	Model	SGAM	ISAM	Weighted Fusion	SCM	Performance (%)		Params (M)	GFLOPs
						$mAP_{0.5}$	$mAP_{0.5:0.95}$		
VisDrone	RTMDet					64.6±0.3	41.2±0.3	**4.1**	**4.7**
	RTMDet	✓				66.5±0.3	42.9±0.3	4.8	6.1
	RTMDet	✓	✓			67.3±0.3	44.0±0.3	5.2	7.4
	RTMDet	✓	✓	✓		67.8±0.3	44.6±0.3	5.6	8.6
	RTMDet	✓	✓	✓	✓	**68.2±0.3**	**49.1±0.3**	6.7	16.1
PUCPR	RTMDet					95.6±0.3	62.9±0.3	**4.1**	**4.7**
	RTMDet	✓				96.0±0.3	63.7±0.3	4.8	6.1
	RTMDet	✓	✓			96.3±0.3	63.9±0.3	5.2	7.4
	RTMDet	✓	✓	✓		96.4±0.3	64.0±0.3	5.6	8.6
	RTMDet	✓	✓	✓	✓	**96.5±0.3**	**64.1±0.3**	6.7	16.1
CARPK	RTMDet					98.2±0.3	64.8±0.3	**4.1**	**4.7**
	RTMDet	✓				98.4±0.3	65.5±0.3	4.8	6.1
	RTMDet	✓	✓			98.5±0.3	66.1±0.3	5.2	7.4
	RTMDet	✓	✓	✓		98.6±0.3	66.7±0.3	5.6	8.6
	RTMDet	✓	✓	✓	✓	**99.0±0.3**	**67.3±0.3**	6.7	16.1

TABLE XIII
ABLATION STUDY OF SGAM ON THE VISDRONE DATASET

Backbone	Global Pooling (1×1)	Local Context Pooling (2×2)	Group Conv	$mAP_{0.5}$	$mAP_{0.5:0.95}$	Params (M)	GFLOPs
CSPNeXt				64.6±0.3	45.6±0.3	**4.1**	**15.4**
CSPNeXt	✓			65.2±0.3	46.2±0.3	4.3	15.9
CSPNeXt	✓	✓		65.8±0.3	46.8±0.3	4.5	16.3
CSPNeXt	✓	✓	✓	**66.8±0.3**	**47.4±0.3**	4.8	16.8

TABLE XIV
ABLATION STUDY OF ISAM ON THE VISDRONE DATASET

Backbone	Average Pooling Stream	Max Pooling Stream	Combined Pooling Fusion	$mAP_{0.5}$	$mAP_{0.5:0.95}$	Params (M)	GFLOPs
CSPNeXt				64.6±0.3	45.6±0.3	**4.1**	**15.4**
CSPNeXt	✓			65.1±0.3	46.2±0.3	4.3	15.5
CSPNeXt	✓	✓		65.6±0.3	46.8±0.3	4.4	15.7
CSPNeXt	✓	✓	✓	**66.1±0.3**	**47.3±0.3**	4.5	15.9

TABLE XV
ABLATION STUDY OF SCM ON THE VISDRONE DATASET WITH EXTENDED METRICS

Backbone	CSP	ELAN	DWConv	$mAP_{0.5}$	$mAP_{0.5:0.95}$	Prec(B)	P/R Ratio	Train Resp	Params(M)	GFLOPs
CSPNeXt				64.3±0.3	47.7±0.3	0.61	0.88	0.046	**4.1**	**15.4**
CSPNeXt	✓			64.7±0.3	48.2±0.3	0.63	0.87	0.044	4.4	15.9
CSPNeXt	✓	✓		64.9±0.3	48.7±0.3	0.64	0.89	0.042	4.7	16.3
CSPNeXt	✓	✓	✓	**65.2±0.3**	**49.1±0.4**	**0.70**	**0.91**	**0.038**	5.0	16.8

highlight ISAM's effectiveness in refining spatial attention and capturing complementary contextual information. Overall, the ablation studies of SGAM and ISAM confirm that each module contributes positively to the detection performance, with controlled increases in computational cost and parameters. SGAM primarily enhances multi-scale feature representation, whereas ISAM effectively captures complementary spatial context, together providing a robust improvement in UAV-based vehicle detection accuracy while maintaining practical efficiency. This analysis underscores the rationale for the module designs and supports their inclusion in the full SICNet architecture.

3) Spatial Cross-Stage Module: Table XV presents the ablation results of the proposed SCM on the VisDrone dataset. To evaluate the contribution of each structural component, CSP connections, the ELAN structure, and DWConv were progressively integrated into the baseline CSPNeXt backbone. The baseline model achieved a $mAP_{0.5}$ of 64.3%, a precision of 0.61, a P/R ratio of 0.88, a training response of 0.046,

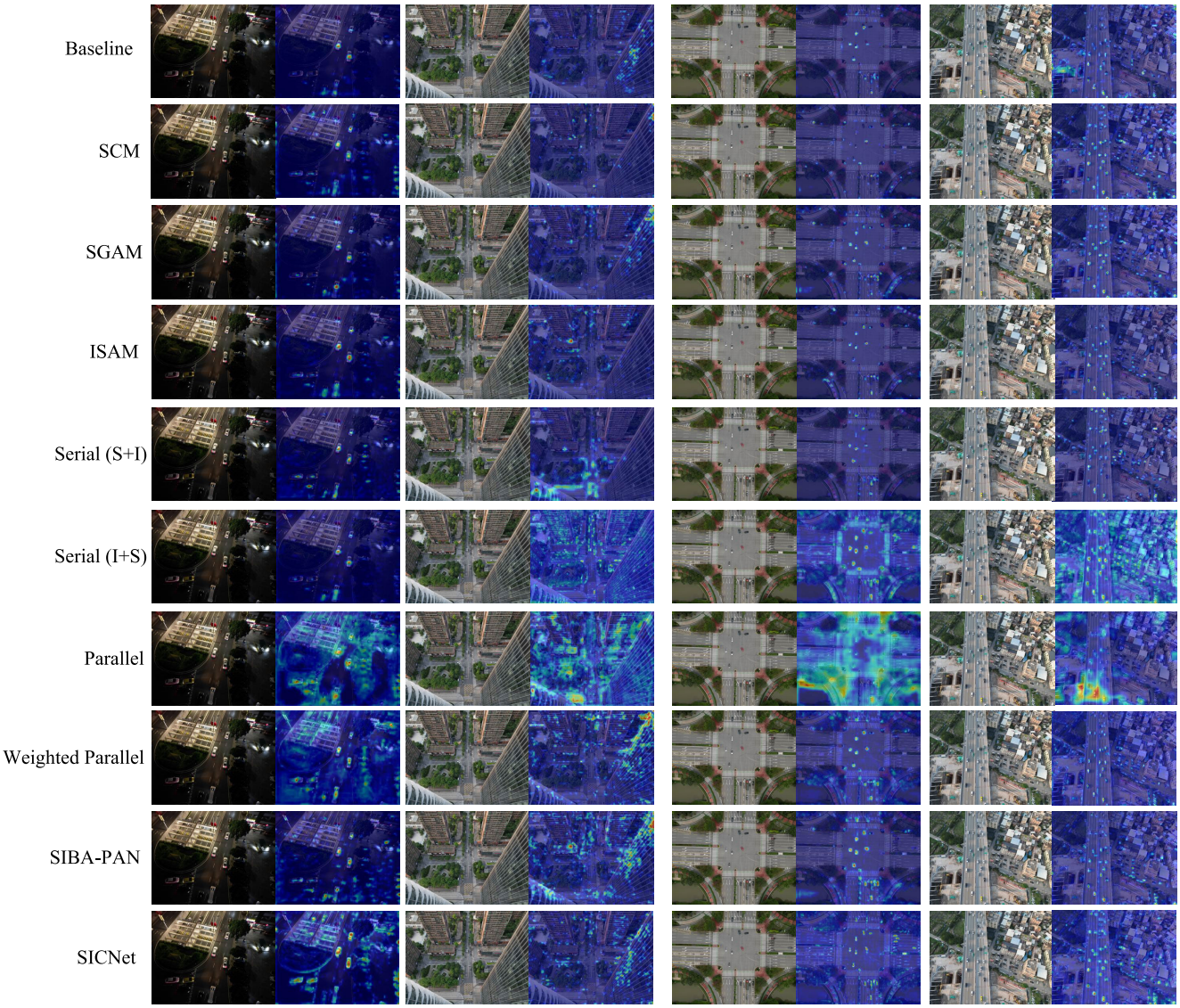

Fig. 6. Heatmap visualization of detection performance across different model variants on the VisDrone dataset, including Baseline, SCM, SGAM, ISAM, Serial (S+I), Serial (I+S), Parallel, Weighted Parallel, SIBA-PAN, and the final SICNet. The figure illustrates the progressive enhancement in spatial focus and activation quality as each proposed component is integrated. Notably, SICNet demonstrates the most concentrated and accurate responses over vehicle regions, highlighting the effectiveness and synergy of the full architecture.

with 4.1M parameters and 15.4 GFLOPs. Introducing CSP connections increased $mAP_{0.5}$ to 64.7%, improved precision to 0.63, and reduced the training response to 0.044, indicating enhanced feature reuse and better foreground–background discrimination. Subsequently, adding the ELAN structure further improved $mAP_{0.5}$ to 64.9%, increased the P/R ratio to 0.89, and decreased the training response to 0.042, demonstrating the effectiveness of hierarchical multi-branch aggregation for enriching multi-scale feature representations. Finally, incorporating DWConv achieved the highest performance, with $mAP_{0.5}$ reaching 65.2%, precision increasing to 0.70, P/R ratio to 0.91, and training response decreasing to 0.038. The total parameter count and GFLOPs increased moderately to 5.0M and 16.8, respectively, maintaining computational efficiency while consistently improving detection accuracy. These results validate the effectiveness of the integrated SCM design. Specifically, CSP preserves low-level spatial information through partial feature propagation, which is critical for small-object representation. ELAN enhances feature diversity and stabilizes gradient flow via hierarchical multi-branch aggregation, improving intermediate feature representations across scales. DWConv strengthens local feature extraction while maintaining efficiency, enabling more discriminative responses under complex background conditions. The coordinated integration of these components forms a complementary mechanism in which fine-grained detail retention, multi-scale feature diversification, and hierarchical context modeling are jointly optimized. Consequently, SCM effectively mitigates feature degradation in UAV imagery and enhances robustness for detecting small and densely distributed vehicles.

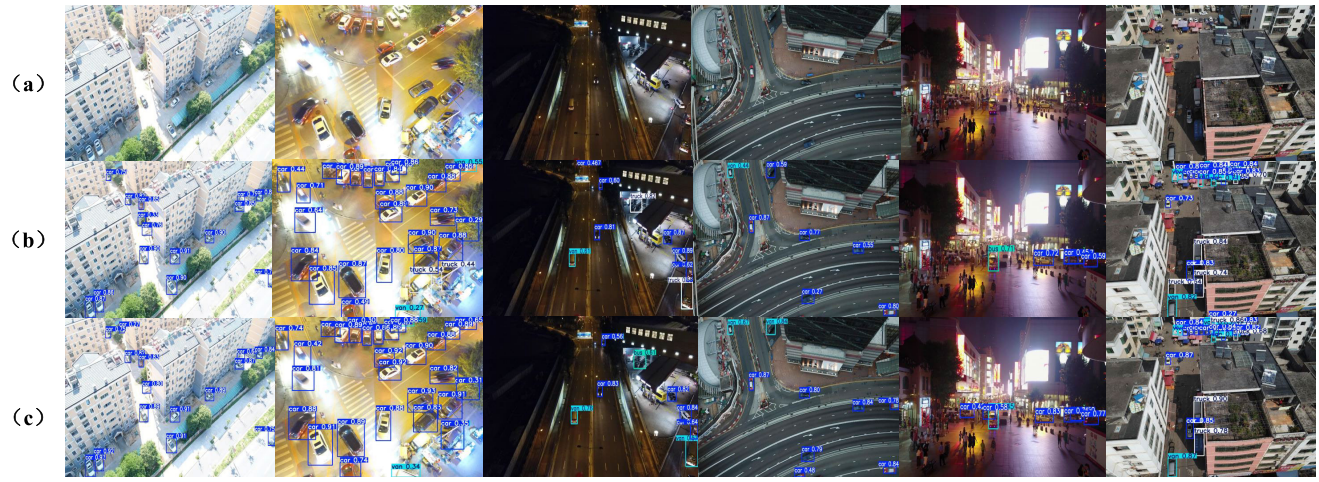

Fig. 7. The qualitative analysis of VisDrone was performed on SICNet. (a) Original images. (b) Detection results from baseline. (c) Detection results from our proposed SICNet.

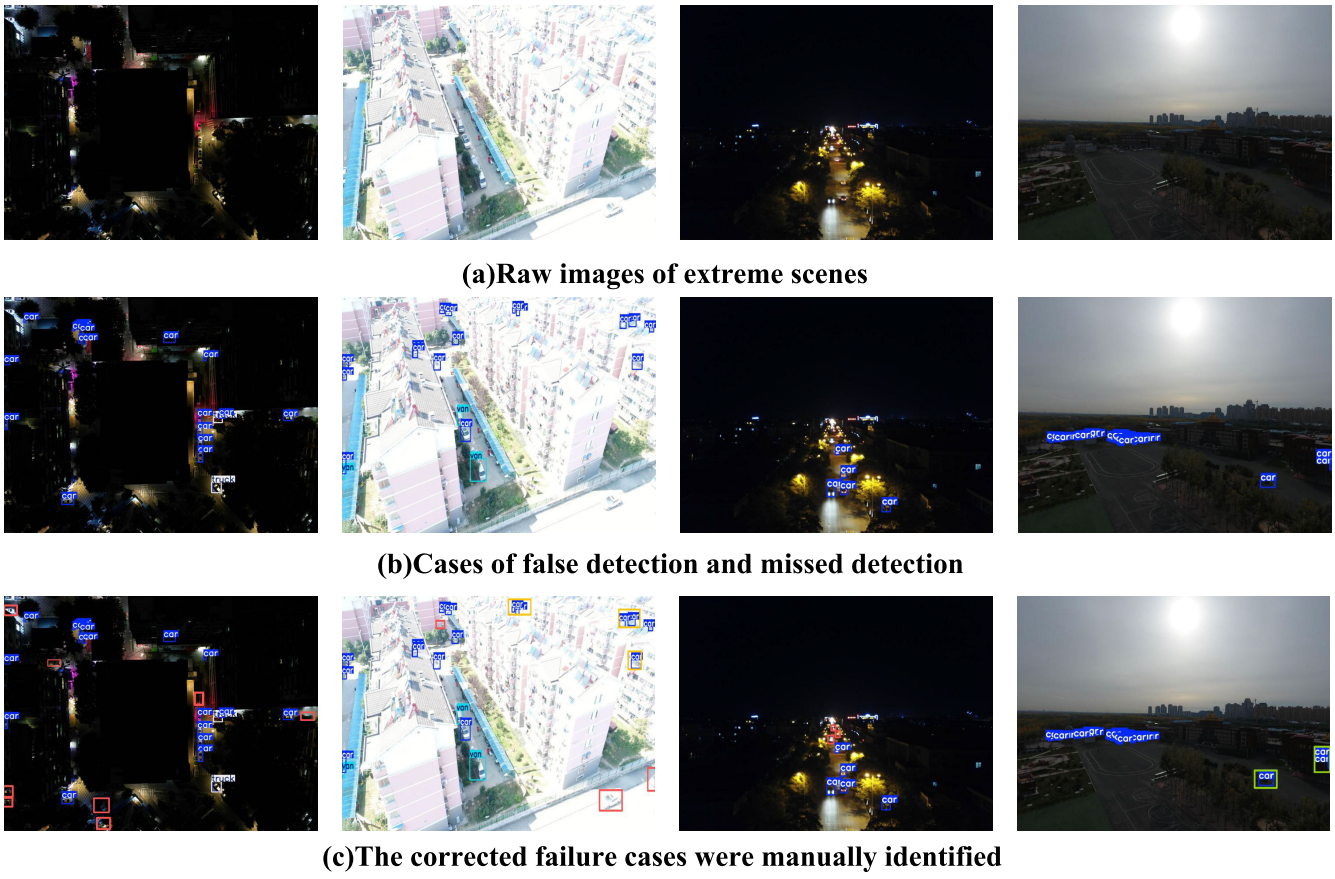

Fig. 8. Qualitative analysis on the VisDrone dataset under challenging UAV scenarios. (a) Raw images of extreme scenes, including severe occlusion under nighttime conditions, strong illumination interference in daytime, feature degradation in low-light environments, and feature ambiguity caused by haze. (b) Representative failure cases of the baseline detector, illustrating false detections and missed detections. (c) Manually corrected annotations based on (b), where yellow bounding boxes denote false positives and red bounding boxes indicate missed detections.

4) Discussion: To comprehensively evaluate the contribution and practical effectiveness of each proposed component, both quantitative and qualitative analyses are conducted. As illustrated in Fig. 6, heatmap-based comparisons among different model variants, including Baseline, SCM, SGAM, ISAM, Serial (S+I), Serial (I+S), Parallel, Weighted Parallel, SIBA-PAN, and the final SICNet, demonstrate that progressive module integration leads to consistent performance improvements. Specifically, early-stage components such as SCM and SGAM yield notable gains by enhancing multi-scale

feature aggregation and structured channel discrimination. The introduction of ISAM further refines spatial responses through multi-statistical pooling and efficient convolutional operations, enabling more accurate localization under feature degradation. Furthermore, different attention combination strategies exhibit varying effectiveness. Serial configurations provide limited improvements due to constrained information flow, whereas parallel designs facilitate complementary feature learning. In particular, the weighted parallel formulation achieves superior performance by enabling adaptive balancing between channel and spatial attention, resulting in more discriminative and robust feature representations. When integrated within the SIBA-PAN framework, this design further enhances feature interaction across hierarchical levels. The final integration of all components into SICNet achieves the highest accuracy, confirming the complementary effectiveness and synergistic interaction of the proposed modules. To further assess practical applicability, Fig. 7 presents qualitative comparisons across six representative UAV scenarios, including partial occlusion by buildings, nighttime intersections, elevated highways under low illumination, daytime overpasses, mixed urban traffic, and densely structured old-town environments. Across all scenarios, SICNet demonstrates superior robustness by producing more accurate localization of small and partially occluded vehicles, reducing false positives in cluttered backgrounds, and maintaining stable performance under varying illumination conditions. In addition, Fig. 8 provides a detailed analysis of typical failure cases, including both false detections and missed detections under challenging conditions such as severe occlusion, illumination interference, low-light environments, and haze. By comparing baseline predictions with manually corrected annotations, it can be observed that SICNet significantly alleviates these issues, particularly in densely distributed and visually ambiguous regions. This improvement is primarily attributed to enhanced feature discrimination and contextual modeling enabled by the structured attention mechanisms. Overall, these results indicate that SICNet not only achieves high detection accuracy but also generalizes effectively across complex UAV scenarios, demonstrating strong potential for reliable aerial vehicle detection in real-world applications.

V. Conclusion and Future Work

In this study, a novel detection framework, SICNet, was proposed to address the challenges of vehicle detection in low-altitude UAV imagery. Multiple attention-driven components, including the structured grouped attention module (SGAM), the integrated spatial attention module (ISAM), and their parallel combination, the structured-integrated balanced attention module (SIBAM), were integrated into the structured integrated balanced attention path aggregation network (SIBA-PAN) to enhance the feature aggregation process. In addition, a spatial cross-stage module (SCM) was introduced into the backbone to efficiently capture multi-scale contextual information. The effectiveness of SICNet was validated through extensive experiments conducted on the VisDrone dataset, where superior performance in detecting small and densely distributed vehicles under complex urban backgrounds was demonstrated. These results confirmed that both feature expressiveness and detection accuracy were significantly improved by the proposed modules while maintaining computational efficiency.

In future work, several directions will be pursued to further enhance the practical value of the proposed framework. Although SICNet achieves competitive accuracy, the attention structures still introduce non-negligible inference delays. Consequently, efforts will focus on optimizing the framework toward hardware-efficient and lightweight architectures to enable real-time performance on resource-constrained UAV platforms. Additionally, the long-tail distribution problem commonly observed in UAV datasets, such as VisDrone, will be addressed through class-balanced learning strategies, data augmentation, and dynamic reweighting mechanisms to improve detection robustness for underrepresented vehicle categories. The framework will also be extended to handle video-based inputs, enabling temporal reasoning for tasks such as vehicle tracking and traffic flow analysis in dynamic scenes. Finally, the incorporation of fine-grained scenario annotations within datasets will be explored to facilitate more comprehensive performance evaluations across diverse urban environments. Collectively, these directions are expected to further improve the reliability, efficiency, and applicability of UAV-based vehicle detection systems in intelligent transportation and urban monitoring applications.

References

[1] R. Feng, H. Cui, Q. Feng, S. Chen, X. Gu, and B. Yao, "Urban traffic congestion level prediction using a fusion-based graph convolutional network," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14695–14705, Dec. 2023.

[2] B.-Z. Li, X.-L. Zhao, X. Chen, M. Ding, and R. Wen Liu, "Convolutional low-rank tensor representation for structural missing traffic data imputation," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 11, pp. 18847–18860, Nov. 2024.

[3] Y. Zhang, X. Kong, W. Zhou, J. Liu, Y. Fu, and G. Shen, "A comprehensive survey on traffic missing data imputation," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 12, pp. 19252–19275, Dec. 2024.

[4] J. Li, C. Liao, S. Hu, X. Chen, and D.-H. Lee, "Physics-guided multi-source transfer learning for network-scale traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 11, pp. 17533–17546, Nov. 2024.

[5] Z. Zeng et al., "Low-rank tensor and hybrid smoothness regularization for spatio-temporal traffic data reconstruction," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 13014–13026, Oct. 2024.

[6] 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.

[7] Z. Li, Z. Zheng, Y. Xia, F. You, X. Ge, and P. Liu, "You only learn one representation: Unified network for multiple tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, May 2022, pp. 2667–2676.

[8] 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. 9626–9635.

[9] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. Cham*, Switzerland: Springer, 2020, pp. 213–229.

[10] 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.

[11] 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.

[12] 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. (CVPR)*, Jun. 2018, pp. 8759–8768.
[13] 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. 10778–10787.
[14] 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. 7029–7038.
[15] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.
[16] D. Du et al., "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 213–226.
[17] Z. Du, Z. Hu, G. Zhao, Y. Jin, and H. Ma, "Cross-layer feature pyramid transformer for small object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5625714.
[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[19] H. Zhang, Y. Wang, F. Dayoub, and N. Sünderhauf, "VarifocalNet: An IoU-aware dense object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8510–8519.
[20] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.
[21] A. Wang et al., "YOLOv10: Real-time end-to-end object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2024, pp. 107984–108011.
[22] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," 2024, *arXiv:2410.17725*.
[23] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric real-time object detectors," 2025, *arXiv:2502.12524*.
[24] C. Lyu et al., "RTMDet: An empirical study of designing real-time object detectors," 2022, *arXiv:2212.07784*.
[25] 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.
[26] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1577–1586.
[27] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[29] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
[30] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[32] L. Yuan et al., "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 538–547.
[33] Y. Zhao, J. Chen, S. Zhang, Y. Cao, J. Wang, and Q. Huang, "RT-DETR: Real-time detection transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 7424–7433.
[34] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet++ for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3509–3521, May 2024.
[35] (2020). YOLOv5. Accessed: Mar. 30, 2026. [Online]. Available: https://github.com/ultralytics/yolov5
[36] M. Zand, A. Etemad, and M. Greenspan, "ObjectBox: From centers to boxes for anchor-free object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2022, pp. 390–406.
[37] 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.
[38] Ultralytics. (2023). *YOLOv8-Ultralytics*. Accessed: Mar. 30, 2026. [Online]. Available: https://github.com/ultralytics/ultralytics
[39] A. Sánchez et al., "Agave crop segmentation and maturity classification with deep learning data-centric strategies using very high-resolution satellite imagery," *Int. J. Remote Sens.*, vol. 44, no. 22, pp. 7017–7032, Nov. 2023, doi: 10.1080/01431161.2023.2275320.
[40] 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.
[41] T. Y. Lin, P. Dollàr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.
[42] Y. Chen et al., "Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases," *Comput. Biol. Med.*, vol. 170, Mar. 2024, Art. no. 107917.
[43] Q. Yang et al., "GraphFPN: Graph feature pyramid network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2800–2809.
[44] 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, doi: 10.1109/TITS.2024.3396915.
[45] 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.

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.

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.

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.

Cuilin Yu received the Ph.D. degree from Sun Yat-sen University, China.

She is currently a Post-Doctoral Researcher with the School of Electronics and Communication Engineering, Sun Yat-sen University. Her research interests focus on multi-source remote sensing data fusion and the cross-disciplinary application of machine learning in geospatial analysis.

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

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

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

He is currently 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.

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

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

Pasquale Coscia (Senior Member, IEEE) received the Ph.D. degree in industrial and information engineering from the Universitá degli Studi di Milano, Italy, in 2019.

From 2019 to 2022, he was a Post-Doctoral Researcher with the Universitá degli Studi di Padova. He has received the 2021 Best Paper Award of the journal *Image and Vision Computing* (Elsevier) for his work on human motion prediction. Additionally, he is a member of the program committees for numerous IEEE international conferences and workshops. Since 2023, he has held the position of the Co-Chair for the Intelligent Measurement Systems Technical Committee (TC-22) within the IEEE Instrumentation and Measurement Society. His main research interests include the areas of the aspects of computational intelligence for signal and image processing, computer vision, machine learning, and probabilistic models applied to multiagents systems.

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

He has been a Post-Doctoral Research Fellow in computer science with the Universitá degli Studi di Milano since 2014. He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. His 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 an Associate Editor of the *Journal of Ambient Intelligence and Humanized Computing* (Springer).

Chun-Lung 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.