

Article

BDNet: A Lightweight YOLOv12-Based Vehicle Detection Framework for Smart Urban Traffic Monitoring

Md Mahibul Hasan ¹, Zhijie Wang ^{1,*}, Hong Fan ², Kaniz Fatima ³, Muhammad Ather Iqbal Hussain ⁴, Rony Shaha ⁵ and Tushar MD Ahasan Habib ^{1,6}

¹ School of Information and Intelligent Science, Donghua University, Shanghai 201620, China; mmhasan@mail.dhu.edu.cn (M.M.H.); 322023@mail.dhu.edu.cn (T.M.A.H.)

² Glorious Sun School of Business and Management, Donghua University, Shanghai 201620, China; hongfan@dhu.edu.cn

³ Institute of Business Administration, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh; kanizulc@gmail.com

⁴ Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China; atheriqbal@fudan.edu.cn

⁵ School of Electronics Information Engineering, Beihang University, Beijing 100191, China; ronyshaha63@buaa.edu.cn

⁶ Electronic Science and Technology of China, Chengdu 610054, China

* Correspondence: wangzj@dhu.edu.cn

Highlights

What are the main findings?

- A lightweight vehicle detection framework (BDNet) achieves a superior accuracy–efficiency trade-off for dense urban traffic, improving detection accuracy, recall, and real-time throughput over recent YOLO baselines while maintaining a compact computational footprint.
- Detail-preserving downsampling, multi-order contextual aggregation, and feature refinement jointly enhance robustness under occlusion, scale variation, and domain shift, enabling effective generalization from ground-level to aerial traffic scenes.

What are the implications of the main findings?

- The proposed architecture provides a practical vehicle perception solution for smart city infrastructures, supporting real-time deployment on edge or roadside devices under strict latency and resource constraints.
- The demonstrated cross-dataset generalization indicates that modular, context-aware lightweight designs can improve reliability of urban traffic monitoring systems across heterogeneous sensing conditions and viewpoints.

Abstract

Accurate and real-time vehicle detection is a fundamental requirement for smart urban traffic monitoring, particularly in densely populated cities where heterogeneous traffic, frequent occlusion, and severe scale variation challenge lightweight vision systems deployed at the edge. To address these issues, this paper proposes BDNet, a lightweight YOLOv12-based vehicle detection framework designed to enhance feature preservation, contextual modeling, and multi-scale representation for intelligent transportation systems. BDNet integrates three complementary architectural components: (i) HyDASE, a hybrid detail-preserving downsampling module that mitigates information loss during resolution reduction; (ii) C3k2_MogaBlock, which strengthens long-range contextual interactions through multi-order gated aggregation; and (iii) an A2C2f_FRFN neck, which refines multi-scale features by suppressing redundancy and emphasizing discriminative responses. To



Academic Editor: Pierluigi Siano

Received: 19 January 2026

Revised: 7 February 2026

Accepted: 9 February 2026

Published: 14 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

support evaluation under realistic developing-region traffic conditions, we introduce the Bangladeshi Road Vehicle Dataset (BRVD), comprising 10,200 annotated images across 13 native vehicle categories captured under diverse urban scenarios, including daytime, nighttime, fog, and rain. On BRVD, BDNet achieves 85.9% mAP_{50} and 67.3% mAP_{50-95} , outperforming YOLOv12n by +1.4 and +0.7 percentage points, respectively, while maintaining a compact footprint of 2.5 M parameters, 6.0 GFLOPs, and a real-time inference speed of 285.7 FPS. Cross-dataset evaluation on VisDrone-DET2019, using models trained exclusively on BRVD, further demonstrates improved generalization, achieving 31.9% mAP_{50} and 17.9% mAP_{50-95} . These results indicate that BDNet provides an effective and resource-efficient vehicle detection solution for smart city-scale urban traffic monitoring.

Keywords: lightweight vehicle detection; smart mobility; intelligent transportation systems; real-time object detection; multi-scale feature representation; cross-domain generalization

1. Introduction

Rapid urbanization and motorization have placed unprecedented pressure on transportation infrastructures, particularly in densely populated metropolitan regions. In smart city ecosystems, real-time vehicle detection plays a central role in enabling intelligent transportation systems (ITS), supporting adaptive traffic signal control, congestion mitigation, road safety analytics, and data-driven urban planning. Camera-based detection systems are especially attractive in this context due to their scalability, cost-effectiveness, and compatibility with existing urban surveillance infrastructure. However, achieving accurate and reliable vehicle detection in real-world smart city environments remains challenging due to severe occlusion, large inter-class and intra-class scale variation, heterogeneous vehicle categories, and adverse illumination conditions. These challenges are further compounded by the strict computational constraints imposed by edge-deployed devices, which limit model capacity and inference complexity in practical deployments [1–3].

In operational smart-city deployments, detection outputs are directly consumed by city traffic management centers to estimate flows, identify bottlenecks, trigger adaptive signal timing, and support safety analytics (near-miss hotspots) and mobility planning indicators. Therefore, robustness and throughput are not only algorithmic metrics but also determinants of downstream decision reliability and service-level latency. Beyond algorithmic benchmarking, vehicle detection is increasingly deployed as a sensing layer for city-scale mobility services, where outputs feed traffic flow optimization, congestion monitoring, and operational dashboards for urban management [4].

Recent Smart Cities studies also emphasize that practical deployments require robust perception under heterogeneous cameras, weather/illumination changes, and constrained edge-computing budgets, making reliability and computational efficiency first-order design objectives for smart mobility pipelines [5–10]. For instance, LNT-YOLO, published recently in Smart Cities, focuses on nighttime traffic light detection under low-illumination conditions, explicitly addressing real-world deployment challenges in smart mobility systems [11]. Adverse weather further amplifies these requirements, as perception degradation can propagate to downstream mobility services, reinforcing the need for robust yet efficient detection pipelines in smart-city deployments [12].

These challenges are particularly pronounced in developing-country megacities, where traffic scenes are highly unstructured and visually complex. In South Asian urban environments, such as Dhaka, Chattogram, Gazipur, and Cumilla in Bangladesh, dense traffic flow, weak lane discipline, mixed vehicle types (e.g., buses, trucks, rickshaws, auto-rickshaws,

motorcycles), and frequent congestion produce conditions that significantly differ from those represented in standard object detection benchmarks. Recent Smart Cities studies have highlighted that detection models designed and validated primarily on structured Western datasets often exhibit degraded robustness when deployed in heterogeneous urban contexts, underscoring the need for region-aware and resource-efficient detection solutions tailored to smart mobility applications [11,13–15].

Deep learning-based object detection has undergone rapid evolution over the past decade, with the You Only Look Once (YOLO) family emerging as a dominant paradigm for real-time visual detection. Successive generations, from early YOLO and YOLO9000 to recent YOLOv8 to YOLOv12 models, have introduced progressively refined backbones, necks, and detection heads, including CSP-based designs, multi-path aggregation, decoupled heads, and attention-centric mechanisms, to improve accuracy–efficiency trade-offs [15–18]. These advances have catalyzed a broad body of lightweight and application-specific detectors targeting intelligent transportation scenarios. Representative examples include LVD-YOLO for efficient vehicle detection in ITS [1], MP-YOLO for dense urban traffic with pruning-based acceleration [19], YOLO-ELWNet and YOLO-PEL for embedded deployment [20,21], and UAV-oriented variants such as LUD-YOLO and LightUAV-YOLO for aerial monitoring [3,22].

Smart-city sensing increasingly combines heterogeneous viewpoints, including roadside cameras and UAV-enabled monitoring, which makes cross-viewpoint robustness a practical requirement rather than a purely academic benchmark. IoT-enabled UAV traffic monitoring has also been studied for real-time mobility analysis, highlighting the importance of resource-aware inference at the edge and reliable perception across changing viewpoints [23]. Similarly, UAV-enabled smart transport networks motivate unified pedestrian–vehicle detection under complex urban conditions, reinforcing the need for lightweight yet robust detectors that generalize beyond a single camera configuration [24].

Despite these advances, several open challenges persist when lightweight YOLO-based detectors are applied to dense, occlusion-heavy smart city traffic environments. First, the long-standing trade-off between speed and accuracy remains difficult to resolve. Many lightweight models achieve high inference speed by aggressively reducing parameters and FLOPs, but this often leads to weakened feature representation and degraded localization performance for small or partially occluded vehicles [20,22]. Conversely, accuracy-oriented designs that incorporate heavy attention modules, transformers, or additional detection heads substantially increase computational complexity, limiting their feasibility for real-time deployment on edge hardware commonly used in urban sensing systems [25,26].

Secondly, multi-scale contextual modeling in dense urban scenes is still insufficient. Overlapping vehicles, extreme scale diversity, and cluttered backgrounds demand richer feature aggregation than what is provided by conventional FPN or PANet structures. Although several studies enhance multi-scale representation through attention mechanisms or refined fusion strategies [27–29], naively stacking such modules may introduce redundancy, attenuate fine-grained cues, or exacerbate optimization instability, particularly in lightweight architectures designed for resource-constrained environments.

Thirdly, domain generalization across heterogeneous urban regions remains a critical limitation. Most mainstream detectors are trained on generic benchmarks such as COCO and PASCAL VOC, which only partially reflect the vehicle distributions, road layouts, and traffic behaviors encountered in developing-world cities [30,31]. Region-specific datasets from Bangladesh and neighboring regions have demonstrated that models trained on generic data often struggle to maintain performance when transferred to locally diverse traffic scenes [32–34]. From a smart cities perspective, this gap undermines the reliability of

detection-driven decision-making systems and motivates the development of models that generalize robustly across domains while remaining computationally efficient.

Motivated by these challenges, this paper proposes BDNet, a lightweight multi-module vehicle detection framework based on YOLOv12, explicitly designed for dense and heterogeneous urban traffic monitoring in smart city environments. BDNet integrates three complementary components: (i) HyDASE, a hybrid downsampling and excitation module that preserves fine-grained structural information during spatial reduction; (ii) C3k2_MogaBlock, which incorporates multi-order gated aggregation to enhance long-range contextual reasoning with modest computational overhead; and (iii) A2C2f_FRFN, an adaptive feature refinement neck that suppresses redundant activations while emphasizing discriminative multi-scale responses.

To support realistic evaluation aligned with smart city deployment scenarios, we introduce the Bangladeshi Road Vehicle Dataset (BRVD), comprising 10,200 annotated images across 13 native vehicle categories captured under diverse day, night, fog, and rain conditions. In addition to in-domain evaluation, cross-dataset experiments are conducted on the VisDrone-DET2019 benchmark to assess generalization across viewpoints and object scales. Extensive experiments demonstrate that BDNet achieves a favorable balance between accuracy, efficiency, and robustness compared to YOLOv8-YOLOv12 and recent lightweight detectors, making it well suited for real-time vehicle detection in resource-constrained smart urban traffic monitoring systems.

The main contributions of this work are summarized as follows:

1. We propose BDNet, a lightweight YOLOv12-based vehicle detection framework tailored for dense and heterogeneous smart city traffic environments.
2. We design three synergistic modules: HyDASE, C3k2_MogaBlock, and A2C2f_FRFN, that jointly enhance detail preservation, contextual reasoning, and multi-scale feature refinement with minimal overhead.
3. We introduce the Bangladeshi Road Vehicle Dataset (BRVD), reflecting realistic South Asian urban traffic conditions, and evaluate cross-domain generalization on VisDrone-DET2019.
4. We conduct comprehensive experiments and ablation studies demonstrating that BDNet consistently outperforms state-of-the-art lightweight YOLO variants in accuracy–efficiency trade-offs, supporting its deployment in smart city ITS applications.

The remainder of this paper is organized as follows. Section 2 reviews related work on lightweight, attention-enhanced, and domain-specific YOLO variants. Section 3 details the proposed BDNet architecture and modules, as well as the datasets and implementation settings. Section 4 reports the experimental setup, quantitative and qualitative results, and ablation analyses. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related Work

2.1. Lightweight YOLO Model for Real-Time Intelligent Transportation Systems

Real-time vehicle detection in intelligent transportation systems (ITS) has motivated extensive research on lightweight variants of the YOLO family, aiming to balance detection accuracy with strict latency and resource constraints. In this context, LVD-YOLO redesigns the backbone using lightweight residual structures to reduce parameter count while maintaining competitive accuracy for vehicle detection in traffic scenes [1]. MP-YOLO introduces multidimensional feature fusion with layer-adaptive pruning, enabling feature reuse across intermediate layers without excessive computational overhead, particularly in dense traffic environments [19]. YOLO-ELWNet further compresses the CSPDarknet–PAN pipeline by

embedding efficient attention mechanisms, improving the accuracy–speed trade-off on embedded platforms commonly used in roadside and on-board sensing units [20].

Complementary lightweight strategies, including channel pruning, depthwise separable convolutions, and simplified attention, have also been explored in GMS-YOLO [29], YOLOv8-QSD [35], and YOLO-PEL [21]. These methods successfully reduce GFLOPs and memory usage while sustaining real-time performance for vehicle and pedestrian detection, making them attractive for smart traffic monitoring applications. Beyond road-based ITS, lightweight YOLO architectures have been adapted to other constrained visual sensing domains. PV-YOLO addresses joint pedestrian–vehicle detection in complex road scenes [2], while LightUAV-YOLO [3] and LUD-YOLO [22] target UAV-based object detection under altitude and viewpoint limitations. LiteFlex-YOLO extends lightweight detection to maritime UAV surveillance, where targets are small and sparsely distributed [36].

In addition to model-centric efficiency improvements, recent studies situate lightweight YOLO-based detectors within smart-city traffic systems, where perception serves as a functional component supporting traffic monitoring and mobility analytics under real deployment constraints [4,11]. These works motivate the need for detectors that balance accuracy, robustness, and computational efficiency, while highlighting that many existing approaches still address architectural bottlenecks, such as downsampling loss, contextual reasoning, or feature redundancy, in isolation rather than within a unified lightweight design.

2.2. Handling Occlusion, Scale Variation, and Adverse Urban Conditions

Urban traffic scenes are characterized by severe occlusion, strong scale variation, and challenging illumination conditions, motivating research into more robust detection architectures. SFFEF-YOLO employs fine-grained feature extraction and fusion to improve small-object recognition in UAV imagery [28], while FOS-YOLO leverages attention-driven multiscale context aggregation to handle complex environments [27]. LAYN introduces a lightweight multi-scale attention mechanism to enhance YOLOv8 for small-object detection [37]. Several works explicitly address adverse environmental conditions common in real-world ITS deployments. GMS-YOLO [29], MLE-YOLO [38], LS-YOLO [39], and I-YOLOv11n [40] incorporate illumination-aware preprocessing, adaptive depth, and enhanced feature fusion to maintain detection quality under fog, rain, and low-light conditions. The importance of deployment-oriented optimization under extreme illumination constraints is further highlighted by SCL-YOLOv11 and related low-light detection frameworks [41].

At the architectural level, more expressive backbones and refinement modules have been explored to improve contextual reasoning. MogaNet introduces multi-order gated aggregation to model long-range dependencies in a computationally feasible manner [42]. Transformer-inspired refinement strategies, such as the Feature Refinement Feed-Forward Network (FRFN), demonstrate how attentive feed-forward design can suppress redundant activations while emphasizing discriminative features [26]. Meanwhile, YOLOv13 incorporates hypergraph-enhanced adaptive correlation (HyperACE) for global context modeling [43], and YOLO-WTB augments the YOLOv12n head with an additional P2 branch to capture fine-grained structures [44]. Although these approaches achieve strong accuracy in their respective domains, their reliance on heavy attention mechanisms, hypergraph reasoning, or transformer blocks often increases model size and inference latency. This limits their practicality for real-time ITS deployments on low-power edge devices, which remain a central requirement in smart city infrastructures.

2.3. Domain-Specific Datasets for Urban Vehicle Detection

Detection performance is strongly influenced by the characteristics of the training dataset. Widely used benchmarks such as COCO [30] and PASCAL VOC [31] provide diverse object categories and backgrounds, but they do not adequately reflect the complexity of traffic in developing urban regions. South Asian traffic environments are particularly challenging due to heterogeneous vehicle types, weak lane discipline, and highly unstructured motion patterns. To address these limitations, several region-specific datasets have been introduced. The Bangladeshi Vehicle Classification Dataset [33] and Poribohon-BD [34] capture native vehicle categories, such as rickshaws, auto-rickshaws, minibuses, and locally modified trucks, primarily for classification tasks.

Despite these efforts, most YOLO-based detectors remain optimized for generic datasets and are not explicitly designed for the extreme occlusion, clutter, and scale diversity present in dense South Asian traffic. Consequently, models trained on COCO or VOC often suffer significant performance drops when deployed directly in such environments. This gap motivates the construction of the Bangladeshi Road Vehicle Dataset (BRVD), which extends earlier classification-focused datasets into a multi-object detection benchmark with 10,200 annotated images across 13 native vehicle classes, captured under day, night, fog, and rain conditions. BRVD provides both contextual realism and fine-grained annotations, enabling systematic evaluation of detectors under realistic smart city traffic conditions. In addition, VisDrone-DET2019 is employed to assess cross-domain generalization from ground-based urban imagery to aerial viewpoints [45].

2.4. Summary and Research Gaps

The reviewed literature indicates that existing approaches rarely address efficiency, robustness, and domain specificity in a unified manner. Efficiency-oriented YOLO variants, such as LVD-YOLO [1], MP-YOLO [19], YOLO-ELWNet [20], YOLO-PEL [21], and GMS-YOLO [29], achieve high throughput and reduced computational cost through pruning and lightweight backbones, but often sacrifice fine-grained spatial detail and contextual richness crucial for dense urban traffic. Accuracy-focused models, including Reb-YOLO [46], FOS-YOLO [27], SFFEF-YOLO [28], LAYN [37], LS-YOLO [47], LSOD-YOLO [25], and YOLO-WTB [44], enhance multi-scale representation and robustness, yet introduce additional parameters and latency that hinder deployment on edge devices. At the same time, region-specific datasets such as the Bangladeshi vehicle datasets and Poribohon-BD [33,34] highlight the necessity of detectors tailored to heterogeneous traffic distributions. Nevertheless, most mainstream YOLO architectures are still developed and validated primarily on generic benchmarks, resulting in suboptimal performance when transferred to dense South Asian urban scenes.

These observations reveal three interconnected research gaps:

1. Efficiency–accuracy trade-off gap, where maintaining both high throughput and robust detection under dense traffic remains challenging [1,19–21,25,29];
2. Architectural integration gap, as prior work typically improves downsampling, context aggregation, or feature refinement in isolation rather than within a unified lightweight topology [26–28,40,46];
3. Domain adaptation gap, since detectors are rarely tailored to region-specific traffic characteristics despite the availability of localized datasets [33,34].

Existing lightweight YOLO variants often improve one aspect in isolation—e.g., lightweight backbone redesign, pruning, or attention-based fusion—while leaving other bottlenecks (detail loss during downsampling, insufficient long-range context under occlusion, or redundant activations in fusion) partially unresolved in dense mixed-traffic scenes [12,23]. In contrast, BDNet targets these three bottlenecks jointly through

(i) detail-preserving downsampling (HyDASE), (ii) efficient long-range contextual aggregation (C3k2_MogaBlock), and (iii) redundancy-suppressing multi-scale refinement (A2C2f_FRFN), producing a unified lightweight topology validated not only in-domain but also under cross-dataset viewpoint shift. Accordingly, BDNet integrates HyDASE, C3k2_MogaBlock, and A2C2f_FRFN within YOLOv12 to jointly address detail loss, occlusion-context modeling, and fusion redundancy.

3. Materials and Methods

This section presents the proposed BDNet architecture and the Bangladeshi Road Vehicle Dataset (BRVD) used for model development and evaluation. BDNet is a lightweight YOLOv12-based framework designed for real-time vehicle detection in dense urban traffic, while BRVD provides a region-specific benchmark reflecting heterogeneous and unstructured road environments common in smart city scenarios.

3.1. Architecture of BDNet

BDNet has been proposed as a lightweight, single-stage detection network specifically designed for real-time vehicle detection in dense and heterogeneous traffic environments. The architecture follows the standard object detection pipeline of Input $\rightarrow$ Backbone $\rightarrow$ Neck $\rightarrow$ Detection Head, while introducing three specialized modules: HyDASE, C3k2_MogaBlock, and A2C2f_FRFN. These modules are jointly designed to address the accuracy–efficiency trade-off and feature-representation limitations identified in Section 2. Figure 1 illustrates the overall architecture of BDNet. All of these elements operate together to preserve spatial detail, strengthen contextual reasoning, and refine multi-scale features and all while ensuring low computational cost.

At the input stage, BDNet takes 640×640 RGB images as input. The backbone has been designed for efficient feature extraction and is composed of initial convolutional layers followed by stacked C3k2_MogaBlocks, which are responsible for multi-order contextual aggregation. HyDASE modules perform detail-preserving downsampling at key stages of the backbone to preserve spatial details while reducing feature map dimensions. The neck, responsible for multi-scale feature fusion, utilizes A2C2f_FRFN modules to refine and enhance the feature maps after each concatenation step, ensuring that only the most discriminative information is propagated forward. This enhances cross-scale information flow, allowing small vehicles such as motorcycles and rickshaws to be represented as effectively as larger objects. By integrating HyDASE, C3k2_MogaBlock, and A2C2f_FRFN within the neck, BDNet achieves a refined balance between lightweight processing and robustness to traffic heterogeneity.

BDNet employs three detection heads operating at different spatial scales to handle targets of varying sizes. Each head uses a decoupled formulation, separating classification from regression to stabilize gradients and improve convergence. Depthwise separable convolutions are introduced where appropriate to maintain efficiency without compromising representational capacity. Bounding-box regression follows the default YOLOv12 design, combining Distribution Focal Loss (DFL) for boundary precision with CIoU and binary cross-entropy for objectness and classification [18,48,49]. We keep the loss settings identical to YOLOv12 to ensure a fair baseline comparison and to isolate the contribution of the proposed architectural modules.

3.2. Hybrid Downsampling and Squeeze-Excitation (HyDASE) Module

Downsampling is an essential component of convolutional detection networks, yet conventional operations such as strided convolutions or max-pooling often degrade the fine structural information needed to recognize small or partially occluded vehicles in dense

traffic. Strided convolutions tend to smooth feature boundaries, while max-pooling may suppress subtle but discriminative cues. These limitations become particularly evident in heterogeneous road environments like BRVD, where object sizes vary widely and occlusion is frequent.

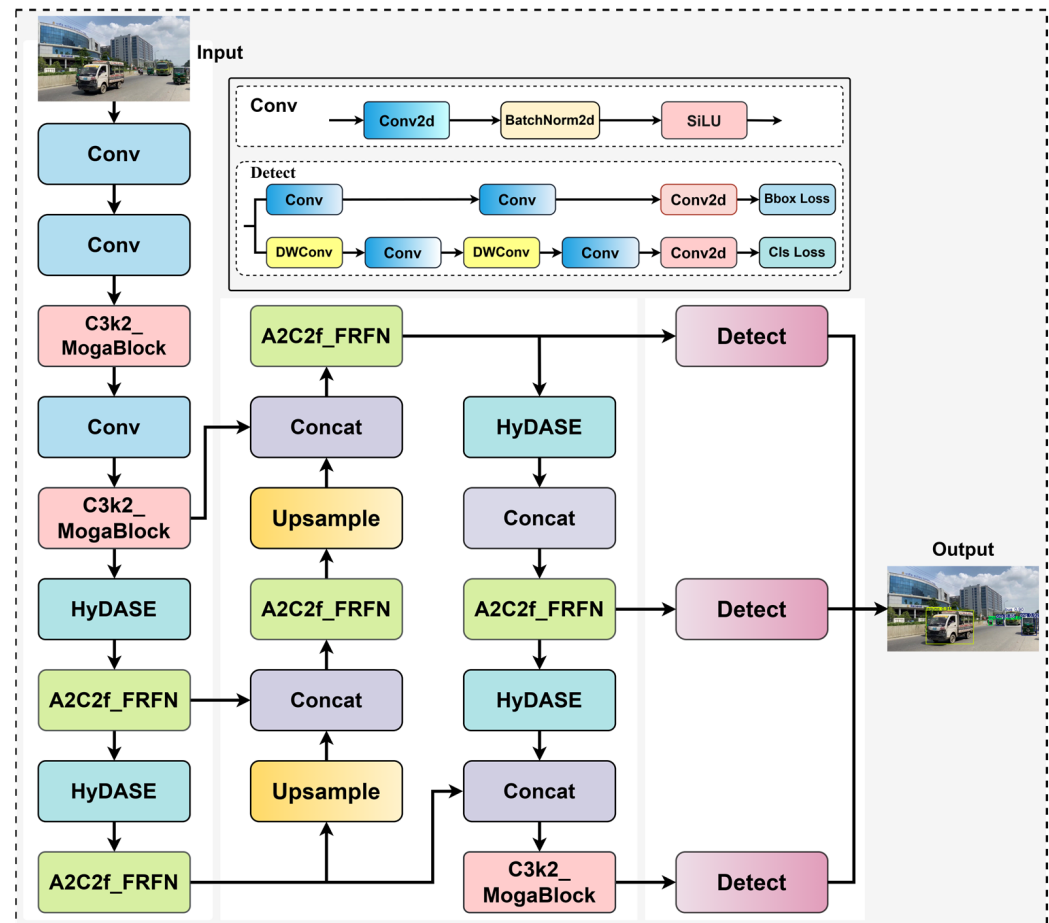


Figure 1. Overall architecture of BDNet. The model follows a single-stage detection pipeline composed of Input, Backbone, Neck, and Detection Head. The backbone integrates Hybrid Downsampling (HyDASE) and C3k2_MogaBlocks for efficient feature preservation and contextual aggregation. The neck employs A2C2f_FRFN modules for multi-scale feature refinement, while three detection heads output predictions across different scales.

To address this challenge, we introduce HyDASE (Hybrid Downsampling with Squeeze-and-Excitation), a lightweight module designed to retain spatial detail while reducing feature-map resolution efficiently. The design incorporates three principles: hybrid pooling for noise-robust preconditioning, parallel convolution-pooling branches for complementary feature extraction, and an SE-based recalibration stage to adaptively emphasize informative channels. The overall structure is shown in Figure 2.

(1) Hybrid Pre-Pooling: The operational flow of the HyDASE module is detailed in Figure 2. The input feature map $X \in R^{C \times H \times W}$ is first normalized through a hybrid pre-pooling layer. This initial step fuses average and max-pooling outputs to create a baseline downsampled feature map that balances prominent activations with broader contextual information:

We first apply a hybrid pre-pooling operation that averages the outputs of max-pooling and average-pooling with equal weights:

$$X' = 0.5 \text{MaxPool}(X) + 0.5 \text{AvgPool}(X), \quad (1)$$

This blended representation preserves prominent activations while embedding broader contextual information, creating a stable foundation for subsequent feature extraction.

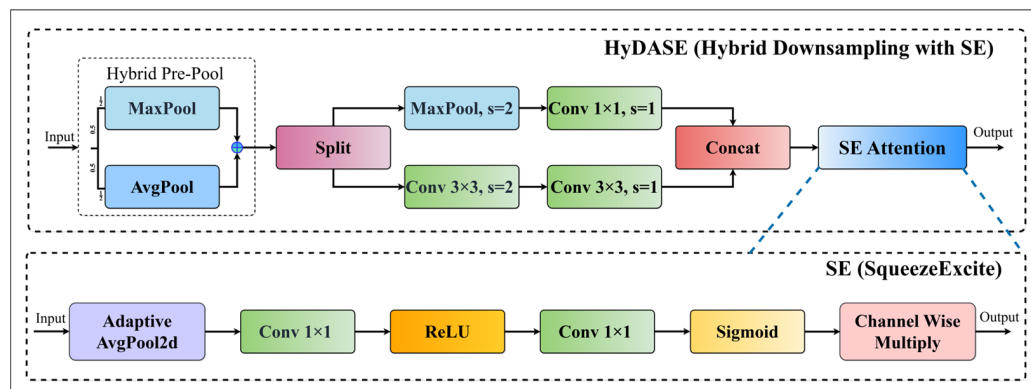


Figure 2. Architecture of the proposed HyDASE module. The block integrates hybrid pre-pooling, parallel convolutional and pooling branches, and a squeeze-and-excitation attention unit to achieve detail-preserving downsampling.

(2) Channel Split and Parallel Paths: The tensor X' is then partitioned channel-wise into two subgroups, X_1 and X_1 . This split allows two distinct processing branches to develop specialized feature representations:

$$X' \rightarrow (X_1, X_2), X_1, X_2 \in \mathbb{R}^{\frac{C}{2} \times H \times W}, \quad (2)$$

Splitting the channels enables two specialized branches to operate independently but efficiently.

Branch A Multi-Scale Convolutional Path: The first branch applies two 3×3 convolutions in sequence. The initial convolution downscales the input using a stride of 2, and the following convolution refines the representation:

$$Y_1 = \text{Conv}_{3 \times 3, s=1}(\text{Conv}_{3 \times 3, s=2}(X_1)), \quad (3)$$

This branch captures fine textures, edges, and shape cues needed for identifying small vehicles such as motorcycles, rickshaws, or distant objects.

Branch B (Pooling and Projection Path): The second branch focuses on robust activation preservation. A 3×3 max-pooling layer (stride = 2) extracts salient regional responses, followed by a 1×1 convolution to align feature dimensionality:

$$Y_2 = \text{Conv}_{1 \times 1, s=1}(\text{MaxPool}_{3 \times 3, s=2, p=1}(X_2)), \quad (4)$$

This path preserves high-confidence activations associated with strongly represented structures such as large vehicles (e.g., buses, trucks).

(3) Concatenation and Fusion: The outputs of the two branches are concatenated to form a unified representation:

$$Y = [Y_1, Y_2], \quad (5)$$

where $Y \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$. This fusion elegantly combines the convolutional branch's fine-scale detail with the pooling branch's salient, noise-resistant features.

(4) Channel Re-calibration via SE Attention: To adaptively prioritize meaningful channels, HyDASE incorporates a lightweight Squeeze-and-Excitation (SE) unit. Global Average Pooling (GAP) generates a compact descriptor, which is passed through a two-layer (Rectified Linear Unit (ReLU) and Sigmoid activations, respectively) excitation network:

$$\alpha = \sigma(W_2 \delta(W_1 \text{GAP}(Y))), \quad (6)$$

where $\delta(\cdot)$ is ReLU, $\sigma(\cdot)$ is Sigmoid, and W_1, W_2 are learnable matrices. The fused feature map is then reweighted channel-wise:

$$Z = Y \odot \alpha, \quad (7)$$

where $\odot$ denotes channel-wise multiplication, yielding a recalibrated output that emphasizes informative channels and suppresses noise.

The HyDASE module provides three key benefits: First, Detail preservation: Hybrid pooling and multi-scale convolutions retain spatial cues critical for detecting small or distant vehicles. Moreover, Efficiency: Only half of the channels pass through the more expensive convolutional operations, reducing computational overhead. Finally, Adaptivity: SE-based channel re-weighting allows the module to adjust dynamically to varying traffic density, lighting, and occlusion levels. By combining these strengths, HyDASE serves as an effective downsampling alternative for real-time detection systems operating in visually complex road environments.

3.3. The C3k2_MogaBlock Module

Detecting vehicles in dense traffic requires the ability to model long-range dependencies and subtle contextual cues, especially when objects are partially occluded or appear at widely varying scales, which is particularly critical for urban traffic monitoring and smart mobility systems. Conventional convolutional blocks, restricted by fixed local receptive fields, struggle to capture these relationships reliably. To address this limitation, BDNet incorporates the C3k2_MogaBlock, a lightweight residual bottleneck that integrates the multi-order gated aggregation (MOGA) operator [42] into a compact YOLO-style topology. This combination enhances contextual reasoning without imposing the computational burden typically associated with attention or transformer-based modules. The structure of the block is shown in Figure 3.

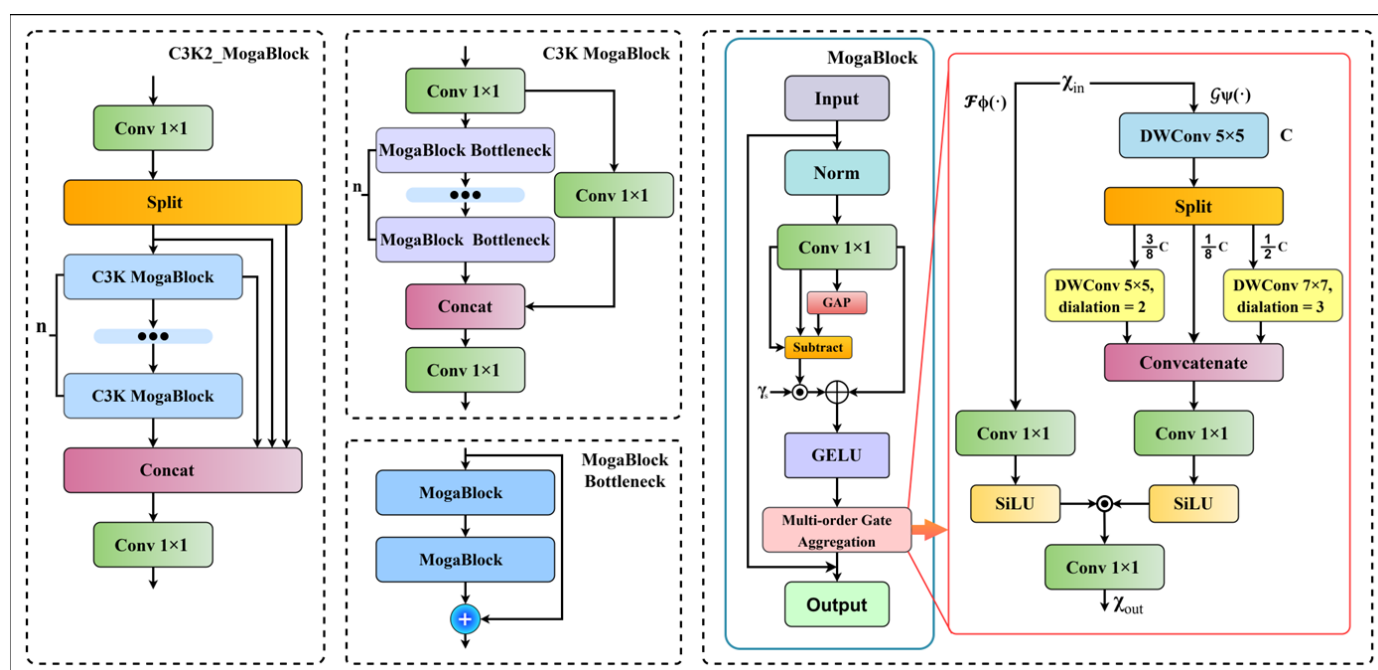


Figure 3. Architecture of the proposed C3k2_MogaBlock module.

Structure and Data Flow: The architecture of the proposed module is illustrated in Figure 3. It processes an input feature map $\mathcal{X} \in \mathbb{R}^{H \times W \times C}$ through the following stages:

- (1) **Input Projection and channel Split:** The input is first compressed by a 1×1 convolution and then divided into two equally sized channel groups:

$$\mathcal{X}_1, \mathcal{X}_2 = \text{Split}(\text{Conv}_{1 \times 1}(\mathcal{X})), \quad (8)$$

where $\mathcal{X}_1, \mathcal{X}_2 \in \mathbb{R}^{\frac{C}{2} \times H \times W}$. This split creates a dedicated pathway for multi-scale context processing ($\mathcal{X}_1$) and a residual shortcut ($\mathcal{X}_2$).

- (2) **Multi-Order Gated Aggregation (Context Branch):** The context branch $\mathcal{X}_1$ is processed by the MOGA operator. As shown in Figure 3, MOGA applies a set of k parallel dilated depthwise convolutional transforms $f_i(\cdot)$ (each with a different dilation rate) and dynamically weights their outputs via learned gating signals $g_i(\cdot)$ before fusion:

$$\mathcal{Y} = \sum_{i=1}^k \sigma(g_i(\mathcal{X}_1)) \cdot f_i(\mathcal{X}_1), \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid activation function. The MOGA output is then combined with the original branch input via a residual connection to preserve gradient flow:

$$\mathcal{Y}_{\text{cnet}} = \mathcal{X}_1 + \mathcal{Y}, \quad (10)$$

ensuring robust gradient propagation and preserving the identity information.

Implementation Details and Feature Decomposition: Following [42], we incorporate a lightweight feature-decomposition mechanism that improves the stability of both the gating and transformation pathways. The input is first projected and decomposed as:

$$\tilde{\mathcal{X}} = \phi(P_1 \mathcal{X}_1 + \gamma \odot [P_1 \mathcal{X}_1 - \text{GAP}(P_1 \mathcal{X}_1)]), \quad (11)$$

where P_1 is a 1×1 projection, γ is a learnable element-wise scale, ϕ is SiLU, $\odot$ is Hadamard product, and GAP is global average pooling. The MOGA operations then proceed as:

$$g = \text{Conv}_{1 \times 1}(\tilde{\mathcal{X}}), \quad (12)$$

$$v = \text{MultiOrderDWConv}(\tilde{\mathcal{X}}), \quad (13)$$

$$\mathcal{Y} = \text{Conv}_{1 \times 1}(\phi(g) \odot \phi(v)), \quad (14)$$

producing a refined multi-order contextual representation.

- (3) **Fusion with the Shortcut Branch:** To blend the contextual features with preserved local information, the outputs of the MOGA branch and shortcut branch are concatenated and projected back to the original dimensionality:

$$\mathcal{Z} = \text{Conv}_{1 \times 1}([\mathcal{Y}_{\text{cnet}}, \mathcal{X}_2]), \quad (15)$$

where $[\cdot, \cdot]$ denotes channel-wise concatenation.

The C3k2_MogaBlock is designed to enrich BDNet's representational capacity while retaining the efficiency required for real-time operation. Several characteristics make it particularly effective for traffic environments such as BRVD: Multi-order aggregation enables the model to respond to objects across a wide range of sizes and positions. Then, Dynamic gating selectively emphasizes informative cues, improving recognition under

occlusion and clutter. After that, Residual coupling and channel split preserve fine spatial details and maintain computational efficiency. Final projection layer harmonizes contextual and local information into a unified feature space.

In summary, C3k2_MogaBlock enhances contextual reasoning by integrating multi-order gated aggregation within a lightweight cross-stage partial structure. By decomposing feature channels into multiple receptive-field orders and selectively aggregating them through gated interactions, the module captures long-range dependencies that are essential for resolving occlusion and scale ambiguity in dense urban traffic scenes. Importantly, this contextual enrichment is achieved with modest computational cost through channel partitioning and residual routing, making C3k2_MogaBlock suitable for real-time deployment in resource-constrained smart-city environments.

3.4. The A2C2f_FRFN Module

Multi-scale detection networks frequently accumulate noisy or redundant activations as features are repeatedly aggregated, transformed, and propagated across layers. This accumulation can obscure subtle patterns, weaken localization accuracy, and make the model especially vulnerable under challenging conditions such as occlusion, visual clutter, or poor illumination. Without an explicit refinement mechanism, irrelevant responses tend to persist, ultimately degrading detection performance.

To mitigate these issues, we draw inspiration from the refine-and-suppress philosophy introduced in the Adaptive Sparse Transformer (AST), particularly its Feature Refinement Feed-Forward Network (FRFN) proposed by Zhou et al. [26]. FRFN selectively amplifies salient activations while dampening redundancy through a lightweight, content-adaptive interaction. Building on this idea, we integrate the essence of FRFN into a YOLO-style A2C2f architecture, forming the A2C2f_FRFN module, a compact refinement bottleneck tailored for efficient multi-scale feature processing.

Figure 4 below illustrates the structure of the proposed module. The input is projected into a hidden representation and passes through a sequence of ABlock_FRFN units, each embedding a refinement pathway derived from FRFN. The refined outputs of these blocks are concatenated, compressed through a 1×1 convolution, scaled using a lightweight modulation operator, and finally merged with the residual shortcut to maintain stable gradient flow.

Let the input feature map be $X \in R^{H \times W \times C}$. Each ABlock_FRFN operates as:

$$F_1 = \text{PConv}(\text{LN}(X)), \quad (16)$$

where LN denotes layer normalization and PConv is partial convolution applied for efficient spatial filtering. The result is projected and split into two components:

$$[F_a, F_b] = \text{split}(\text{Linear}(F_1)), \quad (17)$$

with F_a carrying activation weights and F_b acting as a modulating signal. To capture local spatial cues, F_b is reshaped, processed with a depthwise convolution, and flattened:

$$F'_b = \text{Flatten}(\text{DWConv}(\text{Reshape}(F_b))), \quad (18)$$

The refinement is then performed through an adaptive interaction:

$$R(F) = F_a \otimes F'_b, \quad (19)$$

where $\otimes$ denotes element-wise multiplication. A residual connection ensures stable training:

$$Y = \text{Linear}(R(F)) + X, \quad (20)$$

This interaction selectively emphasizes structure-consistent features while suppressing responses that do not contribute meaningfully to object boundaries or semantics.

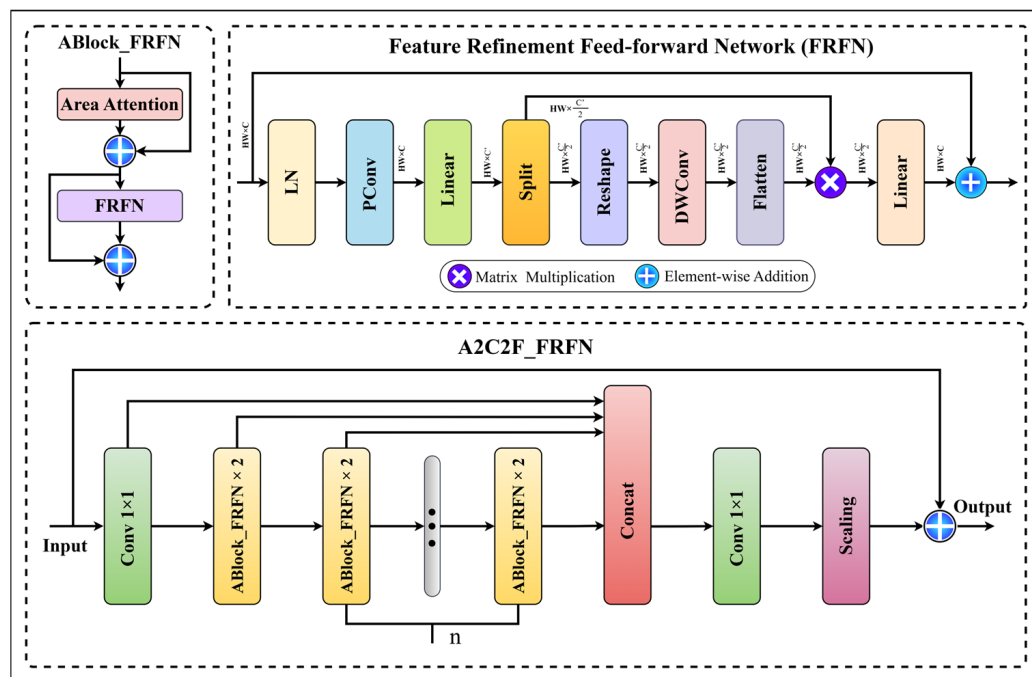


Figure 4. Architecture of our proposed A2C2f_FRFN module. Multiple ABlock_FRFN units, each embedding an FRFN, are stacked within the A2C2f structure. The refined features are concatenated and projected through a lightweight fusion layer, ensuring enhanced feature selectivity while maintaining residual stability.

Following the C2f principle used in YOLO architectures, multiple refinement blocks are stacked.

Let $(A_i)_{i=1}^n$ denote the outputs of n successive ABlock_FRFN units. Their aggregated form is:

$$Y = \mathcal{S}(\text{Conv}_{1 \times 1}(\text{Concat}((A_i)_{i=1}^n))) + X, \quad (21)$$

where Concat denotes channel-wise concatenation, Conv 1×1 compresses the fused representation, and $\mathcal{S}(\cdot)$ is a lightweight scaling operator that stabilizes the fusion.

The A2C2f_FRFN module serves as an efficient feature filter within BDNet. By embedding FRFN-style refinement inside a YOLO-friendly A2C2f layout, the module achieves the following benefits: First, Redundancy suppression: multi-scale aggregation introduces noisy or repetitive activations that are selectively reduced.

In addition, A2C2f_FRFN is introduced to refine multi-scale features by suppressing redundant activations while strengthening discriminative responses across spatial resolutions. Through an attention-guided feed-forward refinement pathway, the module selectively emphasizes informative feature components that contribute to class separation, particularly in scenarios involving partial visibility or background clutter. This refinement process enhances the saliency of subtle visual patterns, such as fragmented silhouettes or boundary cues, while maintaining computational efficiency. As a result, A2C2f_FRFN improves localization stability and classification reliability under dense and heterogeneous traffic conditions.

3.5. Datasets

3.5.1. Bangladeshi Road Vehicle Dataset (BRVD)

To support realistic evaluation of vehicle detection models for smart urban traffic monitoring, we construct the Bangladeshi Road Vehicle Dataset (BRVD), a multi-object vehicle detection benchmark tailored to dense and heterogeneous traffic environments. BRVD contains 10,200 images annotated in YOLO format across 13 native vehicle categories: auto-rickshaw, bicycle, bus, car, CNG, covered van, easy-bike, leguna, motorcycle, pickup, rickshaw, truck, and van. Unlike classification-oriented datasets that primarily include single, centered objects [33], BRVD captures multiple co-existing vehicle instances per image, with frequent inter-class occlusion, pronounced scale variation, and complex background clutter, reflecting realistic traffic patterns in developing-region cities. Figure 5 provides representative instances of changes in occlusion, background clutter, lighting, and traffic density.



Figure 5. Representative images from the BRVD dataset under varying densities, illumination, and weather conditions (day, night, fog, and rain).

BRVD is designed as a deployment-oriented dataset that reflects the operational challenges faced by smart-city traffic perception systems, particularly in regions characterized by mixed motorized and non-motorized traffic, irregular lane discipline, and dense vehicle interactions. Images were collected using handheld smartphones, roadside surveillance cameras, and in-vehicle recording systems across major Bangladeshi cities, including Dhaka, Chittagong, Rajshahi, and Gazipur. Data acquisition covers a range of road contexts, including arterial roads, intersections, and secondary urban corridors, and includes daytime, nighttime, foggy, and rainy conditions, ensuring diversity in illumination, visibility, and traffic density. Both motorized (e.g., buses, trucks, cars, CNG) and non-motorized vehicles (e.g., rickshaws and bicycles) are represented to reflect the mixed-traffic composition typical of South Asian urban mobility systems.

All images were annotated using Labellmg and exported in YOLO format [50]. A two-stage manual verification process was applied to improve annotation consistency and bounding-box accuracy, with particular attention to small and partially occluded targets. During verification, a stratified sampling strategy was applied to ensure coverage across vehicle categories and traffic densities, reducing annotation bias and improving cross-class

consistency. The dataset was split into 70% training, 20% validation, and 10% testing subsets. The instance distribution across the 13 categories is shown in Figure 6.

To comply with ethical and privacy considerations, all collected images were carefully inspected prior to annotation. Personally identifiable information (e.g., faces and license plates) was avoided where possible and anonymized when visible, consistent with common ethical practice for public-space traffic sensing. The dataset does not contain personal identity labels and is intended solely for traffic perception research. At the time of submission, BRVD is not yet publicly released. The dataset will be made available to the research community upon publication of this work, together with a formal license, documentation of data collection and anonymization procedures, and annotation guidelines to support reproducibility and responsible reuse.

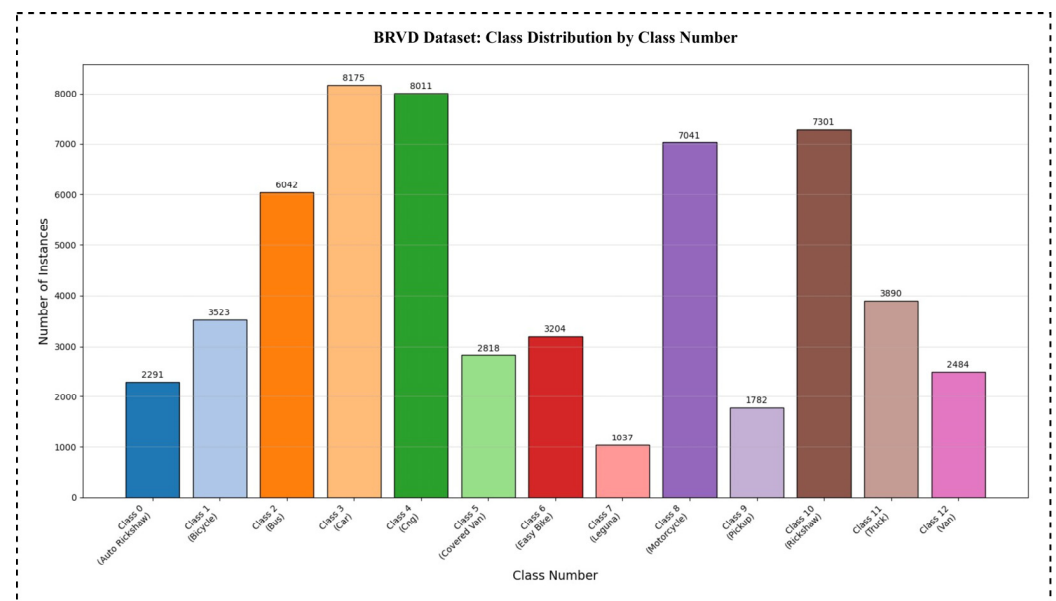


Figure 6. Instance distribution of the BRVD dataset across 13 vehicle categories.

Overall, BRVD is designed as a realistic, region-specific benchmark for evaluating lightweight vehicle detection models under complex smart-city conditions, where dense traffic flow, heterogeneous vehicle morphology, and adverse environmental factors commonly degrade detector performance. Future extensions of BRVD will focus on expanding data coverage across additional urban regions, traffic patterns, and environmental scenarios, while maintaining consistent annotation standards to support large-scale and longitudinal smart-city traffic analysis.

3.5.2. Cross-Dataset Benchmark: VisDrone-DET2019

To assess cross-domain generalization under viewpoint and scale shifts, we additionally evaluate the proposed model on the VisDrone-DET2019 benchmark [45]. VisDrone-DET2019 is a UAV-based object detection dataset containing 6471 training images, 548 validation images, and 1610 test images, annotated across traffic-related categories such as car, bus, truck, van, motor, bicycle, tricycle, awning-tricycle, and pedestrian. In contrast to BRVD, VisDrone-DET2019 features aerial viewpoints, extremely small object sizes, dense layouts, and frequent occlusion, producing a challenging domain shift for models trained on ground-level urban imagery. Representative samples are shown in Figure 7.

The examples illustrate strong scale compression, perspective distortion, and dense object layouts typical of UAV imagery, providing a complementary testbed for evaluating robustness beyond ground-level smart-city traffic scenes. Because BRVD and VisDrone-

DET2019 differ in both imaging perspective and class definitions, no class remapping is applied; results on VisDrone-DET2019 are reported strictly following its standard evaluation protocol. Together, BRVD (ground-level dense urban traffic) and VisDrone-DET2019 (aerial small-object scenarios) enable a complementary assessment of lightweight vehicle detection models under diverse sensing modalities relevant to smart-city monitoring and ITS deployment.

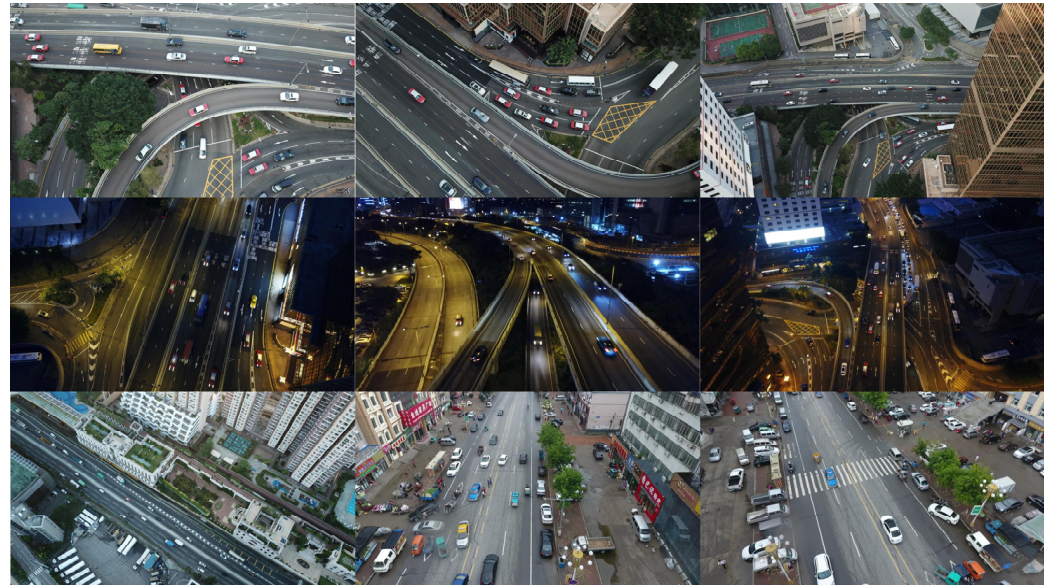


Figure 7. Representative samples from the VisDrone-DET2019 dataset showing aerial viewpoints and small-object challenges.

4. Results and Discussions

4.1. Experimental Setup

All experiments were conducted on two datasets: BRVD (Bangladeshi Road Vehicle Dataset) for ground-level vehicle detection and VisDrone-DET2019 for cross-dataset generalization. BRVD contains 10,200 images annotated across 13 native vehicle categories and is split into 70% for training, 20% for validation, and 10% for testing. VisDrone-DET2019 includes 6471 training images, 548 validation images, and 1610 test images, covering diverse aerial viewpoints and crowding levels.

Training and evaluation were implemented using PyTorch 2.1.2 with CUDA 11.8 on a workstation equipped with an NVIDIA RTX 4090 GPU (24 GB) and an Intel Xeon Platinum 8470Q CPU (20 cores). All models, including BDNet and the baseline YOLO variants, were trained on BRVD under an identical configuration to ensure a fair comparison. The trained models were then evaluated on BRVD and directly tested on VisDrone-DET2019 to assess robustness under domain shift. All networks were trained from random initialization without using any external pre-trained weights.

Each model is trained for 300 epochs with an input resolution of 640×640 pixels and a batch size of 16. We employ Stochastic Gradient Descent (SGD) as the optimizer, with an initial learning rate $\text{lr}_0 = 0.01$, a final learning rate $\text{lrf} = 0.01$, momentum of 0.937, and weight decay of 5×10^{-4} . A linear warm-up is applied over the first three epochs; thereafter, the learning rate remained constant.

Standard YOLO training strategies were employed to ensure stable optimization and reproducibility, including Exponential Moving Average (EMA) of model weights during training. During inference, Non-Maximum Suppression (NMS) was applied using the default IoU and confidence thresholds defined in the YOLO evaluation pipeline. Data

augmentation included Mosaic augmentation, HSV color jittering (hue = 0.015, saturation = 0.7, value = 0.4), and random translation with a maximum offset of 0.1 of the image size. The complete set of hyperparameters is summarized in Table 1. Unless otherwise stated, all reported results correspond to validation performance under this configuration.

Table 1. Hyperparameter settings used for model training.

Parameters	Value
Input image size	640 × 640
Batch size	16
Epochs	300
Optimizer	SGD
Momentum	0.937
Weight decay	0.0005
Initial learning rate (lr0)	0.01
Final learning rate (lrf)	0.01
Warmup epochs	3
Warmup momentum	0.8
Mosaic	1
HSV-Hue	0.015
HSV-Saturation	0.7
HSV-Value	0.4
Translation factor	0.1

4.2. Evaluation Metrics

In order to analyze the performance of the BDNet model, many important metrics were used to evaluate various elements of its performance. These metrics included Precision ($\mathcal{P}_{rec}$), Recall ($\mathcal{R}_{rec}$), Mean Average Precision (mAP) and $\mathcal{F}1$ -score in real-time application. There are three ways to classify each detected bounding box: True Positive (T_p), False Positive (F_p), and False Negative (F_n). If the Intersection over Union (IoU) between a bounding box and the ground truth is more than 50%, the box is considered a T_p . Alternatively, if the IoU is lower than the threshold, it is considered a F_p ; equally, if the ground truth is completely overlooked, it is considered a F_n . The following metrics are used to assess the model's performance based on these areas.

The accuracy of the projected bounding boxes is measured by $\mathcal{P}_{rec}$, which is found by dividing the numbers of T_p , by the sum of T_p , and F_p :

$$\mathcal{P}_{rec} = \frac{T_p}{T_p + F_p}, \quad (22)$$

$\mathcal{R}_{rec}$ evaluates the model's ability to identify all ground truth bounding boxes. It is calculated as the ratio of T_p to the sum of T_p and F_n :

$$\mathcal{R}_{rec} = \frac{T_p}{T_p + F_n}, \quad (23)$$

The $\mathcal{F}1$ -score combines both $\mathcal{P}_{rec}$ and $\mathcal{R}_{rec}$ into a single metric that provides a balanced evaluation. The $\mathcal{F}1$ -score is computed as the harmonic mean of $\mathcal{P}_{rec}$ and $\mathcal{R}_{rec}$:

$$\mathcal{F}_1 - \text{score} = 2 \times \frac{\mathcal{P}_{rec} \times \mathcal{R}_{rec}}{\mathcal{P}_{rec} + \mathcal{R}_{rec}}, \quad (24)$$

Mean Average Precision (mAP) summarizes the model's performance across all object categories. For each class, we compute the Average Precision (AP) as the area under the Precision–Recall curve, which is obtained by varying the detection confidence thresh-

old and evaluating detections against ground truth using an Intersection over Union (IoU) criterion.

The mAP is then obtained by averaging the AP values across all N_c object classes:

$$mAP = \frac{1}{N_c} \sum_{i=1}^{N_c} AP(i), \quad (25)$$

where $AP(i)$ denotes the Average Precision for class i .

An essential statistic that measures a model's processing speed is inference speed, which is expressed in frames per second (FPS). The FPS metric evaluates the model's processing capacity in relation to the time required to process a single frame.

$$FPS = \frac{1000}{t_{pre} + t_{inf} + t_{post}}, \quad (26)$$

where: t_{pre} is the pre-processing time (e.g., image resizing, padding, and channel transformation), t_{inf} is the processing time for generating output, t_{post} is the post-processing time (e.g., metric conversion). FPS helps evaluate the model's overall speed and suitability for real-time applications in dynamic environments. We additionally report parameter count (M) and GFLOPs per 640×640 input, consistent with prior lightweight YOLO research [51–53]. For clarity in comparative analysis, all reported improvements are expressed as absolute percentage-point differences (Δ) unless otherwise stated.

4.3. Quantitative Results on BRVD

4.3.1. Overall Performance on BRVD

The Bangladeshi Road Vehicle Dataset (BRVD) offers a demanding, domain-specific benchmark with 13 native vehicle categories under dense, heterogeneous traffic. Table 2 summarizes the per-class performance of BDNet on the BRVD validation set in terms of precision ($\mathcal{P}_{rec}$), recall ($\mathcal{R}_{rec}$), $\mathcal{F1}$ -score, and mean Average Precision at IoU thresholds of 0.5, 0.75, and 0.5–0.95 (mAP_{50} , mAP_{75} , mAP_{50-95} , respectively). On average, BDNet attains 88.4% precision, 75.3% recall, 81.3% $\mathcal{F1}$ -score, with 85.9% mAP_{50} , 73.5% mAP_{75} and 67.3% mAP_{50-95} . These values indicate that the detector maintains a strong balance between accuracy and computational efficiency across diverse vehicle types (correctly localizing vehicles and avoiding false alarms in crowded scenes). The per-class detection results of BDNet are reported in Table 2, while the training and validation dynamics over 300 epochs are illustrated in Figure 8.

Several categories achieve particularly strong performance. CNG, Truck, Easy-Bike, and Bus reach mAP_{50} values of 89.7%, 88.8%, 88.7%, and 87.0%, respectively, indicating that BDNet effectively captures both large and medium vehicle structures. Bicycle achieves the highest $\mathcal{F1}$ -score (84.7%), demonstrating that the model can still discriminate small, elongated objects despite frequent occlusion and background clutter. Performance drops slightly for a small number of visually ambiguous classes. For example, Leguna exhibits lower recall and an mAP_{50} of 74.9%, which can be attributed to its similarity in shape to covered vans and pickups and to frequent partial occlusion in congested road scenes. Even in these challenging cases, the detector maintains competitive performance, and the overall mAP_{50-95} of 67.3% confirms that BDNet preserves localization quality under stricter IoU thresholds.

The evolution of training and validation metrics over 300 epochs is depicted in Figure 8 below. All loss components (box, classification, and DFL) decrease rapidly in the early stages and stabilize smoothly as training progresses. Precision, recall, and both mAP_{50} and mAP_{50-95} increase steadily without noticeable divergence between training and validation

curves. This behavior suggests effective optimization, good regularization, and no obvious signs of overfitting on BRVD.

Table 2. Per-class detection performance of BDNet on the BRVD validation set, reported in terms of precision, recall, $\mathcal{F}_1$ -score, mAP_{50} , mAP_{75} , mAP_{50-95} .

Class Name	$\mathcal{P}_{rec}$	$\mathcal{R}_{rec}$	$\mathcal{F}_1$ -score	mAP_{50}	mAP_{75}	mAP_{50-95}
Auto-Rickshaw	85.7	77.1	81.2	86.3	80.0	72.9
Bicycle	91.1	79.2	84.7	88.5	62.8	58.6
Bus	91.8	74.5	82.3	87.0	74.7	68.5
Car	91.6	77.3	83.7	87.4	74.1	69.4
CNG	92.2	81.1	86.3	89.7	79.1	72.0
Covered Van	89.3	75.7	81.9	86.7	79.7	73.3
Easy Bike	87.3	78.9	82.9	88.7	82.1	74.1
Leguna	84.4	61.2	71.0	74.9	69.0	62.7
Motorcycle	92.0	74.0	82.0	85.4	64.5	57.4
Pickup	88.7	72.1	79.5	85.2	78.6	71.1
Rickshaw	88.7	75.9	81.8	86.6	72.4	67.6
Truck	85.1	81.5	83.2	88.8	77.4	71.7
Van	81.9	70.3	75.7	81.1	61.6	55.2
All (Mean)	88.4	75.3	81.3	85.9	73.5	67.3

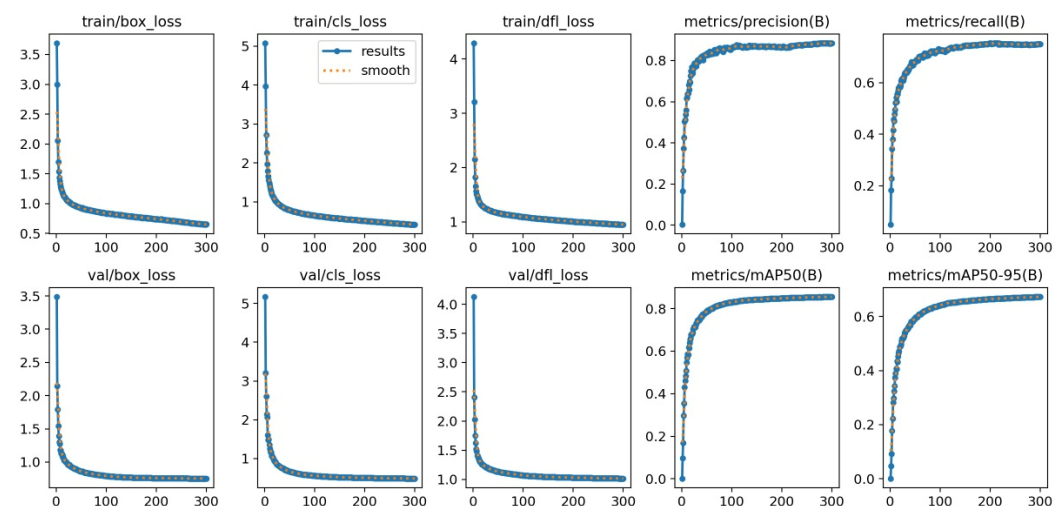


Figure 8. Training and validation performance curves of BDNet on the BRVD validation set. The smooth convergence and stable loss trajectory demonstrate effective optimization and strong generalization.

Collectively, these results demonstrate that BDNet delivers reliable detection accuracy and balanced performance across diverse vehicle classes under complex traffic conditions. The integration of HyDASE enhances detail-preserving downsampling in cluttered scenes, C3k2_MogaBlock strengthens contextual aggregation for multi-scale traffic patterns, and A2C2f_FRFN refines feature representations under occlusion. Together, these components enable BDNet to converge efficiently and provide robust, real-time vehicle detection suitable for smart urban traffic monitoring and intelligent transportation systems.

4.3.2. Comparison with Baseline YOLO Models

To rigorously benchmark the proposed BDNet, we compared it with five recent lightweight YOLO architectures, YOLOv8n [22], YOLOv9t [54], YOLOv10n [55], YOLOv11n [40,41,56], and YOLOv12n [18] on the BRVD dataset. Quantitative results

are presented in Table 3, while Figures 9–11 illustrate the corresponding precision–recall (PR) curves and qualitative detection outcomes.

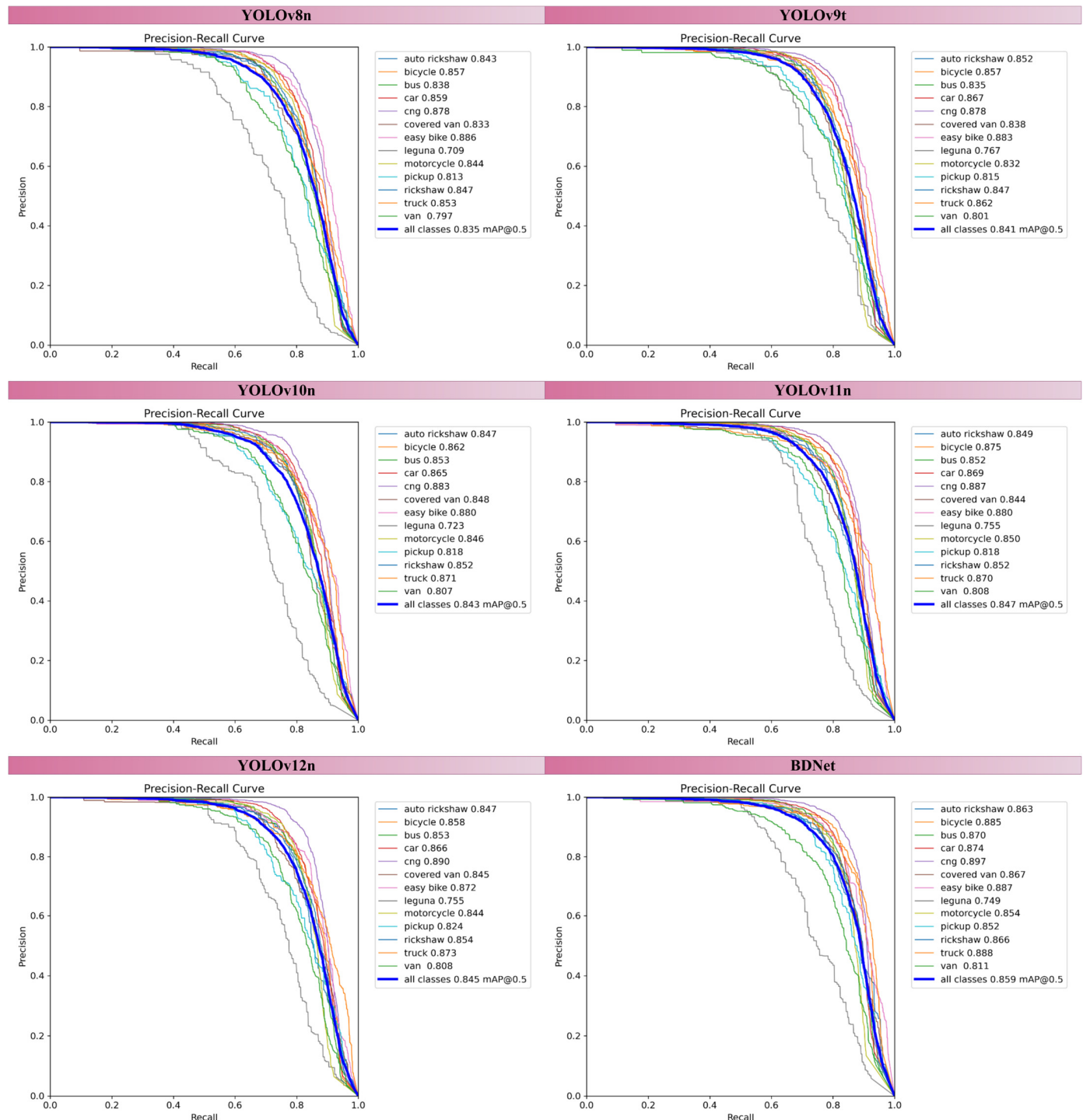


Figure 9. Precision–recall (PR) curves of BDNet and YOLO baseline models (YOLOv8n–YOLOv12n) on the BRVD validation set. BDNet maintains higher precision across the entire recall spectrum, evidencing robustness under dense and occluded traffic.

As shown in Table 3, BDNet achieves the strongest overall detection performance among all evaluated models, obtaining 88.4% precision, 75.3% recall, and an $\mathcal{F}_1$ -score of 81.3%. Compared with the most recent baseline, YOLOv12n, BDNet improves recall by +1.5 percentage points and $\mathcal{F}_1$ -score by +3.3 percentage points, indicating more reliable

detection of vehicles under occlusion and high traffic density. The improvement in recall is particularly important for ITS, where missed detections may propagate downstream to traffic analysis or decision-making modules.

Table 3. Quantitative comparison of BDNet with YOLOv8n–YOLOv12n on the BRVD validation set. Metrics include precision, recall, $\mathcal{F}_1$ -score, and mAP at multiple IoU thresholds, along with model size (M parameters), computational complexity (GFLOPs), and inference speed (FPS).

Model	$\mathcal{P}_{rec}$	$\mathcal{R}_{rec}$	$\mathcal{F}_1$ -score	mAP_{50}	mAP_{75}	mAP_{50-95}	$\mathcal{P}_{params} (\mathcal{M})$	GFLOPs	FPS (f/s)
YOLOv8n	86.1	72.6	78.7	83.5	72.1	64.7	3.2	8.7	217.4
YOLOv9t	87.2	73.7	79.8	84.1	73.1	66.0	2.0	7.7	185.2
YOLOv10n	85.5	74.3	79.4	84.3	72.4	65.4	2.3	6.7	212.8
YOLOv11n	87.0	75.0	80.5	84.7	71.0	65.1	2.6	6.5	222.2
YOLOv12n	87.3	73.8	78.0	84.5	72.6	66.6	2.6	6.5	196.1
BDNet (Ours)	88.4	75.3	81.3	85.9	73.5	67.3	2.5	6.0	285.7

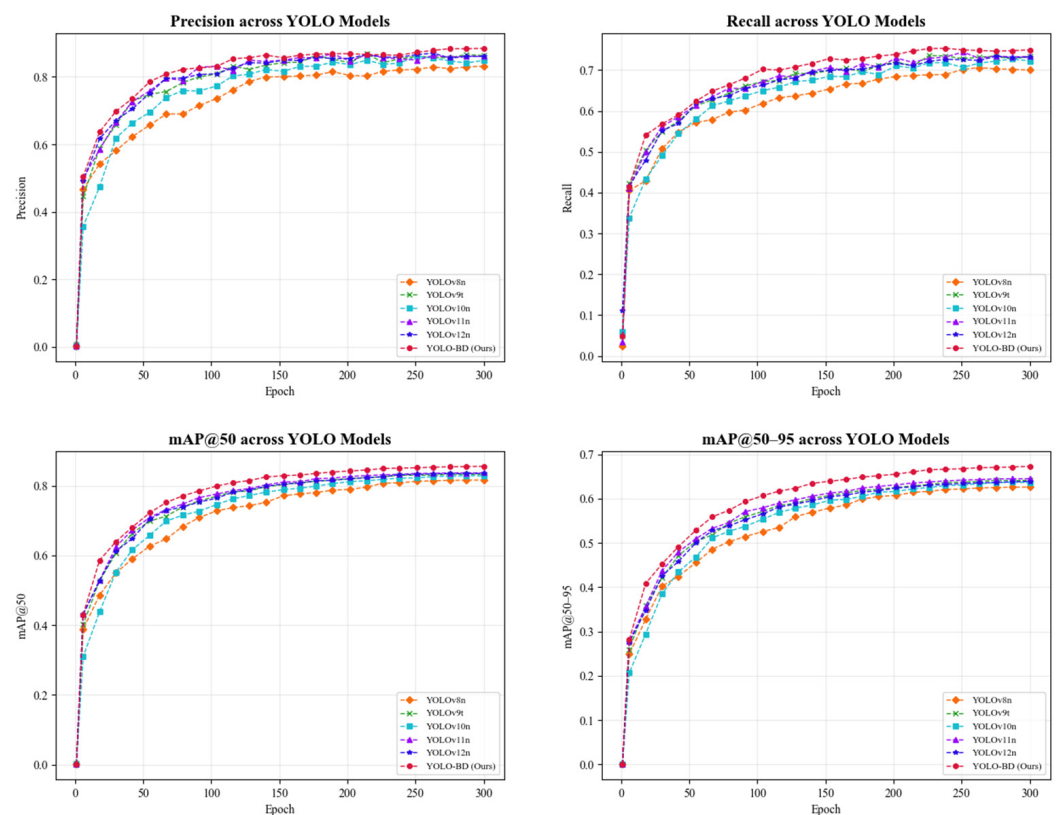


Figure 10. Integrated comparison of precision–recall (PR) curves and mean Average Precision metrics (mAP_{50} and mAP_{50-95}) for BDNet and YOLO baseline models on the BRVD validation set. This comprehensive visualization demonstrates that BDNet consistently outperforms YOLOv8n–YOLOv12n across multiple IoU thresholds, exhibiting superior recall stability and stronger cross-threshold generalization.

In terms of localization accuracy, BDNet reaches 85.9% mAP_{50} , 73.5% mAP_{75} , and 67.3% mAP_{50-95} , consistently outperforming all YOLO baselines across IoU thresholds. While the gains at mAP_{50} are modest relative to YOLOv11n and YOLOv12n, the improvement becomes more pronounced at higher IoU levels, reflecting more accurate bounding-box regression and improved spatial consistency. This behavior suggests that the proposed modules enhance fine-grained localization, which is critical in congested urban scenes with closely spaced vehicles.

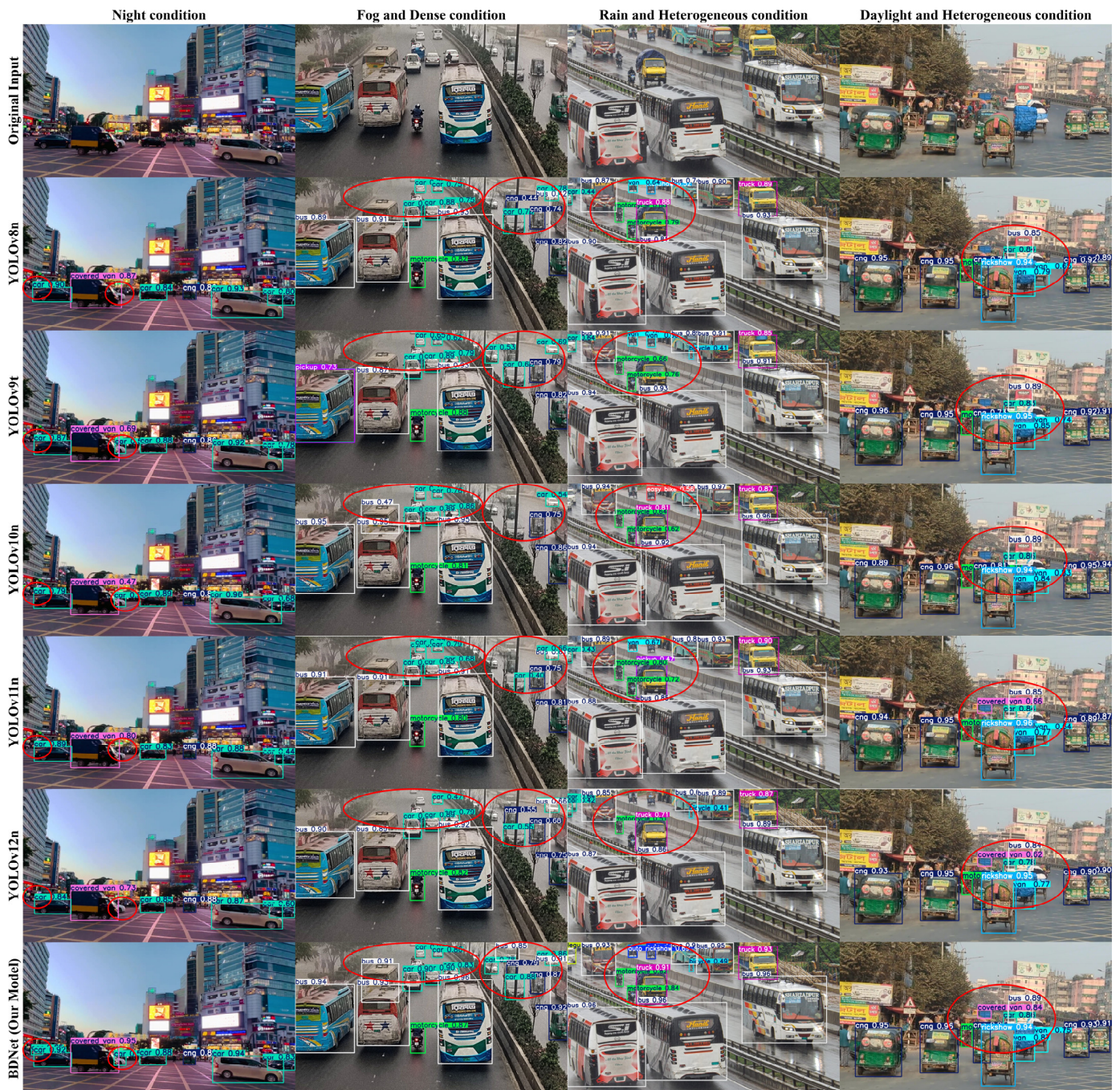


Figure 11. Qualitative detection results for BDNet and YOLO baselines on challenging BRVD scenarios, including night-time, fog, rain, and heterogeneous daytime traffic. BDNet recovers small and partially occluded vehicles, reduces background false positives, and provides more accurate bounding boxes.

From an efficiency perspective, BDNet maintains a compact footprint of 2.5 M parameters and 6.0 GFLOPs, remaining comparable to YOLOv11n and YOLOv12n while delivering higher accuracy. Notably, BDNet achieves the highest inference speed of 285.7 FPS, exceeding YOLOv12n by +89.6 FPS under identical evaluation settings. This substantial speed advantage highlights the suitability of BDNet for real-time deployment on edge or roadside computing platforms commonly used in smart city infrastructures.

Overall, the results in Table 3 demonstrate that BDNet achieves a more favorable accuracy–efficiency trade-off than existing lightweight YOLO variants. By jointly improving

recall, localization precision, and inference speed without increasing model complexity, BDNet provides a robust vehicle detection solution tailored to dense, heterogeneous urban traffic environments. These characteristics make it particularly well suited for real-time smart urban traffic monitoring and intelligent transportation systems operating under strict computational constraints.

4.3.3. Per-Class Performance Comparison on BRVD

As summarized in Table 3, BDNet achieves the strongest overall detection performance among all evaluated models. It reaches 85.9% mAP_{50} , outperforming YOLOv12n (84.5%) by +1.4 percentage points, while also attaining the highest $\mathcal{F}_1$ -score of 81.3% (+3.3 pp over YOLOv12n). Recall improves from 73.8% to 75.3% (+1.5 pp), indicating more complete recovery of vehicle instances in dense scenes. Despite these accuracy gains, BDNet remains lightweight, requiring only 2.5 M parameters and 6.0 GFLOPs. In terms of inference speed, BDNet achieves 285.7 FPS on 640×640 inputs, representing a substantial throughput advantage over YOLOv12n (196.1 FPS) and YOLOv9t (185.2 FPS), confirming its suitability for real-time ITS deployment.

Table 4 presents a per-class comparison of mAP_{50} between BDNet and YOLOv8n–YOLOv12n on the BRVD validation set. BDNet attains the highest mean mAP_{50} (85.9%), exceeding YOLOv12n by +1.4 pp and YOLOv8n by +2.4 pp. Performance gains are consistent across most categories, with notable improvements for CNG (+0.7 pp), Easy-Bike (+1.5 pp), Truck (+1.5 pp), and Pickup (+2.8 pp), reflecting enhanced robustness to scale variation and structural diversity. These results suggest that HyDASE and C3k2_MogaBlock effectively preserve fine-grained details, while A2C2f_FRFN refines discriminative features under occlusion. Although the Leguna class remains challenging due to strong visual similarity with adjacent categories, BDNet maintains competitive performance, indicating that remaining errors are largely attributable to intrinsic class ambiguity rather than architectural limitations.

Table 4. Per-class mAP_{50} comparison between BDNet and YOLO baseline models (YOLOv8n–YOLOv12n) on the BRVD validation set.

Class Name	YOLOv8n	YOLOv9t	YOLOv10n	YOLOv11n	YOLOv12n	BDNet (Ours)
Auto- Rickshaw	84.3	85.2	84.7	84.9	84.7	86.3
Bicycle	85.7	85.7	86.1	87.5	85.8	88.5
Bus	83.8	83.5	85.3	85.2	85.3	87.0
Car	85.9	86.7	86.5	86.9	86.6	87.4
CNG	87.8	87.8	88.3	88.7	89.0	89.7
Covered Van	83.3	83.8	84.8	84.4	84.5	86.7
Easy Bike	88.6	88.3	88.0	88.0	87.2	88.7
Leguna	70.9	76.7	72.3	75.5	75.5	74.9
Motorcycle	84.4	83.2	84.6	85.0	84.4	85.4
Pickup	81.3	81.5	81.8	81.8	82.4	85.2
Rickshaw	84.7	84.7	85.2	85.2	85.4	86.6
Truck	85.3	86.2	87.1	87.0	87.3	88.8
Van	79.7	80.1	80.7	80.8	80.8	81.1
All (Mean)	83.5	84.1	84.3	84.7	84.5	85.9

4.3.4. Comparison with State-of-the-Art Models

To further contextualize the performance of BDNet within the broader object detection landscape, we benchmarked it against a diverse set of state-of-the-art (SOTA) detectors, include classical one-stage and two-stage CNNs as well as recent transformer and lightweight YOLO variants. The compared methods are SSD [57], Faster R-CNN [58], RT-DETR [59], LSOD-YOLO [25], SO-YOLOv8 [60], YOLO-FD [61], SD-YOLO-AWDNet [62],

VP-YOLO [63], EL-YOLO [64], LVD-YOLO [1], and MT-YOLO [65]. All models follow the common training and evaluation protocol described in Section 4.2, and performance is reported on the BRVD validation set.

From Table 5, BDNet achieves the highest overall detection accuracy on the BRVD dataset, with 85.9% mAP_{50} and 67.3% mAP_{50-95} . Compared with lightweight YOLO-style baselines, BDNet consistently improves performance under both loose and strict IoU criteria. In particular, relative to LVD-YOLO (83.1%/64.1%), BDNet yields gains of +2.8 pp in mAP_{50} and +3.2 pp in mAP_{50-95} . When compared with SD-YOLO-AWDNet (81.5%/65.1%), BDNet improves accuracy by +4.4 pp at mAP_{50} and +2.2 pp at mAP_{50-95} . Against VP-YOLO and MT-YOLO, BDNet delivers consistent improvements of +3.1 pp and +3.4 pp in mAP_{50} , respectively, while matching VP-YOLO and surpassing MT-YOLO under the stricter mAP_{50-95} metric. Although LSOD-YOLO attains a strong mAP_{50-95} of 67.5%, BDNet achieves a comparable level (67.3%) with 9.4 percentage points higher mAP_{50} , while using 1.3 M fewer parameters (3.8 M $\rightarrow$ 2.5 M) and 27.9 fewer GFLOPs (33.9 $\rightarrow$ 6.0), corresponding to an approximately 82% reduction in computational cost. These results indicate that BDNet offers a more favorable accuracy–efficiency trade-off for real-time smart-city traffic monitoring.

Table 5. Performance comparison of BDNet with state-of-the-art object detection models on the BRVD dataset. Metrics include mAP_{50} , mAP_{50-95} , parameter count, GFLOPs, and FPS.

Model	mAP_{50}	mAP_{50-95}	Params ($\mathcal{M}$)	GFLOPs	FPS (f/s)
SSD	51.4	45.3	138.0	34.80	96.2
Faster R-CNN	65.1	60.1	41.2	292.3	122.0
RT-DETR	82.9	65.2	41.9	125.7	256.4
LSOD-YOLO	78.5	67.5	3.8	33.9	181.8
SO-YOLOv8	78.2	61.4	69.8	263.0	166.7
YOLO-FD	65.9	30.3	12.0	52.8	208.3
SD-YOLO-AWDNet	81.5	65.1	3.7	8.3	204.1
VP-YOLO	82.8	67.3	66.8	129.7	232.6
EL-YOLO	60.5	42.5	1.1	6.7	185.2
LVD-YOLO	83.1	64.1	3.6	5.7	243.9
MT-YOLO	82.5	65.3	3.5	6.9	238.1
BDNet (Our Model)	85.9	67.3	2.5	6.0	285.7

Overall, the results indicate that BDNet provides a more balanced integration of detection accuracy, computational efficiency, and real-time throughput than existing state-of-the-art models. This advantage is particularly relevant for smart urban traffic monitoring, where high detection reliability must be achieved under strict latency and resource constraints.

4.3.5. Precision–Recall and mAP Curve Analysis

The PR curves in Figure 9 show that BDNet maintains higher precision across almost the entire recall range compared with YOLOv8n–YOLOv12n. In particular, while the baselines suffer noticeable precision drops at high recall, BDNet preserves a smoother decline, indicating better control over false positives under occlusion and illumination variance.

To provide a consolidated view, Figure 10 overlays PR curves with mAP_{50} and mAP_{50-95} trends for all models. BDNet consistently dominates the baselines across IoU thresholds, confirming that the improvements observed in Tables 3 and 4 are not limited to a single operating point but persist across a range of decision thresholds.

4.4. Qualitative Results Analysis on BRVD

This subsection complements the quantitative results by examining how BDNet forms discriminative internal representations on BRVD. We analyze gradient-based attention,

hierarchical feature evolution, and class-wise confusion to elucidate the mechanisms underlying BDNet's improved robustness in dense urban traffic.

4.4.1. Detection Visualization on BRVD

Representative detection examples from the BRVD validation set are presented in Figure 11. The figure covers four representative scenarios: (a) night-time scenes, (b) fog with dense traffic, (c) rain with heterogeneous flows, and (d) daytime urban congestion. For each scenario, rows correspond to detections from YOLOv8n, YOLOv9t, YOLOv10n, YOLOv11n, YOLOv12n, and BDNet (ours).

Annotated regions (red ovals) highlight typical failure modes of the baselines and the corresponding corrections by BDNet. Across these scenes, BDNet (i) recovers small or partially occluded vehicles such as cng, car, van, rickshaws and motorcycles that are often missed by the baselines, (ii) produces tighter bounding boxes with fewer overlaps and background false alarms, and (iii) reduces misclassifications between visually similar classes such as easy-bike and rickshaw. All models are evaluated under the same confidence threshold and Non-Maximum Suppression (NMS) settings to ensure a fair visual comparison.

The qualitative patterns in Figure 11, together with the quantitative evidence in Tables 3 and 4, Figures 9 and 10, confirm that BDNet offers more accurate localization and cleaner detections, particularly for small, occluded, and visually ambiguous vehicles. In summary, BDNet demonstrates a strong lightweight performance advantage on BRVD. It improves mAP , F_1 -score, recall, and FPS simultaneously, while keeping computational cost low. These gains arise from the joint effect of HyDASE's detail-preserving downsampling, C3k2_MogaBlock's multi-order contextual reasoning, and A2C2f_FRFN's feature refinement, making BDNet a strong candidate for real-time ITS in complex urban environments.

4.4.2. Detection Heatmaps

Figure 12 below visualizes Grad-CAM heatmaps for YOLOv8n-YOLOv12n and BDNet using the BRVD validation set [66]. Across all scenes, BDNet concentrates activation on vehicle contours and key parts such as wheels, windshields, and bumpers, including small, distant objects and partially occluded instances highlighted by the red ellipses. Baselines exhibit diffuse responses that spill into road surfaces, vegetation, and signage, which is consistent with their higher false-positive rates and less precise localization. The sharpened focus observed in BDNet suggests that HyDASE preserves fine structure during downsampling, C3k2_MogaBlock injects context to disambiguate overlaps, and A2C2f_FRFN suppresses redundant background features, yielding compact, task-relevant attention maps [66].

4.4.3. Feature-Map Progression

Figure 13 below tracks representative feature maps at three stages (initial, intermediate, final) for each model. In early layers, BDNet retains edge and texture fidelity comparable to baselines. At intermediate depth, baseline activations become fragmented and noisy, especially around clutter and overlapping vehicles. In contrast, BDNet maintains coherent, high-contrast responses along object extents and suppresses repetitive background patterns. In the final stage, BDNet produces compact, high-energy responses aligned with object cores, which translate to tighter bounding boxes and fewer duplicate detections. These observations support the intended roles of the modules: HyDASE preserves scale-sensitive details for small objects, C3k2_MogaBlock aggregates multi-order context to stabilize occlusions, and A2C2f_FRFN refines fused features so that only salient structures propagate to the detection heads.

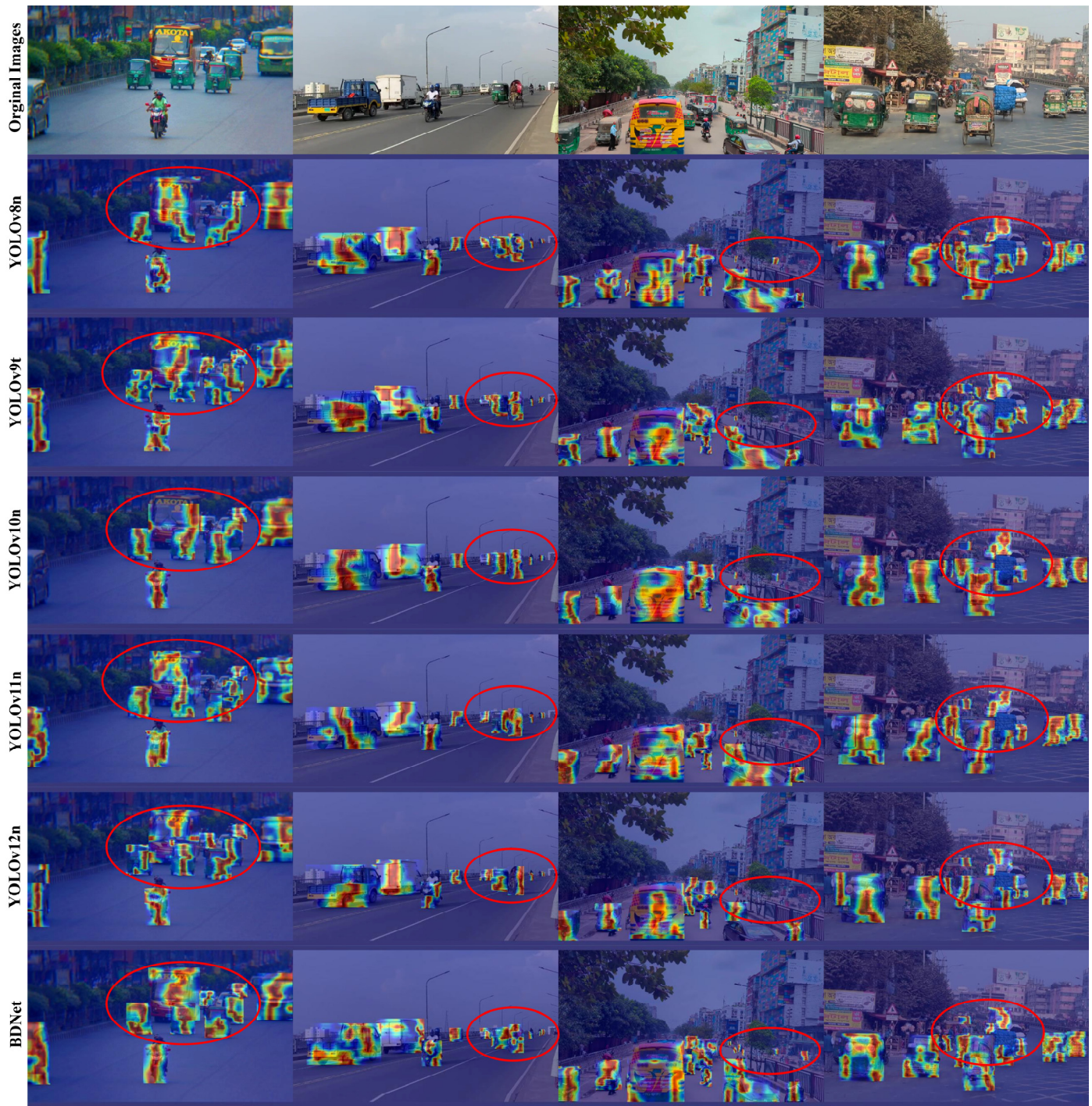


Figure 12. Grad-CAM comparisons on BRVD validation images. BDNet attends tightly to vehicle boundaries and discriminative parts under occlusion and scale variation, whereas baselines display broader, background-biased activations.

4.4.4. Confusion-Matrix Analysis

Figure 14 presents normalized confusion matrices for all baselines and BDNet over the 13 BRVD classes. BDNet shows stronger diagonal dominance with visibly reduced off-diagonal mass. Improvements are most notable for pairs that are visually similar or frequently co-occurring in traffic, such as rickshaw vs. auto-rickshaw and truck vs. covered-van. The reduced confusion indicates better inter-class separation and intra-class

compactness, which is consistent with the refined, class-discriminative representations observed in Figures 12 and 13.

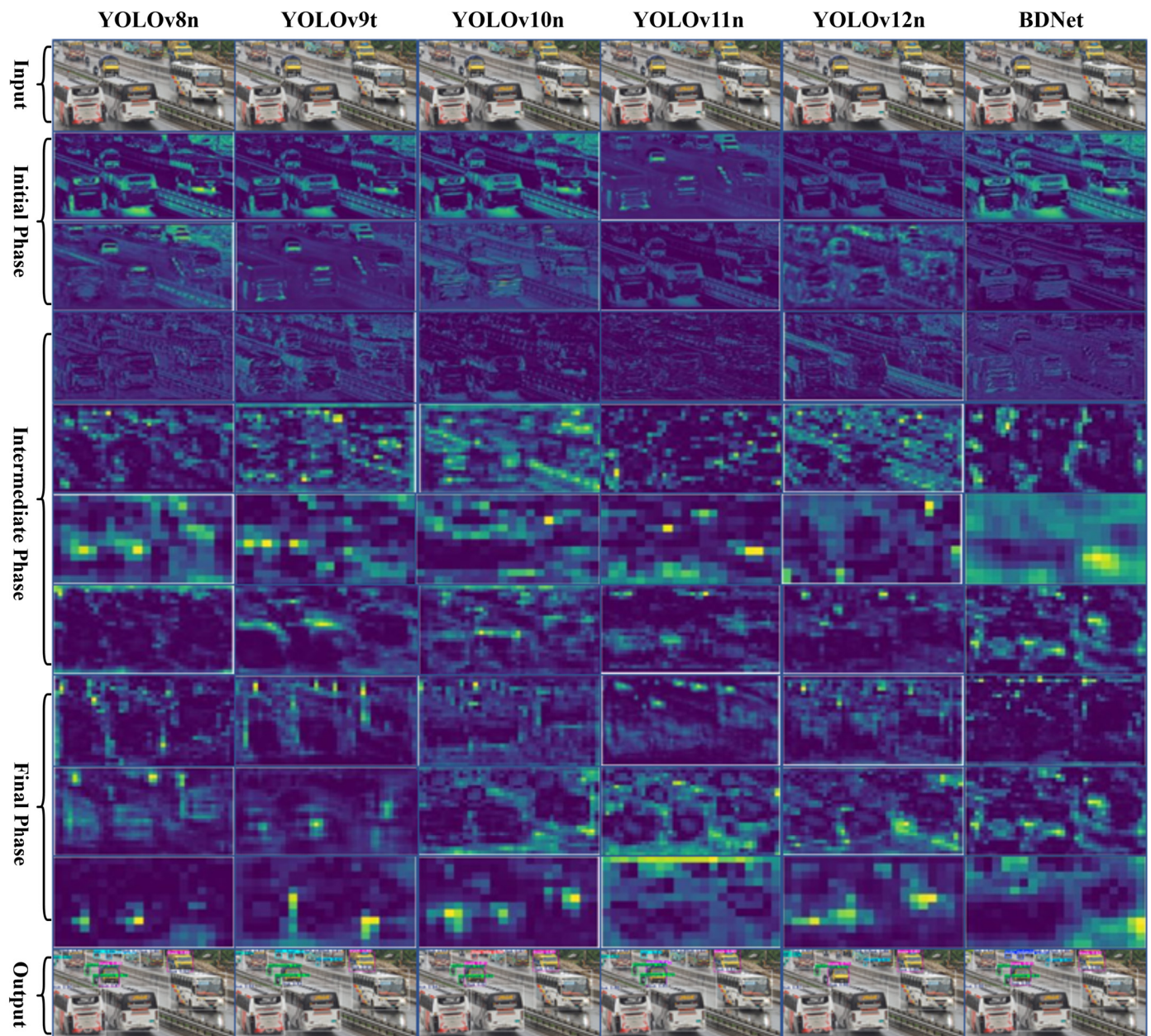


Figure 13. Hierarchical feature evolution for YOLOv8n-YOLOv12n and BDNet on BRVD. BDNet exhibits structured, high-contrast representations at intermediate and final stages, while baselines show scattered or background-dominated activations.

Collectively, these qualitative results provide a mechanistic explanation for BDNet's quantitative gains: attention maps are sharper and object-centric, intermediate features remain structured under clutter and occlusion, and class separability improves at the decision level. The evidence confirms that the combined effect of HyDASE, C3k2_MogaBlock, and A2C2f_FRFN yields representations that are both efficient and discriminative for real-time vehicle detection in dense urban scenes.

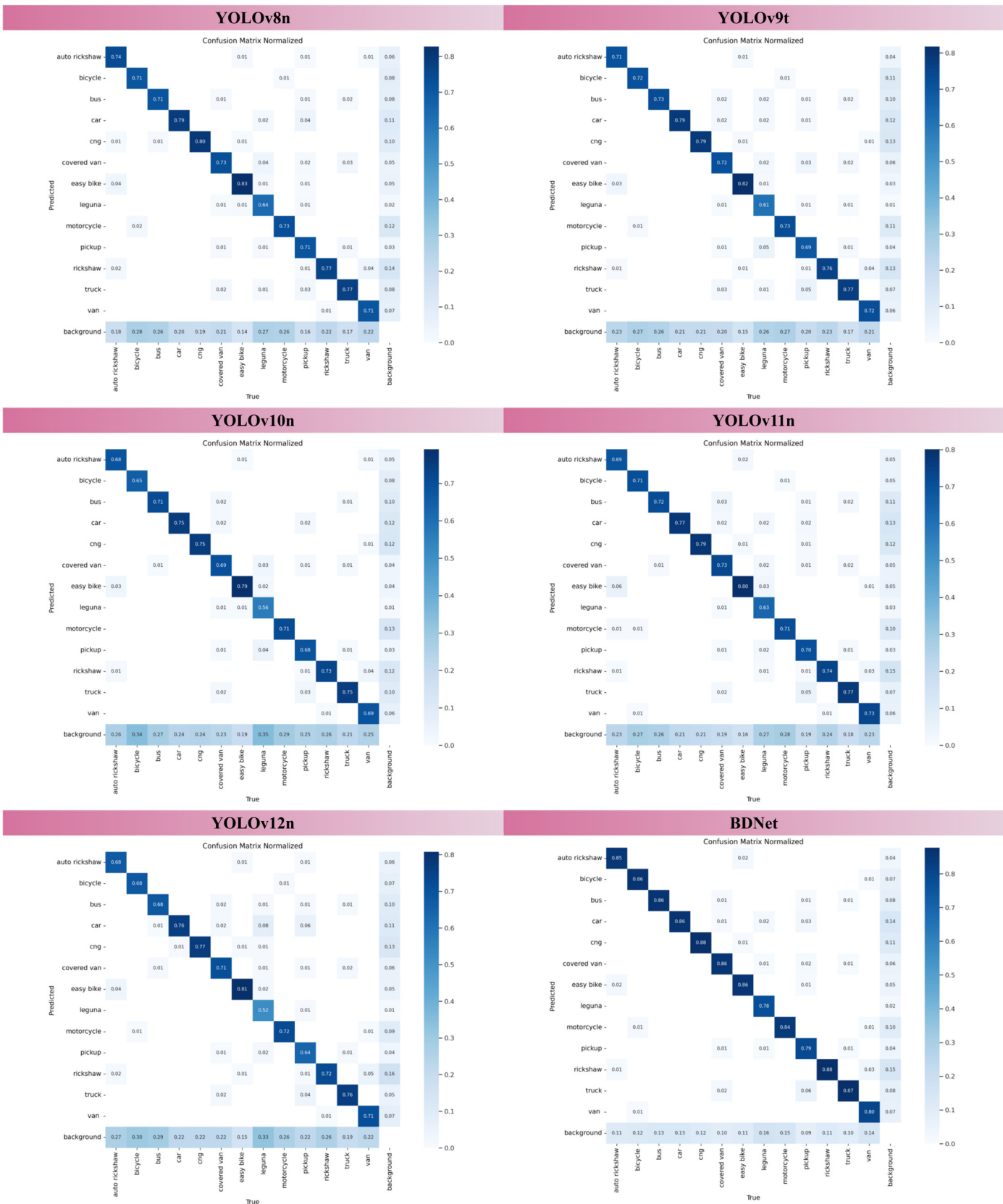


Figure 14. Normalized confusion matrices on the BRVD validation set. BDNet reduces inter-class confusion relative to YOLOv8n-YOLOv12n, particularly for visually similar vehicle types.

4.5. Ablation Study: Module Contribution Analysis

To quantify the impact of each architectural component, we conduct an ablation study on the BRVD validation set under identical training and evaluation settings. Starting from the YOLOv12n baseline, we incrementally integrate HyDASE, C3k2_MogaBlock, and A2C2f_FRFN. The results are summarized in Table 6.

4.5.1. Individual Module Effects

Adding HyDASE with baseline provides the most substantial single-module benefit, improving mAP_{50} from 84.5% to 85.3% and recall from 73.8% to 75.0%, while simultaneously reducing parameters (2.6M $\rightarrow$ 2.3M), FLOPs (6.5G $\rightarrow$ 5.6G), and increasing inference speed to 256.4 FPS. This confirms its effectiveness in preserving spatial detail during downsampling with lower computational cost. Additionally, C3k2_MogaBlock primarily enhances contextual reasoning, increasing precision (87.3% $\rightarrow$ 88.1%) and slightly improving mAP_{50} to 84.9%, with a moderate computational overhead (250.0 FPS). Furthermore, the A2C2f_FRFN module improves recall (74.8%) by suppressing redundant activations, while maintaining a lightweight profile (243.9 FPS), indicating efficient feature refinement.

4.5.2. Combined Module Effects

Pairwise combinations consistently outperform individual modules. For Example, HyDASE combined with C3k2_MogaBlock achieves 85.8% mAP_{50} and 263.2 FPS, demonstrating that efficient downsampling complements contextual aggregation. In addition, HyDASE with A2C2f_FRFN yields the highest speed among two-module configurations (277.8 FPS) while maintaining strong accuracy (85.6% mAP_{50}). Moreover, C3k2_MogaBlock with A2C2f_FRFN provides balanced gains in recall (75.5%) and speed (270.3 FPS), confirming their complementary roles.

Table 6. Ablation study of BDNet on the BRVD validation set. Each configuration is evaluated in terms of precision ($\mathcal{P}_{rec}$), recall ($\mathcal{R}_{rec}$), mAP_{50} , parameter count (Params), FLOPs, and inference speed in frames per second (FPS).

Baseline	HyDASE	C3K2_MogaBlock	A2C2f_FRFN	$\mathcal{P}_{rec}$	$\mathcal{R}_{rec}$	mAP_{50}	Params ($\mathcal{M}$)	FLOPs (G)	FPS (f/s)
✓				87.3	73.8	84.5	2.6	6.5	196.1
✓	✓			88.0	75.0	85.3	2.3	5.6	256.4
✓		✓		88.1	74.4	84.9	2.5	6.2	250.0
✓			✓	86.8	74.8	84.7	2.7	6.3	243.9
✓	✓	✓		88.5	75.3	85.8	2.3	5.8	263.2
✓	✓		✓	88.5	75.2	85.6	2.5	5.8	277.8
✓		✓	✓	88.3	75.5	85.5	2.7	6.4	270.3
✓	✓	✓	✓	88.4	75.3	85.9	2.5	6.0	285.7

4.5.3. Full BDNet Configuration

Integrating all three modules results in the best overall performance. The complete BDNet model achieves 85.9% mAP_{50} , 75.3% recall, and the highest throughput of 285.7 FPS, with only 2.5M parameters and 6.0 GFLOPs. These results demonstrate that the proposed modules act synergistically to improve accuracy, robustness, and real-time efficiency.

Overall, the ablation study verifies that HyDASE ensures cost-effective detail preservation, C3k2_MogaBlock strengthens multi-order contextual modeling, and A2C2f_FRFN refines multi-scale features. Their integration enables BDNet to achieve a favorable accuracy–efficiency balance suitable for real-time vehicle perception in dense smart-city traffic scenarios.

4.6. Scale-Wise Comparison with the YOLOv12 Family

To examine whether the proposed architectural principles generalize across different model capacities, a scale-wise comparison is conducted between the YOLOv12 family (n, s, m, l) and their corresponding BDNNet variants (BDNetn, BDNet s, BDNet m, BDNet l). This analysis evaluates whether the gains achieved through detail-preserving downsampling, contextual aggregation, and feature refinement remain consistent from ultra-lightweight configurations suitable for edge deployment to larger models designed for high-throughput intelligent transportation systems (ITS). Table 7 reports quantitative results on the BRVD validation set, while Figure 15 illustrates the accuracy–efficiency trade-offs in terms of mAP versus model size (Params) and computational cost (GFLOPs). Together, these results provide a comprehensive assessment of BDNNet’s scalability and deployment flexibility.

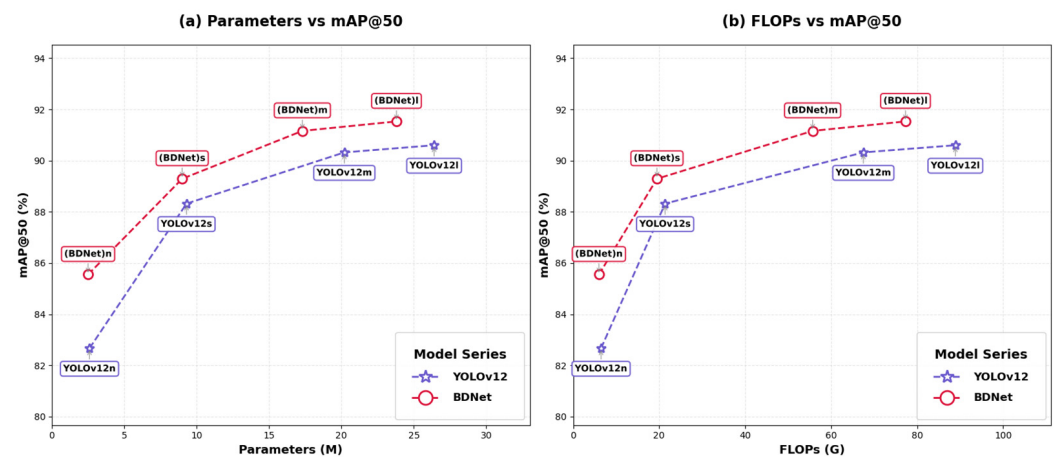


Figure 15. Trade-off analysis (mAP_{50} vs. Params and mAP_{50} vs. GFLOPs) for YOLOv12 family and BDNNet variants. BDNNet consistently achieves higher mAP at equal or lower computational complexity, confirming its scalable efficiency across model sizes.

The results in Table 7 reveal that BDNNet consistently delivers superior accuracy-to-efficiency trade-offs across all scales, confirming that its design principles generalize effectively regardless of network size.

4.6.1. Performance Across Model Scales

Nano and small variants (n, s): For the most resource-constrained configurations, BDNNet demonstrates clear advantages over the corresponding YOLOv12 baselines. Compared with YOLOv12n, BDNetn improves mAP_{50} by +1.4 pp (85.9 vs. 84.5), recall by +1.5 pp, and $F1$ -score by +3.3 pp, while simultaneously reducing model size (2.5 M vs. 2.6 M parameters) and computational cost (6.0 vs. 6.5 GFLOPs). These gains indicate that BDNetn extracts richer contextual information and preserves fine-grained details even under strict capacity constraints, making it well suited for real-time roadside or embedded deployments. In the small-scale setting, BDNet s further improves localization quality under stricter IoU thresholds. Relative to YOLOv12s, BDNet s achieves higher mAP_{75} (+2.2 pp) and mAP_{50-95} (+1.1 pp), despite a slight reduction in precision. This behavior reflects a favorable trade-off toward improved recall and localization robustness, which is particularly important in dense urban traffic where partial occlusion and object overlap are common.

Table 7. Scale-wise comparison between YOLOv12 models and corresponding BDNet variants (n, s, m, l) on the BRVD validation set. Metrics include mean precision, recall, $\mathcal{F}_1$ -score, and mAP at IoU 0.50, 0.75, and 0.50–0.95, along with model size (Params) and computational cost (GFLOPs).

Model	$\mathcal{P}_{rec}$	$\mathcal{R}_{rec}$	$\mathcal{F}_1$ -score	mAP_{50}	mAP_{75}	mAP_{50-95}	Params ($\mathcal{M}$)	GFLOPs
YOLOv12n	87.3	73.8	78.0	84.5	72.6	66.6	2.6	6.5
BDNetn	88.4	75.3	81.3	85.9	73.5	67.3	2.5	6.0
YOLOv12s	91.6	80.1	85.5	88.3	76.3	70.2	9.3	21.4
BDNets	88.5	81.1	84.6	89.3	78.5	71.3	9.0	19.5
YOLOv12m	91.2	84.4	87.7	90.3	80.4	72.2	20.2	67.5
BDNetm	89.1	84.2	86.6	91.2	82.2	74.4	17.3	55.7
YOLOv12l	91.9	84.2	87.9	90.6	79.2	75.2	26.4	88.9
BDNetl	91.0	83.1	86.9	91.5	82.9	75.7	23.8	77.3

Medium and large variants (m, l): As model capacity increases, BDNet continues to deliver a more favorable accuracy-to-efficiency balance. BDNetm improves mAP_{75} by +1.8 pp and mAP_{50-95} by +2.2 pp compared with YOLOv12m, while reducing parameters by approximately 14% and GFLOPs by 18%. These results highlight BDNet’s ability to enhance multi-scale discrimination and localization accuracy without incurring additional computational overhead. At the largest scale, BDNetl maintains this trend. Compared with YOLOv12l, BDNetl achieves higher mAP_{75} (+3.7 pp) and mAP_{50-95} (+0.5 pp), alongside notable reductions in model size and computational cost. The consistent improvements observed even at high capacity confirm that BDNet’s architectural refinements scale effectively and remain computationally efficient. Figure 15 further illustrates these findings, showing that BDNet variants consistently occupy more favorable positions on the accuracy–efficiency frontier, achieving higher mAP at equal or lower parameter counts and GFLOPs across all model sizes.

4.6.2. Summary of Findings

The scale-wise evaluation leads to three key conclusions. First, BDNet achieves higher detection accuracy than the YOLOv12 family at every capacity level, with particularly strong gains under stricter IoU criteria (mAP_{75} and mAP_{50-95}), reflecting improved localization in occluded and cluttered scenes. Second, BDNet consistently requires fewer parameters and lower computational cost across all scales, supporting efficient deployment from lightweight edge devices to high-capacity ITS servers. Third, the benefits of HyDASE, C3k2_MogaBlock, and A2C2f_FRFN are stable and complementary: HyDASE preserves spatial detail during downsampling, C3k2_MogaBlock strengthens multi-order contextual reasoning, and A2C2f_FRFN suppresses redundant activations during multi-scale fusion. Collectively, these results demonstrate that BDNet offers a scalable and robust vehicle detection framework with a consistently superior accuracy-efficiency trade-off, aligning well with the practical requirements of smart urban traffic monitoring systems.

4.7. Cross-Dataset Generalization on VisDrone-DET2019

To evaluate the robustness and transferability of the proposed BDNet beyond ground-level traffic scenes, a cross-dataset generalization experiment was conducted on the VisDrone-DET2019 benchmark [45]. In contrast to BRVD, which predominantly consists of frontal and lateral views captured from roadside cameras, VisDrone-DET2019 contains aerial images acquired at varying altitudes, camera tilts, and illumination conditions. The dataset is characterized by dense object distributions, frequent occlusions, and pronounced scale variation, particularly for small targets, making it a challenging testbed for assessing cross-domain generalization. All models were trained exclusively on BRVD and directly

evaluated on VisDrone-DET2019 without any fine-tuning or domain adaptation, ensuring a strict and fair evaluation of generalization capability.

4.7.1. Quantitative Results

Table 8 below presents a quantitative comparison between BDNet and the YOLOv8n-YOLOv12n baselines on the VisDrone-DET2019 validation set. Despite the substantial domain shift, BDNet achieves the strongest overall performance across all major detection metrics. Specifically, BDNet attains an mAP_{50} of 31.9%, exceeding YOLOv12n (29.3%) and YOLOv11n (31.4%). It also achieves the highest precision (43.6%) and F_1 -score (36.6%), indicating a better balance between detection accuracy and recall in aerial scenes. In addition to accuracy improvements, BDNet demonstrates clear efficiency advantages. With only 2.5 M parameters and 6.0 GFLOPs, it delivers an inference speed of 104.2 FPS, surpassing YOLOv12n (87.7 FPS) while maintaining superior detection accuracy. These results suggest that BDNet learns feature representations that transfer more effectively under the evaluated domain shift from ground-level to aerial imagery, rather than overfitting to dataset-specific visual patterns.

Table 8. Quantitative performance comparison between BDNet and YOLOv8-YOLOv12 baselines on the VisDrone-DET2019 validation set. All models are trained on BRVD and directly evaluated on VisDrone-DET2019 without fine-tuning.

Model	P_{rec}	R_{rec}	F_1 -score	mAP_{50}	mAP_{75}	mAP_{50-95}	$Params$ (M)	GFLOPs	FPS (f/s)
YOLOv8n	38.9	30.6	33.4	29.6	25.6	17.8	3.2	8.7	39.2
YOLOv9t	42.2	31.3	35.1	31.3	26.3	17.9	2.0	7.7	49.3
YOLOv10n	39.0	30.8	33.7	29.5	25.1	17.0	2.3	6.7	59.9
YOLOv11n	42.9	31.4	35.5	31.4	24.8	17.6	2.6	6.5	82.0
YOLOv12n	40.0	30.2	33.6	29.3	25.4	16.5	2.6	6.5	87.7
BDNet (Ours)	43.6	32.6	36.6	31.9	25.9	17.9	2.5	6.0	104.2

4.7.2. Per-Class Performance Analysis

To further analyze cross-domain behavior at the category level, Table 9 reports per-class mAP_{50} results on VisDrone-DET2019. BDNet achieves the highest overall mean mAP_{50} (31.9%) and outperforms all YOLO baselines across most object categories. Performance gains are particularly evident for classes dominated by small or partially occluded objects, such as Pedestrian, Motor, Bicycle, and Truck. These categories are known to be especially challenging in aerial imagery due to reduced spatial resolution and cluttered backgrounds.

Table 9. Per-class mAP_{50} comparison between BDNet and YOLOv8-YOLOv12 models on the VisDrone-DET2019 validation set.

Class Name	YOLOv8n	YOLOv9t	YOLOv10n	YOLOv11n	YOLOv12n	BDNet (Ours)
Pedestrian	30.6	31.9	30.6	32.9	30.2	34.0
People	24.1	25.6	25.9	26.8	24.8	28.0
Bicycle	5.7	7.4	6.6	6.1	5.5	7.5
Car	73.1	74.3	73.2	75.0	73.6	75.1
Van	36.1	37.6	33.3	36.5	34.8	37.3
Truck	24.2	26.3	25.1	24.9	24.0	26.6
Tricycle	18.5	21.6	17.7	19.6	18.4	20.4
Awning-tricycle	10.3	09.9	10.0	12.1	11.1	10.4
Bus	41.3	44.0	40.3	45.1	38.0	42.9
Motor	31.6	34.0	32.8	34.6	32.2	36.9
All	29.6	31.3	29.5	31.4	29.3	31.9

The observed improvements are consistent with BDNet's architectural design. HyDASE preserves fine-grained spatial cues during downsampling, C3k2_MogaBlock aggregates multi-order contextual information to stabilize representations under occlusion, and A2C2f_FRFN enhances feature saliency while suppressing background noise. Together, these components enable more reliable detection of small and visually ambiguous targets that dominate aerial traffic scenes.

4.7.3. Qualitative Evaluation

Figure 16 below presents qualitative detection comparisons across diverse aerial scenarios, including dense daytime traffic captured from high altitude, hazy and low-contrast environments, urban top-view scenes with strong perspective distortion, and night-time traffic with significant illumination variability. Across all scenarios, BDNet produces more stable and compact bounding boxes for small and overlapping objects, recovers subtle instances such as motorcycles and pedestrians that are frequently missed by the baselines, and suppresses false positives on reflective surfaces, rooftops, and street infrastructure.

In contrast, YOLOv9t–YOLOv12n often exhibit fragmented predictions, missed detections, and inflated bounding boxes under challenging lighting and viewpoint conditions. The sharper localization and improved recall achieved by BDNet further highlight the effectiveness of the A2C2f_FRFN refinement pathway in maintaining discriminative representations under severe domain shifts.

While the cross-dataset results demonstrate clear performance advantages, it is important to contextualize these findings within the inherent domain shift between BRVD and VisDrone-DET2019. The two datasets differ substantially in viewpoint (ground-based roadside cameras versus UAV-mounted aerial sensors), object scale distributions (predominantly medium-to-large vehicles versus extremely small and densely packed targets), annotation conventions, and sensor characteristics such as compression artifacts and motion blur. These factors inevitably influence absolute detection performance and contribute to residual errors observed for very small or heavily occluded objects in aerial scenes. Consequently, the reported gains indicate improved robustness under the evaluated ground-to-aerial shift rather than universal generalization across all sensing modalities. This clarification is critical for realistic deployment expectations in smart-city systems, where detector reliability depends not only on architectural design but also on sensor placement, viewpoint geometry, and domain-specific data characteristics.

Overall, the consistent gains observed across quantitative metrics, per-class performance, and qualitative visualizations confirm BDNet's strong cross-dataset generalization capability. Despite being trained solely on BRVD, BDNet transfers effectively to the aerial VisDrone domain while retaining high efficiency. These findings demonstrate that the proposed modular architecture learns transferable and context-aware representations under substantial ground-to-aerial domain shift, supporting its applicability to heterogeneous traffic sensing scenarios when deployment conditions are appropriately considered.

4.8. Discussion

The results demonstrate that BDNet improves the accuracy–efficiency balance of lightweight vehicle detection for dense urban traffic while preserving real-time feasibility across both ground-level and aerial viewpoints. Unlike approaches that rely on modifying the loss or training objective, BDNet achieves these gains through architecture-level refinements that address three practical bottlenecks in smart-city traffic monitoring: (i) detail loss during downsampling, (ii) insufficient contextual reasoning under occlusion, and (iii) redundancy accumulation during multi-scale fusion.

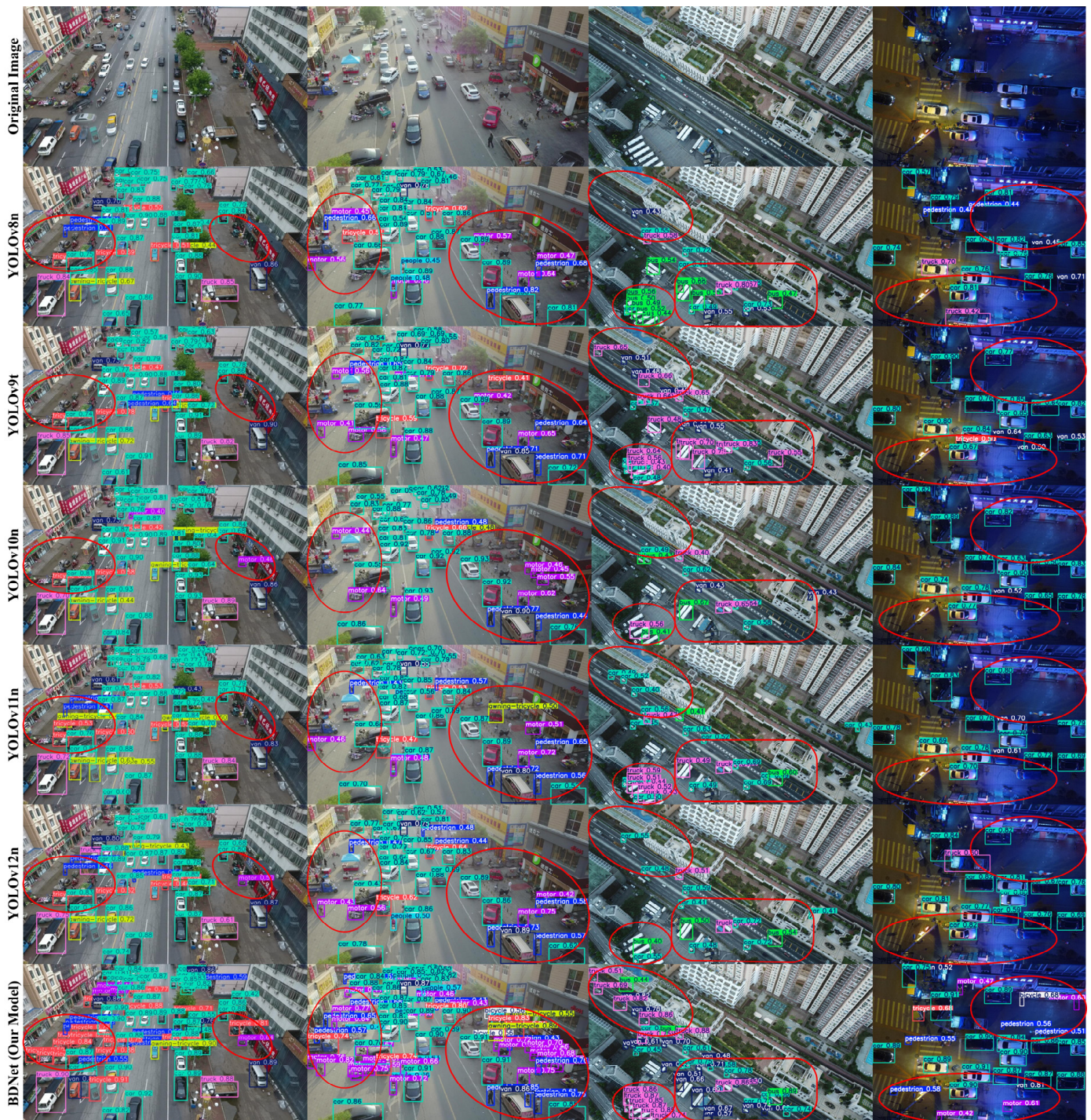


Figure 16. Qualitative detection comparisons on the VisDrone-DET2019 validation set under diverse aerial conditions (daytime, fog/haze, urban top-view, and nighttime). Rows depict YOLOv8n, YOLOv9t, YOLOv10n, YOLOv11n, YOLOv12n, and BDNet. Red ellipses highlight regions where BDNet recovers missed small objects, suppresses false positives, and provides tighter localization.

Urban traffic performance and practical implications: On BRVD, BDNet achieves 88.4% precision, 75.3% recall, and an $\mathcal{F}_1$ -score of 81.3%, with 85.9% mAP_{50} and 67.3% mAP_{50-95} (Tables 2 and 3). Relative to YOLOv12n, BDNet improves recall by +1.5 pp, $\mathcal{F}_1$ -score by +3.3 pp, and mAP_{50} by +1.4 pp, while remaining compact (2.5 M parameters, 6.0 GFLOPs) and faster under the same evaluation setting (285.7 FPS vs. 196.1 FPS) (Table 3). These improvements are particularly relevant for smart-city pipelines because missed detections

can bias vehicle counts and downstream congestion analytics. In operational deployments, such biases can directly affect traffic management center decisions, including adaptive signal timing, congestion mitigation strategies, and safety monitoring dashboards that rely on continuous and accurate perception inputs.

Why the modules work together: The ablation results confirm complementary module contributions (Table 6). HyDASE yields the strongest single-module gain while reducing computation, supporting its role in preserving fine cues during resolution reduction. C3k2_MogaBlock strengthens contextual aggregation with modest overhead, consistent with multi-order gated reasoning, improving robustness under occlusion. A2C2f_FRFN improves feature selectivity by suppressing redundant activations, aligning with refinement-style feed-forward designs. The full integration produces the best trade-off ($mAP_{50} = 85.9\%$, recall = 75.3%, 285.7 FPS), indicating that the modules jointly improve both representation quality and inference efficiency.

Evidence beyond a single metric: Qualitative diagnostics support the quantitative trends. PR curves show BDNet maintaining stronger precision across recall levels (Figures 9 and 10), while visual comparisons indicate fewer background false positives and better recovery of small or occluded vehicles (Figure 11). Grad-CAM and feature progression visualizations suggest more object-centric attention and more structured intermediate features, which correspond to cleaner localization and reduced inter-class confusion (Figures 12–14).

Failure modes and deployment considerations: Despite the overall gains, several error patterns remain relevant for operational use. BDNet can still miss extremely small or distant vehicles in high-density scenes, particularly when objects occupy only a few pixels after multi-stage downsampling; this is most evident for long-range motorcycles and bicycles, as well as partially occluded instances. Confusion also occurs for visually similar categories under occlusion (e.g., pickup vs. covered van, and CNG vs. auto-rickshaw), where shared geometry and truncation reduce discriminative cues. In severe congestion, overlapping instances occasionally lead to duplicate or slightly misaligned boxes, reflecting the difficulty of suppressing near-adjacent detections under class-agnostic post-processing. Low-illumination and low-contrast conditions (night glare, fog) can further destabilize localization for small non-motorized vehicles. From a deployment perspective, these observations clarify the trade-offs between compactness and robustness: BDNet prioritizes real-time feasibility, while the most adverse regimes may benefit from operational tuning (class-aware NMS, adaptive confidence thresholds), stronger low-light augmentation, and region-specific fine-tuning when cameras and traffic patterns differ across cities.

Scalability and cross-domain robustness: BDNet's advantages remain consistent across model sizes: BDNet variants improve stricter IoU metrics while reducing parameters and GFLOPs relative to YOLOv12 at each scale (Table 7), supporting deployment flexibility from edge nodes to centralized ITS platforms. Under the domain shift from BRVD to VisDrone-DET2019, BDNet achieves the highest overall performance among YOLO base-lines ($mAP_{50} = 31.9\%$, precision = 43.6%, F1 = 36.6%) while remaining efficient (104.2 FPS, 2.5 M parameters, 6.0 GFLOPs) (Table 8). This indicates improved transferability to aerial viewpoints dominated by small objects and dense layouts. These results indicate improved transferability under the evaluated ground-to-aerial domain shift, rather than universal robustness across all sensing configurations and traffic domains.

Limitations

Some visually ambiguous categories (e.g., Leguna) remain challenging on BRVD (Table 2). Although cross-dataset performance on VisDrone-DET2019 improves, the absolute accuracy remains constrained by viewpoint, scale, and annotation-domain differences

(Tables 8 and 9). In addition, throughput is hardware-dependent; future work will report latency/energy on representative edge devices and explore lightweight domain adaptation to better support heterogeneous camera deployments.

Overall, BDNet provides a scalable and efficient vehicle detection foundation for smart urban traffic monitoring, combining improved detection reliability, localization robustness, and real-time throughput under practical computational constraints.

5. Conclusions

This study presented BDNet, a lightweight vehicle detection framework designed for dense, heterogeneous urban traffic and real-time intelligent transportation systems. The proposed architecture integrates three complementary components—HyDASE, C3k2_MogaBlock, and A2C2f_FRFN—to address key limitations of existing lightweight detectors by preserving fine-grained spatial detail, strengthening contextual reasoning under occlusion, and refining multi-scale feature representations. Together, these modules enable BDNet to maintain strong detection accuracy while remaining computationally efficient. Extensive experiments on the BRVD dataset show that BDNet achieves a favorable accuracy–efficiency trade-off, reaching 85.9% mAP_{50} and 67.3% mAP_{50-95} with only 2.5 M parameters and 6.0 GFLOPs, while sustaining 285.7 FPS under identical evaluation settings. Compared with YOLOv12n, BDNet improves mAP_{50} by +1.4 percentage points, recall by +1.5 percentage points, and F1-score by +3.3 percentage points while simultaneously delivering substantially higher inference throughput. These gains are achieved without increasing model complexity, highlighting the effectiveness of the proposed architectural refinements.

Cross-dataset evaluation on VisDrone-DET2019 further confirms BDNet’s robustness and transferability. When trained exclusively on BRVD and evaluated without fine-tuning, BDNet attains 31.9% mAP_{50} and 104.2 FPS, outperforming YOLOv12n in both accuracy and speed under a pronounced domain shift from ground-level to aerial viewpoints. Ablation studies verify that HyDASE primarily improves recall and efficiency, C3k2_MogaBlock enhances precision and contextual robustness in occluded scenes, and A2C2f_FRFN suppresses redundant activations during feature fusion, with the full configuration yielding the best overall performance. Qualitative analyses, including detection visualizations, Grad-CAM heatmaps, feature-map evolution, and confusion matrices, provide further insight into BDNet’s behavior. These analyses reveal more object-centric attention, more structured intermediate representations, and reduced inter-class confusion, supporting the observed quantitative improvements in dense and visually complex traffic environments.

Despite these strengths, challenges remain under extreme illumination, severe weather conditions, and for very small or visually ambiguous objects. Future work will extend BRVD by increasing geographic diversity and underrepresented vehicle categories, as well as exploring broader cross-city and cross-country validation to further assess generalization. Although this study focuses on GPU-based evaluation to ensure reproducibility and fair comparison, future work will include deployment-level benchmarking on representative edge devices (e.g., Jetson-class platforms) to further validate real-world efficiency. Overall, BDNet provides an accurate, efficient, and interpretable vehicle detection solution that aligns well with the practical requirements of smart urban traffic monitoring and next-generation ITS applications.

Author Contributions: Conceptualization, M.M.H. and Z.W.; methodology, M.M.H.; software, M.M.H.; validation, M.M.H., R.S., and T.M.A.H.; formal analysis, M.M.H. and M.A.I.H.; investigation, M.M.H. and Z.W.; resources, M.M.H. and K.F.; data curation, M.M.H. and K.F.; writing—original draft preparation, M.M.H.; writing—review and editing, M.M.H., Z.W., M.A.I.H., K.F.; visualization, R.S., T.M.A.H., M.A.I.H. and K.F.; supervision, Z.W.; project administration, Z.W. and H.F.;

funding acquisition, Z.W. and H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China under Grant No. 2019YFC1521300.

Data Availability Statement: Data will be made available upon request.

Acknowledgments: The authors gratefully acknowledge the support provided by the School of Information and Intelligent Science, Donghua University, Shanghai 201620, PR China, for offering an academic environment and computational facilities that enabled the completion of this research. The authors also thank all contributors involved in the collection and annotation of the Bangladeshi Road Vehicle Dataset (BRVD), as well as colleagues who provided technical assistance and constructive feedback during model development and experimental evaluation.

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

Abbreviations

The following abbreviations are used in this manuscript:

ITS	Intelligent Transportation Systems
BRVD	Bangladeshi Road Vehicle Dataset
YOLO	You Only Look Once
HyDASE	Hybrid Downsampling and Squeeze-Excitation
MOGA	Multi-order gated aggregation
FRFN	Feature Refinement Feed-Forward Network
GFLOPs	Giga Floating Point Operations Per Second
mAP	Mean Average Precision
$\mathcal{P}_{rec}$	Precision
$\mathcal{R}_{rec}$	Recall
IoU	Intersection over Union
FPS	Frames per second
ReLU	Rectified Linear Unit
GAP	Global average pooling
SOTA	State-of-the-Art

References

1. Pan, H.; Guan, S.; Zhao, X. LVD-YOLO: An efficient lightweight vehicle detection model for intelligent transportation systems. *Image Vis. Comput.* **2024**, *151*, 105276. [[CrossRef](#)]
2. Liu, Y.; Huang, Z.; Song, Q.; Bai, K. PV-YOLO: A lightweight pedestrian and vehicle detection model based on improved YOLOv8. *Digit. Signal Process.* **2025**, *156*, 104857. [[CrossRef](#)]
3. Lyu, Y.; Zhang, T.; Li, X.; Liu, A.; Shi, G. LightUAV-YOLO: A lightweight object detection model for unmanned aerial vehicle image. *J. Supercomput.* **2025**, *81*, 105. [[CrossRef](#)]
4. Talaat, F.M.; El-Balka, R.M.; Sweidan, S.; Gamel, S.A.; Al-Zoghby, A.M. Smart traffic management system using YOLOv11 for real-time vehicle detection and dynamic flow optimization in smart cities. *Neural Comput. Appl.* **2025**, *37*, 19957–19974. [[CrossRef](#)]
5. Hussain, K.; Moreira, C.; Pereira, J.; Jardim, S.; Jorge, J. A Comprehensive Literature Review on Modular Approaches to Autonomous Driving: Deep Learning for Road and Racing Scenarios. *Smart Cities* **2025**, *8*, 79. [[CrossRef](#)]
6. Toba, A.-L.; Kulkarni, S.; Khallouli, W.; Pennington, T. Long-Term Traffic Prediction Using Deep Learning Long Short-Term Memory. *Smart Cities* **2025**, *8*, 126. [[CrossRef](#)]
7. Tsalikidis, N.; Mystakidis, A.; Koukaras, P.; Ivaškevičius, M.; Morkūnaitė, L.; Ioannidis, D.; Fokaides, P.A.; Tjortjis, C.; Tzouvaras, D. Urban traffic congestion prediction: A multi-step approach utilizing sensor data and weather information. *Smart Cities* **2024**, *7*, 233–253. [[CrossRef](#)]
8. Doha, B.; Khalid, M.; Zouhair, C.; Noreddine, A.; omri Amina, E. A SPAR-4-SLR Systematic Review of AI-Based Traffic Congestion Detection: Model Performance Across Diverse Data Types. *Smart Cities* **2025**, *8*, 143.
9. Tang, Y.; Qu, A.; Jiang, X.; Mo, B.; Cao, S.; Rodriguez, J.; Koutsopoulos, H.N.; Wu, C.; Zhao, J. Robust Reinforcement Learning Strategies with Evolving Curriculum for Efficient Bus Operations in Smart Cities. *Smart Cities* **2024**, *7*, 3658–3677. [[CrossRef](#)]

10. Mystakidis, A.; Koukaras, P.; Tjortjis, C. Advances in traffic congestion prediction: An overview of emerging techniques and methods. *Smart Cities* **2025**, *8*, 25. [\[CrossRef\]](#)
11. Munir, S.; Lin, H.-Y. LNT-YOLO: A Lightweight Nighttime Traffic Light Detection Model. *Smart Cities* **2025**, *8*, 95. [\[CrossRef\]](#)
12. Dasu, B.T.; Reddy, M.V.; Kumar, K.V.; Chithaluru, P.; Ahmed, N.; Abd Elminaam, D.S. A self-attention driven multi-scale object detection framework for adverse weather in smart cities. *Sci. Rep.* **2026**, *16*, 1992. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Porto, J.; Sampaio, P.; Szemes, P.; Pistori, H.; Menyhart, J. Comparative Analysis of Traffic Detection Using Deep Learning: A Case Study in Debrecen. *Smart Cities* **2025**, *8*, 103. [\[CrossRef\]](#)
14. Yuan, X.; Li, H. LLM-Driven Offloading Decisions for Edge Object Detection in Smart City Deployments. *Smart Cities* **2025**, *8*, 169. [\[CrossRef\]](#)
15. Guo, G.; Qiu, X.; Pan, Z.; Yang, Y.; Xu, L.; Cui, J.; Zhang, D. YOLOv10-DSNet: A Lightweight and Efficient UAV-Based Detection Framework for Real-Time Small Target Monitoring in Smart Cities. *Smart Cities* **2025**, *8*, 158. [\[CrossRef\]](#)
16. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
17. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
18. Tian, Y.; Ye, Q.; Doermann, D. Yolov12: Attention-centric real-time object detectors. *arXiv* **2025**, arXiv:2502.12524.
19. Zhou, W.; Wang, J.; Meng, X.; Wang, J.; Song, Y.; Liu, Z. MP-YOLO: Multidimensional feature fusion based layer adaptive pruning YOLO for dense vehicle object detection algorithm. *J. Vis. Commun. Image Represent.* **2025**, *112*, 104560. [\[CrossRef\]](#)
20. Song, B.; Chen, J.; Liu, W.; Fang, J.; Xue, Y.; Liu, X. YOLO-ELWNet: A lightweight object detection network. *Neurocomputing* **2025**, *636*, 129904. [\[CrossRef\]](#)
21. Wang, Z.; Zhang, K.; Wu, F.; Lv, H. YOLO-PEL: The efficient and lightweight vehicle detection method based on YOLO algorithm. *Sensors* **2025**, *25*, 1959. [\[CrossRef\]](#)
22. Fan, Q.; Li, Y.; Deveci, M.; Zhong, K.; Kadry, S. LUD-YOLO: A novel lightweight object detection network for unmanned aerial vehicle. *Inf. Sci.* **2025**, *686*, 121366. [\[CrossRef\]](#)
23. Bakirci, M. Internet of Things-enabled unmanned aerial vehicles for real-time traffic mobility analysis in smart cities. *Comput. Electr. Eng.* **2025**, *123*, 110313. [\[CrossRef\]](#)
24. Tahir, N.U.A.; Kuang, L.; Sinishaw, M.L.; Asim, M. PV3M-YOLO: A triple attention-enhanced model for detecting pedestrians and vehicles in UAV-enabled smart transport networks. *J. Vis. Commun. Image Represent.* **2026**, *115*, 104701. [\[CrossRef\]](#)
25. Wang, H.; Liu, J.; Zhao, J.; Zhang, J.; Zhao, D. Precision and speed: LSOD-YOLO for lightweight small object detection. *Expert Syst. Appl.* **2025**, *269*, 126440. [\[CrossRef\]](#)
26. Zhou, S.; Chen, D.; Pan, J.; Shi, J.; Yang, J. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 2952–2963.
27. He, S.; Yu, W.; Tang, T.; Wang, S.; Li, C.; Xu, E. FOS-YOLO: Multiscale Context Aggregation With Attention-Driven Modulation for Efficient Target Detection in Complex Environments. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 1–13. [\[CrossRef\]](#)
28. Bai, C.; Zhang, K.; Jin, H.; Qian, P.; Zhai, R.; Lu, K. SFDEF-YOLO: Small object detection network based on fine-grained feature extraction and fusion for unmanned aerial images. *Image Vis. Comput.* **2025**, *156*, 105469. [\[CrossRef\]](#)
29. Chen, Y.; Wang, Y.; Zou, Z.; Dan, W. GMS-YOLO: A Lightweight Real-Time Object Detection Algorithm for Pedestrians and Vehicles Under Foggy Conditions. *IEEE Internet Things J.* **2025**, *12*, 23879–23890. [\[CrossRef\]](#)
30. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
31. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [\[CrossRef\]](#)
32. Farid, D.M.; Das, P.K.; Islam, M.; Sina, E. Bangladeshi Vehicle Classification and Detection Using Deep Convolutional Neural Networks With Transfer Learning. *IEEE Access* **2025**, *13*, 26429–26455. [\[CrossRef\]](#)
33. Hasan, M.M.; Wang, Z.; Hussain, M.A.I.; Fatima, K. Bangladeshi Native Vehicle Classification Based on Transfer Learning with Deep Convolutional Neural Network. *Sensors* **2021**, *21*, 7545. [\[CrossRef\]](#)
34. Tabassum, S.; Ullah, S.; Al-Nur, N.H.; Shatabda, S. Poribohon-BD: Bangladeshi local vehicle image dataset with annotation for classification. *Data Brief* **2020**, *33*, 106465. [\[CrossRef\]](#)
35. Wang, H.; Liu, C.; Cai, Y.; Chen, L.; Li, Y. YOLOv8-QSD: An improved small object detection algorithm for autonomous vehicles based on YOLOv8. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–16. [\[CrossRef\]](#)
36. Tang, P.; Zhang, Y. LiteFlex-YOLO: A lightweight small target detection network for maritime unmanned aerial vehicles. *Pervasive Mob. Comput.* **2025**, *111*, 102064. [\[CrossRef\]](#)
37. Ma, S.; Lu, H.; Liu, J.; Zhu, Y.; Sang, P. Layn: Lightweight multi-scale attention yolov8 network for small object detection. *IEEE Access* **2024**, *12*, 29294–29307. [\[CrossRef\]](#)

38. Du, D.; Bi, M.; Xie, Y.; Liu, Y.; Qi, G.; Guo, Y. MLE-YOLO: A Lightweight and Robust Vehicle and Pedestrian Detector for Adverse Weather in Autonomous Driving. *Digit. Signal Process.* **2025**, *168*, 105628. [\[CrossRef\]](#)
39. Ju, C.; Chang, Y.; Xie, Y.; Li, D. LS-YOLO: A lightweight, real-time YOLO-based target detection algorithm for autonomous driving under adverse environmental conditions. *IEEE Access* **2025**, *13*, 118147–118162. [\[CrossRef\]](#)
40. Ma, Y.; Xi, C.; Ma, T.; Sun, H.; Lu, H.; Xu, X.; Xu, C. I-YOLOv11n: A Lightweight and Efficient Small Target Detection Framework for UAV Aerial Images. *Sensors* **2025**, *25*, 4857. [\[CrossRef\]](#)
41. Zhuo, S.; Bai, H.; Jiang, L.; Zhou, X.; Duan, X.; Ma, Y.; Zhou, Z. Scl-yolov11: A lightweight object detection network for low-illumination environments. *IEEE Access* **2025**, *13*, 47653–47662. [\[CrossRef\]](#)
42. Li, S.; Wang, Z.; Liu, Z.; Tan, C.; Lin, H.; Wu, D.; Chen, Z.; Zheng, J.; Li, S.Z. Moganet: Multi-order gated aggregation network. *arXiv* **2022**, arXiv:2211.03295.
43. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733.
44. Nguyen, P.T.; Huynh, D.C.; Ho, L.D.; Dunnigan, M.W. YOLO-WTB: Improved YOLOv12n model for detecting small damage of wind turbine blades from aerial imagery. *IEEE Access* **2025**, *13*, 131257–131270. [\[CrossRef\]](#)
45. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
46. Duan, C.; Guo, Y.; Gong, T.; Sheng, B.; Li, X. Reb-YOLO: An accuracy-improved vehicle detection network under low recognized environments. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 1–13. [\[CrossRef\]](#)
47. Yu, Y.; Huang, M.; Wang, K.; Tang, X.; Bao, J.; Fan, Y. LS-YOLO: A lightweight small-object detection framework with region scaling loss and self-attention for intelligent transportation systems. *Signal Image Video Process.* **2025**, *19*, 1005. [\[CrossRef\]](#)
48. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
49. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
50. Tzutalin, D. Labellmg. *GitHub Repos.* **2015**, *6*, 4.
51. Xu, S.; Cui, K. YOLO-EFM: Efficient traffic flow monitoring algorithm with enhanced multi-level information fusion. *Results Eng.* **2025**, *26*, 105545. [\[CrossRef\]](#)
52. Ramos, L.T.; Casas, E.; Romero, C.; Rivas-Echeverría, F.; Bendek, E. A study of yolo architectures for wildfire and smoke detection in ground and aerial imagery. *Results Eng.* **2025**, *26*, 104869. [\[CrossRef\]](#)
53. Vinoth, K.; Sasikumar, P. VINO_EffiFedAV: VINO with efficient federated learning through selective client updates for real-time autonomous vehicle object detection. *Results Eng.* **2025**, *25*, 103700. [\[CrossRef\]](#)
54. Yang, G.; Wang, Y.; Li, X.; Cui, Q.; Luo, T.; Wang, K. YOLOv9t-DM: A lightweight multi-target detection method for walnut shell kernel materials. *Signal Image Video Process.* **2025**, *19*, 591. [\[CrossRef\]](#)
55. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
56. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [\[CrossRef\]](#)
57. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
58. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [\[CrossRef\]](#)
59. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 16965–16974.
60. Giri, K.J. SO-YOLOv8: A novel deep learning-based approach for small object detection with YOLO beyond COCO. *Expert Syst. Appl.* **2025**, *280*, 127447.
61. Li, X.; Zhao, S.; Chen, C.; Cui, H.; Li, D.; Zhao, R. YOLO-FD: An accurate fish disease detection method based on multi-task learning. *Expert Syst. Appl.* **2024**, *258*, 125085. [\[CrossRef\]](#)
62. Chaudhry, R. SD-YOLO-AWDNet: A hybrid approach for smart object detection in challenging weather for self-driving cars. *Expert Syst. Appl.* **2024**, *256*, 124942.
63. Liu, W.; Qiao, X.; Zhao, C.; Deng, T.; Yan, F. VP-YOLO: A human visual perception-inspired robust vehicle-pedestrian detection model for complex traffic scenarios. *Expert Syst. Appl.* **2025**, *274*, 126837. [\[CrossRef\]](#)
64. Xue, C.; Xia, Y.; Wu, M.; Chen, Z.; Cheng, F.; Yun, L. EL-YOLO: An efficient and lightweight low-altitude aerial objects detector for onboard applications. *Expert Syst. Appl.* **2024**, *256*, 124848. [\[CrossRef\]](#)

65. Dong, L.; Zhu, H.; Ren, H.; Lin, T.-Y.; Lin, K.-P. A novel lightweight MT-YOLO detection model for identifying defects in permanent magnet tiles of electric vehicle motors. *Expert Syst. Appl.* **2025**, *288*, 128247. [[CrossRef](#)]
66. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

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