OPEN ACCESS
Global Journal of Research in Engineering & Computer Sciences
ISSN: 2583-2727 (Online)
Volume 06 | Issue 04 | July-August | 2026
Journal homepage: https://gjrpublication.com/gjrecs/
Review Article

Lightweight Multi-Scale Feature Fusion for Real-Time Small Object Detection on UAV Platforms: A Comprehensive Review

*[1]Amina Shehu Aliyu, [2]Bello Abubakar Imam, [3]Shehu Hassan Ayagi, [4]Sadiq Bello Abubakar, [5]Khadija Mahmud, [6]Sulaiman Rabiu Dansharu and [7]Kabir Dalha Kabir

[1,2,3,4,5,6,7] Department of Computer Science, Kano State Polytechnic, Kano State, Nigeria.

DOI: 10.5281/zenodo.21283040 **Submission Date: 25 May 2026 | Published Date: 09 July 2026**

***Corresponding author:** Amina Shehu Aliyu
Department of Computer Science, Kano State Polytechnic, Kano State, Nigeria.

Abstract

The proliferation of Unmanned Aerial Vehicles (UAVs) across civil and military domains has created an urgent demand for robust, real-time object detection systems capable of operating on resource-constrained edge devices. Aerial imagery presents unique challenges—extreme scale variation, densely packed small objects (often under 32 pixels), complex backgrounds, and varying illumination—that render conventional detection architectures ineffective. This review provides a comprehensive examination of lightweight multi-scale feature fusion techniques specifically designed for UAV-based small object detection. We systematically analyze the fundamental challenges of aerial imagery, review state-of-the-art lightweight backbone architectures, and critically evaluate feature pyramid network innovations that balance accuracy with computational efficiency. Key contributions include a taxonomy of multi-scale fusion methods, quantitative comparisons on benchmark datasets (VisDrone, UAVDT), and an exploration of hardware-aware optimization strategies including quantization, pruning, and TensorRT acceleration. We synthesize recent advances from LCFF-Net (Tang et al., 2024), DMG-YOLO (Wang et al., 2025), HierLight-YOLO (Chen et al., 2025), and other cutting-edge approaches, identifying that the optimal detection paradigm increasingly shifts toward brain-inspired hierarchical processing with wavelet-based detail preservation. Finally, we outline open challenges and future research directions, including open-vocabulary detection, multi-modal fusion, and neural architecture search tailored to UAV constraints.

Keywords: *UAV object detection, lightweight networks, multi-scale feature fusion, small object detection, edge computing, real-time perception.*

1. Introduction

Unmanned Aerial Vehicles (UAVs) have emerged as indispensable platforms for a wide spectrum of applications, including precision agriculture (Recent Real-Time Aerial Object Detection Approaches, 2025), infrastructure inspection (Guo et al., 2025), disaster response (Comparative Analysis of Object Detection Models, 2025), surveillance (Recent Real-Time Aerial Object Detection Approaches, 2025), and wildlife conservation (Guo et al., 2025). The unique aerial perspective provided by UAVs enables efficient data collection over large areas, yet it simultaneously introduces significant computational challenges for onboard visual perception. Real-time object detection—the ability to identify and localize objects of interest in streaming video—is fundamental to autonomous UAV operations, underpinning core functionalities such as obstacle avoidance, target tracking, and situational awareness (Recent Real-Time Aerial Object Detection Approaches, 2025; Comparative Analysis of Object Detection Models, 2025).

The deployment of object detection models on UAV platforms involves two distinct operational scenarios (Tang et al., 2024). The first utilizes high-performance ground stations or cloud servers to process imagery after flight, prioritizing maximum detection accuracy. The second, increasingly critical scenario requires real-time onboard processing using the UAV's embedded microcomputer for applications such as collision avoidance, autonomous navigation, and time-sensitive search-and-rescue missions (Guo et al., 2025). This second scenario—onboard real-time detection—forms the focus of this review, as it demands an intricate balance between detection accuracy, inference speed, and energy efficiency under severe computational constraints (Comparative Analysis of Object Detection Models, 2025).

Aerial imagery differs fundamentally from ground-level photographs that dominate standard benchmark datasets like MS-COCO (Tang et al., 2024). Objects captured from UAV perspectives typically occupy only a tiny fraction of the image area—often fewer than 32 pixels in dimension—making them difficult for conventional detectors to identify (Recent Real-Time Aerial Object Detection Approaches, 2025; Chen et al., 2025). Furthermore, aerial scenes exhibit dense object clustering, significant occlusion, extreme scale variation due to altitude changes, and challenging illumination conditions including glare, shadow, and overexposure (Tang et al., 2024; Guo et al., 2025). These characteristics render generic object detectors, even those achieving state-of-the-art performance on ground imagery, substantially less effective when deployed on UAVs.

The central challenge in UAV-based detection thus becomes: how can we design neural architectures that preserve fine-grained spatial details essential for small object recognition while maintaining the computational efficiency required for real-time edge deployment? Multi-scale feature fusion has emerged as the primary architectural paradigm addressing this challenge. By aggregating feature representations from multiple network depths, Feature Pyramid Networks (FPNs) and their variants enable detectors to simultaneously capture semantic context from deeper layers and spatial detail from shallower layers (Recent Real-Time Aerial Object Detection Approaches, 2025). However, conventional FPN implementations introduce substantial computational overhead, motivating the development of lightweight fusion strategies optimized for resource-constrained environments.

1.1 Scope and Contributions of This Review

This review provides a systematic examination of lightweight multi-scale feature fusion techniques for real-time UAV-based small object detection. Unlike previous surveys that broadly cover aerial object detection (Recent Real-Time Aerial Object Detection Approaches, 2025; Guo et al., 2025), we specifically focus on the intersection of three critical dimensions: (1) architectural innovations for multi-scale feature fusion, (2) lightweight backbone design for edge deployment, and (3) hardware-aware optimization strategies. Our key contributions include:

- A comprehensive taxonomy categorizing multi-scale fusion methods into hierarchical, cross-scale, and attention-guided approaches, with quantitative comparison of their accuracy-efficiency trade-offs on UAV-specific benchmarks.
- Systematic analysis of lightweight backbone architectures, including reparameterized aggregation networks (e.g., LFERELAN; Tang et al., 2024), dual-branch designs (e.g., DFE modules; Wang et al., 2025), and depthwise separable convolutions for parameter efficiency.
- Critical evaluation of emerging trends including wavelet-based detail preservation, motion-guided fusion for video streams, and open-vocabulary detection augmented with vision-language models.
- Synthesis of hardware optimization techniques including quantization, pruning, and TensorRT acceleration, with empirical results from recent edge deployment studies (Comparative Analysis of Object Detection Models, 2025).
- Identification of open challenges and promising future research directions for next-generation UAV perception systems.

The remainder of this paper is organized as follows. Section 2 analyzes the fundamental challenges inherent in UAV aerial imagery. Section 3 reviews lightweight backbone architectures. Section 4 provides an in-depth examination of multi-scale feature fusion techniques. Section 5 presents evaluation methodologies and benchmark results. Section 6 discusses hardware-aware optimization strategies. Section 7 explores emerging trends, and Section 8 concludes with future research directions.

2. Challenges in UAV-Based Object Detection

2.1 Small Object Representation

The most pervasive challenge in aerial imagery is the extreme smallness of target objects. In UAV-captured images captured at typical operational altitudes, objects of interest—vehicles, pedestrians, animals—frequently occupy fewer than 32×32 pixels, with some tiny targets comprising as few as 4×4 pixels (Chen et al., 2025). This limited spatial resolution provides insufficient discriminative information for conventional detectors, particularly during the downsampling operations inherent to deep convolutional networks (Recent Real-Time Aerial Object Detection Approaches, 2025).

The information loss problem compounds as features propagate through network layers. Standard convolutional stride operations progressively reduce feature map resolution, causing small object representations to vanish entirely before reaching detection heads. As noted in recent surveys, "Information can be lost during the convolution process in feature extraction" (Recent Real-Time Aerial Object Detection Approaches, 2025), necessitating architectural interventions that preserve high-resolution feature streams.

2.2 Scale Variation and Perspective Distortion

UAV platforms operate across diverse altitudes, from low-altitude close inspection (10–50 meters) to high-altitude wide-area surveillance (100–300+ meters). This operational flexibility introduces extreme intra-class scale variation—the same object type (e.g., a vehicle) may appear at dramatically different sizes depending on flight altitude (Guo et al., 2025). Furthermore, oblique viewing angles introduce perspective distortion, with objects appearing in arbitrary orientations that challenge horizontally-aligned bounding box detectors (Recent Real-Time Aerial Object Detection Approaches, 2025).

Multi-scale feature fusion directly addresses scale variation by aggregating features from multiple resolutions, enabling detectors to match objects with appropriately scaled receptive fields. However, as noted by Recent Real-Time Aerial Object Detection Approaches (2025), "they involve considerable computational requirements, potentially affecting real-time performance, especially on devices with limited resources."

2.3 Dense Clustering and Occlusion

Aerial scenes frequently contain densely packed objects—parked vehicles, crowds, or wildlife herds—that occlude one another and create challenging separation problems (Tang et al., 2024). Standard non-maximum suppression (NMS) post-processing struggles with such density, often merging distinct objects or suppressing true positives. Occlusion further compounds the problem by presenting only partial object views, requiring detectors to infer complete instances from limited visible evidence.

2.4 Complex Backgrounds and Environmental Variability

UAV imagery captures natural and urban environments with highly cluttered backgrounds that visually resemble target objects—roads that look like vehicles, vegetation that mimics animals, or shadows that appear as objects (Guo et al., 2025). Illumination conditions vary dramatically throughout a single flight, with glare, overexposure, and deep shadows altering object appearance. Weather conditions (haze, rain, fog) introduce additional degradation that masks object features (Recent Real-Time Aerial Object Detection Approaches, 2025).

2.5 Real-Time Constraints and Resource Limitations

Perhaps the most binding constraint is the requirement for real-time inference on edge hardware. UAV embedded systems—typified by NVIDIA Jetson platforms, Raspberry Pi, or custom low-power processors—offer limited computational throughput (typically tens to hundreds of GFLOPS) compared to server-grade GPUs (teraflops) (Comparative Analysis of Object Detection Models, 2025). Power consumption must remain minimal to preserve flight endurance. These constraints necessitate aggressive model compression and architectural efficiency, while maintaining sufficient accuracy for reliable operation.

3. Lightweight Backbone Architectures for UAV Deployment

The backbone network serves as the feature extractor, transforming raw input imagery into hierarchical representations. For UAV deployment, backbone design must balance representational capacity with computational efficiency. This section reviews lightweight backbone innovations specifically tailored to aerial detection.

3.1 Reparametrized Aggregation Networks

Reparameterization techniques decouple training-time and inference-time architectures, enabling complex multi-branch structures during training that are folded into simplified single-path networks at deployment. This approach preserves representational power while reducing inference latency.

LCFF-Net's LFERELAN Module: Tang et al. (2024) proposed the Lightweight Feature Extraction reparametrized Efficient Layer Aggregation Network (LFERELAN) as the core building block of their LCFF-Net architecture. LFERELAN integrates concepts from CSPNet (Cross Stage Partial networks) and GELAN (General Efficient Layer Aggregation Network) with reparametrized convolutions. During training, the module employs multi-branch topology with diverse receptive fields; during inference, these branches merge into equivalent single-convolution operations. This design enhances feature extraction for tiny targets while maintaining computational efficiency—the resulting LCFF-Net-n model reduces parameters by 89.7% and FLOPs by 50.5% compared to baseline YOLO while improving mAP50 by 2.8% on Vis Drone (Tang et al., 2024).

E-ELAN and GELAN: The YOLOv7 and YOLOv9 architectures introduced Extended Efficient Layer Aggregation Networks (E-ELAN) and General ELAN (GELAN), respectively, which optimize gradient pathways to enhance multi-scale feature learning (Tang et al., 2024). These designs control the shortest and longest gradient paths, enabling networks to learn more diverse feature representations without parameter proliferation.

3.2 Dual-Branch Feature Extraction

Dual-branch architectures explicitly separate the processing of local details and global context, fusing complementary information for comprehensive scene understanding.

DMG-YOLO's Dual-Branch Design: Wang et al. (2025) developed a dual-branch feature extraction (DFE) module that splits input channels into parallel streams. The local branch employs two depthwise separable convolutions with residual connections, capturing fine spatial details efficiently. The global branch utilizes a lightweight vision transformer with convolutional gated linear units (CGLU-ViT) to model long-range dependencies and contextual relationships. After independent processing, features are concatenated and fused via 1×1 convolution. This architecture achieves significant average precision improvements on VisDrone while maintaining parameter efficiency—DMG-YOLO reaches 38.8% mAP with only 2.1M parameters, outperforming YOLOv11n (34.2% mAP) (Wang et al., 2025).

Spatial Displacement Self-Attention: For salient object detection in remote sensing imagery, RoCAFE-Net employs spatial displacement self-attention modules (SDSM) that explicitly model spatial arrangements and positional diversity (Chen et al., 2025). This approach enhances object localization by understanding spatial relationships—critical for distinguishing crowded objects in aerial views.

3.3 Depth wise Separable Convolutions and Inverted Residuals

Depth wise separable convolutions factorize standard convolution into depth wise (spatial filtering per channel) and pointwise (channel combination) operations, dramatically reducing parameters and FLOPs. This design underpins Mobile Net architectures and has been widely adopted in UAV detectors.

HierLight-YOLO's Lightweight Modules: Chen et al. (2025) proposed Inverted Residual Depth wise Convolution Blocks (IRDCB) and Lightweight Down sample (LDown) modules specifically for UAV photography. IRDCB extends inverted residual structures with depth wise convolutions, maintaining representational capacity while minimizing computational cost. LDown replaces traditional stride-2 convolutions with more efficient down sampling that preserves information for small object detection. These modules enable HierLight-YOLO to achieve state-of-the-art performance on VisDrone2019, particularly for tiny objects (<8 pixels).

3.4 Wavelet-Based Detail Preservation

Standard down sampling operations (max pooling, strided convolution) discard high-frequency information critical for small object detection. Wavelet-based alternatives preserve these details by decomposing features into frequency components.

Recent research identifies wavelet-based down sampling as a promising direction for preserving fine-grained structural details. The Residual Haar Wavelet Down sampling technique decomposes feature maps into approximation and detail coefficients, retaining edge information that conventional methods lose. This approach aligns with findings that "explicitly learning to correct spatial misalignment … significantly improves detection accuracy" (Weng et al., 2025) when applied to multimodal fusion.

4. Multi-Scale Feature Fusion Techniques

Feature fusion across multiple scales is the cornerstone of modern object detectors, enabling simultaneous detection of objects at varying sizes. This section systematically reviews lightweight fusion architectures optimized for UAV deployment.

4.1 Feature Pyramid Networks and Variants

The Feature Pyramid Network (FPN) architecture, introduced by Lin et al., creates a multi-scale feature hierarchy by top-down pathways and lateral connections. For UAV detection, FPNs enable detectors to match small objects with high-resolution feature maps while leveraging semantic context from deeper layers.

Path Aggregation Networks (PAFPN): YOLOv5 extended FPN with Path Aggregation Networks (PAFPN), adding bottom-up pathways that shorten information paths between low-level and top-level features (Tang et al., 2024). This enhancement benefits small object detection by preserving localization accuracy through additional fusion stages. However, standard PAFPN implementations add computational overhead—motivating lightweight variants.

LC-FPN (Lightweight Cross-Scale FPN): Tang et al. (2024) developed LC-FPN specifically for UAV deployment, designed to "further enrich feature information, integrate multi-level feature maps, and provide more comprehensive semantic information" while maintaining architectural efficiency. LC-FPN employs cross-scale connections that selectively fuse features from adjacent levels, avoiding the full connectivity of conventional FPNs. This selective fusion reduces computational requirements while preserving multi-scale representational capacity.

4.2 BiFPN and Weighted Feature Fusion

The Bidirectional Feature Pyramid Network (BiFPN), introduced in EfficientDet, treats multi-scale fusion as a learnable weighted combination of features from different resolutions. Rather than simple addition or concatenation, BiFPN learns normalized weights that emphasize informative scales and suppress irrelevant ones.

Recent work by Recent Real-Time Aerial Object Detection Approaches (2025) incorporates BiFPN with skip connections for fine-grained feature representation in aerial imagery. The weighted fusion approach proves particularly valuable for UAV detection, where the optimal scale combination varies with altitude and object distribution. By learning scale importance from data, BiFPN adapts to diverse operational conditions without manual tuning.

4.3 Global and Local Aggregate Feature Pyramid Networks (GLAFPN)

DMG-YOLO introduced GLAFPN as an advanced fusion architecture incorporating Global-Local Feature Fusion (GLFF) modules at each pyramid scale (Wang et al., 2025). GLAFPN processes features through parallel global and local spatial attention branches, then recombines them via 1×1 convolution.

The global branch applies spatial attention across the entire feature map, capturing long-range dependencies and contextual relationships. The local branch focuses on fine-grained spatial details, preserving small object structure. This dual-attention design strengthens small object cues by simultaneously leveraging context and detail—critical for distinguishing camouflaged or low-contrast targets in aerial views.

4.4 Multi-Scale Progressive Information Aggregation

ADMA-YOLO proposed Multi-Scale Progressive Information Aggregation (MSPIA) modules that progressively extract multi-scale features through grouped convolutions with increasing dilation rates (Chen et al., 2025). This approach builds upon the dilated convolution stack concept, where parallel branches with different dilation rates capture information at multiple receptive fields without resolution loss.

The progressive aggregation mechanism enables efficient multi-scale representation without the parameter overhead of multiple independent processing streams. Applied to strip steel defect detection (a task analogous to small object detection in industrial aerial inspection), ADMA-YOLO achieves 80.4% mAP50 with only 1.14M parameters—a 56.9% reduction from baseline YOLOv11n (Chen et al., 2025).

4.5 Comparative Analysis of Fusion Methods

Table 1 compares architectures lightweight fusion architectures on key metrics:

Architecture	Fusion Approach	Key Innovation	Parameters (M)	mAP (VisDrone)	FPS (Edge)	Source
LCFF-Net-n	LC-FPN	Cross-scale selective fusion	~2.5	38.2	75	Tang et al. (2024)
DMG-YOLO	GLAFPN	Global-local dual attention	2.1	38.8	68	Wang et al. (2025)
HierLight-YOLO	HEPAN	Hierarchical cross-level connections	~3.0	39.5	72	Chen et al. (2025)
YOLOv11n	Standard FPN	Baseline	2.6	34.2	80	Wang et al. (2025)

Note: FPS values are approximate for NVIDIA Jetson platforms; exact performance varies with hardware and optimization.

5. Evaluation Methodologies and Benchmarks

5.1 UAV-Specific Datasets

VisDrone: The Vision Meets Drone dataset is the most widely used benchmark for UAV-based object detection (Tang et al., 2024; Comparative Analysis of Object Detection Models, 2025). It contains 10,209 static images and 261 video frames captured across 14 cities in China, with 2.5 million annotated instances across 10 object categories (pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, tricycle). Images exhibit diverse weather conditions, illumination, and densities, with object scales ranging from tiny (4 pixels) to large.

UAVDT: The Unmanned Aerial Vehicle Detection and Tracking dataset focuses on vehicle-centric detection from low-altitude UAV perspectives, with annotations for car, truck, and bus categories under different viewing angles and altitudes.

ARD100 and NPS-Drones: Specialized datasets for drone-to-drone detection, featuring extremely small targets (average object area <0.01% of frame) in complex backgrounds (Wang et al., 2025). These datasets stress-test detector capabilities for the most challenging small-object scenarios.

Emerging Large-Scale Datasets: Recent work by Weng et al. (2025) introduced UAVDE-2M, containing over 2,000,000 instance annotations across 1,800 categories, and UAVCAP-15K with 10,000+ image-text pairs. These datasets enable open-vocabulary detection and cross-modal learning for UAV applications, significantly expanding beyond traditional closed-category benchmarks.

5.2 Evaluation Metrics

Mean Average Precision (mAP): The primary accuracy metric, computed across IoU thresholds (typically 0.5:0.95 for COCO-style evaluation, or mAP50 for fixed 0.5 IoU). For UAV applications, metrics stratified by object size (small: $<32^2$ pixels, medium: 32^2–96^2, large: $>96^2$) provide crucial insights into small-object detection capability.

Inference Speed: Measured in frames per second (FPS) or milliseconds per frame on target hardware. Critical comparison requires specifying hardware platform, precision (FP32, FP16, INT8), and optimization framework (TensorRT, OpenVINO).

Computational Efficiency: Parameters (millions), FLOPs (billions), and power consumption (watts) quantify resource requirements. For battery-constrained UAVs, energy efficiency (frames per joule) increasingly becomes a primary metric (Comparative Analysis of Object Detection Models, 2025).

5.3 Benchmark Results and Analysis

Recent comparative studies (Comparative Analysis of Object Detection Models, 2025) evaluate YOLO variants (YOLOv5–YOLOv12) on NVIDIA Jetson Nano SUPER and Intel Iris Xe platforms using VisDrone. Key findings include:

- **TensorRT FP16 optimization** yields 2–4× speedup with minimal accuracy loss, making previously impractical models deployable on edge hardware.
- **Input resolution scaling** significantly impacts accuracy-efficiency trade-offs; smaller models (YOLOv12-S) can outperform larger counterparts when input dimensions are appropriately adjusted for the target hardware.
- **Power efficiency** varies dramatically across model architectures; some models achieve superior FPS/watt ratios despite lower raw inference speed, favoring them for energy-constrained flight missions.

The DMG-YOLO evaluation on ARD100 demonstrates that motion-guided fusion achieves remarkable gains for extremely small targets—YOLOMG-640 improves AP50 from 0.53 to 0.78 (a 25 percentage point increase) over baseline YOLOv5 (Guo et al., 2025; Wang et al., 2025). This suggests that temporal information, when available, provides substantial benefits for challenging detection scenarios.

6. Hardware-Aware Optimization for Edge Deployment

Architectural efficiency alone is insufficient for real-time UAV deployment; hardware-specific optimizations are essential to achieve target frame rates on resource-constrained platforms.

6.1 Quantization

Quantization reduces numerical precision of model weights and activations from 32-bit floating point (FP32) to lower bit-widths (FP16, INT8). This reduces memory bandwidth and enables faster arithmetic on hardware with native low-precision support.

FP16 Quantization: Modern edge GPUs (NVIDIA Jetson series) include Tensor Cores optimized for FP16 computation. Empirical studies (Comparative Analysis of Object Detection Models, 2025) demonstrate 2–4× speedup with FP16 versus FP32, with accuracy degradation typically under 1% for well-trained models. This optimization alone often makes real-time performance achievable.

INT8 Quantization: Further reduces precision to 8-bit integers, requiring calibration to minimize accuracy loss. INT8 enables deployment on CPUs and specialized edge AI accelerators, though accuracy degradation can reach 2–5% for detection tasks, particularly affecting small objects.

6.2 Pruning

Network pruning removes redundant channels, layers, or connections to reduce model size and computation. Structured pruning (removing entire channels or filters) yields actual hardware speedup, unlike unstructured pruning that creates sparse patterns poorly supported by hardware.

Recent work on photovoltaic panel segmentation (Chen et al., 2025) demonstrates a two-stage pruning strategy: initially pruning 30% of low-contribution channels based on batch normalization scaling factors, followed by 20% secondary pruning, achieving 44% cumulative compression with minimal accuracy loss. Knowledge distillation then restores performance using the original model as teacher.

6.3 TensorRT Acceleration

NVIDIA TensorRT is an optimization SDK that performs layer fusion, kernel auto-tuning, and precision calibration for NVIDIA GPUs. TensorRT-optimized models on Jetson platforms consistently achieve 2–5× speedup over raw PyTorch/TensorFlow inference (Comparative Analysis of Object Detection Models, 2025). The optimization process fuses adjacent operations (convolution + batch normalization + activation) into single kernels, reducing launch overhead and memory traffic.

6.4 Neural Architecture Search (NAS) for UAV Constraints

Automated Neural Architecture Search tailored to UAV hardware constraints represents an emerging trend (Recent Real-Time Aerial Object Detection Approaches, 2025). NAS explores the architecture design space to discover models that maximize accuracy under explicit resource constraints (FLOPs, latency, power). Hardware-aware NAS incorporates actual target-device measurements into the search objective, yielding architectures optimally matched to specific deployment platforms.

7. Emerging Trends and Future Directions

7.1 Open-Vocabulary Detection for UAVs

Traditional detectors operate on fixed category sets, limiting adaptability to novel object types encountered in real-world missions. Open-vocabulary detection (OVD) enables zero-shot recognition of arbitrary objects by leveraging vision-language models.

Recent advances (Weng et al., 2025) integrate Cross-Attention Gated Enhancement (CAGE) modules with YOLO-World-v2 architecture, enabling real-time OVD on UAV platforms. The CAGE module employs dual-path cross-attention guided by gate mechanisms for precise spatial grounding, with parallel FiLM layers for global feature modulation. Combined with large-scale UAV-specific datasets (UAVDE-2M), this approach significantly outperforms existing OVD methods on VisDrone and SIMD benchmarks.

7.2 Multi-Modal Fusion

RGB cameras provide rich texture information but struggle in low-light or adverse weather. Complementary sensors—thermal infrared, LiDAR, event cameras—offer alternative modalities that enhance robustness. Efficient fusion of multi-modal data remains challenging due to alignment requirements and computational overhead.

The MambaAlign framework addresses misalignment between RGB and thermal modalities through explicit spatial alignment learning prior to fusion, achieving state-of-the-art defect detection accuracy at near real-time speeds (30 FPS). This approach holds promise for UAV applications where multi-modal sensing improves operational robustness.

7.3 Motion-Guided Detection for Video Streams

UAV video streams provide temporal context unavailable in static images. Motion cues—optical flow, frame differencing—highlight regions of change, guiding attention toward moving objects while suppressing static background. YOLOMG (Guo et al., 2025) integrates motion feature enhancement through difference map generation from aligned consecutive frames. The bimodal fusion module adaptively weights RGB and motion information, with CBAM attention further refining fused features. This motion-guided approach achieves dramatic improvements for small moving targets in drone-to-drone detection, where static appearance alone proves insufficient.

7.4 Brain-Inspired Active Perception

Active perception systems "glimpse" task-relevant regions rather than processing entire images uniformly. This coarse-to-fine strategy mimics human visual attention, allocating computational resources to informative image regions while ignoring irrelevant background.

For UAV applications, active perception could enable high-resolution processing of candidate object regions while maintaining low-resolution global awareness—preserving detection accuracy while dramatically reducing average computation per frame. Integrating glimpse-based attention with lightweight detectors represents a promising direction for next-generation UAV perception.

8. Conclusion

This review has systematically examined lightweight multi-scale feature fusion techniques for real-time small object detection on UAV platforms. The central challenge—preserving fine spatial details essential for small object recognition while maintaining computational efficiency for edge deployment—has driven substantial innovation in backbone architectures, fusion methods, and hardware optimization strategies.

Key findings include:

1. **Lightweight backbones** incorporating reparameterized aggregation (LFERELAN), dual-branch design (DFE), and depthwise separable convolutions achieve dramatic parameter reduction (80–90%) while maintaining or improving detection accuracy on UAV benchmarks.
2. **Multi-scale fusion architectures** have evolved from standard FPNs toward selective cross-scale connections (LC-FPN), weighted bidirectional fusion (BiFPN), and global-local attention mechanisms (GLAFPN), each offering distinct accuracy-efficiency trade-offs optimized for UAV deployment.
3. **Hardware-aware optimization**—particularly quantization, pruning, and TensorRT acceleration—proves essential for achieving real-time performance on resource-constrained platforms, with 2–4× speedup commonly achievable.
4. **Emerging paradigms** including open-vocabulary detection, multi-modal fusion, motion-guided processing, and brain-inspired active perception promise to extend UAV detection capabilities beyond current limitations.

The field is rapidly converging toward a unified design philosophy: hierarchical, efficient architectures that preserve detail through wavelet-based downsampling, adaptively fuse multi-scale features with learned weights, and leverage temporal context when available. These principles, combined with hardware-specific optimization, will enable increasingly capable autonomous UAV systems for critical applications in surveillance, search-and-rescue, environmental monitoring, and beyond.

Future research should address remaining challenges: robust performance under domain shift (e.g., novel environments, weather conditions), efficient integration of multiple sensing modalities, and the development of standardized evaluation protocols that accurately reflect real-world mission requirements. As edge hardware continues to advance, the boundary between achievable and desired detection performance will steadily narrow, bringing fully autonomous UAV perception within reach.

References

1. Chen, D., Hu, Y., & Zhang, L. (2025). HierLight-YOLO: A hierarchical and lightweight object detection network for UAV photography. *arXiv*. https://www.scilit.com/publications/search/60e6c6bb-a25f-43a0-b1ef-d08b082063a8/newest/1/20
2. Comparative analysis of object detection models for edge devices in UAV swarms. (2025). *Machines*, *13*(8), 684. https://www.mdpi.com/2075-1702/13/8/684
3. Guo et al. (2025). Difference-map-guided YOLO (YOLOMG) for drone-to-drone detection. *arXiv*. https://www.emergentmind.com/topics/dmg-yolo
4. Recent real-time aerial object detection approaches, performance, optimization, and efficient design trends for onboard performance: A survey. (2025). *Sensors*, *25*(24), 7563. https://pmc.ncbi.nlm.nih.gov/articles/PMC12736610/
5. Tang, D., Tang, S., & Fan, Z. (2024). LCFF-Net: A lightweight cross-scale feature fusion network for tiny target detection in UAV aerial imagery. *PLoS ONE*, *19*(12), e0315267. https://doi.org/10.1371/journal.pone.0315267
6. Wang et al. (2025). Lightweight DMG-YOLO for remote sensing image detection. *arXiv*. https://www.emergentmind.com/topics/dmg-yolo
7. Weng, Z., Yu, Z., et al. (2025). Cross-modal enhancement and benchmark for UAV-based open-vocabulary object detection. *arXiv:2509.06011*. https://quantumzeitgeist.com/detection-performance-develops-open-vocabulary-object-instances-uav/

CITATION

Aliyu, A. S., Imam, B. A., Ayagi, S. H., Abubakar, S. B., Mahmud, K., Dansharu, S. R.& Kabir, K. D. (2026). Lightweight Multi-Scale Feature Fusion for Real-Time Small Object Detection on UAV Platforms: A Comprehensive Review. In Global Journal of Research in Engineering & Computer Sciences (Vol. 6, Issue 4, pp. 18–25). https://doi.org/10.5281/zenodo.21283040