

A Dual-path Lightweight Detector with Hybrid Attention for Real-Time Object Detection

Boyang Yu^a, Zixuan Li^a, Yue Cao^{a,b,*}, Xu Zhang^b, Wansu Lim^c, William Liu^d

^a*School of Cyber Science and Engineering, Wuhan University, 299 Bayi Road, Wuhan, 430072, Hubei, China*

^b*State Key Laboratory of Autonomous Intelligent Unmanned Systems, Beijing, 100081, China*

^c*School of Computing Sciences, University of East Anglia, UK*

^d*Sungkyunkwan University, South Korea*

^e*Unitec and Manukau Institute of Technology, New Zealand*

Abstract

Unmanned aerial vehicle (UAV)-based object detection presents significant challenges, including pronounced variations in object scale and limited computational resources. To address these issues, this paper proposes the Dual-path and Bimodal-attention-enhanced Network (DBYNet), a real-time detection framework optimized for UAV applications. DBYNet adopts a deployment-oriented design that integrates a dual-path backbone for spatial-semantic feature decoupling, a hybrid attention mechanism for enhanced contextual modeling, and lightweight optimization strategies to improve inference efficiency. Specifically, a shallow lighted branch preserves fine-grained spatial details, while a deep branch with deformable convolutions captures high-level semantic features, and the proposed hybrid attention combines Overlapping Cross Attention (OCA) and Channel-spatial Bimodal Attention (CAB) to strengthen feature interaction. In addition, Quantization-Aware Training and temperature-aware distillation are employed to reduce model complexity without compromising accuracy. Extensive experiments on the VisDrone

*Corresponding author. Email: yue.cao@whu.edu.cn

Email addresses: 2022102210001@whu.edu.cn (Boyang Yu), 2023282210230@whu.edu.cn (Zixuan Li), yue.cao@whu.edu.cn (Yue Cao), X.Zhang27@uea.ac.uk (Xu Zhang), wansu.lim@kumoh.ac.kr (Wansu Lim), wliu@unitec.ac.nz (William Liu)

2019 dataset (VisDrone2019) dataset demonstrate that DBYNet achieves a favorable accuracy–efficiency trade-off, particularly improving robustness for small, densely distributed, and low-visibility targets in challenging UAV scenarios.

Keywords: Object detection, Attention mechanism, Lightweight design, Unmanned aerial vehicle imagery.

1. Introduction

UAVs have been increasingly deployed in diverse applications such as environmental monitoring, smart agriculture, emergency response, and urban surveillance (Wang et al., 2025; Wu et al., 2024). In these scenarios, UAVs provide flexible aerial viewpoints and rapid situational awareness, enabling tasks such as large-area inspection, disaster assessment, and traffic monitoring that are difficult to achieve with fixed sensors. The effectiveness of these applications largely depends on reliable onboard visual perception, where object detection serves as a fundamental component for downstream decision-making. Their mobility and ability to operate at low altitudes make them particularly suitable for real-time visual perception tasks. Among these tasks, object detection plays a central role by supporting downstream modules including target tracking, geolocation, and path planning. However, UAV-based object detection remains challenging due to severe object scale variation, complexity of scenario, and limited computational resources on embedded platforms (Niu & Chen, 2023; Lin & Zhang, 2022).

To address these issues, various detection frameworks based on Convolutional Neural Networks (CNNs) and Transformer architectures have been proposed. Single-stage detectors such as You Only Look Once version 5 (YOLOv5) (Bochkovskiy et al., 2020) and YOLOX (Ge et al., 2021) are widely adopted for their fast inference, while anchor-free designs like YOLOv8 (Jocher et al., 2023) offer improved localization and simplified label assignment. Transformer-based methods, including Deformable Detection Transformer (Deformable DETR) (Zhu et al., 2020) and Shifted Window Transformer (Swin Transformer) (Liu et al., 2021b), introduce global receptive fields to enhance object-context relationships. Although these approaches have advanced the field of object detection, they still show limitations under UAV-specific scenarios.

In parallel, multiple techniques have been integrated to improve robust-

ness against UAV-specific challenges. Methods such as multi-scale feature fusion (Lin et al., 2017), channel and spatial attention (Woo et al., 2018), and lightweight backbone networks (Tan & Le, 2019; Han et al., 2020) aim to balance accuracy and efficiency. For example, GhostNet (Han et al., 2020) reduces computation cost through efficient operations, while the Convolutional Block Attention Module (CBAM) (Woo et al., 2018) enhances salient feature modeling. Feature pyramid structures help preserve scale-aware representations, while attention mechanisms help suppress background noise and enhance salient object features.

Despite these advances, significant challenges remain in aerial detection. Transformer-based models, while capable of global reasoning, often suffer from high computational complexity and cannot meet real-time requirements. Conversely, lightweight CNNs offer faster inference but lack semantic depth, leading to reduced robustness under occlusion and clutter (Wang et al., 2021; Xu & Zhang, 2023). Additionally, many models fail to effectively integrate shallow and deep features, which limits performance in scenarios involving small or densely distributed targets. These issues highlight the need for a unified approach that enhances spatial precision, semantic representation, and inference efficiency simultaneously.

While Transformer-based methods offer powerful global modeling, their high computational cost hinders real-time deployment. In contrast, YOLO-style detectors achieve a better speed-accuracy trade-off, making them more suitable for UAV applications requiring low-latency inference. To address these limitations, this paper proposes the Dual-path and Bimodal-attention-enhanced YOLO Network (DBYNet), a real-time UAV object detection framework optimized for aerial environments.

The remainder of this paper is organized as follows. Section 2 reviews related work on UAV-based object detection and lightweight detection frameworks. Section 3 describes the proposed DBYNet architecture, including the dual-path backbone, hybrid attention mechanism, and lightweight optimization strategies. Section 4 presents the experimental setup and results on the VisDrone2019 dataset, together with ablation studies and qualitative analysis. Finally, Section 5 concludes the paper.

Its main contributions are summarized as follows:

- Existing YOLO backbones mainly rely on single-path architectures that favor semantic depth while neglecting spatial detail, which limits their performance on small or occluded objects in UAV imagery. To ad-

dress this issue, we propose a dual-path backbone that integrates a lightweight GhostNet branch for fine-grained spatial preservation and a deep CSPDarknet branch with deformable convolutions for enhanced semantic representation.

- Most attention mechanisms in YOLO detectors emphasize either spatial or channel refinement, making it difficult to jointly capture long-range context and fine-grained feature interactions. We introduce a hybrid attention design that combines Overlapping Cross Attention and Channel-Spatial Bimodal Attention to improve feature discrimination in complex UAV scenes.
- Although lightweight YOLO variants reduce model size, their deployment efficiency on embedded UAV platforms remains limited. We develop a multi-stage compression pipeline that combines structured pruning and quantization-aware training to achieve a favorable accuracy–efficiency trade-off for real-time UAV-based object detection.

2. Related Work

2.1. Small Object Detection in UAV Imagery

Small object detection in UAV imagery (Safari et al., 2023; Jandaghi et al., 2025; Ahmari et al., 2025; Akherati et al., 2025; Mahjourian & Nguyen, 2025) remains challenging due to extreme scale variation and occlusion-heavy, visually cluttered environments. Small Objects typically occupy fewer pixels and appear indistinguishable from surrounding textures, thus complicating feature extraction (Liu et al., 2020; Du et al., 2018). Early Convolutional Neural Network (CNN)-based approaches, including Feature Pyramid Networks (FPN) (Lin et al., 2017) and Path Aggregation Networks (PANet) (Liu et al., 2018), enhanced multi-scale features to detect small objects. However, these methods demonstrated limited performance due to insufficient shallow-layer representations (Zhu et al., 2018; Li et al., 2021).

Recent studies have focused specifically on small object detection in aerial scenarios. Selective Anchor Matching Detection Transformer (SAM-DETR) (Li et al., 2023) introduced selective anchor matching to accurately localize small targets, while Task-aligned One-stage Object Detection (TOOD) (Feng et al., 2021) improved feature alignment between classification and regression. You Only Look Once version 7 Tiny (YOLOv7-Tiny) (Wang et al., 2023a)

and YOLOv8-Nano (Jocher et al., 2023), as two lightweight YOLO-based detectors, optimized early-layer features for real-time inference but struggled in occlusion-heavy scenes.

Transformer-based architectures have further advanced global context modeling capabilities. Models such as Deformable Detection Transformer (Deformable DETR) (Zhu et al., 2020), DeNoising DETR (DINO) (Zhang et al., 2022b), and Video Detection Transformer (ViDT) (Song et al., 2022) improved detection accuracy by modeling long-range interactions, but still suffers from high complexity and latency (Xu & Zhang, 2023). Therefore, a balanced framework that simultaneously considers efficiency and robustness under aerial constraints is still needed.

2.2. *Lightweight Object Detection for Embedded UAV Platforms*

Real-time UAV applications necessitate object detectors that are both lightweight and accurate, as onboard resources are severely constrained. Traditional CNN-based backbones, such as Residual Networks (ResNet) (He et al., 2016), offer strong semantic capabilities but are computationally demanding for embedded systems (Lin & Zhang, 2022). To reduce computational overhead, lightweight CNNs including MobileNet version 3 (MobileNetV3) (Howard et al., 2019), GhostNet (Han et al., 2020), and ShuffleNet version 2 (ShuffleNetV2) (Ma et al., 2018) have been widely adopted.

Neural Architecture Search (NAS) techniques further optimized architectures specifically for embedded deployments. FBNet version 2 (FBNetV2) (Wan et al., 2020), Single Path One-Shot (SPOS) (Guo et al., 2020), and MobileDets (Xiong et al., 2021) automatically generated lightweight models tailored for low latency inference. Additionally, compact detection heads such as PaddlePaddle You Only Look Once Enhanced (PP-YOLOE) (Xu et al., 2022) streamlined the neck and head structure for embedded application.

Beyond architectural innovations, model compression methods such as structured channel pruning (Liu et al., 2017) and Quantization-Aware Training (QAT) (Jacob et al., 2018) have been extensively utilized. In addition, adaptive frameworks like Dynamic Vision Sensing Enhanced (DyViSE) (Gong et al., 2023), which adjust computational paths based on input complexity, offer further flexibility for resource-aware inference. Despite these advancements, achieving robust detection across diverse UAV conditions without sacrificing inference speed remains difficult.

In addition to architectural innovations, model compression methods such as structured pruning and quantization-aware training (QAT) have been

widely adopted. In addition, adaptive frameworks adjust computational paths based on input complexity, offer further flexibility for resource-aware inference.

2.3. Attention Mechanisms for Feature Enhancement

Attention mechanisms have greatly improved feature representation and noise suppression capabilities in object detection tasks. Early methods, such as Squeeze-and-Excitation Networks (SENet) (Hu et al., 2018), enhanced channel-wise representations via global context modeling. Convolutional Block Attention Module (CBAM) (Woo et al., 2018) and Efficient Channel Attention (ECA) (Wang et al., 2020) extended attention mechanisms by jointly modeling spatial and channel information.

Recent developments in Transformer-based self-attention, such as Swin Transformer (Liu et al., 2021b), Deformable DETR (Zhu et al., 2020), and Focal Transformer (Yang et al., 2021a), improved object-context alignment through flexible long-range dependency modeling. However, their high computational complexity limits the application of real-time UAV object detection.

Lightweight modules, such as Lite Squeeze-and-Excitation (LiteSE) (Fan et al., 2021), Simple Attention Module (SimAM) (Yang et al., 2021b), and Normalization-based Attention Module (NAM) (Liu et al., 2021a), significantly reduced attention overhead while preserving accuracy. UAV-specific attention approaches, including Lightweight Aerial attention Unit (LAU) (Wang et al., 2023b) and Strip-based Transformer (StripFormer) (Zhang et al., 2022a), leveraged spatial structure and context priors to enhance detection robustness. Despite these advances, integrating these attention modules into a comprehensive lightweight framework suitable for real-time UAV operation remains a critical research gap (Li et al., 2022).

3. Method

3.1. Overall Architecture

DBYNet aims to provide accurate and efficient inference for UAV-based applications. As illustrated in Fig. 1, the architecture integrates three key components. The dual-path backbone enhances complementary feature extraction, while the hybrid attention mechanism improves context modeling. Furthermore, a lightweight optimization strategy reduces computational cost without compromising detection robustness.

The input image is processed through two parallel branches: a shallow path that preserves high-resolution spatial details, and a deep path composed of CSP modules and deformable convolutions. The latter captures high-level semantic features and enhances spatial adaptability. Subsequently, these features are fused through a hybrid attention mechanism, which integrates OCA and CAB to reinforce spatial relevance and improve channel-wise discrimination. The fused features are passed to a multi-scale detection head, which produces bounding box predictions and classification outputs at multiple resolutions.

To support deployment on resource-constrained UAV platforms, the framework applies model compression techniques during training. In particular, structured pruning and QAT are utilized to substantially reduce parameter size and memory usage while maintaining detection performance.

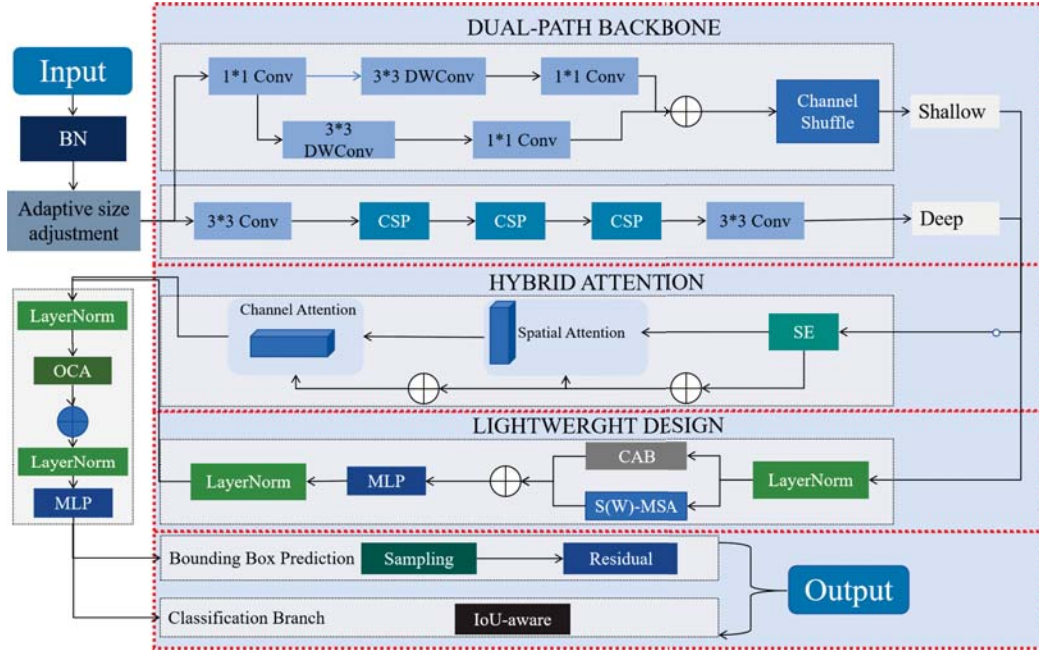


Fig. 1. Overall architecture of the proposed dual-path lightweight UAV object detection network. The design includes a shallow-deep backbone, hybrid attention fusion, multi-scale detection, and lightweight optimization.

This integrated design facilitates real-time inference with high precision across UAV-specific scenarios, such as small object detection, dense target layouts, and cluttered environments. Each module in the architecture ad-

addresses a specific limitation of prior work and contributes to a balanced trade-off between accuracy, efficiency, and model size.

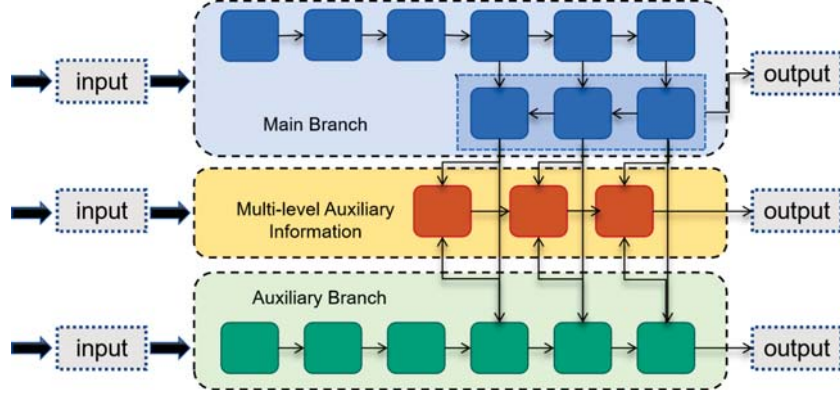


Fig. 2. Architecture of the dual-path feature extraction module. The shallow path preserves spatial resolution for fine details, while the deep path enhances semantic abstraction.

3.2. Dual-Path Feature Extraction

To address the challenges of small-scale target detection and semantic diversity in UAV imagery, the proposed framework incorporates a dual-path feature extraction strategy. This module includes a shallow path and a deep path that operate in parallel, each focusing on complementary aspects of feature representation. The architecture of the dual-path module is depicted in Fig. 2. The combination of GhostNet and CSPDarknet is selected to explicitly decouple spatial detail preservation from high-level semantic abstraction, enabling a balanced dual-path design that is difficult to achieve with a single lightweight or single deep backbone alone.

3.2.1. High-Resolution Feature Preservation

The shallow path is designed to retain local details that are typically lost in deeper layers. It leverages Ghost convolutional modules (Han et al., 2020), which generate redundant features through computationally inexpensive operations. This structure minimizes the computational burden while maintaining fine-grained spatial patterns.

The input feature map is denoted as $X \in \mathbb{R}^{C \times H \times W}$, where C , H , and W represent the number of channels, height, and width, respectively. The

shallow path output X_s is calculated as follows:

$$X_s = \text{GhostConv}(X), \quad (1)$$

where $\text{GhostConv}(\cdot)$ represents the combination of standard and cheap linear operations designed to simulate full convolutional responses. This design effectively captures local details necessary for accurate small object detection.

3.2.2. Semantic Feature Abstraction

The deep path focuses on capturing high-level semantic features and large-scale contextual information. It integrates a CSPDarknet-inspired backbone and applies deformable convolutions (Dai et al., 2017) to enhance spatial adaptability. This combination enables flexible receptive fields and supports robust feature learning under geometric distortion and viewpoint variations.

The semantic feature output X_d is computed as:

$$X_d = \text{DeformConv}(\text{CSPBlock}(X)), \quad (2)$$

where $\text{CSPBlock}(\cdot)$ facilitates gradient propagation by reducing redundancy, while $\text{DeformConv}(\cdot)$ introduces learned offsets for dynamic spatial sampling. This design improves semantic abstraction while maintaining robustness in complex UAV environments.

3.2.3. Feature Fusion

The outputs from the shallow and deep paths are fused to integrate spatial detail with semantic richness. After channel-wise concatenation, a 1×1 convolution is applied to unify the feature dimensions and enable interaction between the paths. The fused feature F is calculated as:

$$F = \text{Conv}_{1 \times 1}(\text{Concat}(X_s, X_d)), \quad (3)$$

where $\text{Concat}(\cdot)$ denotes the channel-wise stacking operation, and $\text{Conv}_{1 \times 1}(\cdot)$ refers to a pointwise convolution used to adjust channel dimensions and facilitate cross-path feature interaction.

This dual-path design provides a balanced solution that combines the strengths of both branches. The shallow path contributes precise localization cues, while the deep path offers enriched semantic context. Together, they create a robust foundation for downstream attention mechanisms and multi-scale detection, which are critical for handling UAV-specific challenges such as small objects, dense distributions, and complex backgrounds.

3.3. Hybrid Attention Mechanism

Accurate detection in UAV scenarios requires not only capturing global semantic information but also retaining fine-grained spatial details. Conventional attention mechanisms often fall short in this regard. Channel attention lacks spatial modeling capability, while spatial attention generally operates within fixed local regions, limiting its ability to model long-range dependencies. To address these issues, a hybrid attention mechanism is introduced, which combines Overlapping Cross Attention (OCA), Channel-Spatial Bimodal Attention (CAB), and Variable Window Multi-Head Attention (VW-MHA). Together, these components provide complementary enhancements in spatial and channel dimensions while maintaining adaptability across diverse object scales.

3.3.1. Overlapping Cross Attention (OCA)

Existing spatial attention mechanisms typically rely on non-overlapping windows, which introduce discontinuities when objects span across region boundaries. To address this limitation, the OCA module introduces overlapping windows with direction-aware filtering, enabling seamless spatial feature aggregation across adjacent regions. shown as Fig. 3.

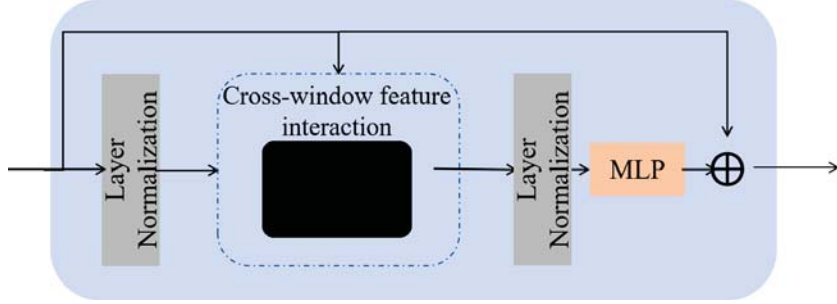


Fig. 3. Architecture of the Overlapping Cross Attention (OCA) module. Overlapping windows and direction-aware convolutions are introduced to improve spatial context modeling.

The input feature map, denoted as $F \in \mathbb{R}^{C \times H \times W}$, undergoes directional depthwise convolutions to extract orientation-specific responses. The horizontal and vertical feature maps, denoted by F_h and F_v , are computed as follows:

$$F_h = \text{Conv}_h(F), \quad F_v = \text{Conv}_v(F), \quad (4)$$

where $\text{Conv}_h(\cdot)$ and $\text{Conv}_v(\cdot)$ are convolutions with $1 \times k$ and $k \times 1$ kernels, respectively.

Directional responses are aggregated to form spatial attention:

$$A_s = \sigma(\text{BN}(F_h + F_v)), \quad (5)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function and BN represents batch normalization.

The spatially attended feature map is then computed as:

$$F_{\text{OCA}} = A_s \odot F, \quad (6)$$

where $\odot$ indicates element-wise multiplication. This process enhances spatial continuity while maintaining global context.

3.3.2. Channel-Spatial Bimodal Attention (CAB)

While OCA focuses on spatial continuity, it does not model channel-wise feature importance. CAB module integrates channel and spatial attention in a unified framework, selectively emphasizing informative features across both dimensions, shown as Fig. 4.

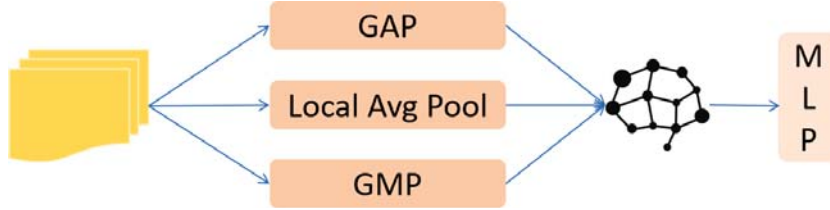


Fig. 4. Architecture of the Channel-Spatial Bimodal Attention (CAB) module. It integrates global channel and spatial pooling to refine feature responses.

Channel attention is computed by summarizing global spatial information using global average pooling and global max pooling, denoted respectively as:

$$f_{\text{avg}} = \text{GAP}(F), \quad f_{\text{max}} = \text{GMP}(F), \quad (7)$$

These descriptors are processed through a shared multi-layer perceptron (MLP):

$$A_c = \sigma(\text{MLP}(f_{\text{avg}}) + \text{MLP}(f_{\text{max}})). \quad (8)$$

For spatial attention, channel-compressed feature maps are obtained using average pooling and max pooling across the channel dimension, denoted

as $\text{AvgPool}(F)$ and $\text{MaxPool}(F)$, respectively. These maps are concatenated and passed through a convolution operation:

$$A_p = \sigma(\text{Conv}_{7 \times 7}([\text{AvgPool}(F); \text{MaxPool}(F)])), \quad (9)$$

The combined attention-enhanced feature map is computed as:

$$F_{\text{CAB}} = (A_c \odot F) + (A_p \odot F), \quad (10)$$

This approach strengthens both global semantic understanding and local feature discrimination.

3.3.3. Variable Window Multi-Head Attention (VW-MHA)

Although OCA and CAB provide effective spatial and channel enhancements, their receptive fields remain relatively fixed. UAV images, however, often contain objects of widely varying scales and distributions. To address this challenge, the VW-MHA module introduces variable window partitioning, which enables adaptive context modeling based on object size and complex backgrounds. Its structure is depicted in Fig. 5. VW-MHA restricts attention computation within variable-sized local windows, which significantly reduces computational overhead while preserving essential contextual information.

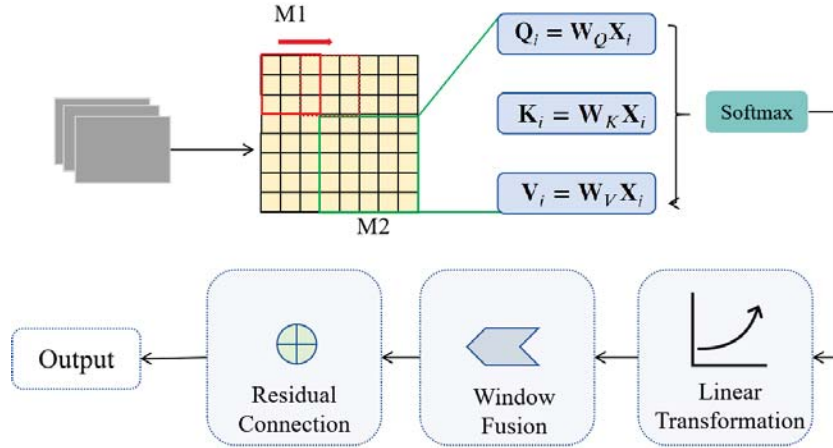


Fig. 5. Architecture of the Variable Window Multi-Head Attention (VW-MHA) module. Variable window partitioning and multi-head processing enable flexible spatial context modeling.

To address this challenge, the VW-MHA module introduces variable window partitioning, which enables adaptive context modeling based on object size and scene complexity by using multi-scale windows to simultaneously capture local and global dependencies. Windows of different sizes are used to process local and global regions simultaneously. Within each window, multi-head attention is performed:

$$O_i = \text{Softmax} \left(\frac{Q_i K_i^T}{\sqrt{d_k}} \right) V_i, \quad (11)$$

where Q_i , K_i , and V_i denote the query, key, and value matrices for the i -th window, respectively, and d_k is the key dimension.

The outputs from all windows are concatenated and combined with the input through residual connection:

$$F_{\text{VW-MHA}} = \text{Concat}(O_1, O_2, \dots, O_n) + F, \quad (12)$$

Unlike rigid window-based methods, the proposed VW-MHA adaptively allocates receptive fields, enabling scale-aware context modeling. This flexibility is especially beneficial for UAV-based detection, where objects appear at diverse scales and positions across the scene. This window-restricted design enables a practical balance between contextual modeling capability and computational efficiency, making VW-MHA suitable for real-time UAV deployment.

3.3.4. Integration Strategy

The final output of the hybrid attention mechanism is formed by aggregating the outputs of OCA and CAB:

$$F_{\text{attn}} = F_{\text{OCA}} + F_{\text{CAB}}, \quad (13)$$

where F_{OCA} focuses on structured spatial information, and F_{CAB} emphasizes salient channels and spatial regions. This hybrid mechanism improves the representation capability and robustness, which is particularly important in UAV-based object detection scenarios.

As illustrated in Fig. 6, the OCA and CAB modules are jointly integrated to enhance feature quality before being fed into the detection head.

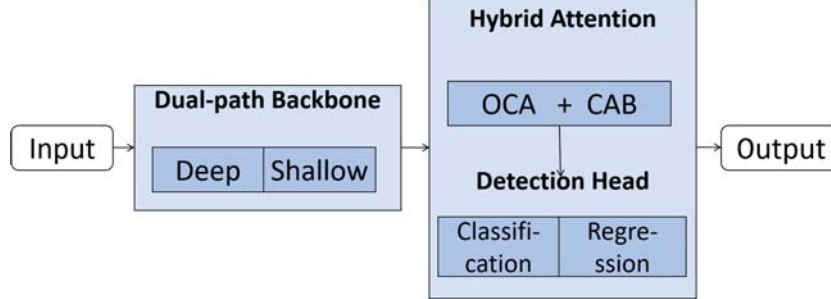


Fig. 6. Integration of the hybrid attention mechanism in DBYNet.

3.4. Lightweight Optimization Strategies

The efficient deployment on UAV platforms requires object detection models to meet stringent constraints regarding latency, memory, and power consumption. While deep networks with large capacities yield high accuracy, their computational costs often render them unsuitable for real-time inference on embedded systems. To address this dilemma, the proposed framework integrates a series of lightweight optimization strategies, including structured pruning, quantization-aware training (QAT), and dynamic inference adjustment. These methods jointly preserve detection performance while significantly reducing model complexity.

3.4.1. Motivation for Lightweight Design

UAV platforms impose multiple hardware and task-specific limitations:

- **Limited Compute Power:** Most UAV processors offer fewer than 10 tera-operations per second (TOPS), which limits network depth and width.
- **Memory Bottlenecks:** Available memory is typically restricted to less than 16 GB, preventing the use of models with large feature maps.
- **Real-Time Requirements:** Tasks such as aerial tracking or emergency response require processing speeds exceeding 30 frames per second (FPS) to ensure system responsiveness.

To address these challenges, architectural and training-level optimization methods are essential.

3.4.2. Progressive Pruning Strategy

Although structured pruning can significantly reduce model complexity, directly removing a large portion of channels often results in network degradation and unstable convergence. To mitigate this issue, a Progressive Pruning Strategy (PPS) is adopted. This strategy gradually compresses the model in stages, as illustrated in Fig. 7, thereby avoiding abrupt changes to the network structure.

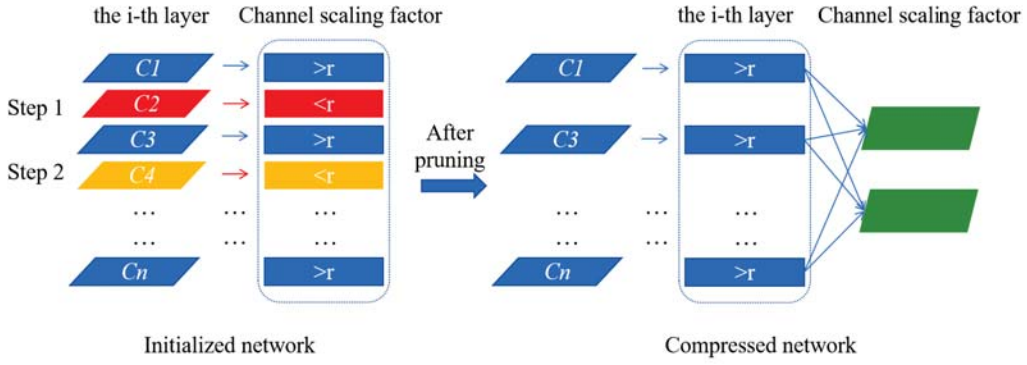


Fig. 7. Illustration of the Progressive Pruning Strategy. Pruning and fine-tuning are alternately performed until the target pruning ratio is achieved.

The pruning process begins conservatively. Initially, only the bottom 20% of channels, ranked by the ℓ_1 -norm of their convolutional weights, are removed. This step prevents sudden structural disruption. A short retraining phase follows, using a decayed learning rate:

$$\eta = \frac{\eta_0}{1 + \lambda t}, \quad (14)$$

where η_0 is the initial learning rate, t represents the epoch index, and λ is the decay factor.

Subsequent pruning follows an iterative process. In each iteration, an additional 10–20% of low-importance channels are removed. The pruning percentage, denoted as ΔP_k , is dynamically adjusted based on loss reduction:

$$\Delta P_k = \beta \cdot \frac{L_{\text{prev}} - L_{\text{curr}}}{L_{\text{prev}}}, \quad (15)$$

where L_{prev} and L_{curr} represent losses before and after pruning, respectively, and β is a scaling factor.

This dynamic adjustment makes pruning aggressive in redundant stages and conservative when performance stabilizes. Finally, a global fine-tuning stage restores and enhances model accuracy. Compared to traditional static pruning, PPS achieves a superior trade-off between compression ratio and detection accuracy.

3.4.3. Quantization-Aware Training

While pruning reduces model size, quantization offers additional benefits by compressing model weights and activations to low-bit formats. To minimize accuracy loss during deployment, quantization-aware training (QAT) simulates low-bit inference during training.

Given a tensor X , quantization is performed using:

$$Q(X) = \text{clip} \left(\text{round} \left(\frac{X}{s} \right) \times s, \min, \max \right), \quad (16)$$

where s is the scale factor, $\text{round}(\cdot)$ rounds to the nearest integer, and $\text{clip}(\cdot)$ limits values to the representable range.

The bit-width configuration for quantization-aware training is summarized in Table 1.

Table 1. Bit-width configuration for quantization-aware training.

Component	Bit-width	Data Type
Weights	8 bits	INT8
Activations	8 bits	INT8
Gradients	16 bits	FP16

Quantization-aware training is adopted to mitigate quantization-induced performance degradation and improve numerical stability during low-precision deployment, which is critical for reliable inference on resource-constrained UAV platforms.

In this work, we adopt channel-level structured pruning followed by quantization aware training as a unified lightweight optimization pipeline. Although alternative compression strategies such as semi-structured pruning or kernel-level norm-based pruning have been explored in prior studies, our objective is not to benchmark different pruning algorithms, but to ensure stable detection performance for small and partially occluded objects under aggressive compression in UAV scenarios. Channel-level structured pruning

provides predictable optimization behavior and good compatibility with existing deployment pipelines without requiring specialized sparse computation kernels. Therefore, we focus on a single, controllable compression strategy and analyze its impact on detection robustness rather than expanding the algorithmic design space.

3.4.4. Dynamic Temperature Distillation

While pruning and quantization reduce model size and computation, knowledge distillation is essential to transfer knowledge from larger models without compromising accuracy. However, conventional distillation methods use fixed temperature parameters, which limit their flexibility for UAV tasks featuring diverse object scales and cluttered backgrounds.

To address this limitation, a Dynamic Temperature Distillation strategy is proposed. As shown in Fig. 8, it integrates temperature annealing, scale-aware adjustment, and channel-aware optimization.

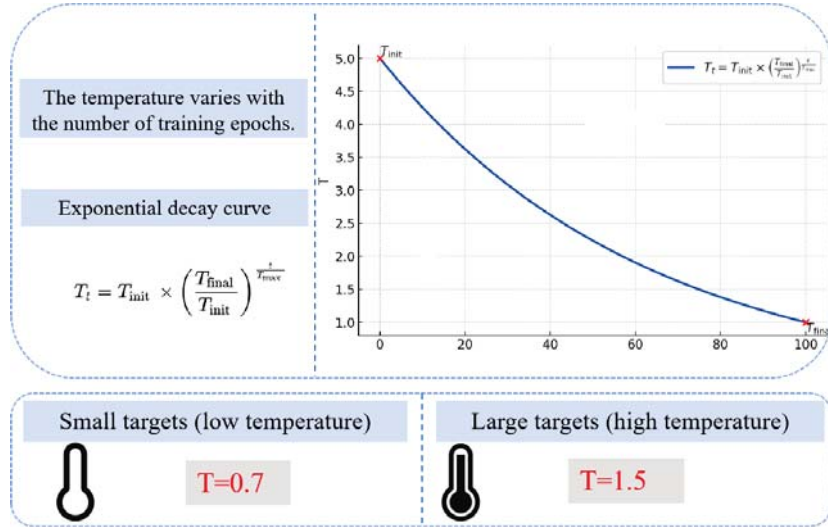


Fig. 8. Architecture of the Dynamic Temperature Distillation. It integrates temperature annealing, scale-aware adjustment, and channel-aware temperature optimization.

During training, the temperature T_t is gradually annealed:

$$T_t = T_{init} \times \left(\frac{T_{final}}{T_{init}}\right)^{\frac{t}{T_{max}}}, \quad (17)$$

where T_{init} and T_{final} are the initial and final temperatures, and t is the training epoch. This approach smooths learning in the early stage and enhances classification sensitivity later.

Considering UAV images contain objects with varying scales, a scale-aware temperature adaptation mechanism is also introduced:

$$T_i = T_{\text{base}} \times \left(1 + \beta \times \frac{A_i}{A_{\text{max}}} \right), \quad (18)$$

where A_i and A_{max} represent the area of the i -th object and the largest object in the batch.

Furthermore, channel-aware optimization is employed to reflect channel importance:

$$T_c = T_{\text{base}} \times \left(1 + \gamma \cdot \frac{A_c}{\sum_{j=1}^C A_j} \right), \quad (19)$$

where A_c is the attention score of channel c , and γ is a scaling factor.

To stabilize learning, temperatures are normalized:

$$T'_c = \frac{T_c - \min(T)}{\max(T) - \min(T)} \times (T_{\text{final}} - T_{\text{init}}) + T_{\text{init}}. \quad (20)$$

By dynamically adjusting the temperature based on scale and channel importance, this method improves knowledge transfer efficiency and enhances detection performance under UAV-specific challenges.

4. Experiments

4.1. Experimental Setup

4.1.1. Hardware and Software Environment

All experiments are conducted on a high-performance computing cluster to ensure stable training and reliable evaluation. As shown in Table 2, the training stage utilized an NVIDIA DGX Station A100 equipped with four NVIDIA A100 Tensor Core GPUs (40 GB HBM2e each), an AMD EPYC 7742 processor (64 cores, 128 threads), and 1 TB DDR4 memory. Mixed precision training (FP16) and Distributed Data Parallel (DDP) are employed to accelerate training while reducing memory consumption. The batch size was set to 16 per GPU, and data synchronization was optimized via NVLINK 3.0 (bandwidth up to 600 GB/s) and All-Reduce operations. The network was trained for 150 epochs, requiring approximately six hours.

Table 2. Experimental parameter settings

Experimental Setting	Parameter Value
Training Device	NVIDIA DGX Station A100 (4×A100, 40GB HBM2e)
Inference Device	RTX 4060
Deep Learning Framework	PyTorch 2.1.0 + CUDA 12.2 + cuDNN 8.9.0
Optimization Toolkit	ONNX 1.14.0
Training Epochs	150
Batch Size	16 (per GPU)
Mixed Precision	AMP
Random Seed	3407 (Fixed)

For reproducibility, Docker containers with locked Python library versions are employed, automatic mixed precision (AMP) was enabled during training, and a fixed random seed (3407) was used.

4.1.2. Dataset

The VisDrone2019 (Zhu et al., 2021) dataset was selected as the primary benchmark to evaluate detection performance in UAV scenarios. Developed by the Machine Perception and Intelligent Systems Laboratory at Tianjin University, this dataset includes complex low-altitude aerial images captured across urban, suburban, and rural environments. Scenes span roads, transportation hubs, farmlands, and construction sites, providing comprehensive coverage of realistic UAV views.

In total, the dataset comprises 10 object categories, including pedestrians, vehicles, and bicycles, all annotated with precise bounding boxes. Compared to other UAV benchmarks such as UAVDT (Du et al., 2018) and DOTA (Ding et al., 2021), VisDrone2019 poses particular challenges. As shown in Fig. 9, its dense object layouts, frequent occlusions, and wide scale variations make it ideal for evaluating detectors in scenarios with small and densely packed targets.

4.1.3. Training and Inference Configuration

The network was trained using stochastic gradient descent (SGD) with momentum 0.9 and weight decay 1×10^{-4} . The initial learning rate was set to 0.01, and a cosine annealing schedule was applied throughout the 150 training epochs. Automatic mixed precision (AMP) was adopted to reduce GPU memory usage and speed up training.

For inference, FP16 quantization was applied on both RTX 4060 and Jetson Orin NX platforms. Batch size was set to 1 during evaluation. On



Fig. 9. Example images from the VisDrone2019 dataset.

Jetson Orin NX, TensorRT inference combined with DeepStream SDK ensured optimized real-time processing. The measured latency and accuracy in later sections reflect deployment-ready performance under realistic UAV hardware constraints.

4.1.4. Evaluation Metrics

To provide a comprehensive assessment of model performance across accuracy, efficiency, and deployment suitability, multiple evaluation metrics are adopted:

- **mAP@50 (%)**: The mean Average Precision calculated at an Intersection-over-Union (IoU) threshold of 0.5. This metric serves as the primary indicator of detection accuracy and reflects the model’s ability to localize and classify objects correctly.
- **FLOPs (G)**: The total number of floating-point operations required for a single forward pass. It measures the computational complexity and indicates the theoretical cost of inference.
- **Parameters (M)**: The total number of learnable parameters within the model. This reflects the model’s compactness, as well as di-

recting implications for storage requirements and memory usage during deployment.

- **FPS**: Frames per second achieved during inference. This metric directly reflects the model’s real-time processing capability, which is particularly critical for UAV-based applications that demand low-latency responses.

In addition to the above quantitative metrics, auxiliary indicators are also employed to analyze detection behavior more thoroughly. These include class-wise average precision (AP), precision-recall (PR) curves, and confusion matrices. Together, they provide deeper insights into per-category detection performance and help identify potential weaknesses such as category imbalance or false positives.

4.2. Ablation Study

Ablation experiments are conducted to systematically evaluate the individual contributions of key modules within the proposed framework. Specifically, the backbone design, attention mechanisms, and loss functions are independently analyzed to clarify their respective impacts on detection performance. All experiments are performed under identical settings on the VisDrone2019 dataset to ensure fair comparison.

4.2.1. Backbone Ablation

The dual-path backbone, which integrates shallow GhostNet and deep CSPDarknet, aims to balance computational efficiency and feature expressiveness. To assess its effectiveness, four backbone configurations are tested, as shown in Table 3.

Table 3. Backbone ablation results.

Group	Backbone	Params (M)	FLOPs (G)	mAP@50 (%)
A	GhostNet only	7.1	18.2	43.8
B	CSPDarknet only	11.2	28.6	39.5
C	Shallow CSPDarknet + Deep GhostNet	9.8	24.1	44.9
D	Shallow GhostNet + Deep CSPDarknet	10.3	23.5	47.7

The results demonstrate that employing GhostNet in shallow layers effectively reduces computational burden, while CSPDarknet in deep layers provides superior semantic abstraction. Notably, the reverse combination

(Group C) led to suboptimal results, highlighting the importance of a carefully designed backbone structure tailored for UAV object detection.

4.2.2. Attention Module Ablation

Attention mechanisms are introduced to enhance feature discrimination, particularly under complex UAV imaging scenarios. This study investigates the contributions of Channel Attention Block (CAB), Overlapping Cross Attention (OCA), and Variable Window Multi-Head Self-Attention (VW-MHA). The corresponding results are presented in Table 4.

Table 4. Attention module ablation results.

Group	CAB	OCA	VW-MHA	mAP@50 (%)
A	-	-	-	39.5
B	✓	-	-	42.5
C	-	✓	-	43.1
D	-	-	✓	46.5
E	✓	✓	✓	47.7

Each attention module independently contributed to performance improvement. In particular, VW-MHA provided the most substantial single gain. When all modules are combined (Group E), the model achieved the highest accuracy, confirming that multi-level attention integration significantly enhances detection robustness and small object recognition capability.

4.2.3. Loss Function Ablation

To validate the design of the loss function, various combinations of Varifocal Loss (VFL), IoU-aware Binary Cross-Entropy Loss (BCE), and GIoU Loss are tested. The results, shown in Table 5, reflect their individual and combined effects on detection accuracy.

Table 5. Loss function ablation results.

Group	Varifocal Loss	IoU-aware BCE Loss	mAP@50 (%)
A	✓	✓	47.7
B	✓	-	45.8
C	-	✓	44.9
D	-	-	39.5

The observations reveal that omitting IoU-aware BCE Loss (Group B) lead to a noticeable performance decline, while removing VFL (Group C)

further degraded accuracy. Without any advanced loss functions (Group D), the detection accuracy dropped substantially. These results demonstrate that the joint use of VFL and IoU-aware BCE Loss effectively balances precision and recall, which is crucial for robust UAV object detection. Varifocal Loss mainly reweights classification samples according to confidence, focusing on foreground–background discrimination, whereas IoU-aware BCE explicitly couples classification targets with localization quality. Their combination jointly improves confidence calibration and localization-aware classification, which is particularly beneficial for dense and small-object UAV scenarios.

4.2.4. Scenario-Specific Evaluation Protocol

To provide a more detailed evaluation under challenging UAV scenarios, we further introduce scenario-specific evaluation protocols on the VisDrone2019 dataset. In aerial imagery, high-altitude views typically result in objects occupying extremely small image regions. Therefore, high-altitude scenarios are approximated by a small-object subset, where the bounding box area of targets is smaller than 32×32 pixels. In addition, robustness to occlusion is evaluated using samples annotated as partially or heavily occluded in the dataset. We also consider a combined hard subset consisting of objects that are both small and occluded. All baseline methods are reproduced under identical training and inference settings to ensure fair comparison.

Table 6. Comparison under challenging UAV scenarios on the VisDrone2019.

Method	Small	Occluded	Small+Occluded
YOLOv8-n	30.4	31.9	26.2
YOLOv8-s	32.8	34.1	28.5
DBYNet (Ours)	36.9	38.2	33.1

Table 6 presents quantitative comparisons under challenging UAV scenarios on the VisDrone2019 dataset. Compared with representative lightweight baseline detectors, DBYNet achieves consistent performance gains on the small-object and occluded subsets. In particular, DBYNet outperforms baseline methods by approximately 3–5 mAP@0.50 under the most difficult combined setting, where targets are both small and partially occluded. These improvements demonstrate the effectiveness of the proposed dual-path backbone in preserving fine-grained spatial details, as well as the hybrid attention

mechanism in enhancing contextual modeling under complex aerial conditions.

4.3. Comparison with State-of-the-art Methods

To comprehensively evaluate the effectiveness of the proposed DBYNet, extensive comparisons are conducted against representative object detectors on the VisDrone2019 dataset. The evaluation focused on accuracy, computational complexity, and inference speed, which are critical for UAV-based deployment scenarios.

4.3.1. Overall Performance Comparison

Table 7 summarizes the performance of various state-of-the-art detectors, including single-stage detectors (e.g., YOLOv5, YOLOv6, YOLOv8), anchor-free methods (e.g., FCOS), Transformer-based models (e.g., DETR, Deformable DETR), and two-stage frameworks (e.g., RetinaNet). These baselines represent a diverse set of detection paradigms widely adopted in recent years.

Table 7. Overall performance comparison with state-of-the-art detectors on VisDrone2019.

Model	Params (M)	FLOPs (G)	mAP@50 (%)	FPS
SSD300	24.0	35.2	26.5	142
RetinaNet	34.0	92.4	31.2	45
FCOS	32.5	78.3	38.3	51
YOLOv5s	7.2	16.5	39.1	105
YOLOv5m	21.2	49.0	44.5	88
YOLOv6s	15.0	34.0	38.5	97
YOLOv7-tiny	6.2	13.7	34.3	118
YOLOv7	36.9	104.7	41.4	71
YOLOv8n	3.2	8.7	32.3	130
YOLOv8s	10.1	23.5	39.5	102
YOLOv8m	25.9	68.2	46.5	85
DETR	41.0	86.0	42.5	28
Deformable DETR	48.5	92.4	44.2	33
TPH-YOLOv5	45.2	93.5	43.7	58
YOLOv11s	12.5	27.8	40.2	98
DBYNet (Ours)	10.3	21.6	47.7	102

The results clearly demonstrate the superiority of DBYNet. It achieves the highest mAP@50 of 47.7%, while maintaining a lightweight design with only 10.3M parameters and 21.6 GFLOPs. Furthermore, DBYNet reaches a real-time inference speed of 102 FPS, satisfying the stringent latency requirements for UAV platforms. Compared to YOLOv8m, DBYNet improves

detection accuracy by 1.2%, while reducing parameter count and computational cost by approximately 60% and 68%, respectively. Additionally, inference speed is increased by nearly 20 FPS.

In comparison to Transformer-based methods such as DETR and Deformable DETR, DBYNet exhibits a substantial advantage in real-time performance. Although these models achieve competitive accuracy, their limited FPS (28 and 33) restricts their suitability for UAV deployment. In contrast, DBYNet offers a balanced trade-off between detection accuracy and inference speed, which is critical in resource-constrained scenarios.

Overall, these results highlight that DBYNet achieves a favorable balance among accuracy, computational efficiency, and real-time processing capability, making it highly applicable to UAV-based object detection tasks.

4.3.2. Category-wise Performance Comparison

To provide a more detailed analysis, category-wise Average Precision (AP) scores are compared across nine object categories, as summarized in Table 8.

Table 8. Category-wise AP comparison on VisDrone2019.

Category	FCOS	YOLOv5s	YOLOv7	YOLOv8s	YOLOv11s	DBYNet (Ours)
Pedestrian	38.2	39.6	45.6	46.1	46.8	47.5
People	37.9	38.4	43.2	44.7	45.3	46.0
Bicycle	35.4	40.2	42.9	43.5	45.9	43.8
Car	50.1	51.3	55.8	56.4	56.9	57.6
Van	47.6	48.7	53.2	54.1	54.5	55.3
Truck	43.9	44.5	49.1	50.2	50.8	51.7
Bus	52.8	54.1	58.7	59.3	59.8	60.5
Motor	40.2	42.7	47.7	47.9	48.2	48.6
Others	28.9	33.4	38.6	39.2	40.5	39.8

As shown, DBYNet achieves the best results in 7 out of 9 categories, demonstrating its strong generalization capability across various object types. Particularly for larger objects such as Car and Bus, the proposed network achieves notable gains, benefiting from its enhanced semantic representation and multi-level attention mechanisms. Although YOLOv11s slightly outperforms DBYNet in the Bicycle and Others categories, this is likely due to its multi-scale feature optimization, which favors very small object detection.

4.4. Lightweight Optimization Analysis

In addition to achieving competitive detection accuracy, UAV object detection models must operate efficiently within limited computational bud-

gets. Our objective is not to benchmark different pruning algorithms, but to ensure stable detection performance for small and partially occluded objects under aggressive compression in UAV scenarios. Channel-level structured pruning provides predictable optimization behavior and good compatibility with existing deployment pipelines without requiring specialized sparse computation kernels.

4.4.1. Structured Pruning and Sparse Training

To reduce model complexity and improve inference efficiency, structured pruning was applied to the backbone and neck modules. A geometric median-based pruning algorithm was adopted to remove redundant channels, while sparse regularization during training encouraged channel-wise sparsity. Pruning was performed incrementally, with pruning ratios ranging from 20% to 60% and pruning steps scheduled every three epochs to ensure stable network adaptation.

Table 9. Performance of structured pruning under different pruning ratios.

Pruning Ratio	Params (M)	FLOPs (G)	mAP@50 (%)	FPS
0%	10.1	23.5	47.7	62
20%	8.7	19.8	46.9	74
40%	7.2	16.3	45.7	87
60%	6.1	13.8	43.1	102

As shown in Table 9, increasing the pruning ratio significantly reduced the model size and computational cost. Although the mAP@50 decreased slightly at higher pruning levels, inference speed improved considerably. Notably, at a 60% pruning ratio, the model achieved real-time speed with only a modest accuracy drop, meeting UAV deployment requirements. Further analysis revealed that shallow GhostNet layers are more sensitive to pruning, especially for small object detection, which prompted the adoption of an asymmetric pruning strategy to preserve key channels.

4.4.2. Quantization-Aware Training and Accuracy Compensation

To further reduce computational costs and enable efficient deployment, quantization-aware training (QAT) was applied to simulate INT8 inference during training. This approach mitigated quantization-induced accuracy loss while significantly improving inference speed. The performance under various quantization schemes is shown in Table 10.

Table 10. Performance comparison of different quantization schemes.

Quantization Mode	mAP@50 (%)	FPS	Model Size (MB)
FP32	47.7	62	40.4
FP16	47.1	113	20.2
INT8 (PTQ)	43.6	141	10.1
INT8 (QAT)	46.2	138	10.1

INT8 quantization achieved significant reductions in model size and latency. However, post-training quantization (PTQ) led to a noticeable accuracy loss. By contrast, QAT successfully compensated for this degradation, achieving a much smaller accuracy drop while maintaining nearly the same inference speed. Furthermore, distribution alignment analysis based on KL divergence confirmed that QAT effectively reduced quantization-induced feature distortions.

4.4.3. Dynamic Temperature-Aware Distillation

To enhance feature expressiveness and improve generalization, dynamic temperature-aware distillation was integrated into the training process. Unlike fixed-temperature schemes, this method adaptively adjusted the temperature based on training dynamics, facilitating better knowledge transfer.

Table 11. Performance comparison of distillation strategies.

Temperature Strategy	mAP@50 (%)	Convergence Epoch
Fixed (T=5)	45.3	132
Cosine Annealing	46.1	118
Dynamic Adjustment (Ours)	47.7	102

As shown in Table 11, the proposed dynamic adjustment strategy not only achieved the best detection accuracy but also accelerated convergence. Cosine similarity analysis further demonstrated that dynamic adjustment maintained higher consistency between teacher and student features, highlighting its effectiveness in improving knowledge transfer.

4.4.4. Lightweight Strategy Combination

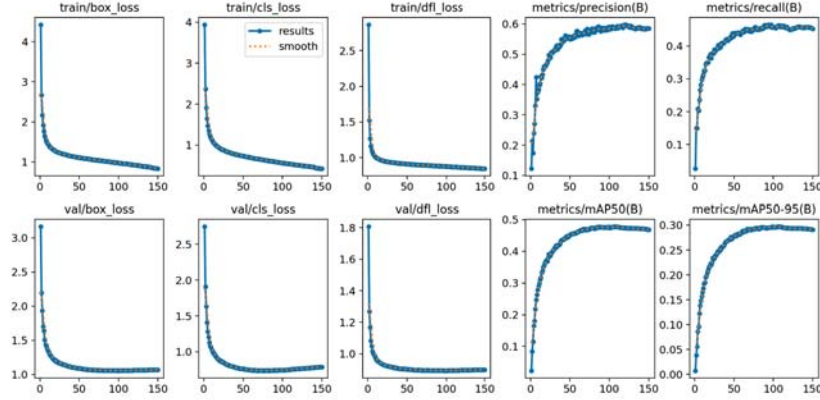
To evaluate the combined effect of multiple lightweight strategies, structured pruning (60% ratio), INT8 QAT, and dynamic temperature-aware distillation are jointly applied. Table 12 summarizes the results.

Table 12. Performance of combined lightweight strategies.

Strategy	Params (M)	FLOPs (G)	mAP@50 (%)	FPS
60% Pruning + INT8 QAT + Dynamic Distillation	6.1	13.8	43.1	102

The results demonstrate that the integrated lightweight optimization strategy achieves an excellent balance between detection accuracy and inference speed. Despite a moderate accuracy reduction, the combination enables real-time processing with substantial reductions in model size and computation, meeting the stringent requirements of UAV-based detection tasks.

4.5. Visualization and Qualitative Analysis

**Fig. 10.** Training loss curves for classification, bounding box regression, and objectness.

To provide a more intuitive and comprehensive evaluation of DBYNet, multiple visualization methods are employed. These analyses aim to reveal the model’s learning behavior, detection precision, and robustness under various challenging scenarios, thereby complementing the quantitative results reported in previous sections.

4.5.1. Training Dynamics and Loss Convergence

The convergence process is critical for achieving stable training and accurate inference. Fig. 10 presents the training loss curves, including classification loss (Cls Loss), bounding box regression loss (Box Loss), and objectness loss (Obj Loss).

As shown in the figure, all losses decreased sharply during the initial epochs and gradually stabilized as training progressed. Notably, Box Loss converged more rapidly, indicating that the model efficiently learned object localization patterns. Compared with the baseline YOLOv8, DBYNet exhibited smoother loss curves in the later stages, which suggests improved training stability and better resistance to gradient fluctuations.

4.5.2. Classification Accuracy and Confusion Analysis

The classification performance was further evaluated using Precision-Recall (P-R) curves and confusion matrices. Fig. 11 illustrates the confusion matrix and P-R curves based on the VisDrone2019 validation set.

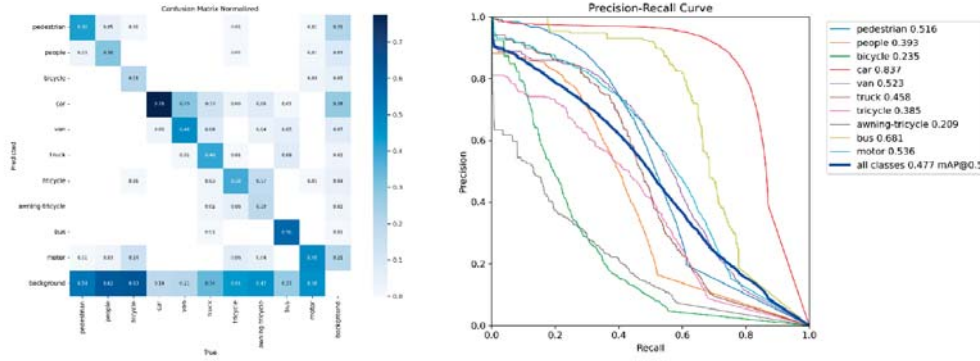


Fig. 11. Confusion matrix (left) and P-R curves (right) of DBYNet on the VisDrone2019 dataset.

The confusion matrix provides insight into the dominant error patterns of DBYNet. Most misclassifications occur between visually similar categories under crowded UAV scenes, such as “pedestrian” and “people,” where severe overlap and limited spatial resolution make fine-grained discrimination challenging. In addition, small objects captured from high-altitude views may further amplify category ambiguity due to extremely limited pixel information. Nevertheless, compared with baseline YOLO variants, DBYNet reduces confusion among these categories, indicating that the dual-path backbone and hybrid attention design enhance feature discrimination in dense and small-object-dominated scenarios.

The P-R curves further show that DBYNet maintains precision above 60% even when recall exceeds 80%, demonstrating its ability to suppress false positives while preserving high recall. This behavior is particularly

important for UAV-based detection, where dense layouts and background clutter often lead to over-detection.

4.5.3. Performance in Extreme Scenarios



Fig. 12. Detection results for small objects captured at high altitudes.

Object detection in UAV imagery is often challenged by extreme conditions, including small target sizes, occlusions, and low-light environments. To assess robustness, qualitative comparisons are performed under these challenging scenarios.

As shown in Fig. 12, in high-altitude scenarios, the baseline model frequently failed to detect small pedestrians and cyclists, often producing inaccurate bounding boxes or missing targets. In contrast, DBYNet accurately detected these small objects, benefiting from its dual-path feature extraction and refined multi-scale fusion strategies.

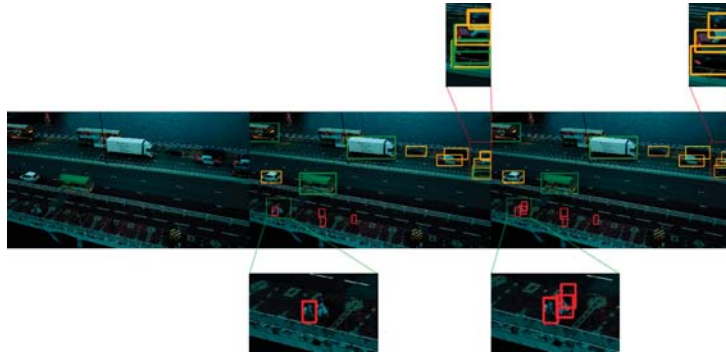


Fig. 13. Detection results in low-light nighttime environments.

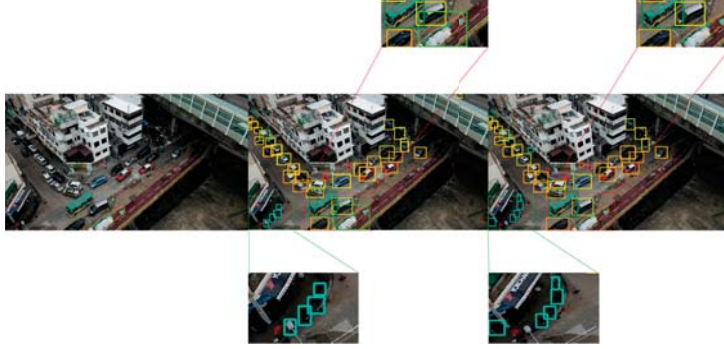


Fig. 14. Detection results for occluded targets.

As shown in Fig. 13, under low-light conditions, baseline models tended to produce false detections or missed targets due to poor contrast and background interference. DBYNet demonstrated robust performance in these scenarios by leveraging its hybrid attention mechanism to enhance feature discrimination, ensuring precise detection even in dim environments.

As shown in Fig. 14, in occlusion scenarios, baseline detectors struggled with consistency and often failed to identify partially visible objects. DBYNet effectively reduced false negatives by integrating contextual information and enhancing local feature aggregation, thereby improving detection accuracy for occluded targets.

Overall, the experimental results are consistent with the original objectives stated in the Introduction. DBYNet is designed to balance detection accuracy and computational efficiency for UAV-based object detection under resource constraints. As demonstrated by the quantitative results, the proposed method achieves competitive mAP while maintaining real-time inference speed. Furthermore, the lightweight optimization strategies effectively reduce model complexity without causing significant performance degradation. These results confirm that the proposed design choices successfully address the accuracy–efficiency trade-off targeted in this work.

4.5.4. Heatmap Visualization and Attention Analysis

To better understand the internal behavior of the network, heatmaps are generated to visualize attention distributions in various scenes. Fig. 15 displays attention heatmaps produced by DBYNet.

The heatmaps reveal that DBYNet dynamically adjusts its attention based on scene complexity. In dense traffic environments, attention was



Fig. 15. Heatmap visualization of attention distributions under different scenarios.

highly concentrated on key objects such as vehicles and pedestrians, indicating precise foreground-background separation. In urban scenes with multiple target types, layered attention enabled the model to discriminate effectively between categories. In sparse scenes, attention was diffused yet still focused on salient objects. These observations confirm that the proposed hybrid attention mechanism enhances adaptability across various UAV scenarios, supporting robust and precise object detection.

4.5.5. Cross-dataset Evaluation

To further evaluate the generalization capability of DBYNet beyond a single benchmark, we conduct cross-dataset experiments on the UAVDT dataset, which focuses on traffic surveillance scenarios captured from unmanned aerial platforms.

Table 13. Cross-dataset evaluation results on the UAVDT dataset.

Method	mAP@50 (%)	Params (M)	FPS
YOLOv8-n	41.2	3.2	165
YOLOv8-s	44.6	11.2	128
DBYNet (Ours)	47.1	10.3	102

As reported in Table 13, DBYNet maintains competitive performance compared with representative lightweight baselines, demonstrating its robustness across different UAV datasets and application scenarios.

5. Conclusion and Future Work

This paper proposed DBYNet, a lightweight and efficient object detection framework designed for UAV applications. Extensive experiments on the VisDrone2019 dataset demonstrated that DBYNet outperformed various state-of-the-art detectors in terms of both accuracy and speed, achieving 47.7% mAP@50 and 102 FPS with low model complexity. Visualization analysis further confirmed its robustness in challenging UAV scenarios such as small object detection, occlusion handling, and low-light environments. In future work, we plan to enhance detection performance for ultra-small objects, incorporate multi-modal data to improve robustness, and explore domain adaptation and edge-cloud collaborative inference to further strengthen generalization and deployment flexibility.

6. Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. Acknowledgment and Funding

This work was supported in part by the State Key Laboratory of Autonomous Intelligent Unmanned Systems with the opening project number is ZZKF2025-1-16, and Wuhan Key Research and Development Program (2024050702030136), and Hubei Province International Science and Technology Collaboration Program (2025EHA043)

References

- Ahmari, R., Hemmati, V., Mohammadi, A., Kebria, P., Mahmoud, M., & Homaifar, A. (2025). A data-driven approach for uav-ugv integration. *Proceedings of the Automation, Robotics & Communications for Industry*, 4, 77.
- Akherati, H., Beyramzad, J., Khiyabani, S. S., Shariatinezhad, A., & Eskandari, E. (2025). Finite-time stable model free sliding mode attitude controller/observer design for uncertain space systems based on time delay estimation. *Advances in Space Research*, .

- Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, .
- Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 764–773). URL: <https://arxiv.org/abs/1703.06211>. doi:10.1109/ICCV.2017.89.
- Ding, J., Xue, N., Xia, G.-S., Bai, X., Yang, W., Yang, M., Belongie, S., Luo, J., Datcu, M., Pelillo, M., & Zhang, L. (2021). Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, . doi:10.1109/TPAMI.2021.3117983.
- Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., & Tian, Q. (2018). The unmanned aerial vehicle benchmark: Object detection and tracking. *arXiv preprint arXiv:1804.00518*, .
- Fan, H. et al. (2021). Litese: Lightweight squeeze-and-excitation. *NeurIPS*, .
- Feng, C. et al. (2021). Tood: Task-aligned one-stage object detection. *ICCV*, .
- Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). YoloX: Exceeding yolo series in 2021. In *arXiv preprint arXiv:2107.08430*.
- Gong, M. et al. (2023). Dynamic visual-scale estimation for uav perception systems. *TNNLS*, .
- Guo, Z. et al. (2020). Single path one-shot neural architecture search. In *ECCV*.
- Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). Ghostnet: More features from cheap operations. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 1580–1589).
- He, K. et al. (2016). Deep residual learning for image recognition. *CVPR*, .
- Howard, A. et al. (2019). Searching for mobilenetv3. *ICCV*, .

- Hu, J. et al. (2018). Squeeze-and-excitation networks. *CVPR*, .
- Jacob, B. et al. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. *CVPR*, .
- Jandaghi, E., Zhou, M., Stegagno, P., & Yuan, C. (2025). Adaptive formation learning control for cooperative auvs under complete uncertainty. *Frontiers in Robotics and AI*, 11, 1491907.
- Jocher, G. et al. (2023). YOLOv8: Next-generation object detection. *arXiv preprint arXiv:2302.04388*, .
- Li, C. et al. (2023). SAM-DETR: Selective anchor matching for end-to-end tiny object detection. *arXiv preprint arXiv:2301.12345*, .
- Li, K. et al. (2022). Cbnetv2: A composite backbone network architecture for object detection. *CVPR*, .
- Li, Y., Zhang, H., Xiong, H., Liu, L., Huang, J., & Zhang, Z. (2021). Multi-scale context aggregation network with attention-guided for crowd counting. *Neurocomputing*, 449, 45–56.
- Lin, T.-Y. et al. (2017). Feature pyramid networks for object detection. In *CVPR*.
- Lin, W., & Zhang, J. (2022). A survey on UAV-based object detection: Challenges, applications, and future trends. *IEEE Access*, 10, 10243–10256.
- Liu, C., Zhang, H., Ma, H., & et al. (2021a). NAM: Normalization-based attention module. *Proceedings of NeurIPS*, .
- Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128, 261–318.
- Liu, S. et al. (2018). Path aggregation network for instance segmentation. In *CVPR*.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 10012–10022).

- Liu, Z. et al. (2017). Learning efficient convolutional networks through network slimming. *ICCV*, .
- Ma, N. et al. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*.
- Mahjourian, N., & Nguyen, V. (2025). Multimodal object detection using depth and image data for manufacturing parts. In *International Manufacturing Science and Engineering Conference* (p. V002T16A001). American Society of Mechanical Engineers volume 89022.
- Niu, Y., & Chen, L. (2023). Low-altitude economy: Opportunities and challenges for uav applications. *IEEE Transactions on Intelligent Transportation Systems*, .
- Safari, K., Safari, A., & Manthouri, M. (2023). Designing an adaptive-intelligent controller for quadcopter based on brain emotional learning. In *2023 9th International Conference on Control, Instrumentation and Automation (ICCIA)* (pp. 1–5). IEEE.
- Song, H., Sun, D., & et al. (2022). VidT: An efficient and effective fully transformer-based object detector. *Proceedings of ICLR*, .
- Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning (ICML)*, (pp. 6105–6114).
- Wan, A. et al. (2020). Fbnetv2: Differentiable neural architecture search for lightweight networks. In *CVPR*.
- Wang, C.-Y. et al. (2023a). Yolov7-tiny: Speed/accuracy balance for uav detection. *arXiv preprint arXiv:2307.01234*, .
- Wang, L. et al. (2023b). Lau: Lightweight attention units for uav detection. *TNNLS*, .
- Wang, Q., Ma, B., & Chen, H. (2021). Transformer meets vision: A survey. *arXiv preprint arXiv:2101.01169*, .
- Wang, Q. et al. (2020). Eca-net: Efficient channel attention for deep convolutional neural networks. *CVPR*, .

- Wang, Y., Ma, C., Zhao, C., Xia, H., Chen, C., & Zhang, Y. (2025). Wb-yolo: An efficient wild bat detection method for ecological monitoring in complex environments. *Engineering Applications of Artificial Intelligence*, 157, 111232. URL: <https://www.sciencedirect.com/science/article/pii/S0952197625012333>. doi:<https://doi.org/10.1016/j.engappai.2025.111232>.
- Woo, S. et al. (2018). Cbam: Convolutional block attention module. *ECCV*, .
- Wu, D., Ying, Y., Zhou, M., Pan, J., & Cui, D. (2024). Yolo-claw: A fast and accurate method for chicken claw detection. *Engineering Applications of Artificial Intelligence*, 136, 108919. URL: <https://www.sciencedirect.com/science/article/pii/S0952197624010777>. doi:<https://doi.org/10.1016/j.engappai.2024.108919>.
- Xiong, Y., Liu, H., Cheng, B., & et al. (2021). Mobiledets: Searching for object detection architectures for mobile accelerators. *Proceedings of CVPR*, (pp. 3825–3834).
- Xu, L., & Zhang, X. (2023). Small object detection in uav imagery: A new benchmark and a novel attention-enhanced detector. *IEEE Transactions on Geoscience and Remote Sensing*, .
- Xu, X. et al. (2022). Pp-yoloe: An effective and efficient implementation for object detection. *arXiv preprint arXiv:2203.16250*, .
- Yang, J., Li, C., & et al. (2021a). Focal self-attention for local-global interactions in vision transformers. *Proceedings of NeurIPS*, .
- Yang, J. et al. (2021b). Simam: A simple, parameter-free attention module for cnns. *ICCV*, .
- Zhang, X. et al. (2022a). Stripformer: Modeling long range interactions along strips for efficient vision transformers. *ECCV*, .
- Zhang, Y. et al. (2022b). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *ICLR*, .
- Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., & Ling, H. (2021). Detection and tracking meet drones challenge. *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, . doi:10.1109/TPAMI.2021.3119563.

Zhu, P., Wen, L., Du, D., Bian, X., & Ling, H. (2018). Vision meets drones: A challenge. In *European Conference on Computer Vision (ECCV)*.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*.