

Deep Learning-Based Small-Object Detection in UAV Imagery: A Symmetry-Informed Review of Scale Variation and Scenario Constraints

Ling Wen ^{1,2} , Anmin Gong ^{2,*} , Caihong Ma ² , Shengzhou Ma ²  and Yanzhe Zhang ³ 

¹ Key Laboratory of CTC & IE, Engineering University of PAP, Xi'an 710086, China; lingwen_525@163.com

² College of Information Engineering, Engineering University of PAP, Xi'an 710086, China; rainbowma78922@163.com (C.M.); mashengzhou2025@163.com (S.M.)

³ College of Cryptography Engineering, Engineering University of PAP, Xi'an 710086, China; zyz20011222@163.com

* Correspondence: gonganmincapf@163.com

Abstract

As UAV platforms are increasingly employed in military reconnaissance, disaster monitoring, precision agriculture, and other remote-sensing applications, small-object detection in UAV imagery is a critical yet challenging task. Existing reviews primarily address either general small-object detection or general object detection in UAV imagery, whereas reviews specifically devoted to small-object detection in UAV imagery remain scarce. This review examines existing deep learning-based research on small-object detection in UAV imagery from a problem-driven and deployment-oriented perspective. A symmetry-informed perspective was also adopted to analyze how variations in object-scale, viewpoint, imaging quality, and scene conditions affected feature representation and detection stability. Building on five major challenges in UAV small-object detection, a three-tier framework encompassing detection paradigms, key technical strategies, and scenario-specific constraints was established. Representative UAV datasets were further examined to characterize object-scale distributions, while existing methods were reviewed in terms of multiscale representation, feature enhancement, training optimization, and lightweight deployment. The applicability of different technical approaches was then discussed across five typical UAV scenarios. In addition, under relatively consistent experimental conditions, representative detection models based on two-stage, one-stage, and Transformer-based paradigms were analyzed in terms of detection accuracy, computational efficiency, and deployment feasibility. Finally, current limitations and future directions were summarized for scale-robust, scenario-adaptive, and resource-aware small-object detection in UAV imagery.

Keywords: unmanned aerial vehicles; small-object detection; deep learning; scenario constraints



Academic Editor: Zhixun Su

Received: 8 August 2026

Revised: 11 September 2026

Accepted: 14 September 2026

Published: 16 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

1. Introduction

Unmanned aerial vehicles (UAVs) have become an important platform for remote sensing because of their flexibility, easy deployment, high accessibility, and multi-perspective imaging capability [1–4]. They are widely applied in fields like smart transportation, border patrol, and disaster assessment. However, constrained by factors such as flight altitude, shooting angle, and sensor resolution, objects in aerial images often exhibit typical small-object characteristics, including low pixel proportion, severe scale variation, and severe

dense occlusion. Simultaneously, flight jitter, low illumination, and weather conditions can easily induce motion blur and detail degradation, further weakening the texture and edge representation capability of objects [5–8]. This leads to a severe degradation in the localization accuracy and robustness of traditional detection methods, making it difficult to satisfy simultaneously the requirements of high precision and inference efficiency in practical applications. Therefore, efficient and accurate detection of small objects in complex UAV imagery scenes remains a key obstacle to the practical deployment of UAV visual perception systems.

In recent years, deep learning-based object detection methods have achieved remarkable progress in general visual tasks, offering new possibilities for small-object detection in UAV imagery. To address challenges such as small-object scales, strong background interference, feature degradation, and limited onboard computing resources, researchers have proposed a wide range of improvements. Existing approaches have gradually evolved from direct adaptations of general-purpose detectors to specialized optimizations tailored for UAV scenario characteristics. However, despite the increasing number of related studies, several common problems remain, including inconsistent dataset choices, large differences in experimental settings, varying evaluation priorities, and unclear application scenarios. These issues limit the comparability of existing studies, weaken their generalizability, and reduce their value for practical deployment.

With the rapid advancement of this field, numerous studies have reviewed the latest developments in deep learning-based small-object detection and general-purpose object detection in UAV imagery. As shown in Table 1, existing reviews can be broadly divided into two categories. The first focuses on general small-object detection, primarily discussing the evolution of methods, taxonomies of techniques, and commonly used benchmark datasets [9–13]. The second concentrates on general-purpose object detection in UAV imagery, placing greater emphasis on the challenges posed by aerial scenarios and the distinctive characteristics of UAV platforms [14–17].

Existing reviews have systematically examined research progress in small-object detection and UAV-based object detection. They have made important contributions by summarizing representative detection methods, technological developments, datasets, and application-related challenges, thereby laying a solid foundation for further research and advancement in the field. Nevertheless, relatively few studies have integrated these aspects into a unified analytical framework specifically designed for small-object detection in UAV imagery.

This review presents a problem-oriented and deployment-focused review of small-object detection in UAV imagery. By establishing a three-level taxonomy, it organizes, categorizes, and conducts an in-depth analysis of existing approaches. In addition to qualitative analysis, this study performs a quantitative experimental evaluation of representative detection models on widely used aerial-image datasets.

Table 1. Comparison of existing reviews and this work across key analytical dimensions.

Review Scope	Ref.	Method	Datasets	Technical Taxonomy	Imaging Conditions	Scenario Constraints	Deployment Issues	Model Evaluation
General	[9]	Systematic	COCO, VOC, DOTA, VisDrone	CNN→Transformer, class imbalance	×	×	△	△
	[10]	Narrative	COCO, VOC, WIDER FACE	Multi-scale, data augmentation, context, GAN	×	×	×	△
	[11]	Survey	-	Small/tiny object detection perspectives	×	×	×	△
	[12]	Survey + empirical evaluation	Four well-known small-object datasets	Resolution boosting, scale-aware training, context	×	×	×	△
	[13]	Survey	State-of-the-art SOD datasets	Four pillars-multiscale, context, super-resolution, region proposal	△	×	△	△
UAV	[14]	Review	DOTA, VisDrone, etc.	Deep learning methods for aerial SOD	△	×	△	△
	[15]	Narrative	VisDrone, UAVDT	Feature enhancement, multi-scale fusion, lightweight	△	△	△	△
	[16]	Survey	UAV imagery	Backbone-neck-head architectures	△	△	△	×
	[17]	Review	Low-altitude UAV datasets	Two-stage, one-stage, advanced detectors	△	△	△	△
	This work	Structured + benchmark	VisDrone2019, UAVDT, DOTA, CARPK, AU-AIR	Three-tier: paradigms, techniques, scenario constraints	✓	✓	✓	✓

Note: ✓ indicates that the dimension is explicitly and substantially discussed; △ indicates partial or limited discussion; × indicates that the dimension is not discussed or receives little attention. These symbols describe the analytical emphasis of the selected reviews and do not imply exhaustive coverage of the entire literature.

This review adopts a structured narrative review approach rather than a formal systematic review with meta-analysis, as the primary objective is to construct a problem-driven taxonomy and analyze representative technical routes rather than to perform statistical effect-size aggregation. To ensure transparency and reproducibility, the following literature retrieval and screening protocol was established:

- Databases and Period: IEEE Xplore, Web of Science Core Collection, and arXiv. The search spanned from January 2021 to August 2026, covering the most recent advances in deep learning-based UAV small-object detection.
- Search Keywords: The core query combined terms related to platforms (“UAV” OR “unmanned aerial vehicles”) and tasks (“small object detection” OR “small target detection”) with methodological filters (“deep learning”).
- Inclusion Criteria: Studies explicitly addressing small-object/target detection in UAV-captured or aerial imagery; studies grounded in deep learning methodologies with quantitative experimental results; studies evaluated on at least one mainstream UAV/aerial benchmark (e.g., VisDrone, UAVDT, DOTA).
- Exclusion Criteria: Studies focusing solely on generic object detection in natural images without adaptation to UAV scenes; non-English publications or those without accessible full text.
- Screening: A total of 913 articles were initially retrieved across the three databases after removing duplicates. Screening was performed by title and abstract, yielding approximately 143 candidate papers. After full-text reading and applying the inclusion/exclusion criteria, 93 core papers were selected for in-depth review. Both the final three-tier classification framework and the experimental analysis were developed based on this curated literature database.

The main contributions of this review are summarized as follows:

1. Problem-driven challenge analysis. This review analyzes the core challenges of small-object detection in UAV imagery from the perspectives of UAV imaging characteristics and practical mission requirements.
2. Three-tier analytical framework. A progressive analytical framework covering detection paradigms, key technical strategies, and scenario constraints is developed to organize and discuss representative studies in this field.
3. Scenario-aware method analysis. The applicability of representative methods is examined across five selected UAV imaging and application scenarios in terms of detection accuracy, robustness, and computational efficiency.
4. Partially unified evaluation and outlook. Under partially unified experimental settings, representative detectors from different paradigms were experimentally evaluated and analyzed on the VisDrone2019 and UAVDT datasets. The current limitations and potential directions for future research are also discussed.

2. Definition and Core Challenges of Small Objects in UAV Imagery

This section mainly defines the relative scale and absolute scale of small-object in aerial scenarios and analyzes the core challenges of small objects detection from five dimensions: object attributes, imaging conditions, algorithm adaptability, task constraints, and data annotation.

2.1. Definition of Small Objects

In object detection research, there is no universally agreed-upon definition for small objects. Different studies usually define small objects from the perspectives of application scenarios, image resolution, and dataset statistics. As shown in Figure 1, current definitions

can be broadly categorized into two categories—relative-scale-based and absolute-scale-based [18]—which are described in the following:

1. Definition based on relative scale. As shown in Figure 1a, this definition uses the ratio of the pixel area of an object's bounding box to the total pixel area of the image as the criterion. The International Society for Optical Engineering (SPIE) defines a small-object as one that occupies less than 80 pixels in a 256×256 image, meaning its area constitutes less than 0.12% of the total image area. Chen et al. [19] categorize objects as small if the median ratio of their bounding-box area to the image area falls within the range of 0.08% to 0.58%. Since this relative-scale definition describes object size as a proportion of the whole image, this definition is more convenient for comparing objects across images with different resolutions. However, its results are susceptible to variations in image size, limiting its interpretative power in cross-domain evaluations.
2. Definition based on absolute scale. As shown in Figure 1b, this definition directly uses the pixel dimensions or pixel area of the object's bounding box as the criterion. Absolute definitions for small objects are not entirely consistent across different datasets. For example, DOTA [20] considers objects with scales within the range of [10, 50] pixels as small objects; SODA-A [21] and MS-COCO [22] typically classify instances with a bounding-box area less than 1024 pixels as small objects.

As shown in Figure 1c,d, compared with the relative-scale definition based on overall image dimensions, the absolute-scale definition based on bounding-box pixel area corresponds more directly to the network's receptive field, feature-map resolution, and scale-assignment mechanism, making it more suitable for comparative analyses of small-object detection methods. In this review, 32×32 pixels is adopted as the threshold for distinguishing small objects, primarily based on the following three considerations. Human visual perception: Torralba et al. [23] demonstrated that effective human scene recognition requires an object in a color image to occupy at least 32×32 pixels; objects below this size are difficult even for humans to identify. Network architecture: Mainstream backbone networks typically have a maximum downsampling factor of 32. Accordingly, a 32×32 -pixel object in the input image corresponds to exactly one feature unit on the deepest feature map. Below this threshold, spatial features are severely diluted through successive downsampling stages, resulting in inadequate feature representations. Established convention: The MS-COCO dataset defines objects with an area smaller than 32×32 pixels as small objects. This threshold has been widely used in COCO-based evaluation protocols, including COCO-style evaluations of datasets such as VisDrone and DOTA. Based on the above analysis, all instances with a bounding-box area smaller than 32×32 pixels are hereafter uniformly defined as small objects.

This definition is based on bounding-box area, i.e., an area of less than 32×32 pixels, rather than on either width or height alone. For objects with extreme aspect ratios, bounding-box area accounts only for the total number of pixels while disregarding the pixel extent along the shorter axis. Some elongated objects may have an area greater than 32×32 pixels, while their shorter side spans only a few pixels. After successive network downsampling, their features along the shorter axis may be entirely eliminated, making them as difficult to detect as small objects. Conversely, some compact, approximately square objects with smaller areas may retain richer spatial feature information. Therefore, an area-based definition may not fully capture an object's perceptual size, which remains a noteworthy challenge in small-object detection.

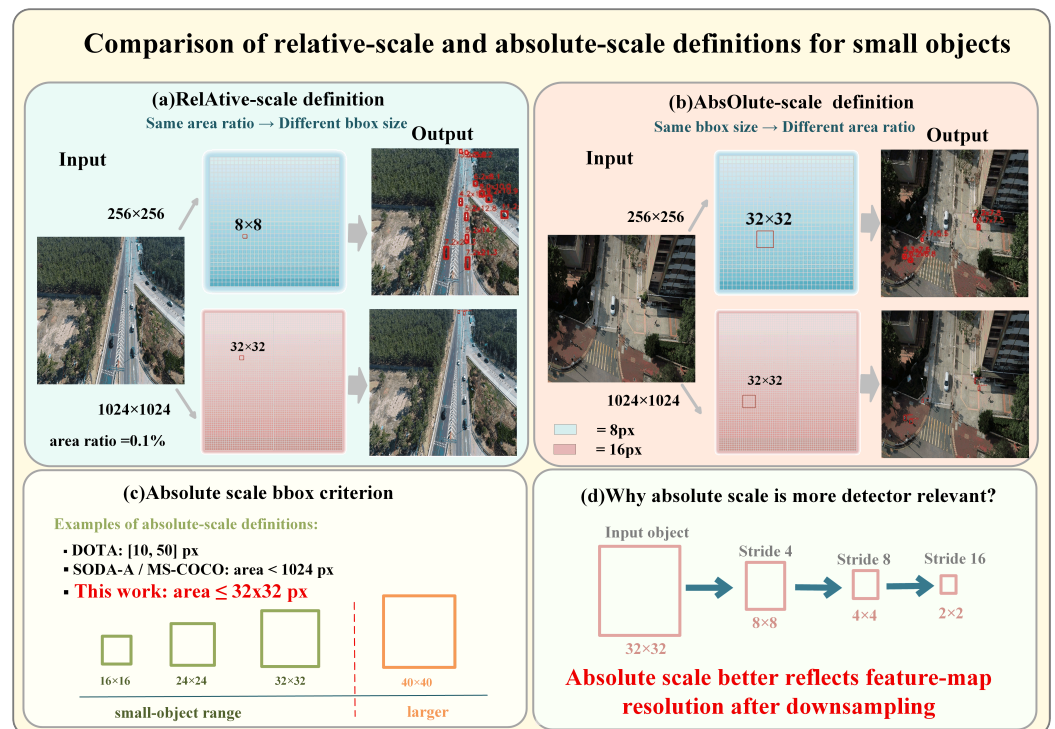


Figure 1. Schematic diagram of relative-scale and absolute-scale definitions for small objects. The boxes annotated in red as 8×8 and 32×32 are specific illustrative examples used to demonstrate the small-object scale under given image resolutions. They are not intended to represent a universal small-object scale applicable to all width, height, or area conditions.

To assess how the choice of scale thresholds influences the research conclusions, we further calculated the proportion of small objects in each dataset under the two thresholds of 16×16 and 40×40 , as presented in Table 2. The results indicate that although the complete ranking of the five datasets is not entirely consistent across different thresholds, the overall scale-based grouping remains relatively stable. Under both the original-resolution setting and the setting in which all images are uniformly resized to 640×640 , VisDrone2019 and UAVDT generally exhibit higher proportions of small objects, whereas CARPK and AU-AIR generally exhibit lower proportions. Because the small-object proportion in DOTA is particularly sensitive to the input resolution, uniform resizing substantially alters statistics based on absolute pixel dimensions. Accordingly, this study uses statistics obtained at the original resolution to characterize the inherent scale distribution of each dataset while employing statistics obtained at a unified resolution to analyze the effective scale distribution within the model's input space.

When the threshold is reduced from 32×32 to 16×16 , the proportion of objects categorized as small decreases substantially. For example, at the original resolution, the small-object proportions of VisDrone2019, UAVDT, DOTA, CARPK, and AU-AIR are 26.0%, 12.4%, 10.9%, 0.1%, and 1.2%, respectively. Therefore, 16×16 is more suitable for characterizing an extremely small-object subset and should not be directly adopted as a general criterion for defining small objects. After multiple stages of downsampling, such objects retain only limited spatial information, making them considerably more difficult to detect.

Table 2. Proportion of small objects in the dataset under three thresholds.

Type	Threshold	VisDrone2019	UAVDT	DOTA	CARPK	AU-AIR
Original Resolution	16×16	26.0	12.4	10.9	0.1	1.2
	32×32	60.4	61.7	44.5	2.1	11.7
	40×40	70.5	80.4	60.7	4.4	19.4
640×640	16×16	57.9	24.0	76.4	0.6	15.5
	32×32	85.3	75.0	91.6	10.6	43.5
	40×40	90.8	86.4	94.3	63.6	55.2

Note: The 32×32 threshold follows the COCO evaluation protocol.

When the threshold is relaxed to 40×40 , the proportion of small objects increases across all datasets. The additional instances primarily comprise objects with dimensions satisfying $32 \leq \sqrt{A} < 40$, which are classified as medium-scale objects under the 32×32 criterion. At the original resolution, the small-object proportions of VisDrone2019, UAVDT, DOTA, CARPK, and AU-AIR increase by 10.1, 18.7, 16.2, 2.3, and 7.7 percentage points, respectively. These findings suggest that a larger threshold may incorporate a greater number of medium-scale instances into the small-object category, thereby reducing the category's ability to distinguish detection performance on extremely small objects. Whether changes in the threshold lead to corresponding variations in specific evaluation metrics must be verified in conjunction with detection performance across different object-scale ranges. Since 16×16 and 40×40 represent relatively stringent and permissive operational definitions, respectively, these results demonstrate that the threshold substantially affects both the proportion and the composition of the small-object category. Consequently, 32×32 may serve as a reference standard for cross-dataset comparisons, but it should not be regarded as a universally applicable definition for all UAV-based scenarios.

2.2. Analysis of Major Challenges

The challenges of UAV imagery small-object detection do not stem from a single factor, but from five tightly coupled factors: intrinsic object limitations, imaging degradation, insufficient adaptation of detector architectures, scenario uncertainty in practical missions, and dataset construction as well as annotation limitations. As illustrated in Figure 2, these five dimensions interact with and amplify each other, making the detection task substantially harder than the sum of its individual components. This section analyzes their causes and impacts to provide a problem-oriented basis for the subsequent review of methods.

From a symmetry-informed perspective, the core difficulty of small-object detection in UAV imagery lies in the deviation of model feature representations from the transformation symmetry properties required by an effective detector. In UAV imagery scenes, objects are frequently subject to simultaneous variations in scale, position, pose, viewpoint, and appearance, giving rise to composite geometric transformations such as scaling, rotation, and affine distortion, which impose stringent symmetry constraints on model feature learning. This review provides a clear definition and distinction among three core concepts: robustness refers to the model's fundamental ability to withstand perturbations such as noise and minor displacements while maintaining stable predictions; transformation invariance serves object classification, ensuring that category predictions remain constant under variations in scale, viewpoint, and pose; transformation equivariance applies to object localization, guaranteeing that the predicted bounding boxes adapt consistently with the geometric transformation of the target. Together, these three constitute the fundamental symmetry priors for models to handle dynamic scenes. Conventional detectors rely solely

on network architectural biases, limited training data, and routine augmentation to adapt to scene variations, achieving only limited transformation adaptability and falling short of satisfying strict invariance and equivariance constraints. This symmetry deficiency is particularly pronounced for small objects with weak features and extremely low pixel occupancy, ultimately resulting in insufficient domain adaptability and severe degradation of cross-scene generalization performance.

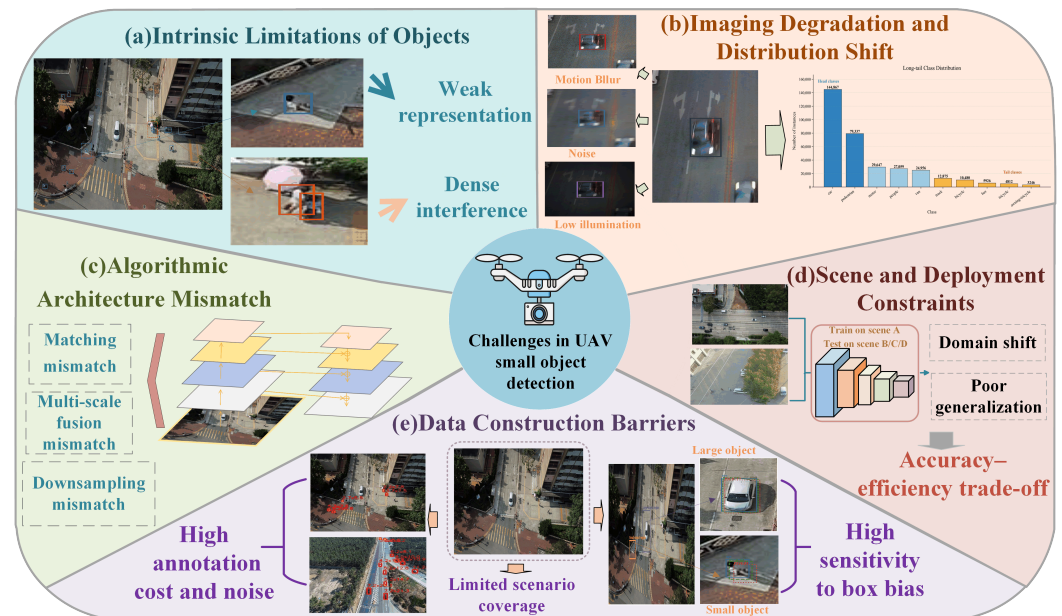


Figure 2. Challenges in small-object detection in UAV imagery.

As illustrated in Figure 2a, small objects in aerial images usually occupy only a very limited number of pixels and often exhibit weak texture, blurred edges, and insufficient discriminative cues. Their apparent scales may change substantially as flight altitude, camera focal length, and viewing geometry vary, even for objects from the same category. During repeated downsampling in deep networks, these already fragile details are further diluted, which easily leads to instability in both classification and localization. Although multiscale feature pyramids and feature fusion can alleviate this problem, they generally provide approximate scale robustness rather than exact scale equivariance across the continuous range of scales encountered in UAV imagery. In addition, densely distributed aerial objects often overlap with each other and are embedded in cluttered backgrounds, making instance separation more difficult and reducing the effectiveness of conventional non-maximum suppression (NMS), thereby increasing the risks of missed detections and duplicate detections.

Figure 2b illustrates that UAV imagery can be affected by platform vibration, attitude changes, illumination variation, and harsh weather. Changes in platform attitude and view-point variation may introduce geometric transformations, such as rotation, scale variation, and perspective distortion. Meanwhile, vibration and adverse environmental conditions can cause motion blur, localized defocus, intensified noise, and uneven illumination, further diminishing the visibility of small objects. These degradations weaken the visual cues on which the detector relies and reduce the robustness to geometric and photometric transformations learned from the training data. When texture and edge details are lost, detectors are more likely to be distracted by the background, resulting in degraded classification confidence, localization accuracy, and detection stability in complex environments. In practical applications, this challenge is often further aggravated by foreground–background confusion and class imbalance.

Figure 2c further shows that most existing detection frameworks are built for general natural-image settings, so they often struggle when applied directly to UAV imagery small-object detection tasks. Problems such as excessive downsampling, weak interaction across scales, and imperfect sample matching can easily cause small-object information to be lost during feature extraction and training. Moreover, stride-based downsampling and discrete feature sampling may reduce the stability of localization under small translations and scale changes, particularly when an object is represented by only a few pixels. This does not imply that conventional detectors completely lack transformation robustness; rather, their robustness is limited and often depends on the training scale range, feature resolution, and effectiveness of multiscale fusion. For this reason, UAV imagery small-object detection usually demands better feature retention, more effective scale modeling, and greater training stability than standard object detection tasks.

As shown in Figure 2d, UAV missions in real-world settings often differ greatly in flight altitude, camera viewpoint, background complexity, object density, and surrounding conditions. These variations jointly alter object-scale, orientation, appearance, and contextual relationships, creating distribution shifts that exceed the transformation range represented in the training data. Consequently, models may fail to maintain stable detection performance across scenarios, even when the object categories remain unchanged. In addition, onboard deployment has to deal with limited computing power, restricted energy supply, bandwidth constraints, and latency requirements. For this reason, it is still difficult for practical UAV systems to achieve a good balance between detection accuracy, robustness, and inference efficiency.

Finally, high-quality annotated data largely determine the performance ceiling of detection models. As shown in Figure 2e, existing UAV imagery datasets often suffer from high annotation cost, annotation noise, limited scene diversity, and insufficient cross-domain coverage. For small objects in particular, even slight deviations in bounding-box annotation may significantly affect intersection over union (IoU), training supervision quality, and evaluation reliability. These issues further restrict stable model optimization and reduce generalization ability across datasets and scenarios.

2.3. Object-Scale Distribution in Mainstream Aerial Datasets

To analyze the object-scale distribution across datasets, we computed the effective size of each annotated object using its bounding-box area. We define the object size as $\sqrt{A}$, where 'A' is the bounding-box area in pixels. This formulation is adopted because the square root of area provides a more intuitive one-dimensional representation of object-scale and is consistent with common practice in scale-aware detection analysis. The area 'A' is calculated according to the annotation format used by each dataset. For YOLO-format datasets, namely, VisDrone2019 and UAVDT, the normalized bounding-box width and height are converted into pixel dimensions using the corresponding original image dimensions. For CARPK, whose annotations provide pixel-coordinate bounding boxes (x1,y1,x2,y2), the pixel width and height are computed directly. For DOTA, the polygon annotations are converted into horizontal bounding box (HBB) before the corresponding dimensions are calculated. For AU-AIR, the width and height are extracted directly from the COCO-style JSON annotations. For DOTA, objects marked as difficult (difficult = 1) are excluded to avoid biased scale statistics.

For all datasets, annotations with zero or negative width/height are filtered out. The statistics are computed using all annotated objects from the training splits of each dataset (or the full dataset when training splits are not separately defined) to ensure sufficient sample sizes. The analysis is performed by aggregating all object categories within each dataset, as the primary goal is to characterize dataset-level scale bias rather than category-

specific scale patterns. There are differences in scale distribution across different categories, for example, vehicle targets such as buses and trucks typically occupy a larger area than pedestrians and cyclists. However, an aggregated statistical perspective is sufficient to meet the research objectives of this review, namely, to provide an overall assessment of the difficulty of datasets for small-object detection tasks. Although an in-depth analysis by category is valuable for the design of category-specific detectors, it falls outside the scope of this review and is left for future work. All calculations are performed on the original annotations at each dataset's native resolution, before any resizing or preprocessing, as stated in the Figure 3 caption.

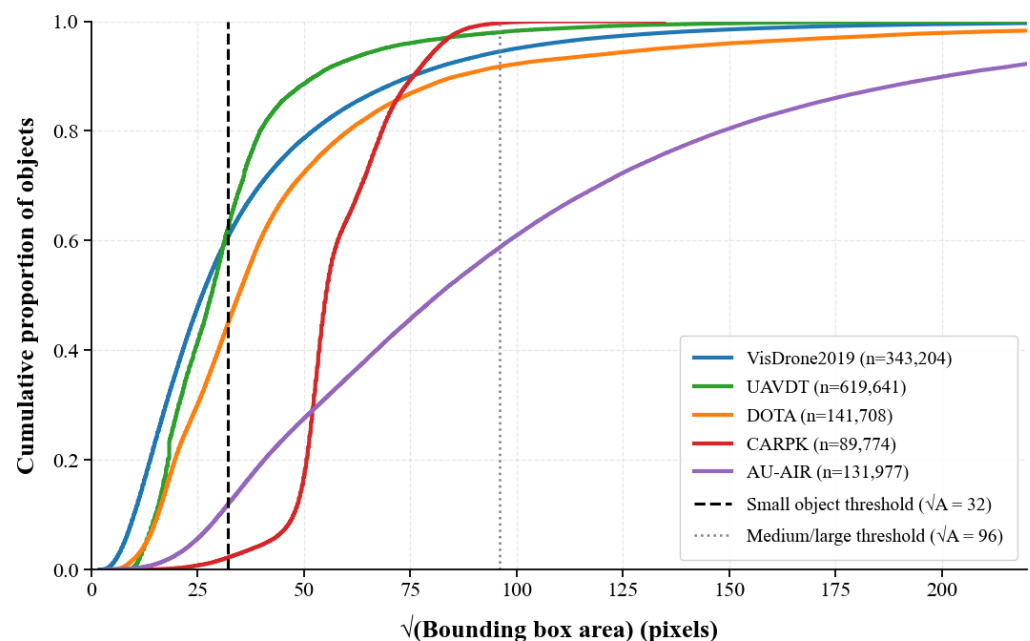


Figure 3. Cumulative distribution of object sizes across five aerial datasets. Statistics are computed at the original native resolution of each dataset before any resizing or preprocessing. The vertical dashed line marks the small-object threshold ($\sqrt{A} = 32$ pixels, corresponding to area $< 32 \times 32$ pixels). Datasets with curves closer to the left have a higher proportion of small objects.

The cumulative distribution of object sizes reveals substantial cross-dataset variability in scale distribution. Datasets such as VisDrone2019, UAVDT, and DOTA exhibit a strong skew toward small objects, with cumulative proportions at the 32-pixel threshold reaching approximately 60.4%, 61.7%, and 44.5%, respectively. CARPK and AU-AIR show significantly lower small-object proportions, at approximately 2.1% and 11.7%, respectively, reflecting their distinct acquisition scenarios and object categories. These results demonstrate that while small objects constitute a substantial proportion in most mainstream UAV imagery datasets, the degree of scale bias varies considerably across different data sources. This also implies that a detector's overall performance is heavily influenced by its ability to handle small instances, particularly when applied to datasets with high small-object proportions.

During the experimental phase of Section 5, all input images were uniformly resized to 640×640 . Since resizing alters the effective pixel area of each object, it also changes the distribution of small objects actually encountered by the detector. Consequently, the proportions of small objects reported in Figure 3 reflect the intrinsic characteristics of the datasets at their native resolutions rather than the object-scale distributions faced by the detector during evaluation. To accurately assess the detection difficulty of each dataset, both

the original resolution and the effective object scales after resizing should be considered during evaluation.

This analysis further indicates that aerial object detection is not merely a direct extension of generic object detection. It is a task characterized by a heavy predominance of small objects in most scenarios, further complicated by scene complexity and deployment constraints. For this reason, the following review focuses on how existing methods differ in their ability to represent small objects, adapt to scale variation, and balance scene robustness with deployment efficiency.

3. Recent Advances in Small-Object Detection in UAV Imagery

In recent years, to address the key challenges of small-object detection in aerial imagery from UAV, most efforts have been directed toward redesigning the detector architecture, key module optimization, and scenario adaptation strategies. To provide a clear and systematic overview of the current research landscape, as shown in Figure 4, this review presents the field from three aspects: the detection paradigm layer, the key techniques layer, and the scenario constraints layer.

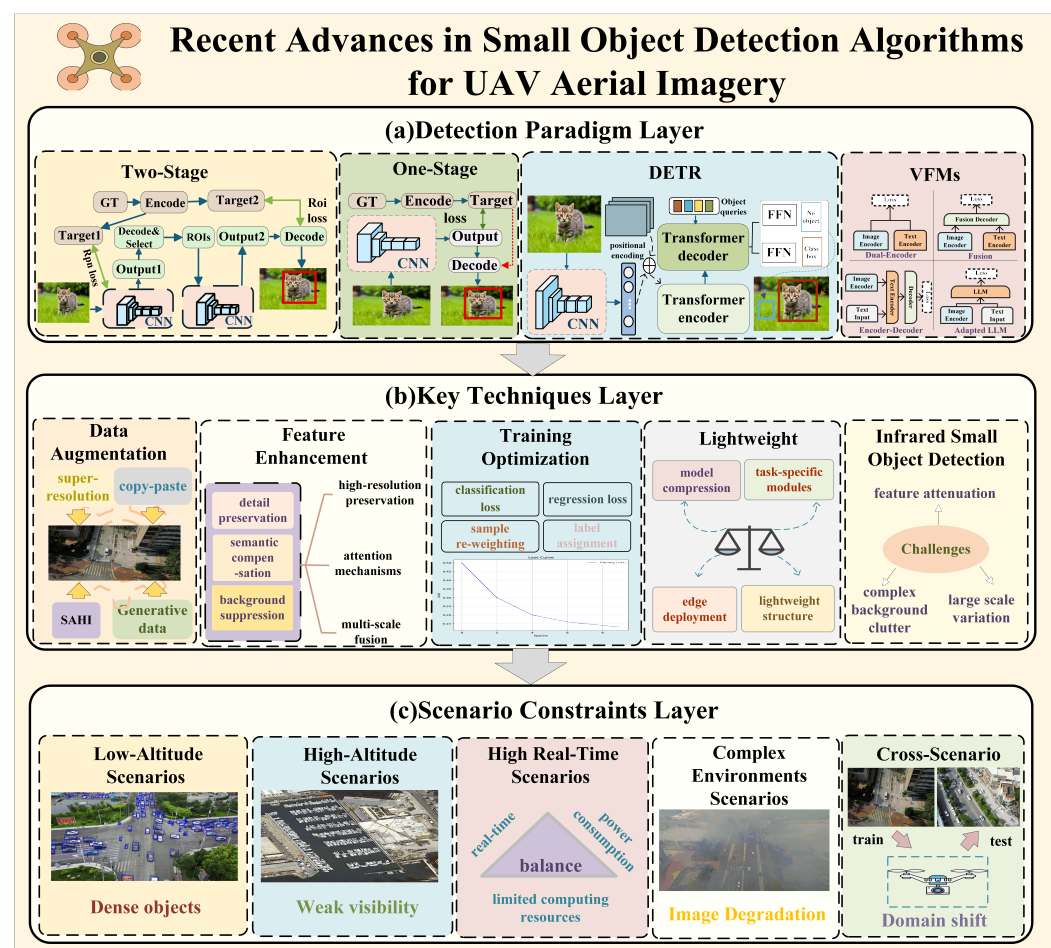


Figure 4. Three-tier review framework.

3.1. Detection Paradigm Layer

At the detection paradigm layer, we compare existing frameworks across four fundamental design dimensions: region proposal generation, feature representation, sample assignment, and prediction decoding. As categorized in Figure 4a, current approaches can be divided into three core detection architectures: two-stage detectors, one-stage detectors, and Transformer-based detectors, each embodying distinct trade-offs between accuracy

and efficiency. In addition to these conventional architectures, Visual Foundation Models (VFM) have recently emerged as a complementary semantic enhancement layer. Rather than functioning as standalone detectors, VFMs provide auxiliary semantic priors, open-vocabulary reasoning, or zero-shot adaptability that can be integrated with any of the three core architectures to improve detection performance. The following subsections first examine the three core architectures and then discuss the role of VFMs as an emerging enhancement component.

3.1.1. Two-Stage Detection Paradigm

Two-stage detectors first generate region proposals and then perform classification and bounding-box refinement on those proposals. This paradigm originated with R-CNN [24], which introduced deep convolutional features to object detection. Fast R-CNN [25] later added RoI Pooling and multi-task learning, while Faster R-CNN [26] replaced handcrafted proposals with a learnable Region Proposal Network (RPN), establishing the standard two-stage framework. In UAV imagery, Xu et al. [27] were among the first to apply Faster R-CNN to low-altitude vehicle detection, paving the way for subsequent aerial studies.

The main advantage of two-stage detectors for UAV small-object detection is their proposal-filtering and refinement mechanisms, which in many conventional settings provide more accurate localization and stronger background robustness compared to one-stage detectors [28]. However, this advantage is not universal; it depends on factors such as the quality of the region proposal network, the feature resolution, and the training data distribution. This is especially valuable for small, partially occluded, or objects in cluttered scenes. As a result, Faster R-CNN and its variants have long served as strong baselines in aerial detection benchmarks.

However, this high localization accuracy comes with high computational cost. The sequential pipeline introduces significant latency and resource demands, making two-stage detectors difficult to deploy on resource-constrained UAV platforms that require inference efficiency. Thus, despite their accuracy advantages in offline settings, their use in real-time UAV applications remains limited by the precision–efficiency trade-off.

3.1.2. One-Stage Detection Paradigm

One-stage detectors omit the explicit proposal generation step and directly predict object classes and locations from dense feature maps, making inference much more efficient. According to whether predefined anchor boxes are used, one-stage detectors can be broadly divided into two categories: anchor-based methods and anchor-free methods.

Anchor-based one-stage detectors, represented by SSD [29], RetinaNet [30], and classic anchor-based YOLO variants [31,32], predefine numerous anchor boxes with varying scales and aspect ratios on dense feature maps, reformulating detection as per-anchor classification and offset regression. SSD pioneered multi-scale prediction across feature maps of different resolutions, enabling scale-adaptive detection without explicit proposals. RetinaNet introduced the focal loss to alleviate the severe foreground-background imbalance inherent in dense anchor prediction, shifting training focus toward hard examples. Subsequent YOLO versions further integrated FPN and PAN structures to achieve a well-regarded balance between speed and accuracy. Despite these advances, anchor-based methods suffer from three fundamental limitations: manually engineered hyperparameters (anchor scales, aspect ratios, and matching thresholds) that generalize poorly across domains; diluted fine-grained features during repeated downsampling, which undermines small-object localization; and heavy reliance on NMS post-processing due to massive redundant predictions. As a result, later research gradually shifted toward more flexible

anchor-free detection methods to reduce the constraints imposed by predefined anchor design and sample matching.

Anchor-free one-stage detectors, including CenterNet [33], FCOS [34], and recent YOLO variants [35,36], eliminate anchor design by adopting alternative regression formulations. CenterNet models objects as center points with size and offset regression, simplifying the pipeline and removing anchor-dependent heuristics, yet it often suffers from center ambiguity in extremely small or densely packed aerial scenes. FCOS reformulates detection as per-pixel regression of distances to the four bounding-box sides, offering greater flexibility across scale variations. However, its performance remains sensitive to feature resolution and scale assignment, especially when small objects contain only weak visual cues. Recent YOLO versions, such as YOLOv8 [37] and later models, have adopted anchor-free prediction while retaining the framework's efficiency, simplifying label assignment and decoding for real-time deployment.

One-stage detectors are widely favored in UAV applications for their architectural simplicity, inference speed, and deployment convenience. Nevertheless, their application in dense small-object scenarios remains constrained by a coupled triad of challenges: fragile positive-negative sample assignment, insufficient fine-grained feature representation, and persistent scale misalignment. Collectively, these issues manifest as elevated false-positive rates and missed detections in complex aerial backgrounds, revealing that the efficiency gains of one-stage paradigms often come at the cost of discriminative power under challenging conditions such as dense small-object scenes. However, this trade-off is not absolute; recent advances in anchor-free designs and improved training strategies have substantially narrowed the performance gap, and the actual behavior depends on the specific architecture, feature resolution, and training configuration.

3.1.3. Transformer-Based Detection Paradigm

Transformer-based detectors have emerged as a prominent research direction in object detection, fundamentally departing from anchor-based paradigms by framing detection as a set prediction problem that eliminates manual anchor design and NMS post-processing. DETR [38] was the first to introduce this framework, using Hungarian matching to build one-to-one correspondence between predictions and ground-truth labels, making end-to-end detection possible.

The original DETR's reliance on global self-attention over the full feature map incurs prohibitive computational costs and inefficient multi-scale feature exploitation, rendering it impractical for resource-constrained UAV platforms. Subsequent advances have sought to mitigate these bottlenecks. Deformable DETR [39] introduced deformable attention to sparsely sample key spatial locations, reducing complexity while enhancing multi-scale interaction, which is a particularly relevant improvement for small objects. DINO [40] further refined optimization through contrastive denoising training and hybrid query selection, boosting both convergence and accuracy. Meanwhile, RT-DETR [41] prioritized runtime efficiency by hybridizing Transformer global context with CNN-based multi-scale fusion, narrowing the speed gap with one-stage detectors.

Despite these advances, Transformer-based detectors remain challenging to apply to small-object detection in UAV imagery. Their strong reliance on global semantic representations can, in some cases, come at the cost of local fine-grained details, which are especially important when objects are small and densely distributed. This limitation is not inherent to all Transformer-based detectors; it is influenced by the design of the attention mechanism, feature resolution, and the availability of multi-scale representations. Some recent variants have explicitly addressed this issue through sparse attention mechanisms. In addition, they also tend to require more careful optimization strategies, longer training schedules,

and higher deployment costs than conventional one-stage detectors. Therefore, under current UAV deployment conditions, Transformer-based detectors are better regarded as a complement to traditional detection methods than as a full replacement.

3.1.4. VFMs as Complementary Enhancement Components

Beyond the three core detection architectures discussed above, VFMs have attracted increasing attention as a complementary enhancement component for UAV small-object detection. Unlike conventional detectors that are trained end-to-end for closed-set detection, VFMs are typically pre-trained on large-scale data via self-supervised, weakly-supervised, or multimodal learning, and they can be adapted to downstream tasks through prompting or fine-tuning [42]. Their primary value lies not in replacing existing detectors, but in providing auxiliary semantic priors, zero-shot or few-shot transferability, and open-vocabulary recognition capabilities that can be integrated with two-stage, one-stage, or Transformer-based detectors to enhance feature representation or compensate for limited training data [43,44].

However, directly transferring general-purpose VFMs to UAV small-object scenarios still faces significant challenges. Most general-purpose foundation models are pre-trained primarily on natural images, where training samples usually contain medium-to-large-scale objects, whereas UAV imagery objects often occupy only a few pixels and suffer from weak texture, blurred boundaries, dense distribution, severe occlusion, and strong background clutter [45]. This scale discrepancy, compounded by domain shift, severely undermines the transferability of generic visual representations [46]. Beyond this macro-level mismatch, each VFM family exhibits task-specific deficiencies that are particularly detrimental in aerial scenes. Vision–language models such as CLIP primarily perform image-level semantic alignment and have limited region-level localization capabilities, leading to coarse spatial responses and unstable small-object localization in dense aerial scenes [47,48]. Promptable segmentation models like SAM, while possessing strong interactive segmentation capabilities, rely heavily on the quality of input prompts, and their segmentation ability is not equivalent to bounding-box-level object detection performance [49,50]. Self-supervised models like DINO offer relatively fine-grained spatial features, yet still demand task-specific adaptation through detection heads, feature fusion, or knowledge distillation to bridge the gap to object detection [51,52]. Compounding these functional limitations, VFMs typically carry prohibitive parameter sizes and inference costs, directly conflicting with UAV platforms' stringent requirements for low latency, low power consumption, and real-time operation [53]. Future research should focus on aerial domain adaptation, small-object-aware prompt learning, parameter-efficient fine-tuning, knowledge distillation, and lightweight integration for onboard deployment so as to fully leverage the generalization capability and open-vocabulary recognition advantages of foundation models while maintaining the localization accuracy and inference efficiency of specialized detectors.

Based on the above analysis, the three core architectures exhibit different general tendencies: two-stage detectors typically pursue accuracy at the cost of efficiency; one-stage detectors generally favor speed but may sacrifice discriminative power in dense scenes; Transformer-based detectors often excel at global context yet may struggle with local details under certain configurations. It should be emphasized that these are general tendencies rather than absolute rules. The actual performance of any detector family is highly dependent on the specific architecture variant, training strategy, input resolution, dataset characteristics, and model scale. These three core architectures constitute the core methodological backbone of UAV small-object detection. VFMs are best understood as a complementary enhancement layer rather than a standalone detection architecture. They can be integrated with any of the three core detector types to provide additional

semantic priors, open-vocabulary reasoning, or domain adaptability. Future research should focus on aerial-domain adaptation, parameter-efficient fine-tuning, knowledge distillation, and lightweight integration for onboard deployment so as to fully leverage the generalization capability of foundation models while maintaining the localization accuracy and inference efficiency of specialized detectors. Table 3 compares the core ideas, strengths, and limitations of these four detector categories.

Table 3. Comparison of three core detection architectures and an emerging complementary enhancement component.

Category	Core Design	Advantages	Limitations
Two-Stage	Region proposal and RoI refinement	Accurate localization; robust in complex scenes	High computation; slow inference
One-Stage	Dense feature-map prediction	Fast inference; simple design; easy deployment	Limited fine-grained features; prone to missed detections in dense small-object scenes
Transformer-based	Query-based set prediction	Strong global modeling; unified pipeline	High training cost; insufficient small-object details
VFM	Large-scale pre-training	Zero-shot generalization and open-vocabulary recognition	Large number of parameters and high inference cost

Note: VFMs are not standalone detection architectures but complementary enhancement layers that can be integrated with any of the three core detector types. The advantages and limitations listed in this table reflect general tendencies observed in the literature and in our experiments. They are not absolute and may vary depending on specific architecture designs, training configurations, datasets, and deployment constraints.

3.2. Key Techniques Layer

This section focuses on how existing methods address the core technical challenges of small-object detection in UAV imagery. As shown in Figure 5, current studies can generally be divided into five main technical directions.

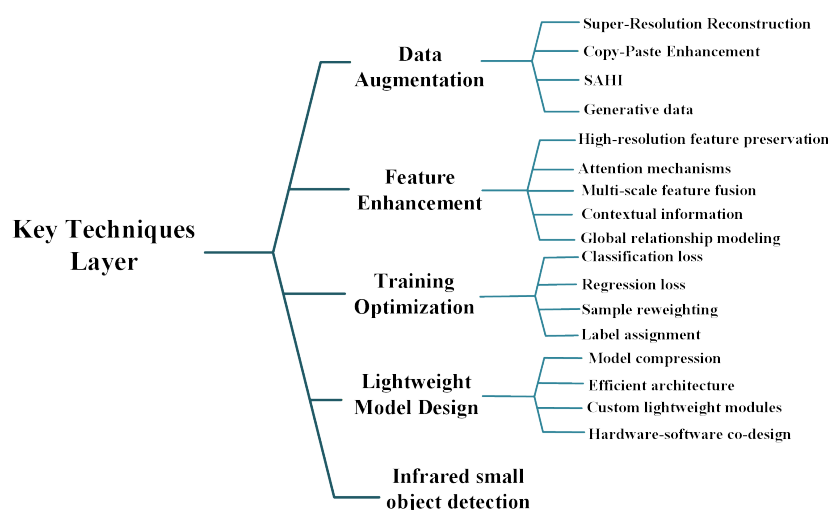


Figure 5. Five major technical routes.

3.2.1. Data Augmentation

Data augmentation addresses the inherent scarcity and scale bias of UAV small objects from the data perspective. Because aerial small objects often occupy only a few pixels and are easily affected by scale compression, background clutter, and insufficient sample diversity, directly improving the training data has become an important way to enhance detector

robustness. Existing studies have mainly developed four representative augmentation strategies as follows.

Super-resolution reconstruction improves small-object discernibility by restoring edge, texture, and structural details from low-resolution imagery. Research in this direction broadly falls into two categories: standalone pre-processing before detection [54] and joint optimization of super-resolution with the detection network via collaborative reconstruction and adaptive feature fusion [55]. Wang et al. [56] further integrated lightweight super-resolution with attention mechanisms and feature pyramids for infrared small-object scenarios. While super-resolution enhances boundary clarity and object-background separability, it introduces pseudo-details, misalignment between reconstruction and detection objectives, extra computational overhead, as well as limited cross-domain robustness.

Copy-paste augmentation alleviates sample scarcity and class imbalance by duplicating small instances and pasting them into new backgrounds [57]. For UAV datasets with severe long-tail distributions, Li et al. [58] proposed dynamic class-balanced copy-paste (DCC), which resamples tail classes around cluster centers and maintains dynamic memory banks per class. This strategy is straightforward and framework-compatible, yet it risks introducing spatial-semantic inconsistencies, unrealistic layouts, and training distribution shifts if contextual coherence is not carefully preserved.

Slicing Aided Hyper Inference (SAHI) augmentation improves recall by dividing high-resolution images into multiple local patches, allowing small objects to occupy a larger proportion of each patch [59]. Combined with real-time detectors like RT-DETR-X [60], SAHI has demonstrated clear performance gains on VisDrone. However, it requires multiple inference passes and additional merge operations, entailing a trade-off between accuracy improvement and computational overhead in practical deployment.

Diffusion-based generation has recently emerged as a powerful paradigm for synthesizing diverse, realistic, and semantically controllable training samples [61]. Fang et al. [62] proposed a controllable diffusion pipeline with CLIP guidance for object detection augmentation, while Wu [63] further refined this approach with weighted fine-tuning for small-object regions, automated scene-prompt generation, and CLIP-based similarity screening. Despite their ability to enrich rare categories, diffusion-based methods remain constrained by generation costs, annotation consistency, and residual domain shifts in synthetic samples.

Current work on data augmentation has progressively shifted from traditional heuristic transformations toward task-oriented and generative strategies moving from passive sample enrichment to active data synthesis. Despite their diversity and increasing sophistication, these methods remain data-level compensation strategies. They improve input quality and sample distribution but cannot reverse the irreversible dilution of fine-grained details during deep feature extraction. Consequently, subsequent research increasingly shifts toward discriminative enhancement at the feature level, aiming to achieve a tighter coupling between input compensation and feature reinforcement.

3.2.2. Discriminative Feature Enhancement

Discriminative feature enhancement aims to improve the distinctiveness and robustness of small-object representations through network refinement or auxiliary mechanisms. Its central challenge lies in balancing four conflicting objectives: detail preservation, semantic compensation, background suppression, and global relation modeling. Existing studies can be broadly divided into five progressively related directions.

High-resolution feature retention. To alleviate detail loss caused by repeated downsampling, existing studies mainly explore continuous high-resolution representations [64], optimized downsampling and low-level feature retention [65,66], cross-layer high-resolution

detail aggregation [67], and high-resolution prediction branches [68]. The main advantage of this route is its ability to preserve spatial structure and local texture. However, in UAV imagery, many small objects occupy only 10–20 pixels, and their shallow features are often dominated by weak and low-frequency signals. Excessive high-resolution retention may therefore introduce redundant background noise instead of truly useful object details. Future work should pay more attention to lightweight design, adaptive scale modeling, and the joint optimization of detail enhancement and noise suppression.

Multi-scale feature fusion is one of the most important directions in feature enhancement. Relevant methods have gradually evolved into various more sophisticated derivative structures starting from Feature Pyramid Network (FPN) [69] and Path Aggregation Network (PAN) [70]: scale-selective feature sharing mechanism [71], cross-layer aggregation structure [72], bidirectional dense fusion scheme [73], spatial correction module with full-scale interaction [74], and lightweight collaborative framework [75,76]. Despite these advances, multi-scale feature fusion in UAV imagery still faces an important problem. Variations in flight altitude can cause significant scale changes, which often weaken the spatial correspondence assumed by FPN/PAN-style architectures and lead to misalignment during cross-layer fusion. Future work must prioritize feature alignment, geometric consistency, and background suppression under extreme scale variation.

Attention mechanisms improve small-object detection by emphasizing informative responses and suppressing irrelevant background noise [77]. Although channel attention is easy to integrate into convolutional networks, it lacks explicit spatial modeling. By comparison, spatial attention is more effective for highlighting key regions and improving localization, though it may be less effective in capturing channel-wise semantic differences. Self-attention [78] improves global context modeling by capturing long-range dependencies, which is helpful for detecting small objects in cluttered scenes, although it usually comes with higher computational cost and weaker preservation of local details. In addition, hybrid designs, such as interactive feature pyramid network (IFPN) [79], channel-spatial cross-correlation [80], and multi-scale weighted fusion [81] are attempts to combine complementary strengths. Although much progress has been made in the research on attention mechanisms, it is still necessary to clarify that the essence of attention lies in filtering valid information. Small objects contain very limited extractable information, and critical feature signals are highly likely to be lost during the filtering process. Further improvements are still needed in efficiency, fine-detail preservation, and robustness in complex scenes.

Contextual modeling therefore introduces neighborhood cues, scene layout, and semantic dependencies to compensate for weak object signals [82]. Early studies mainly enlarged the receptive field through dilated convolutions and spatial pyramid pooling. Current research, such as MS-MRFNet [83], still mainly enhances context at the feature-fusion level, while explicit interactions between local details and global contextual cues remain insufficiently explored. To improve context utilization, later studies moved toward attention-guided local-global contextual modeling, such as MSA-DETR's local-global joint modeling [84], semantic associations and spatial dependencies modeling [85], and GA-FPN's global-aware selective fusion [86]. The core difficulty is that contextual modeling relies on reliable object-background correlations, yet UAV small objects are often sparsely scattered across complex backgrounds where such correlations are too weak to establish robust statistical associations. This renders context exploitation inherently fragile, particularly in scenes with high background variability.

Global relation modeling focuses on capturing long-range dependencies among objects, backgrounds, and different image regions, which helps overcome the local limitations of convolutional receptive fields. Tao et al. [87] proposed a multi-level feature enhancement and global relation modeling framework for small-object detection. Wu et al. [88]

further introduced a Mamba-based vision module to improve global dependency modeling with linear complexity. To better address dense scenes and large scale variation, Hu et al. [89] enhanced spatial correlation modeling through hybrid spatial–frequency feature extraction and a position relation decoder that integrates IoU and relative angle priors. Tong et al. [90] improved cross-scale collaboration by dynamically adjusting fusion weights across different scales, and Zhou et al. [91] introduced a context-spatial alignment module based on RT-DETR to correct both contextual inconsistency and spatial misalignment. Although global attention treats all spatial positions equally, UAV small objects occupy sparse, low-signal regions that demand prioritized focus. This conflicts fundamentally with the efficiency–accuracy trade-off, particularly under onboard resource constraints.

Table 4 summarizes the advantages, limitations, and applicable scenarios of these directions. Discriminative feature enhancement has evolved from a local detail compensation method into a global-oriented multi-dimensional enhancement collaborative framework. Future research should further incorporate aerial-scene priors, object-scale distribution patterns, and deployment constraints to better balance fine-grained representation, background suppression, and inference efficiency.

Table 4. Summary of five discriminative feature enhancement methods.

Technical Approach	Advantages	Limitations	Applicable Scenarios
High-resolution feature preservation	Preserves spatial details and local textures	Prone to introducing shallow background noise	Low-resolution scenarios where detail loss is pronounced
Multi-Scale feature fusion	Reduces cross-level semantic gaps	Spatial misalignment; structural complexity is relatively high	Scenarios with significant scale variation
Attention mechanisms	Highlights key features and suppresses background interference	Limited global modeling or high computation cost	Complex backgrounds and weak object saliency
Context modeling	Introduces surrounding semantic cues and object-background relationships	May introduce irrelevant context or background bias	Scenes with occlusion, clutter, or weak local appearance
Global relation modeling	Captures long-range dependencies and improves discrimination	High training and deployment cost	Dense scenes with strong object relations

3.2.3. Loss Function and Training Strategy Optimization

Optimizing the loss function and training strategy is another important route for improving UAV imagery small-object detection. Compared with natural-scene detection, aerial small objects are more likely to suffer from extreme foreground–background imbalance, high sensitivity to localization errors, and inconsistency between classification and regression supervision during training. Accordingly, existing studies have evolved from simply replacing individual loss terms to a more collaborative optimization framework involving classification loss, regression loss, sample re-weighting and optimization dynamics, and label assignment.

Classification loss in UAV scenarios faces a severe information scarcity bottleneck. Small objects occupy few pixels under complex backgrounds, dense layouts, weak textures, and high inter-class similarity, causing training to be dominated by easy negative samples while hard, informative ones remain under-emphasized. Standard cross-entropy may thus be insufficient. To address this, Ahmad et al. [92] proposed Dynamic Focal Loss that adapts sample contributions based on data distribution and prediction con-

fidence. Liu et al. [93] enhanced hard-sample supervision via dynamic weighting. For fine-grained remote sensing detection with hierarchical labels and high inter-class similarity, Cheng et al. [94] introduced balanced hierarchical contrastive loss with decoupled learning for fine-grained remote sensing. Di et al. [95] incorporated the focal-learning idea into classification supervision to improve robustness in degraded scenes. Despite these refinements, a fundamental limitation persists: classification loss reweighting cannot create discriminative information that does not exist in the feature space. For objects occupying fewer than 20 pixels, even optimal weighting cannot compensate for the sheer absence of distinctive texture or shape cues.

Compared with classification, bounding-box regression loss is more task-specific in aerial small-object detection. The reason is that small objects are far more sensitive to localization deviations: An offset that is tolerable for large objects may cause sharp IoU fluctuations for small ones, thereby destabilizing regression supervision. To alleviate this problem, Wang et al. [96] modeled bounding boxes as two-dimensional Gaussian distributions and proposed the Normalized Wasserstein Distance (NWD), reducing IoU's over-sensitivity to minor offsets. Qu et al. [97] integrated NWD with GIoU in AMMFN for UAV remote sensing. More recently, Guan et al. [98] further introduced Gaussian Combined Distance (GCD) to address scale invariance limitations. However, a deeper issue remains: NWD/GCD measures distributional similarity, whereas IoU measures geometric overlap. These two objectives are not inherently aligned. These are not inherently aligned. Optimizing for distributional proximity does not guarantee precise boundary alignment, and the tension between these two objectives becomes more acute at smaller scales. Thus, the shift from IoU to distribution-based metrics, while valuable, trades one form of sensitivity for another rather than resolving the underlying regression ambiguity.

Although sample reweighting, hard sample mining, and gradient stabilization can address the unequal contribution of different training samples, this issue is amplified by image blur, dense occlusion, and inconsistent annotations. Uniform supervision tends to weaken learning signals for hard samples and aggravate localization instability. Liao et al. [99] introduced the WIoU method, which applies differentiated penalties to samples of varying quality. Han et al. [100] integrated difficulty-aware gradients into localization. Sun et al. [101] proposed Uncertainty-Aware Gradient Stabilization (UGS), which reformulates object localization as a classification task and uses non-uniform label discretization to generate bounded, confidence-driven gradients. Cheng et al. [102] enhanced training stability in dense scenes through a dynamic non-monotonic focusing mechanism and auxiliary bounding boxes. These methods share a common constraint: they operate within fixed network capacity, adjusting how training signals are distributed rather than increasing the amount of usable information. The fundamental cause of small-object learning difficulty is feature-level information loss during downsampling. Therefore, no matter how sophisticated the reweighting strategy is, it cannot recover the feature information that has been permanently lost.

Label assignment critically influences small-object performance. Static assignment strategies based on fixed intersection over union (IoU) thresholds often lead to an insufficient number of positive samples or poor matching performance. To address the limitations of IoU in handling minor offsets and cases where boxes do not overlap but are similar, recent research has primarily explored adaptive assignment thresholds [103], the reconstruction of similarity metrics [104], and dynamic sample assignment [105]. Shi et al. [106] further proposed a dynamic adaptive label assignment (DALA) framework. This framework integrates similarity distance, positive sample ratio allocation, and dynamically weighted loss without extra hyperparameters. Whether the threshold is fixed or adaptive, the Intersection over Union (IoU) threshold itself is arbitrarily set by humans. For extremely small objects, even "dynamic" thresholds are merely empirically defined artificial rules. When

ground-truth bounding boxes only occupy a handful of pixels, any threshold-based method cannot fully define what constitutes a valid match. This indicates that disruptive alternative approaches such as point-based detection schemes and set prediction architectures deserve more research attention.

3.2.4. Model Lightweight Design

Onboard UAV platforms impose stringent constraints on computing power, memory, and energy budgets, rendering most general-purpose detection networks unsuitable for real-time deployment. Unlike generic lightweight tasks, UAV imagery small-object detection faces a unique and fundamental tension: the compression of lightweight models will lead to loss of fine-grained features of objects. Accordingly, existing studies have gradually formed four main routes: model compression, lightweight structural design, task-specific module optimization, and hardware–software co-deployment.

Model compression reduces network redundancy through knowledge distillation, sparsification, and channel pruning, aiming to minimize performance degradation while lowering complexity. Fan et al. [107] applied distillation atop optimized feature fusion to transfer discriminative knowledge from teacher to student. To reduce the inference burden caused by high-resolution feature preservation, Belmouhcine et al. [108] proposed dual-training distillation where an early-upsampling network guides a late-upsampling detector. Ma et al. [109] further combined response alignment, feature imitation, and structure preservation, and they introduced channel pruning to preserve channels sensitive to small objects. More recently, Wang et al. [110] developed ADD-Net by integrating a dynamic sampling backbone, sparse global context modeling, a dual adaptive feature pyramid, and hybrid knowledge distillation. Despite these refinements, structural contradictions still persist in model compression: knowledge distillation and pruning are essentially means of regulating information loss rather than lossless information preservation. For small objects with inherently sparse discriminative feature signals, any compression operation will further weaken the already faint object responses, and the model's information loss cannot be eliminated through parameter adjustment. Such an inherent trade-off means that optimization schemes relying solely on compression techniques have an upper performance limit.

Lightweight structural design reduces redundancy by redesigning backbones, feature propagation paths, and detection heads. Wang et al. [111] removed high-level branches serving large objects and combined inverted bottlenecks with semantic-detail fusion. Tan et al. [112] proposed a Lightweight Detail-Sensitive Network (LDSNet), which preserves spatial details during downsampling through anti-aliasing smoothing and dual-path feature extraction, and it reduces the computational cost of high-resolution prediction with a deeply decoupled grouped head. Xiao et al. [113], from the perspective of edge deployment, reduced the reliance on group convolutions and element-wise operations with high memory-access cost. Li et al. [114] proposed an enhanced RT-DETR framework AGE-DETR that integrates an attention-gated backbone, an efficient small-object pyramid, and context-aware multi-scale fusion to improve small-object representation while reducing parameters and computational overhead. The current design of lightweight network structures has evolved from single operator replacement to a systematic optimization approach featuring multi-module collaboration. Looking ahead, while lightweight architectures reduce parameter counts and accelerate inference, they must fully account for architectural adaptation constraints imposed by practical deployment environments such as onboard hardware of unmanned aerial vehicles.

Task-specific module optimization tailors lightweight components to UAV-specific degradation issues such as blur, noise, and low contrast. Wu et al. [115] enhanced the feature restoration capability for degraded small objects by introducing a fast deblurring

module, a dual-domain feature aggregation mechanism, and a lightweight multi-scale fusion structure. To further improve detection under low-quality remote sensing conditions, Lu et al. [116] proposed LEGNet with an Edge-Gaussian Aggregation (EGA) module to sharpen weak structural details, suppress noise, and improve foreground–background separation. Similarly, Wang et al. [117] proposed a dual-domain attention mechanism to recover discriminative representations for small objects occluded or blurred. Yu et al. [118] employed adaptive feature selection, multi-layer feature fusion, and partial ghost convolutions to dynamically filter and compress multi-scale information, which reduced the number of parameters while strengthening the representation of key regions and detailed features. Beyond convolution-based lightweight modules, Xia et al. [119] proposed EFSI-DETR, which integrates dynamic frequency-spatial guidance and an efficient semantic feature concentrator to enhance multi-scale feature fusion and fine-grained detail retention. Task-specific lightweight optimization is shifting toward degradation-aware, detail-preserving, and frequency–spatial adaptive designs for UAV small-object detection. Future research should focus on improving their robustness across diverse degradation conditions while maintaining real-time efficiency on onboard platforms.

Hardware-software co-design has become a critical step in transitioning lightweight models from laboratory validation to practical embedded deployment. Unlike optimization approaches that focus solely on the number of parameters and computational complexity, this line of research places greater emphasis on aligning the model architecture with hardware characteristics and balancing computational cost with real-time inference performance. Guo et al. [120] proposed a lightweight adaptation of RT-DETR for embedded UAV platforms, using wavelet convolution, spatial attention, and efficient feature fusion to reduce the computational burden of Transformer-style token modeling. Wu et al. [121] designed efficient downsampling modules and multi-core aggregation structures for edge platforms, while optimizing feature-pyramid fusion to better balance extremely small-object detection and inference efficiency. The lightweight deployment-oriented design is evolving from pure model reduction to joint optimization of structure, computation pattern, and hardware-resource constraints.

In summary, model lightweighting for UAV imagery small-object detection has evolved from early parameter compression to a multi-level optimization system that combines compression, structural reconfiguration, task customization, and deployment synergy. Future research should pay more attention to the joint trade-off among lightweighting, detection accuracy, and deployment efficiency, especially in small-object detail preservation, edge-friendly operator design, the coordinated use of compression and architectural redesign, and end-to-end co-design for embedded platforms.

3.2.5. Infrared Small-Object Detection in UAV Imagery

Visible-light aerial imagery is susceptible to environmental interference, whereas infrared imaging relies on the intrinsic thermal radiation emitted by objects themselves, which is independent of ambient illumination and effectively complements the discriminative information for small and weak objects while suppressing background false alarms [122]. In recent years, small-object detection in UAV-based infrared aerial imagery has evolved from traditional local contrast mechanisms toward network architectures characterized by superior adaptability, higher efficiency, and enhanced global perception capabilities [123]. This field primarily faces three core challenges: feature attenuation, complex background clutter interference, and significant object-scale variation.

Typical CNN architectures lack sufficient multi-scale interaction during the propagation of shallow features to deeper layers, resulting in a severe loss of texture information for dim small objects. Although traditional attention modules can guide the model to focus

more on small-object regions and suppress irrelevant background interference, they inevitably introduce redundant computational overhead [124]. To address this bottleneck, Li et al. designed a lightweight attention module with cross-layer dual-feature fusion, which dynamically adjusts channel and spatial weights to enhance responses, effectively improving the feature representation capability for weak objects [125]. Complex background clutter interference is a critical difficulty in low signal-to-noise ratio (SNR) scenarios. Fang et al. introduced a temporal difference convolution (TDC) re-parameterization module that employs parallel branches at short-, mid-, and long-term temporal scales to capture motion context. By leveraging temporal features as queries for cross-attention computation, this approach effectively suppresses static thermal noise and complex dynamic background interference, significantly reducing the false alarm rate [126]. Similarly, the UAMF model proposed by Cui et al. aggregates effective information across consecutive frames through an adaptive memory fusion mechanism, substantially enhancing the SNR of thermal infrared objects [127]. To address the challenge of significant object scale variation, Nan et al. designed the cross-stage adaptive mixer (CSAM) module, which integrates multi-branch, multi-receptive-field dynamic spatial feature mixing and visual adapters with various kernel sizes, achieving precise multi-scale perception with extremely low computational cost [128].

Recent advances in UAV-based infrared small-object detection have made notable progress in feature enhancement, background suppression, temporal information exploitation, multi-scale adaptation, and lightweight design, effectively compensating for the previous overemphasis on visible-light imagery. Nevertheless, future efforts remain indispensable to further improve cross-scenario generalization, detection capability under extremely low SNR conditions, and false-alarm resistance in complex thermal backgrounds. Furthermore, exploring the end-to-end joint optimization of temporal motion cues, visible-infrared complementary features, and lightweight global modeling is a critical avenue for breaking through the constraints of extreme onboard performance in UAV platforms.

3.2.6. Emerging Research Trends

Based on the five technical directions outlined above, an analysis of existing relevant research reveals six emerging research trends in the field of small-object detection using UAVs. These technological advancements primarily address the five core challenges identified in Section 2.2.

Small-object-specific feature preservation. Because repeated downsampling can irreversibly weaken the fine-grained cues of pixel-sparse targets, recent methods increasingly prioritize shallow, high-resolution, and frequency-sensitive representations. Representative directions include dedicated small-object detection heads [64], detail-preserving backbone designs [66], and frequency-domain feature enhancement [89]. For example, FDL-YOLO introduces frequency dynamic convolution to improve sensitivity to high-frequency object details, while CY-DETR employs a small-object-oriented feature pyramid to preserve spatial information through adaptive feature enhancement. These methods primarily address Challenge 1 by retaining boundary, texture, and localization cues that are easily lost during deep feature extraction. They may also improve Challenge 3 by adapting the representation to the object-scale distribution of UAV imagery. However, high-resolution and frequency-sensitive features may amplify background noise and increase memory access and computational cost.

Adaptive scale selection and cross-level alignment. Fixed feature-pyramid structures are often insufficient for UAV imagery because changes in flight altitude and imaging resolution produce substantial variations in object-scale. Recent work therefore moves beyond uniform multi-scale fusion toward scale-selective and input-adaptive mechanisms.

Representative approaches include adaptive fusion weights in cross-scale detection transformers [90], scale-selective feature pyramids [71], and context-spatial alignment modules that reduce semantic and spatial misalignment [91]. MAFSNet, FDL-YOLO, and RSMamDet further illustrate the use of selective gating, high-resolution feature levels, and dimension-aware cross-scale integration. These developments mainly address Challenges 1 and 4 by improving the preservation of weak object cues under changing scales and heterogeneous mission conditions. Nevertheless, dynamic fusion introduces additional architectural complexity, and its effectiveness may depend on whether the training distribution adequately covers the scale variations encountered after deployment.

Domain-aware augmentation and restoration. Recent data-centric research has shifted from generic image transformation toward the generation and restoration of samples that reflect the target UAV domain. Diffusion-based methods use domain adaptation, region-weighted fine-tuning, or language-guided generation to synthesize samples with more realistic object scales, backgrounds, and imaging conditions [61,63]. For example, region-weighted diffusion fine-tuning emphasizes small-object areas, while controllable diffusion pipelines can reduce the domain gap between synthetic and UAV imagery while preserving object annotations. In parallel, deblurring and dual-domain feature aggregation methods jointly recover detection-relevant structures and perform object detection under degraded imaging conditions [115–117]. These approaches primarily address Challenge 2 and Challenge 5 by improving robustness to blur, low contrast, and other imaging degradations while alleviating data scarcity and annotation limitations. However, synthetic data may introduce unrealistic layouts or annotation inconsistencies, and restoration modules may recover visually plausible but detection-irrelevant details.

Efficient Transformer-based and Mamba-based modeling. Recent architectural research aims to retain the global modeling capability of Transformers while reducing their computational and deployment costs. Deformable attention sparsely samples informative spatial locations [39], whereas state-space models such as Mamba provide an alternative route to long-range dependency modeling with approximately linear computational scaling [88,89]. CY-DETR combines Mamba-based state-space modeling with grouped convolution, and RSMamDet replaces quadratic self-attention in RT-DETR with linear state-space scanning and a MobileMamba backbone. DCAM-DETR further explores state-space modeling for RGB–infrared fusion. These methods address Challenge 3 by improving the efficiency of global context modeling and may support Challenges 1, 2, and 4 by strengthening scale-aware, contextual, and multimodal representation. However, efficient global modeling does not automatically guarantee better small-object localization. Its practical benefit still depends on the preservation of local high-resolution details, the quality of multi-scale fusion, and validation on the target hardware.

Foundation-model adaptation. Vision–language and segmentation foundation models are increasingly incorporated into UAV detection through open-vocabulary recognition, prompt-based adaptation, and knowledge distillation [43,44,49,52]. CLIP- and SAM-based approaches can provide semantic priors, novel-category recognition, and zero-shot or few-shot transfer, which are particularly relevant when UAV datasets contain limited annotations or rapidly changing target categories. These methods primarily address Challenge 5 and may also mitigate Challenge 4 by improving adaptation to unfamiliar scenes and domains. However, the transfer of foundation models from natural images to UAV imagery remains constrained by differences in object-scale, texture, viewpoint, and spatial density. Their localization accuracy for extremely small objects, robustness under severe imaging degradation, and compatibility with onboard computation therefore require further validation.

Deployment-aware optimization. Recent lightweighting research has increasingly come to recognize that parameter count and desktop GPU throughput are insufficient indicators of UAV deployment suitability. Current directions include hardware–software co-design [120], efficient downsampling and multi-core aggregation for edge platforms [121], pruning and quantization, knowledge distillation, and hardware-aware architecture search [107]. For example, recent YOLO-based methods combine structured channel pruning, post-training quantization, and distillation to reduce computation while preserving small-object performance. ADD-Net integrates dynamic sampling, sparse context modeling, adaptive feature pyramids, and hybrid distillation, while FDL-YOLO applies structured pruning to reduce redundant channels. These efforts directly address Challenge 3 by balancing representation quality with latency, memory, and energy constraints. They also support Challenge 4 by enabling adaptation to heterogeneous edge platforms, although cross-hardware generalization should not be assumed without direct embedded-device evaluation.

Recent research has moved from isolated detector modifications toward coordinated optimization across data generation, detail preservation, adaptive scale modeling, efficient global reasoning, semantic transfer, and hardware deployment. The central challenge is that these objectives remain coupled: preserving small-object details and improving robustness often increase memory and computational demands, whereas aggressive compression may further weaken already sparse object signals. Future studies should therefore evaluate recent methods using scenario-aligned metrics, including APS, small-object recall, cross-domain retention, degradation robustness, end-to-end latency, memory usage, and energy consumption, rather than relying solely on overall mAP or desktop GPU speed.

3.3. Scenario Constraints Layer

Compared with the technique-oriented taxonomy in Section 3.2, this section analyzes the applicability boundaries and onboard deployment suitability of different methods from the perspective of real mission requirements. In this review, UAV imagery small-object detection tasks are organized into five representative scenario constraints: low-altitude dense object distribution constraints, high-altitude object visibility constraints, hardware resource deployment constraints, imaging quality degradation constraints, and cross-domain generalization constraints.

The following analysis provides qualitative discussions of scenario-specific considerations that may inform model selection and evaluation design. These discussions are intended to highlight practical factors beyond conventional accuracy metrics, rather than to propose formal evaluation metrics. The quantitative implementation of such factors is an important direction for future work.

3.3.1. Low-Altitude Dense Object Distribution Constraints

UAVs often fly at relatively low altitudes in tasks such as traffic monitoring, security inspection, and crowd activity analysis to obtain higher spatial resolution. Low-altitude imaging also increases the number of objects within a unit field of view, intensifies mutual occlusion, and introduces more background clutter. In these scenarios, the central difficulty is not object visibility alone, but whether densely distributed neighboring instances can be reliably separated and accurately localized.

Recent studies have addressed this issue through multi-scale detail enhancement [102], contextual modeling [129], and localization optimization for dense objects [130]. These strategies help preserve fine-grained boundary cues, exploit surrounding contextual relationships, and improve box regression in crowded regions. Nevertheless, their effectiveness is still constrained by severe inter-object overlap, background interference, duplicate predic-

tions, and the limited discriminability of small-object boundaries. Therefore, low-altitude dense scenes require not only stronger feature representation but also more reliable instance separation and localization mechanisms.

The performance evaluation of models designed for such scenarios cannot rely solely on the overall mean average precision (mAP), since average scores may obscure dense-region errors such as missed detections, duplicate boxes, and unstable localization under occlusion. In such scenarios, practical evaluation may benefit from more detailed error analysis, including recall in crowded regions, false-positive suppression, duplicate detection control, and localization stability, to better reflect real-world deployment performance. In practical UAV missions, a detector with slightly lower overall mAP but fewer missed or duplicate detections in dense regions may be more valuable than a benchmark-oriented model optimized mainly for higher average scores.

3.3.2. High-Altitude Object Visibility Constraints

In large-scale missions such as maritime surveillance, border patrol, and disaster search and rescue, UAVs often fly at high altitudes to ensure flight safety and expand area coverage. However, high-altitude imaging sharply reduces the projected size of objects, causing many objects to occupy only a few pixels. Their texture, boundary, and contrast cues are consequently weakened. Therefore, the central challenge in high-altitude observation is not only whether objects are visible, but whether they retain sufficient discriminative representation for reliable detection.

To cope with weak visibility, existing methods mainly rely on cross-scale representation enhancement [90], joint spatial–frequency domain modeling [89], and spatial relationship compensation [131]. These strategies aim to preserve weak object responses, recover scale-compressed details, and compensate for the loss of discriminative cues under long-range observation. However, their effectiveness may still be limited by background noise amplification, unstable contextual priors, and the additional computational cost introduced by fine-grained feature recovery. Compared with low-altitude dense scenes, high-altitude scenarios depend less on separating adjacent instances and more on preserving or reconstructing weak object evidence before it is suppressed by the background.

For high-altitude scenarios, evaluation may place greater emphasis on the recall rate of small objects, the confidence stability of weak-texture objects, and detection consistency across different flight altitudes and imaging scales. Additional analysis along these dimensions, such as scale-wise recall, altitude-wise performance degradation, and confidence calibration, can provide more informative evidence of practical reliability. In long-range search and monitoring missions, recall usually has a higher operational priority, although precision remains important for controlling false alarms and reducing downstream verification costs.

3.3.3. Resource-Constrained Deployment Constraints

For tasks such as autonomous cruising, airborne early warning, and online tracking, UAV systems require that detection results be directly applied to flight control, path planning, and mission scheduling. This application demand imposes strict constraints on the detection model, necessitating stable operation under limited computational resources. For small UAVs, the core challenge in practical model deployment is no longer pursuing ultimate offline detection accuracy, but rather achieving continuous and reliable environmental perception with low latency and high resource efficiency under stringent hardware conditions.

To address this practical deployment need, recent research has shifted from relying solely on model compression techniques toward the joint optimization of lightweight

network architectures, efficient computational operators, and local feature enhancement. For instance, inspired by the mechanisms of the biological visual cortex V2 area, Zhang et al. [132] designed a lightweight feature extraction module, a hierarchical feature fusion structure, and an adaptive sparse attention mechanism. This approach significantly reduces computational cost while effectively enhancing the model's ability to represent features such as edges and texture details of small objects. Wan et al. [133] improved low-level information for small objects by combining receptive-field attention with dynamic attention-based upsampling, while also simplifying the overall network structure. Guo et al. [120] introduced wavelet convolution, spatial attention, and a bidirectional dynamic feature fusion pyramid, thereby reducing the deployment overhead of Transformer-based frameworks on embedded platforms.

For resource-constrained deployment, practical model selection may also consider end-to-end latency, memory footprint, energy efficiency, thermal stability, hardware compatibility, and long-term detection stability, in addition to conventional accuracy and complexity metrics. In practical UAV deployment, a model with slightly lower benchmark accuracy but more predictable latency and lower resource consumption may be more valuable than one optimized mainly for peak offline performance.

3.3.4. Imaging Degradation Constraints

In scenarios such as nighttime surveillance, foggy patrol, backlight observation, high-speed maneuvering, and infrared search, imaging quality is often degraded by low illumination, atmospheric scattering, platform vibration, and relative motion. These factors reduce the contrast of objects in aerial images, blur object boundaries, and obscure their fine-grained texture information, thereby directly compromising the stability of detection results. As a result, models trained on clean and well-exposed data often exhibit substantial performance degradation after deployment under such conditions.

To address this problem, recent work has gradually shifted from focusing only on detector improvement to jointly considering image restoration and detection. For instance, Wu et al. [115] integrated fast deblurring module, dual-domain feature aggregation module, and lightweight detection backbone into an end-to-end framework, which improves real-time detection performance in blurred aerial imagery. For infrared UAV images affected by motion blur, Wang et al. [134] adopted a clear/blur dual-branch architecture with shared weights and further introduced feature-level supervision, frequency-structure guidance, and consistency constraints to enhance discrimination under degraded conditions. More importantly, these methods are designed not merely to improve subjective image quality, but to recover the structural information most useful for detection and discrimination.

In general, imaging degradation scenarios rely more heavily on degradation-aware feature modeling, frequency-domain compensation, and restoration–detection joint optimization. Evaluation in such scenarios may emphasize performance drop under degradation, robustness across adverse environments, and false alarm control under impaired observations, in addition to clean-condition accuracy.

3.3.5. Cross-Domain Generalization Constraints

In real deployment, UAV detection models often need to be transferred across different regions, weather conditions, seasons, flight altitudes, and sensor specifications. Consequently, there is often a substantial distribution gap between the training domain and the object application domain. At the same time, large-scale reannotation for every new environment is usually impractical in engineering settings. Under such conditions, cross-domain generalization becomes a key factor determining whether a detector is truly reusable over time.

Existing research on cross-domain transfer mainly focuses on two directions: robust semantic representation and data-level compensation. Park et al. [135] improved adaptability to viewpoint, scale, and scene changes through visual–language joint modeling, a visual semantic reasoner, and a relational learning loss. Qilong et al. [63] adapted a diffusion model to the UAV domain and combined it with automated prompt construction and CLIP-based filtering to generate high-quality image–label pairs, providing a novel data compensation strategy for scenarios where samples are extremely scarce and domain shift exists. Compared with methods optimized only for a single benchmark domain, these studies pay greater attention to how detection capability can be preserved when the environment, device, or observation condition changes.

For cross-domain scenarios, model assessment can be complemented by cross-dataset retention rate, few-shot adaptation efficiency, and retraining cost, beyond source-domain performance alone. From an engineering perspective, low transfer cost, rapid adaptation, and stable environmental robustness are often more valuable than achieving the best performance in a single aspect.

4. Mainstream Aerial Imagery Datasets and Evaluation Metrics

4.1. Mainstream Aerial Imagery Datasets

UAV imagery refers to images captured by unmanned aerial vehicles during low-altitude flight, typically characterized by dynamic viewpoints, continuous motion, and onboard sensor constraints. Aerial imagery is a broader category encompassing all top-down or oblique images captured from airborne platforms, including both UAVs and manned aircraft. Remote-sensing imagery refers to images acquired by satellites or high-altitude aircraft for large-scale Earth observation, often with higher altitudes, larger coverage, and fixed viewing geometries.

While these categories overlap in practice, the datasets selected in this review serve different reference purposes. VisDrone2019 and UAVDT are standard low-altitude UAV video datasets that closely resemble real deployment scenarios. DOTA is included as a high-resolution aerial remote-sensing benchmark that provides oriented bounding boxes and large-scale scene coverage, offering insights into rotation-robust detection. CARPK focuses on dense vehicle counting in parking-lot scenes, while AU-AIR provides multi-modal UAV data with flight-state metadata. Although DOTA and CARPK are not typical low-altitude UAV video datasets, their inclusion enriches the review by covering additional imaging geometries, scene structures, and annotation characteristics that are relevant to UAV small-object detection.

This section provides a detailed elaboration of five widely used public datasets from the perspectives of data source, acquisition scenario, UAV imaging characteristics, object-scale distribution, annotation format, task suitability, and engineering relevance.

VisDrone2019 [136] is one of the most widely used low-altitude UAV video benchmarks for general UAV vision tasks. It was collected from low-altitude urban and suburban scenes under diverse weather, illumination, altitude, and viewing-angle conditions. The dataset contains multiple traffic-related categories, such as pedestrians, cars, buses, bicycles, and tricycles, and is characterized by dense object distributions, frequent occlusions, background clutter, and large intra-class scale variations. In particular, its high proportion of small objects makes it a core benchmark for evaluating small-object representation, dense-scene robustness, and tracking stability in complex UAV urban environments.

DOTA [20] is a representative high-resolution aerial remote sensing dataset emphasizing large-field-of-view top-down imaging and arbitrary-oriented object detection. Different from low-altitude UAV video datasets, DOTA mainly focuses on overhead aerial and remote sensing imagery, where objects often appear with large scale differences and arbitrary

orientations. With polygon or oriented bounding-box annotations, it is more suitable for rotated detection, remote sensing scene modeling, and orientation-sensitive localization analysis. However, because it lacks continuous low-altitude flight sequences and onboard motion effects, it is less representative of dynamic UAV deployment scenarios.

UAVDT [137] mainly focuses on urban traffic monitoring and is built from UAV video sequences collected across multiple cities, viewpoints, flight altitudes, and illumination conditions. Compared with general UAV datasets, it focuses more on vehicle-related categories and structured road scenes, while still retaining typical UAV challenges such as camera motion, viewpoint changes, scale variation, motion blur, partial occlusion, and dense traffic flow. Owing to its sequential and deployment-oriented characteristics, it is widely used for vehicle detection, tracking, counting, and temporal stability evaluation in UAV traffic surveillance.

CARPK [138] is a UAV-captured vehicle dataset focused on dense car counting in parking-lot scenarios. It contains a large number of densely arranged cars with relatively regular spatial layouts, making it valuable for evaluating dense object localization, occlusion handling, and counting consistency under aerial views. Although it contains only one category and has limited scene diversity, its controlled acquisition setting provides a useful benchmark for analyzing detector performance in dense vehicle counting and top-down UAV object localization.

AU-AIR [139] is a low-altitude UAV traffic monitoring dataset that provides multi-modal data, including flight-related information and sensor metadata, making it particularly useful for multimodal perception, navigation-aware detection, and studies integrating visual features with flight-state cues. The dataset covers road traffic scenes with multiple object categories and reflects typical low-altitude UAV challenges such as viewpoint variation, object-scale changes, partial occlusion, and background interference. Compared with purely visual datasets, AU-AIR is more suitable for exploring UAV perception methods that combine image content with platform-state information.

As shown in Table 5, annotated samples with bounding-box areas smaller than 32×32 pixels in five mainstream datasets were counted based on the absolute scale judgment criteria. The statistical results reveal that small objects constitute a substantial proportion in most UAV imagery datasets, yet there are remarkable discrepancies among different datasets. As shown by Small object proportion in Table 5, the proportion of small objects exceeds 60% in the VisDrone2019 and UAVDT datasets, while it accounts for 44.5% in DOTA, 11.7% in AU-AIR, and only 2.1% in CARPK. This reflects the differences among various datasets in terms of collection scenarios, object categories, and imaging conditions.

A dataset consists of original images together with manually annotated ground truth. A benchmark is established on the basis of a dataset by defining fixed training–test splits, evaluation protocols, and performance metrics, thereby enabling fair comparisons among different algorithms. Benchmark bias arises when a benchmark’s scene distribution and statistical characteristics of target scales fail to represent those encountered in real-world deployment. For example, an algorithm extensively optimized for dense urban environments, such as those represented by VisDrone2019, may excel primarily at detecting small objects in complex settings. However, when transferred to scenes with distinctive urban characteristics, it may overfit and consequently experience performance degradation in structured road environments, as represented by UAVDT, or in top-down parking-lot views, as represented by CARPK. Conversely, a detector fine-tuned for large, high-resolution targets in UAVDT may suffer a substantial decline in performance when confronted with the densely arranged, minuscule objects characteristic of VisDrone2019.

Table 5. Mainstream aerial imagery datasets.

Dataset	Source	Scenario	Classes	Small-Object Proportion	Main Use
VisDrone2019	AISKYEYE Team, TJU	Urban low-altitude UAV imagery	10	60.4%	General UAV detection and tracking
DOTA	WHU	Aerial/remote sensing imagery	15/18	44.5%	Rotated detection and remote sensing analysis
UAVDT	UCAS	Urban road and traffic monitoring	3	61.7%	UAV traffic detection and tracking
CARPK	NTU and GE-GRC	Parking lot aerial view	1	2.1%	Dense vehicle detection and counting
AU-AIR	Adelaide University	Low-altitude traffic monitoring	8	11.7%	Multimodal UAV perception

Note: Statistics are computed at the original native resolution of each dataset. Small objects: area $< 32 \times 32$ pixels (< 1024 pixels²).

The foregoing analysis demonstrates that performance on public benchmarks cannot be equated directly with performance in real-world UAV deployment. Actual UAV missions vary considerably in flight altitude, scene structure, target categories, and environmental conditions. Because no single benchmark dataset can comprehensively encompass the full range of real-world circumstances, benchmark performance must not be treated as synonymous with deployment performance.

Evaluation methods should be aligned with the intended application scenario. Although a detector may exhibit strong performance after being trained on a particular benchmark dataset, this does not guarantee comparable effectiveness in more complex real-world deployment environments. Accordingly, the detector's practical detection performance should be assessed through cross-validation across multiple benchmarks and datasets, rather than inferred from a single benchmark alone. This approach can more fully reveal performance fluctuations under different scene distributions. In addition to general-purpose evaluation metrics, task-specific measures should be incorporated; for example, small-object detection tasks should place particular emphasis on metrics such as APS and small-object recall.

4.2. Evaluation Metrics

This section summarizes the commonly used performance evaluation metrics for small-object detection, as shown in Table 6. Intersection over Union (IoU) quantifies the degree of overlap between the predicted bounding box and the ground-truth bounding box, as illustrated in Figure 6a. The confusion matrix, shown in Figure 6b, provides a basis for analyzing classification errors, missed detections, and false detections. It also serves as the interpretive foundation for core metrics, such as precision, recall, and accuracy, and mainly consists of four parameters.

Table 6. Evaluation metrics for object detection models.

Metric	Formula	Definition	Purpose
Precision	$\frac{TP}{TP + FP}$	Correct detections among all predicted positives	Measures prediction reliability
Recall	$\frac{TP}{TP + FN}$	Detected objects among all ground-truth objects	Measures detection completeness and missed detections

Table 6. Cont.

Metric	Formula	Definition	Purpose
F1-score	$\frac{2TP}{2TP + FP + FN}$	Harmonic mean of precision and recall	Balances false positives and missed detections
AP	Area under the PR curve	Detection accuracy for a single class	Single-class detection evaluation
mAP	Mean AP over all classes	Overall detection accuracy across classes	Overall detection evaluation
mAP@0.5	mAP at IoU = 0.5	Detection performance under standard IoU threshold	Benchmark testing
mAP@[0.5:0.95]	mAP under multiple IoU thresholds	Performance across different localization accuracy	Localization accuracy and robustness evaluation
APS	Area $< 32 \times 32$	Small-object AP	Evaluates small-object detection capability
APM	$32 \times 32 \leq \text{area} < 96 \times 96$	Medium-object AP	Evaluates medium object detection capability
APL	Area $\geq 96 \times 96$	Large-object AP	Evaluates large object detection capability
FPS	$\frac{\text{Total frames}}{\text{Total inference time}}$	Frames per second	Measures inference speed and real-time performance

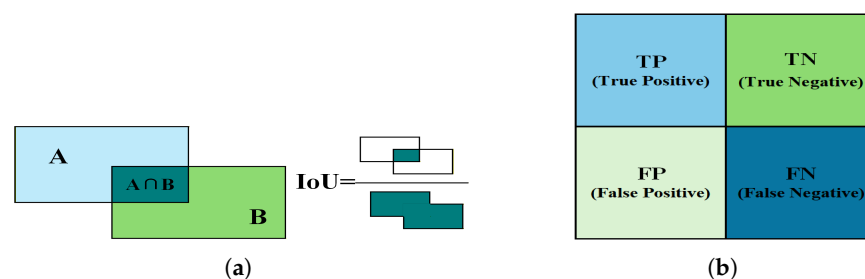


Figure 6. (a) IoU calculation process. (b) Confusion matrix. A represents the area of the predicted bounding box, and B represents the area of the ground-truth bounding box.

5. Experimental Results and Analysis

The experiments in this section are designed for illustrative and reference purposes. They evaluate representative detectors under their respective default configurations on UAV datasets with a unified input resolution, thereby complementing the qualitative analyses in the preceding sections. The aim is to provide practical, configuration-specific performance references that help readers understand what to expect when adopting these detectors with their standard training recipes. This is not a controlled comparative study, nor is it intended to rank detector families or draw universal conclusions about the superiority of any particular paradigm. Given that this study specifically focuses on small-object detection, we treat APS (Average Precision for objects with area $< 32 \times 32$ pixels) as the primary evaluation metric for assessing detector performance. Overall mAP@0.5 and mAP@0.5:0.95 are reported as complementary metrics to provide context on general detection capability. These experiments aim to investigate the following questions:

1. How do detectors from different paradigms perform in UAV imagery small-object detection tasks under the present experimental configuration?
2. Are improvements in overall detection accuracy consistent with improvements in small-object detection performance?

3. How do different UAV scene characteristics, such as dense object distribution, occlusion, background complexity, and structured traffic scenes, influence model performance?
4. What accuracy–efficiency trade-offs are observed among the evaluated detectors?

The VisDrone2019 dataset contains complex urban scenes with dense small objects, severe occlusions, and a large amount of complex background interference. The UAVDT dataset corresponds to structured traffic monitoring scenarios. Due to the significant differences in data distribution between the two datasets, they can be complementarily used to evaluate the performance of the detection models, thereby testing the models' adaptability to different scenarios. In this experiment, the VisDrone2019 and UAVDT datasets were selected as the benchmark test datasets. From the three core architectures described in Section 3.1, representative models were selected for quantitative comparative experiments.

5.1. Experimental Environment

To provide empirical reference points for representative detectors on UAV datasets, we adopt a partially unified experimental protocol that standardizes data-level and evaluation-level conditions while preserving each framework's native training characteristics. As shown in Table 7, we unify the input resolution, datasets, evaluation metrics, and hardware/software environments.

Table 7. Experimental environment.

Hardware and Software Environment		Training Hyperparameters	
Environment Name	Version	Parameter	Value
Operating System	Ubuntu 22.04.1 LTS	Images	640 × 640
CPU	Intel(R) Xeon(R) Platinum 8352V	Epochs	100
GPU	RTX 4090 (24 GB)	-	-
CUDA	11.8	-	-
Python	Python 3.11.11	-	-
Deep Learning Framework	PyTorch 2.2.2	-	-
Development Environment	PyCharm 2024	-	-

This study does not enforce unified training hyperparameters. Instead, we adopt each detector's framework-recommended or commonly used configuration. This design choice is intentional: forcing a single optimizer or augmentation policy across all architectures would unfairly favor some detectors while disadvantaging others that were designed with specific training dynamics in mind. Since our goal is to illustrate how these detectors behave under their standard usage practices, this design provides practically relevant reference points while avoiding the artificial constraints of a fully uniform training setup.

During the data preprocessing stage, the annotation files of VisDrone2019 and UAVDT were processed and converted into YOLO and COCO formats according to the input requirements of the corresponding detection frameworks. The YOLO format was used for the YOLO-series models, whereas the COCO format was used for Faster R-CNN, Transformer-based detectors, and the COCO-protocol scale evaluation metrics, including APS, APM, and APL.

During the training stage, we followed the standard practices of each model family. Faster R-CNN used an ImageNet-pretrained ResNet-50 backbone, while YOLOv5n [140] and YOLOv7-tiny used COCO-pretrained weights. SSD, FCOS, CenterNet, YOLOv8n [37], YOLOv13n, RT-DETR, and Deformable-DETR were trained from scratch. This asymmetric use of pre-training may influence the observed results. Each model adopted the data augmentation strategies provided by its corresponding framework. YOLO-series models used

the relatively strong augmentation policies provided by the Ultralytics framework, whereas the other detectors used basic horizontal flipping. No dataset-specific augmentation tuning was performed. Worker counts were set to 0, 4, or 8 according to the requirements of the corresponding implementations to ensure stable training on the 24 GB RTX 4090 GPU, as detailed in Table 7.

According to the optimization characteristics of the selected model implementations, commonly used optimizer settings were retained. Faster R-CNN used SGD with a learning rate of 0.001. SSD and FCOS used SGD with their corresponding framework configurations, while the YOLO-series models used SGD with the settings provided by their implementations. Deformable-DETR used AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, a weight decay of 10^{-5} , and a step-decay schedule with $lr_drop = 30$. CenterNet used Adam with a ReduceLROnPlateau scheduler. Detailed learning-rate schedules are provided in Table 8.

Table 8. Per-model detailed hyperparameter configurations.

Model	Optimizer	Learning Rate	Weight Decay	LR Schedule	Batch/Workers
Faster R-CNN	SGD	0.001	1×10^{-4}	Cosine (step 50→100)	16/8
SSD	SGD	0.001	5×10^{-4}	Step decay (cfg defined)	8/4
FCOS	SGD	0.001	1×10^{-4}	Warmup (501) + Step-wise	16/4
CenterNet	Adam	0.001	5×10^{-4}	ReduceLROn Plateau	16/8
YOLOv5n	SGD	0.01	5×10^{-4}	Linear + Warmup (3 ep)	16/8
YOLOv7-tiny	SGD	0.01	5×10^{-4}	Cosine + Warmup (3 ep)	16/8
YOLOv8n	SGD	0.01	5×10^{-4}	Cosine + Warmup (3 ep)	16/8
YOLOv13n	SGD	0.01	5×10^{-4}	Cosine + Warmup (3 ep)	16/8
Deformable-DETR	AdamW	0.001	1×10^{-5}	Step decay ($lr_drop = 30$)	4/4
RT-DETR	SGD	0.01	5×10^{-4}	Cosine + Warmup (3 ep)	16/8

For the backbone networks, ResNet-50 was used for Faster R-CNN and Deformable DETR, whereas the YOLO models used their native compact backbones. The YOLO variants considered in this study are nano-scale or tiny-scale models and therefore have substantially fewer parameters than several of the other evaluated detectors. This difference in model capacity should be considered when interpreting the accuracy–efficiency trade-offs.

In the model evaluation stage, this study conducts a comprehensive multi-dimensional assessment encompassing detection accuracy, small-object detection capability, model complexity, and inference efficiency, thereby demonstrating the overall performance and adaptability of various detection models across different UAV scenarios. The complete set of evaluation metrics employed in this study, along with their specific definitions, is detailed in Table 6.

The performance differences reported in Tables 9 and 10 reflect not only the architectural characteristics of the evaluated detectors but also the combined effects of asymmetric pre-training, non-uniform augmentation policies, optimization configurations, and model scaling. The conclusions should therefore be interpreted as empirical observations under the present partially unified but framework-aware configuration. They provide reference information for model comparison and preliminary model selection, rather than definitive claims about the inherent superiority or inferiority of any detection paradigm.

Table 9. Detection performance of representative models on VisDrone2019 under framework-recommended configurations.

Paradigm	Type	Model	P	R	F1	mAP@0.5	mAP@0.5:0.95	Params ($\times 10^6$)	GFLOPs	FPS	APS	APM	APL
Two-stage detector	Region-proposal-based	Faster R-CNN	28.6	21.8	24.7	20.5	10.2	28.3	474.1	26.2	2.8	18.5	38.6
Classical anchor-based one-stage	Anchor-based	SSD	26.1	48.2	33.9	16.9	9.6	24.9	138.8	42.6	5.6	15.9	17.2
	Anchor-free	FCOS	37.1	22.2	27.8	24.5	12.4	32.1	80.7	65.3	5.8	23.6	49.7
	Anchor-free	CenterNet	7.0	26.7	11.1	13.2	5.5	32.7	54.8	198.5	0.25	6.2	26.7
YOLO-based real-time detector	Anchor-based YOLO	YOLOv5n	42.5	31.1	35.9	31.0	17.7	2.5	7.1	196.2	6.9	23.4	37.8
	Anchor-based YOLO	YOLOv7-tiny	35.2	30.5	32.7	23.5	12.3	6.0	13.3	261.2	6.2	16.6	24.0
	Anchor-free YOLO	YOLOv8n	42.7	32.2	36.7	31.8	18.2	3.0	8.1	204.0	7.2	23.1	37.2
	Anchor-free YOLO	YOLOv13n	41.9	31.5	35.9	30.8	17.4	2.4	6.2	56.1	6.9	22.7	37.3
Transformer-based detector	End-to-end DETR	Deformable-DETR	33.6	24.8	28.5	18.2	8.6	50.1	203.3	32.1	3.8	16.4	29.7
	Real-time DETR	RT-DETR	32.3	19.6	24.4	10.8	5.2	41.9	125.7	37.3	3.6	9.3	10.8

Note: Results are reported under the specific experimental configuration described in Section 5.1. Performance rankings may differ under other configurations. These results are intended as reference observations rather than generalizable conclusions about any detectors.

Table 10. Detection performance of representative models on UAVDT under framework-recommended configurations.

Paradigm	Type	Model	P	R	F1	mAP@0.5	mAP@0.5:0.95	Params ($\times 10^6$)	GFLOPs	FPS	APS	APM	APL
Two-stage detector	Region-proposal-based	Faster R-CNN	50.8	44.6	47.5	48.2	26.8	28.3	474.1	26.0	13.8	38.5	63.2
Classical anchor-based one-stage	Anchor-based	SSD	52.3	57.8	54.9	42.3	35.4	24.0	137.3	45.0	32.4	38.5	43.2
	Anchor-free	FCOS	31.4	26.9	28.9	20.3	9.0	32.1	80.6	66.9	7.6	19.1	26.8
	Anchor-free	CenterNet	8.9	46.0	14.9	8.2	3.8	32.7	54.8	237.8	1.8	7.6	12.4
YOLO-based real-time detector	Anchor-based YOLO	YOLOv5n	43.4	33.6	37.9	26.5	13.1	2.5	7.1	178.5	20.5	21.7	27.4
	Anchor-based YOLO	YOLOv7-tiny	49.5	49.3	49.4	30.2	27.6	6.0	13.0	242.3	14.1	29.2	30.5
	Anchor-free YOLO	YOLOv8n	35.3	32.1	33.6	26.2	18.9	3.0	8.1	222.2	12.4	20.0	28.4
	Anchor-free YOLO	YOLOv13n	42.7	33.4	37.5	26.8	18.9	2.4	6.2	57.8	11.3	19.2	28.8
Transformer-based detector	End-to-end DETR	Deformable-DETR	62.3	48.7	54.7	24.5	12.8	50.1	203.3	32.7	6.2	21.4	38.7
	Real-time DETR	RT-DETR	43.2	39.5	41.3	29.1	21.8	41.9	125.6	41.6	15.8	26.5	30.3

Note: Results are reported under the specific experimental configuration described in Section 5.1. Performance rankings may differ under other configurations. These results are intended as reference observations rather than generalizable conclusions about any detectors.

5.2. Analysis of Experimental Results

5.2.1. Experimental Results on the VisDrone2019 Dataset

The experimental results in Table 9 show that, in complex UAV scenarios, there is a significant gap between the APS and the overall mAP, and detecting small objects remains challenging. For example, YOLOv8n achieves an mAP@0.5 of 31.8% but an APS of only 7.2%, while its APM and APL reach 23.1% and 37.2%, a discrepancy of over 16 percentage points between small-object and medium-object performance. YOLOv5n and YOLOv13n exhibit APS values of 6.9% and 6.9% compared to APM values above 22%. This gap is even more pronounced for Faster R-CNN and RT-DETR. These results show that overall detection accuracy is not necessarily consistent with small-object detection performance. The considerable gap between APS and APM/APL suggests that APS should be considered when evaluating UAV imagery small-object detection.

Under the present experimental configuration, the evaluated YOLO implementations exhibit a favorable measured accuracy–efficiency trade-off. Among the models evaluated, YOLOv8n yields the highest numeric mAP@0.5, mAP@0.5:0.95, and APS values. It should be emphasized that this observation is configuration-specific; different hardware platforms, training protocols, or resolution settings may lead to different results. YOLOv5n obtains an mAP@0.5 of 31.0%, with 2.5 million parameters, 7.1 GFLOPs, and 196.2 FPS. YOLOv13n obtains an mAP@0.5 of 30.8%, with 2.4 million parameters and 6.2 GFLOPs. In comparison, Faster R-CNN requires 474.1 GFLOPs and obtains an mAP@0.5 of 20.5% and an APS of 2.8%. Deformable-DETR contains 50.1 million parameters and requires 203.3 GFLOPs, while its mAP@0.5 and APS values are 18.2% and 3.8%, respectively. Thus, under the present configuration, the lightweight YOLO implementations show a more favorable measured balance among accuracy, model complexity, and inference speed. This observation should not be interpreted as evidence of the intrinsic superiority of the YOLO paradigm, because the evaluated implementations use different initialization, augmentation, optimization, and model-capacity settings.

The observations in Table 9 are broadly consistent with the challenges discussed in the preceding sections, including dense object distributions, occlusion, complex backgrounds, scale variation, and limited fine-grained information. These factors may contribute to false positives, missed detections, and reduced localization quality, although the present comparison does not isolate the contribution of any individual architectural component. The results therefore support the need for task-specific techniques, including small-object feature enhancement, multi-scale feature fusion, dense-object handling, occlusion robustness, and lightweight model design, rather than the direct transfer of general-purpose detectors without further adaptation.

From the perspective of the measured desktop-GPU results, YOLOv8n provides the strongest reported accuracy among the evaluated lightweight models on VisDrone2019, whereas YOLOv5n and YOLOv13n provide lower computational complexity. These observations may be useful for selecting preliminary baselines for UAV-oriented research. However, actual onboard deployment decisions require evaluation on the target edge hardware, since the FPS values reported here were obtained on an RTX 4090 GPU and do not directly represent the latency, memory, or power consumption of a UAV platform.

5.2.2. Experimental Results on the UAVDT

As on VisDrone2019, the UAVDT results in Table 10 reveal a consistent gap between overall mAP and APS, although the gap is less pronounced due to the larger object scales in UAVDT. For instance, YOLOv5n achieves an mAP@0.5 of 26.5% but an APS of 20.5%, while its APL reaches 27.4%, a more modest but still meaningful gap. Faster R-CNN, which achieves the highest overall mAP of 48.2%, exhibits an APS of only 13.8%, far below its

APM of 38.5% and APL of 63.2%. This indicates that even detectors that perform well on medium and large objects may struggle with small-object localization. The UAVDT results further confirm that APS is a more discriminating metric than overall mAP for evaluating small-object detection performance, and it should therefore be given primary consideration in detector selection for UAV applications.

The YOLO-series models provide relatively high measured inference speeds on the RTX 4090 platform. YOLOv7-tiny and YOLOv8n reach 242.3 FPS and 222.2 FPS, respectively, while YOLOv5n reaches 178.5 FPS. However, the highest mAP@0.5 among the evaluated YOLO models is 30.2%, achieved by YOLOv7-tiny, and the highest APS is 20.5%, achieved by YOLOv5n. These results show that high measured inference speed does not necessarily correspond to the highest detection accuracy on UAVDT. Conversely, Faster R-CNN and SSD obtain higher reported accuracy or small-object performance at a substantially higher computational cost.

The ranking differences between UAVDT and VisDrone2019 suggest that the performance of the evaluated implementations is dataset-dependent. UAVDT mainly represents structured traffic-monitoring scenes with relatively concentrated object categories and regular background structures, whereas VisDrone2019 contains more diverse urban scenes, dense objects, severe occlusion, and stronger background interference. Under the present configuration, Faster R-CNN and SSD produce higher numeric accuracy on UAVDT, while the lightweight YOLO implementations show a more favorable measured accuracy-efficiency trade-off on VisDrone2019. This contrast is consistent with the scenario-dependent nature of UAV imagery object detection. However, these observations are limited to the specific configurations, datasets, and hardware evaluated in this study; they should not be generalized to other settings without further validation. Nevertheless, because the current comparison does not control pre-training, augmentation, optimization schedules, or model capacity, the observed differences should not be interpreted as general evidence that one detector family is inherently better suited to UAVDT.

The UAVDT results are also consistent with the broader observation that detection performance is jointly affected by scene composition, object-scale distribution, category characteristics, training configuration, and computational cost. They further suggest that no single accuracy or efficiency metric is sufficient for model selection. Instead, the selection of a reference detector should depend on the target scene and the relative importance assigned to overall accuracy, small-object recall, computational complexity, and inference speed.

From the perspective of the present desktop-GPU measurements, Faster R-CNN can be used as an accuracy-oriented reference model for UAVDT-like scenes, whereas SSD provides a compromise between accuracy and measured computational cost. YOLOv5n, YOLOv7-tiny, and YOLOv8n provide higher measured inference speeds with lower parameter counts and GFLOPs. These observations can provide preliminary references for subsequent UAV-oriented optimization. However, they should not be regarded as direct deployment recommendations, because actual edge or onboard deployment requires additional testing on the target hardware and under the target application conditions.

In the preceding subsections, we have conducted quantitative experiments and analyses on the VisDrone2019 and UAVDT datasets, supplemented these with qualitative visualizations across five typical scenarios, and performed cross-domain quantitative evaluations. These results provide empirical support for the relevant discussions throughout the paper. To establish a clear mapping between technical solutions, scenario constraints, and experimental evidence, Table 11 systematically maps the core challenges and scenario constraints defined in the framework to the technical solutions, datasets, evaluation metrics, and experimental findings discussed in this study.

Table 11. Framework—experiment mapping matrix.

Challenge/Scenario	Relevant Technical Solutions	Dataset/Experimental Scenario	Metrics Quantified in This Review	Potential Metrics for Future Work	Experimental Evidence	Key Observations
Small-object attributes (weak texture, detail loss from down-sampling)	High-resolution feature preservation, shallow feature enhancement, cross-layer aggregation, super-resolution, detection head optimization	VisDrone2019, UAVDT; high-altitude low-visibility scenes	APS, small-object Recall, mAP@0.50:0.95	Localization stability	Tables 9 and 10, Section 5.3.2	Overall mAP does not fully reflect small-object detection capability; APS should be prioritized
Extreme scale variation and cross-layer semantic misalignment	Multi-scale feature fusion, adaptive scale selection, deformable attention, dynamic fusion	VisDrone2019, DOTA; varying scale distributions	APS, APM, APL, AP under different thresholds	Scale-wise recall, altitude-wise performance degradation, confidence calibration	Tables 2, 9 and 10	Dataset-scale distribution alters model rankings; no single detector dominates across scales
Dense distribution and occlusion	Two-stage RoI refinement, contextual modeling, instance separation, attention mechanisms, scenario-aware augmentation	VisDrone2019, CARPK	Recall, Precision, F1	Recall in crowded regions, false-positive suppression, duplicate detection control, localization stability	Section 5.3.1, qualitative results	Faster R-CNN achieves higher recall but produces redundant boxes in dense regions
Complex background and clutter	Spatial/channel attention, frequency-domain modeling, context suppression, template modeling, background modeling	VisDrone2019, infrared scenes, UAVDT	Precision, F1	False-positive suppression, background-region FP	Section 5.3.3	Trade-off between false-positive suppression and detection completeness varies across paradigms
Imaging degradation (blur, low contrast, low SNR)	Deblurring, super-resolution, dual-domain feature aggregation, infrared temporal fusion, degradation-aware training	UAVDT, infrared UAV data, degradation scenarios	APS, clean-condition accuracy	Performance drop under degradation, robustness across adverse environments, false alarm control under impaired observations	Section 5.3.3	Faster R-CNN: high recall but many redundant boxes; YOLOv13n: balanced; RT-DETR: less sensitive to blurred small objects

Table 11. Cont.

Challenge/Scenario	Relevant Technical Solutions	Dataset/Experimental Scenario	Metrics Quantified in This Review	Potential Metrics for Future Work	Experimental Evidence	Key Observations
Domain shift and data distribution gaps	Diffusion-based augmentation, domain adaptation, class-balanced augmentation, knowledge distillation, foundation model adaptation	VisDrone2019, UAVDT, DOTA, AU-AIR	cross-dataset retention rate, APS degradation	Few-shot adaptation efficiency, retraining cost, low transfer cost, rapid adaptation, stable environmental robustness	Tables 2 and 12; cross-dataset analysis	Benchmark performance does not directly equate to general UAV deployment capability
Resource-constrained deployment	Lightweight backbone, pruning, distillation, low-memory operators, Mamba, efficient Transformer	UAVDT, VisDrone2019	FPS, Params, GFLOPs	End-to-end latency, memory footprint, energy efficiency, thermal stability, hardware compatibility, long-term detection stability	Tables 9 and 10	Lightweight models offer superior speed and resource efficiency, but compression may sacrifice small-object features

Note: Metrics listed under “Metrics Quantified in This Review” were quantitatively calculated and reported in this review. Metrics listed under “Potential Metrics for Future Work” are proposed as promising evaluation directions but were not systematically quantified in the present experiments. They are discussed qualitatively in the corresponding sections.

5.3. Visualization Experiment of Typical Scenarios

To further reveal the applicable boundaries of different detection paradigms in complex UAV scenarios, this study selects representative models from three mainstream detection paradigms for comparative analysis, including the two-stage detector Faster R-CNN, the one-stage detector YOLOv13n, and the Transformer-based detector RT-DETR. Based on the VisDrone2019 and UAVDT datasets, the experiments select image samples covering typical scenarios such as dense low-altitude objects, weakly visible high-altitude objects, image degradation, and cross-domain transfer, and they conduct qualitative evaluations of each model in terms of small-object feature representation, scale adaptability, localization consistency, background interference suppression, and scene generalization ability.

Among them, the blue boxes in subgraph (a) of the following figures represent the ground truth serving as the benchmark, and subgraphs (b), (c) and (d) show the detection results of the Faster R-CNN, YOLOv13n and RT-DETR models in sequence.

5.3.1. Experiment on Scenarios with Dense Low-Altitude Objects

In scenarios with dense low-altitude objects, partial occlusion, densely clustered objects, and complex road background, interference impose stringent requirements on the detector's recall capability for small objects and discrimination performance for crowded objects. As illustrated in Figure 7b, Faster R-CNN delivers strong responses to some small objects and can detect a considerable number of objects. Nevertheless, its region-proposal-based detection mechanism tends to generate overlapping predictions in dense regions, resulting in abundant redundant bounding boxes in detection outputs. Meanwhile, it suffers from severe missed detections for extremely small objects and objects located at image edges. As shown in Figure 7c, YOLOv13n produces generally neater bounding boxes with almost no obvious redundant boxes, indicating excellent compactness of its detection outputs. Compared with Faster R-CNN, it exhibits weaker responses to certain small objects, leading to substantial missed detections in heavily congested vehicle areas and faraway small-object regions, alongside a small quantity of false positives. As depicted in Figure 7d, RT-DETR achieves relatively accurate detection of large-scale objects with distinct appearances and produces well-aligned bounding box localization. Yet it lacks sufficient sensitivity to small objects, particularly densely packed pedestrians, cyclists, and distant vehicles, resulting in prominent missed detections.

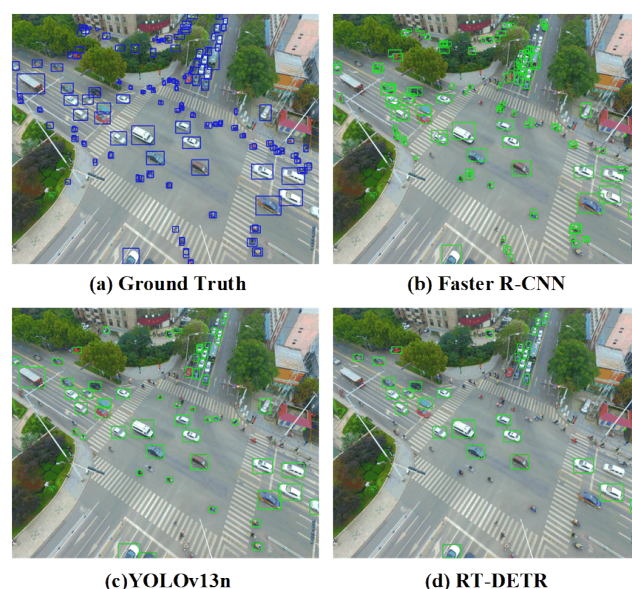


Figure 7. Visualization diagram of detection results under the scenario of dense low-altitude objects.

Experiments show that for such dense low-altitude scenes, Faster R-CNN achieves high recall yet suffers from severe redundant predictions; YOLOv13n generates more concise outputs but lacks adequate recall for small objects; RT-DETR provides stable localization but possesses limited detection performance for clustered small objects.

5.3.2. High-Altitude Visibility Experiment

In scenarios with limited visibility at high altitudes, object scales are further reduced, visual details of vehicles are significantly degraded, and some objects only appear as low-resolution blobs, posing great challenges to detectors in perceiving small objects and achieving precise localization. As illustrated in Figure 8b, Faster R-CNN can detect most vehicle objects and exhibits strong responses to small-scale vehicles, yet it generates numerous overlapping bounding boxes at road intersections and in densely trafficked areas. This indicates that its region proposal mechanism tends to introduce redundant predictions in scenes with small objects featuring weak textures. As shown in Figure 8c, YOLOv13n produces relatively concise bounding boxes with drastically reduced redundant ones and delivers favorable localization performance for major vehicle objects. Constrained by excessively small-object sizes and insufficient detailed information, some distant vehicles and objects in marginal areas fail to be detected, resulting in non-negligible missed detections. As depicted in Figure 8d, RT-DETR yields stable detection results, outperforming YOLOv13n in localization accuracy for certain vehicles and generating more regular bounding boxes. Nevertheless, it still suffers from missed detections for vehicles with extremely small scales or indistinct visual features.

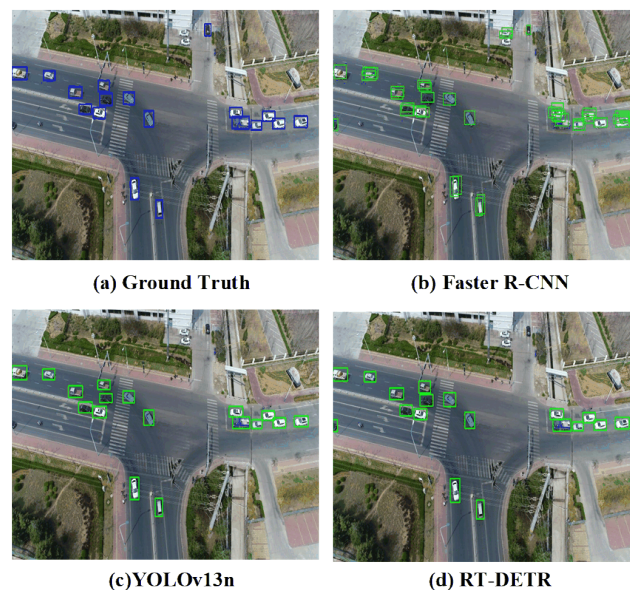


Figure 8. Visualization diagram of detection results under scenarios with limited high-altitude visibility.

Experiments show that under high-altitude low-visibility conditions, Faster R-CNN achieves a high recall rate but generates abundant redundant predictions; YOLOv13n outputs streamlined results with stable localization; RT-DETR boasts advantages in bounding-box localization quality. All three models, however, are constrained by the degraded features of small objects.

5.3.3. Imaging Degradation Scene Experiment

In imaging degradation scenarios, as shown in Figure 9a, the discriminability between vehicles and roads, shadows, as well as surrounding environments is reduced, significantly increasing detection difficulty. As shown in Figure 9b, Faster R-CNN still delivers strong

responses to most vehicle objects and can detect a large number of small-scale objects. However, overlapping or duplicate detection boxes tend to appear in densely packed object areas and regions with complex background textures, indicating that its region proposal generation process is sensitive to degraded images and certain redundancies exist in detection results. As shown in Figure 9c, YOLOv13n generates generally more compact detection boxes with drastically reduced redundant detections and achieves favorable recognition performance for major vehicle objects. Nevertheless, affected by blurred imaging and weak boundaries, some distant small objects and low-contrast objects are missed, while a small number of false detections emerge in partial regions. As shown in Figure 9d, RT-DETR produces relatively conservative detection outputs with fewer bounding boxes than the previous two models and reduced false detections, yet it exhibits inadequate responses to blurred small objects and extremely distant vehicles, leading to more prominent missed detections.

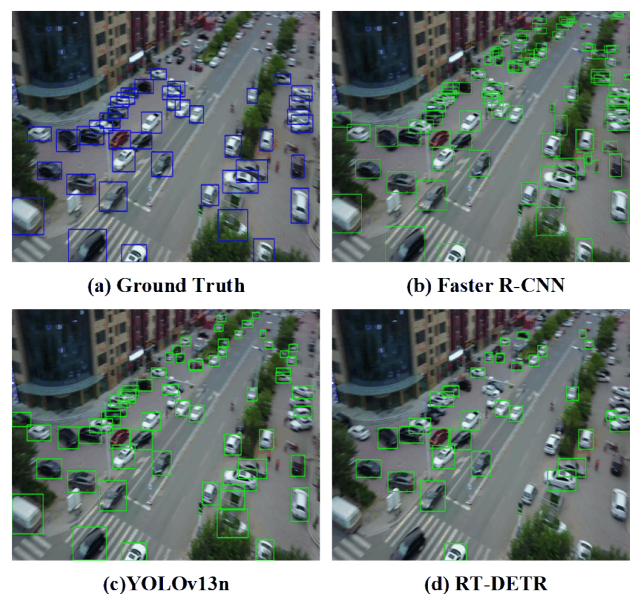


Figure 9. Visualization diagram of detection results under image degradation scenarios.

The experiments show that under imaging degradation conditions, Faster R-CNN boasts strong recall performance but generates abundant redundant boxes; YOLOv13n strikes a relatively balanced trade-off between detection integrity and output conciseness; RT-DETR achieves effective false detection suppression, yet its sensitivity to small objects with faint textures still needs improvement.

5.3.4. Cross-Domain Generalization Experiment

As shown in Figure 10, the first row presents the cross-domain experimental results of Faster R-CNN, the second row presents those of YOLOv13n, and the third row presents those of RT-DETR. The first and fourth columns correspond to the experimental results on the VisDrone2019 dataset, while the second and third columns correspond to the experimental results on the UAVDT dataset.

For the models trained on the VisDrone2019 dataset, the in-domain detection results are shown in Figure 10a,e,i. Faster R-CNN and YOLOv13n achieve good detection accuracy in urban scenes with complex backgrounds, dense objects, and a high proportion of small objects, whereas RT-DETR suffers from a large number of missed detections. After these models are directly transferred to the UAVDT dataset, as shown in Figure 10b,f,j, they can still detect the main vehicle objects, but their confidence stability, scale adaptability, and localization consistency decrease in structured traffic monitoring scenarios, with many

false positives observed. This indicates that the features learned from dense small-object scenes cannot fully adapt to the object-scale distribution and scene prior information of the UAVDT dataset.

The model trained on UAVDT shows more regular and stable in-domain detection behavior, as shown in Figure 10c,g,k. This is primarily attributed to UAVDT's focus on traffic monitoring as its core scenario, where the image structure is relatively clear, background interference is minimal, and the proportion of medium- and large-scale vehicle objects is relatively higher. Under such conditions, the model can more readily leverage stable scene structures and complete object appearance features for discrimination. However, when the model is transferred to the VisDrone dataset for detection, as shown in Figure 10d,h,l, missed detections increase significantly in dense regions, the response capability for small objects weakens, and the model's robustness notably declines in the presence of complex backgrounds and frequent occlusions.

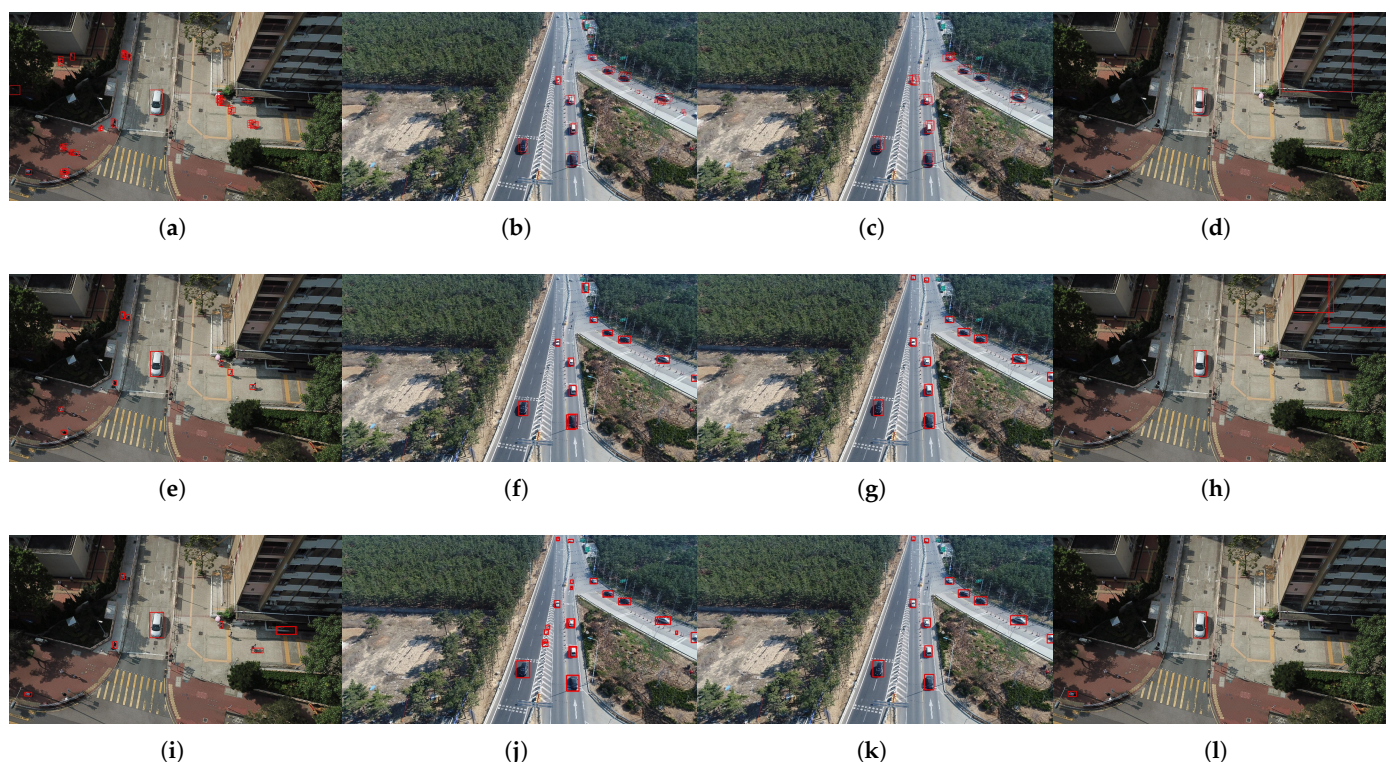


Figure 10. Cross-domain comparative experiments on the VisDrone2019 and UAVDT datasets.

To evaluate the cross-domain object detection performance of different models across UAV datasets, this study conducts quantitative experiments using the validation sets of the VisDrone2019 and UAVDT datasets. To ensure the fairness and comparability of the experimental results, all models were evaluated under identical inference configurations. The input image size was uniformly set to 640×640 pixels, while both the confidence threshold and the IoU threshold for NMS were set to 0.5. When calculating metrics such as precision and recall, the IoU threshold for matching predicted bounding boxes with ground-truth boxes was likewise set to 0.5. To address differences in category definitions and category indices between the two datasets, the relevant objects were consistently mapped to three categories (car, bus, and truck), and a unified category-indexing scheme was adopted for evaluation. The evaluation metrics were calculated according to the COCO protocol and primarily include $mAP@0.5:0.95$, $mAP@0.5$, precision, and recall.

Table 12 presents the results of cross-domain quantitative experiments. Based on the experimental data in the table, the following three conclusions can be drawn:

Transferring a model trained on a complex domain to a simpler domain can lead to performance gains on coarse metrics, but at the cost of precision and localization accuracy. Taking Faster R-CNN as an example, after being trained on VisDrone2019, the model achieved an mAP@0.5 of 42.5% on UAVDT, substantially surpassing its performance under in-domain evaluation. This finding suggests that features learned from challenging aerial-view scenes can be effectively transferred to less demanding scenarios. However, this improvement entails a pronounced trade-off: cross-domain precision declined to 30.9%, while the APS metric only slightly increased from 6.1% to 7.2%. Although recall and mAP@0.5:0.95 improved, the substantial drop in precision indicates a sacrifice in output reliability rather than localization accuracy.

Dataset complexity imposes a fundamental constraint on the performance of object detection models. This effect is particularly pronounced in small-object detection and is clearly reflected in the asymmetric results of the cross-domain experiments. When models trained on the less complex domain are directly transferred to the more complex domain, all detectors exhibit substantial performance degradation. For example, in the UAVDT to VisDrone2019 transfer setting, the mAP@0.5 of YOLOv13n drops sharply from its in-domain performance of 20.0% to 7.0%, while its APS declines dramatically from 7.7 to 0.1. Similarly, the APS of RT-DETR decreases from 13.5 to 1.3. These results clearly demonstrate that complex datasets characterized by densely distributed small objects, severe occlusion, and substantial background clutter create a critical performance bottleneck for object detectors, one that cannot be overcome through domain transfer alone. By contrast, the reverse transfer from VisDrone2019 to UAVDT achieves a markedly higher performance retention rate, further confirming that the disparity in difficulty between domains is inherently asymmetric and primarily determined by the scene complexity of the target domain itself.

While in-domain training generally provides a strong and reliable baseline, the cross-domain results reveal a more nuanced picture. For some detectors, transferring from a complex domain (VisDrone2019) to a simpler domain (UAVDT) can yield higher mAP@0.5 due to improved recall, but this often comes at the cost of reduced precision and only marginal gains in small-object detection. Conversely, transferring from a simpler domain (UAVDT) to a more complex one (VisDrone2019) leads to catastrophic degradation in both overall accuracy and APS. Therefore, in-domain training tends to provide more consistent and reliable performance, particularly for small-object detection, although specific metrics such as mAP@0.5 may improve under certain cross-domain transfer settings. In practical deployment, cross-domain transfer may offer recall-oriented benefits in less demanding scenarios, but this comes at the expense of output reliability and small-object localization quality.

Table 12. Cross-domain quantitative experiment.

Detector	Train Data	Test Data	P	R	mAP@0.5	mAP@0.5:0.95	APS	APS Retention	APS Degradation
Faster R-CNN	VisDrone2019	VisDrone2019	52.2	48.9	28.2	13.3	6.1	118.0	−18.0
	VisDrone2019	UAVDT	30.9	68.6	42.5	19.6	7.2		
	UAVDT	UAVDT	38.0	68.5	12.9	4.1	1.8	5.6	94.4
	UAVDT	VisDrone2019	40.4	26.8	6.3	2.3	0.1		

Table 12. Cont.

Detector	Train Data	Test Data	P	R	mAP@0.5	mAP@0.5:0.95	APS	APS Retention	APS Degradation
YOLOv13n	VisDrone2019	VisDrone2019	60.5	50.5	30.2	22.0	6.4	143.8	−43.8
	VisDrone2019	UAVDT	44.5	35.2	28.2	16.2	9.2		
	UAVDT	UAVDT	40.2	54.1	20.0	11.6	7.7	1.3	98.7
	UAVDT	VisDrone2019	28.9	20.3	7.0	4.2	0.1		
RT-DETR	VisDrone2019	VisDrone2019	76.3	42.5	13.9	9.5	4.7	178.7	−78.7
	VisDrone2019	UAVDT	58.0	64.7	19.0	10.7	8.4		
	UAVDT	UAVDT	48.2	70.4	24.7	24.7	13.5	9.6	90.4
	UAVDT	VisDrone2019	36.6	22.4	8.5	5.5	1.3		

Note: APS Retention = APS_{cross} / APS_{in} ; APS Degradation = $(APS_{in} - APS_{cross}) / APS_{in}$. All results in this table use a unified three-class protocol (car/bus/truck, 640×640 , conf = 0.5, NMS = 0.5, COCO metrics) and are not directly comparable to Tables 9 and 10, which use original category definitions and framework-recommended settings.

6. Conclusions and Future Work

6.1. Conclusions

Different from existing reviews of general small-object detection or general UAV object detection, this review provides a task-specific review of deep learning-based small-object detection in UAV imagery from problem-driven, deployment-oriented, and symmetry-informed perspectives. In this review, the symmetry-informed perspective refers to the need to maintain stable detection performance under scale variation and transformation-related changes caused by UAV imaging, including changes in viewpoint, object orientation, image quality, and scene conditions. By integrating challenge analysis, scenario-aware method review, and partially unified benchmark evaluation, this work provides a practical reference for method selection and deployment in UAV small-object detection.

The review and experimental results indicate that the performance of deep learning-based UAV small-object detectors is not determined by the detection framework alone. It is jointly influenced by object-scale distributions, imaging conditions, background complexity, training strategies, and onboard computational-resource constraints. Substantial scale variation and transformation-related appearance changes can weaken feature representation and reduce detection stability for small objects. Experiments based on the VisDrone2019 and UAVDT datasets further demonstrate that the performance rankings of different detectors vary across different datasets, indicating that existing methods remain highly dependent on dataset characteristics and specific application scenarios. There is no single detector capable of maintaining optimal performance across all UAV scenarios.

Under the specific experimental configuration evaluated in this review, the lightweight YOLO implementations exhibit a measured trade-off among detection accuracy, inference efficiency, and deployment cost that may be advantageous for resource-constrained applications. This observation is configuration-specific and should not be interpreted as a general claim about the inherent superiority of the YOLO family. In different scenarios, particularly those with different hardware platforms, training protocols, or dataset characteristics, other paradigms such as Faster R-CNN or SSD may offer complementary strengths.

It should be emphasized that detector selection involves inherent trade-offs among accuracy, speed, and computational cost. For instance, on UAVDT, Faster R-CNN achieves the highest mAP@0.5 but requires 474.1 GFLOPs, whereas YOLOv7-tiny attains 30.2% mAP with only 13.0 GFLOPs, a 36-fold reduction in computation. On VisDrone2019, YOLOv8n (3.0M parameters, 8.1 GFLOPs) outperforms Faster R-CNN (28.3M parameters,

474.1 GFLOPs) by 11.3 percentage points in mAP. These varying results across datasets indicate that the optimal choice depends on the specific accuracy, speed, and resource constraints of the target application.

The experimental results of this review also reveal an important methodological insight: some detectors that perform well on comprehensive evaluation metrics, such as overall mAP, may not deliver satisfactory results in small-object detection tasks. Therefore, when selecting performance evaluation metrics for small-object detectors used in UAV applications, APS should be prioritized rather than relying solely on overall mAP.

In contrast, while two-stage detectors and Transformer-based detectors offer certain advantages under specific conditions, their performance is more dependent on the degree of alignment between task characteristics, data distribution, and training configurations, and they face significant challenges under current deployment conditions with limited onboard resources. Therefore, when designing practical UAV imagery object detection systems, the focus should not be solely on achieving optimal single-point performance; rather, the design should be task-driven, balancing accuracy, cross-domain robustness, and hardware compatibility. This study has several limitations:

1. This evaluation of detector performance is limited to the VisDrone2019 and UAVDT datasets, which primarily reflect urban traffic and low-altitude surveillance scenarios. Adverse weather conditions, nighttime environments, disaster scenarios, and a broader range of drone applications are not fully covered.
2. Although the experiment demonstrated the performance of detectors under different paradigms, it did not include some of the latest methods specifically optimized for UAV imagery datasets.
3. This performance evaluation employs the training configurations recommended by each framework, rather than a strictly standardized training configuration. Although this asymmetry was intentionally adopted to respect the standard practices of each framework, it introduces confounding factors that prevent direct comparisons across detectors under controlled conditions. Therefore, the reported results should be regarded as reference observations specific to the evaluated configurations, rather than a universally applicable ranking of frameworks.
4. The experiments are restricted to RGB-based object detection and do not include systematic multimodal evaluations using infrared imagery or LiDAR.
5. The performance of models trained on the UAVDT dataset is limited by the 100-epoch setting. Although 100 epochs of training are sufficient for VisDrone2019, this may not be enough to achieve full convergence for the larger UAVDT dataset (approximately 23,000 images). Therefore, the in-domain performance of UAVDT reported in this review may be underestimated, and the cross-domain results with UAVDT as the target domain should be considered preliminary observations that require further training to validate;
6. Deployment feasibility is assessed mainly through model complexity, inference speed, and computational cost, whereas real-world tests on embedded UAV hardware are lacking. The deployment analysis in this review is based solely on desktop GPU test results and can only provide a preliminary reference for deployment efficiency. During actual deployment, further consideration must be given to issues such as power consumption, memory access, communication latency, and integration with the flight control system.

6.2. Future Work

In view of the above limitations and recent research progress, future work on deep learning-based UAV small-object detection should focus on the following directions:

1. Fine-grained and transformation-robust representation for small objects. In low-altitude scenes with densely distributed objects, detectors still struggle with weak textures, low resolution, blurred boundaries, and substantial scale variation. This indicates that current models have limited representational capacity for preserving small-object details under changing viewpoints, scales, and imaging conditions. Future work should improve shallow-detail preservation, high-resolution feature reconstruction, multiscale fusion, and robustness to transformation-related appearance variation, while exploring super-resolution, generative enhancement, and foundation models for small-object detection.
2. Introducing multi-modal data sources to enhance robust perception in complex environments. At present, factors such as nighttime scenes, rain, fog, backlighting, and motion blur still create major difficulties for detection systems that rely only on visible-light images. Future work should integrate modalities such as infrared, LiDAR, and platform status information, while addressing issues including cross-modal alignment, modal conflicts, missing modalities, and reliability assessment, in order to improve detection stability under complex conditions.
3. Lightweight deployment for airborne platforms. Lightweight YOLO models have demonstrated strong engineering adaptability, but they are still constrained by the limited computational resources of UAV airborne platforms. Future efforts should strengthen pruning, quantization, distillation, reparameterization, dynamic inference, and hardware-software co-design to enhance model deployment capability on embedded airborne platforms.
4. Spatiotemporal modeling for dynamic tasks. Since single-frame detection struggles to fully utilize the temporal information collected during continuous flight and is prone to performance degradation in cases of occlusion, blur, and short-term object disappearance, future work should strengthen cross-frame association, motion compensation, temporal contextual modeling, and joint optimization of detection and tracking. Adaptability design should also be carried out for tasks such as traffic monitoring, disaster search and rescue, and border patrol.
5. Dataset construction and evaluation system improvement for cross-domain applications. Existing public datasets have insufficient coverage of extreme weather, nighttime scenarios, disaster environments, and cross-platform differences, limiting the generalization ability of current detection methods under distribution shift conditions. Future efforts should focus on constructing datasets that more closely resemble real-world tasks and establishing a multidimensional evaluation system encompassing robustness, temporal stability, inference efficiency, energy consumption, and deployment costs.
6. Multi-dimensional collaborative optimization. The performance of UAV small-object detection models is constrained by multiple factors, including data, training, hardware, imaging conditions, and task environments. Relying solely on local architectural improvements to gain detection benefits has become highly limited. Future research should promote end-to-end collaborative design encompassing the task environment, detection model, and onboard deployment, forming a closed-loop optimization framework oriented towards real-world tasks.

Author Contributions: Conceptualization, L.W., C.M., A.G., S.M. and Y.Z.; methodology, L.W., A.G., S.M. and Y.Z.; validation, L.W., A.G. and C.M.; writing—original draft preparation, L.W., A.G. and C.M.; writing—review and editing, L.W., C.M., S.M. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Research Project On UAV Swarm Task Distribution And Trajectory Planning For Mountain Counter-Terrorism No. CTC&IE20240301.

Data Availability Statement: VisDrone2019: <https://github.com/VisDrone/VisDrone-Dataset> (accessed on 13 September 2026); UAVDT: <https://sites.google.com/site/daviddo0323/projects/uavdt> (accessed on 13 September 2026).

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

References

1. Zhang, Z.; Zhu, L. A Review on Unmanned Aerial Vehicle Remote Sensing: Platforms, Sensors, Data Processing Methods, and Applications. *Drones* **2023**, *7*, 398. [\[CrossRef\]](#)
2. Ke, S.; Yang, K.; Zhan, C.; Liao, S.; Ma, Y.; Shang, H.; Shen, E.; Li, Z.; Zhang, Z.; Chen, Z. Advances in earth observation using unmanned aerial vehicles: A bibliometric and content analysis (2000–2024). *Geocarto Int.* **2025**, *40*, 2600779. [\[CrossRef\]](#)
3. Kulsoom, F.; Nazeer Chaudhry, H.; Mirza, A.; Mehmood, Z.; Arshad, F. Integrating Unmanned Aerial Vehicle Remote Sensing with Internet of Things Frameworks: A Survey of Methods and Applications. *Cureus J. Comput. Sci.* **2026**, *3*, es44389-025-00076-6. [\[CrossRef\]](#)
4. Chen, Y.; Liu, X.; Zhu, B.; Zhu, D.; Zuo, X.; Li, Q. UAV Image-Based 3D Reconstruction Technology in Landslide Disasters: A Review. *Remote Sens.* **2025**, *17*, 3117. [\[CrossRef\]](#)
5. Gu, Z.; Wang, S.; Yu, K.; Wang, Y.; Zhang, X. UEBNet: A Novel and Compact Instance Segmentation Network for Post-Earthquake Building Assessment Using UAV Imagery. *Remote Sens.* **2025**, *17*, 3530. [\[CrossRef\]](#)
6. Jia, X.; Huang, B.; Jia, B. FAD-YOLOv11: Fence Anomaly Detection with Wavelet Convolution and Trajectory Verification. In Proceedings of the 2025 International Conference on Virtual Reality and Visualization (ICVRV), Los Alamitos, CA, USA, 19–21 December 2025; pp. 630–635. [\[CrossRef\]](#)
7. Ye, J.; Kyrkou, C. A Review of Vision-Based Vehicle Detection for UAV-Based Traffic Monitoring: Experimental Insights and Future Directions. *IEEE Trans. Intell. Transp. Syst.* **2026**, *27*, 10514–10532. [\[CrossRef\]](#)
8. Yang, F.; He, M.; Liu, J.; Jin, H. RMH-YOLO: A Refined Multi-Scale Architecture for Small-Target Detection in UAV Aerial Imagery. *Sensors* **2025**, *25*, 7088. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Muzammul, M.; Li, X. Comprehensive Review of Deep Learning-Based Tiny Object Detection: Challenges, Strategies, and Future Directions. *Knowl. Inf. Syst.* **2025**, *67*, 3825–3913. [\[CrossRef\]](#)
10. Tong, K.; Wu, Y.; Zhou, F. Recent Advances in Small Object Detection Based on Deep Learning: A Review. *Image Vis. Comput.* **2020**, *97*, 103910. [\[CrossRef\]](#)
11. Tong, K.; Wu, Y. Deep Learning-Based Detection from the Perspective of Small or Tiny Objects: A Survey. *Image Vis. Comput.* **2022**, *123*, 104471. [\[CrossRef\]](#)
12. Feng, Q.; Xu, X.; Wang, Z. Deep Learning-Based Small Object Detection: A Survey. *Math. Biosci. Eng.* **2023**, *20*, 6551–6590. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Chen, G.; Wang, H.; Chen, K.; Li, Z.; Song, Z.; Liu, Y.; Chen, W.; Knoll, A. A Survey of the Four Pillars for Small Object Detection: Multiscale Representation, Contextual Information, Super-Resolution, and Region Proposal. *IEEE Trans. Syst. Man Cybern. Syst.* **2022**, *52*, 936–953. [\[CrossRef\]](#)
14. Hua, W.; Chen, Q. A Survey of Small Object Detection Based on Deep Learning in Aerial Images. *Artif. Intell. Rev.* **2025**, *58*, 162. [\[CrossRef\]](#)
15. Wu, Y.; Tong, K. Research advances on deep learning-based small object detection in UAV aerial images. *Acta Aeronaut. Astronaut. Sin.* **2025**, *46*, 181–207. [\[CrossRef\]](#)
16. Tang, G.; Ni, J.; Zhao, Y.; Gu, Y.; Cao, W. A Survey of Object Detection for UAVs Based on Deep Learning. *Remote Sens.* **2024**, *16*, 149. [\[CrossRef\]](#)
17. Mittal, P.; Singh, R.; Sharma, A. Deep Learning-Based Object Detection in Low-Altitude UAV Datasets: A Survey. *Image Vis. Comput.* **2020**, *104*, 104046. [\[CrossRef\]](#)
18. Jian, B.; Qu, R.; Li, Y.; Li, C. Object detection in UAV imagery based on deep learning: Review. *Acta Aeronaut. Astronaut. Sin.* **2021**, *42*, 137–151. [\[CrossRef\]](#)
19. Chen, C.; Liu, M.Y.; Tuzel, O.; Xiao, J. R-CNN for Small Object Detection. In *Computer Vision—ACCV 2016; Lecture Notes in Computer Science*; Springer: Taipei, Taiwan, 2017; Volume 10115, pp. 214–230. [\[CrossRef\]](#)

20. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983. [\[CrossRef\]](#)
21. Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards Large-Scale Small Object Detection: Survey and Benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13467–13488. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014*; Springer International Publishing: Cham, Switzerland, 2014; Volume 8693, pp. 740–755. [\[CrossRef\]](#)
23. Torralba, A.; Fergus, R.; Freeman, W.T. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1958–1970. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [\[CrossRef\]](#)
25. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [\[CrossRef\]](#)
26. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Xu, Y.; Yu, G.; Wang, Y.; Wu, X.; Ma, Y. Car Detection from Low-Altitude UAV Imagery with the Faster R-CNN. *J. Adv. Transp.* **2017**, *2017*, 2823617. [\[CrossRef\]](#)
28. Bakshi, A.P. A Comprehensive Review of Object Detection: From Handicraft Features to Deep Convolutional and Transformer-based Architectures. *Int. J. Comput. Appl.* **2026**, *187*, 41–49. [\[CrossRef\]](#)
29. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2016; Volume 9905, pp. 21–37. [\[CrossRef\]](#)
30. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767v1.
32. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475. [\[CrossRef\]](#)
33. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850v2.
34. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9626–9635. [\[CrossRef\]](#)
35. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524v1.
36. Le, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733v2.
37. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. Version 8.0.0. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 13 September 2026).
38. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Computer Vision—ECCV 2020*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12346, pp. 213–229. [\[CrossRef\]](#)
39. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2021**, arXiv:2010.04159v4.
40. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *arXiv* **2022**, arXiv:2203.03605v4.
41. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974. [\[CrossRef\]](#)
42. Liu, C.; Xiang, X.; Duan, Z.; Li, W.; Fan, Q.; Gao, Y. DON'T NEED RETRAINING: A Mixture of DETR and Vision Foundation Models for Cross-Domain Few-Shot Object Detection. In *Proceedings of the Advances in Neural Information Processing Systems*; Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2025; Volume 38, pp. 146347–146371. [\[CrossRef\]](#)
43. Tao, L.; Wei, G.; Wang, Z.; Qi, Z.; Li, Y.; Zhang, H. UAV-OVD: Open-Vocabulary Object Detection in UAV Imagery via Multi-Level Text-Guided Decoding. *Drones* **2025**, *9*, 495. [\[CrossRef\]](#)

44. Choi, S.; Go, K.; Jang, T. ZERO: Industry-ready Vision Foundation Model with Multi-modal Prompts. *arXiv* **2025**, arXiv:2507.04270v4.
45. Li, N.; Ye, M.; Zhou, L.; Li, S.; Tang, S.; Ji, L.; Zhu, C. Multimodal Causal Reasoning for UAV Object Detection. In Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems, San Diego, CA, USA, 2–7 December 2025.
46. Xu, R.; Liu, Y.; Tang, Y.; Bai, S.; Ye, X.; Yu, B.; Guo, Y.; Lu, J.; Zhou, J. DisDop: Distillation with Domain Priors for Open-Vocabulary Aerial Object Detection. *arXiv* **2026**, arXiv:2605.24639.
47. Zhang, P.; Zhang, Y.; Wu, H.; Liu, X.; Hou, Y.; Wang, L. Language-Guided Object Localization via Refined Spotting Enhancement in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5621315. [[CrossRef](#)]
48. Salehi, A.; Karami, E.; Noey, S.; Noey, S.; Yamada, M.; Hosseini, R.; Sabokrou, M. TIPS Over Tricks: Simple Prompts for Effective Zero-shot Anomaly Detection. *arXiv* **2026**, arXiv:2602.03594v1.
49. Wang, H.; Jia, T.; Wang, Q.; Zuo, W. WS-SAM: Generalizing SAM to Weakly Supervised Object Detection with Category Label. *IEEE Trans. Image Process.* **2025**, *34*, 4052–4066. [[CrossRef](#)] [[PubMed](#)]
50. Li, Y.; Guo, W.; Yang, X.; Liao, N.; Zhang, S.; Yu, Y.; Yu, W.; Yan, J. Exploiting Unlabeled Data with Multiple Expert Teachers for Open Vocabulary Aerial Object Detection and Its Orientation Adaptation. *Int. J. Comput. Vis.* **2026**, *134*, 156. [[CrossRef](#)]
51. Lavoie, M.A.; Mahmoud, A.; Waslander, S.L. Large Self-Supervised Models Bridge the Gap in Domain Adaptive Object Detection. *arXiv* **2025**, arXiv:2503.23220v1.
52. Caron, M.; Touvron, H.; Misra, I.; Jegou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9630–9640. [[CrossRef](#)]
53. Xu, K.; Wu, S. Tiny-YOLO SAM: Fast Hybrid Image Segmentation. *arXiv* **2025**, arXiv:2512.22193v1.
54. Ferdous, S.N.; Mostofa, M.; Nasrabadi, N.M. Super resolution-assisted deep aerial vehicle detection. In *Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*; Pham, T., Ed.; International Society for Optics and Photonics, SPIE: Bellingham, WA, USA, 2019; Volume 11006, p. 1100617. [[CrossRef](#)]
55. Wang, D.; Gao, Z.; Fang, J.; Li, Y.; Xu, Z. Improving UAV Aerial Imagery Detection Method via Superresolution Synergy. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 3959–3972. [[CrossRef](#)]
56. Wang, W.; Xu, J.; Zhang, R. Optimized Small Object Detection in Low Resolution Infrared Images Using Super Resolution and Attention Based Feature Fusion. *PLoS ONE* **2025**, *20*, e0328223. [[CrossRef](#)] [[PubMed](#)]
57. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for Small Object Detection. *arXiv* **2019**, arXiv:1902.07296v1.
58. Li, Z.; Lian, S.; Pan, D.; Wang, Y.; Liu, W. AD-det: Boosting Object Detection in UAV Images with Focused Small Objects and Balanced Tail Classes. *Remote Sens.* **2025**, *17*, 1556. [[CrossRef](#)]
59. Akyon, F.C.; Altinuc, S.O.; Temizel, A. Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 966–970. [[CrossRef](#)]
60. Muzammul, M.; Algarni, A.; Ghadi, Y.Y.; Assam, M. Enhancing UAV Aerial Image Analysis: Integrating Advanced SAHI Techniques with Real-Time Detection Models on the VisDrone Dataset. *IEEE Access* **2024**, *12*, 21621–21633. [[CrossRef](#)]
61. Tang, D.; Cao, X.; Wu, X.; Li, J.; Yao, J.; Bai, X.; Jiang, D.; Li, Y.; Meng, D. AeroGen: Enhancing Remote Sensing Object Detection with Diffusion-Driven Data Generation. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 3614–3624. [[CrossRef](#)]
62. Fang, H.; Han, B.; Zhang, S.; Zhou, S.; Hu, C.; Ye, W.M. Data Augmentation for Object Detection via Controllable Diffusion Models. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024; pp. 1246–1255. [[CrossRef](#)]
63. Wu, Q. Domain-adapted Diffusion for Synthetic Data Augmentation in UAV Small Object Detection. In *2025 22nd International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*; IEEE: Chengdu, China, 2025; pp. 1–6. [[CrossRef](#)]
64. Xing, W.; Cui, Z.; Qi, J. HRCTNet: A Hybrid Network with High-Resolution Representation for Object Detection in UAV Image. *Complex Intell. Syst.* **2023**, *9*, 6437–6457. [[CrossRef](#)]
65. Zhang, H.; Sun, W.; Sun, C.; He, R.; Zhang, Y. HSP-YOLOv8: UAV Aerial Photography Small Target Detection Algorithm. *Drones* **2024**, *8*, 453. [[CrossRef](#)]
66. Wang, R.; Lin, C.; Li, Y. RPLFDet: A Lightweight Small Object Detection Network for UAV Aerial Images with Rational Preservation of Low-Level Features. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 5013514. [[CrossRef](#)]
67. Huo, T.; Liu, Z.; Zhang, S.; Wu, J.; Yuan, R.; Duan, S.; Wang, L. CDNet: Object Detection Based on Cross-Level Aggregation and Deformable Attention for UAV Aerial Images. *Vis. Comput.* **2025**, *41*, 4603–4621. [[CrossRef](#)]
68. Hu, L.; Xue, P.; Guo, B.; Chen, Y.; Zha, W.; Tian, J. STAIR-DETR: A Synergistic Transformer Integrating Statistical Attention and Multi-Scale Dynamics for UAV Small Object Detection. *Sensors* **2025**, *25*, 7681. [[CrossRef](#)] [[PubMed](#)]

69. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [\[CrossRef\]](#)
70. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768. [\[CrossRef\]](#)
71. Hong, M.; Li, S.; Yang, Y.; Zhu, F.; Zhao, Q.; Lu, L. SSPNet: Scale Selection Pyramid Network for Tiny Person Detection from UAV Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8018505. [\[CrossRef\]](#)
72. Zhang, Y.; Wu, C.; Guo, W.; Zhang, T.; Li, W. CFANet: Efficient Detection of UAV Image Based on Cross-Layer Feature Aggregation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5608911. [\[CrossRef\]](#)
73. Jiang, L.; Yuan, B.; Du, J.; Chen, B.; Xie, H.; Tian, J.; Yuan, Z. MFFSODNet: Multiscale Feature Fusion Small Object Detection Network for UAV Aerial Images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5015214. [\[CrossRef\]](#)
74. Zhang, F.; Ji, H.; Zhang, Y.; Su, Z. Consistent and Comprehensive Scale Aggregation Network for Drone-View Small Object Detection. *Neural Netw.* **2026**, *195*, 108250. [\[CrossRef\]](#) [\[PubMed\]](#)
75. Wang, Y.; Bai, X.; Hu, B.; Xu, C.; Chen, H.; Chung, V.; Li, T.; Chen, X. MGDFIS: Multi-scale Global-detail Feature Integration Strategy for Small Object Detection. *arXiv* **2025**, arXiv:2506.12697v2.
76. Bai, X.; Wang, Y.; Xu, C.; Hu, B.; Han, K.; Pan, R.; Niu, X.; Guan, X.; Fu, L.; Ye, P. CollabOD: Collaborative Multi-Backbone with Cross-Scale Vision for UAV Small Object Detection. *arXiv* **2026**, arXiv:2603.05905v1.
77. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [\[CrossRef\]](#)
78. Bai, J.; Hu, H.; Liu, X.; Zhuang, S.; Wang, Z. UAV Image Object Detection Based on Self-Attention Guidance and Global Feature Fusion. *Image Vis. Comput.* **2024**, *151*, 105262. [\[CrossRef\]](#)
79. Zhang, J.; Song, L.; Gao, M.; Li, W.; Wei, Z. Interactive feature pyramid network for small object detection in UAV aerial images. *Displays* **2026**, *92*, 103352. [\[CrossRef\]](#)
80. Li, J.; Zheng, C.; Chen, P.; Zhang, J.; Wang, B. Small Object Detection in UAV Imagery Based on Channel-Spatial Fusion Cross Attention. *Signal Image Video Process.* **2025**, *19*, 302. [\[CrossRef\]](#)
81. Yuan, Z.; Gong, J.; Guo, B.; Wang, C.; Liao, N.; Song, J.; Wu, Q. Small Object Detection in UAV Remote Sensing Images Based on Intra-Group Multi-Scale Fusion Attention and Adaptive Weighted Feature Fusion Mechanism. *Remote Sens.* **2024**, *16*, 4265. [\[CrossRef\]](#)
82. Ren, B.; Xu, B.; Yang, X.; Pu, Y.; Wang, J.; Deng, Z. Context-aware Aerial Object Detection: Leveraging Inter-Object and Background Relationships. *arXiv* **2024**, arXiv:2404.04140v2.
83. Li, H.; Cui, Z. MS-MRFNet: A Multi-scale and Multi-receptive Field Network for UAV Aerial Object Detection. In *Proceedings of the MultiMedia Modeling*; Lokoč, J., Peška, L., Zahálka, J., Rudinac, S., Kastner, M., Chen, J., Hu, M.C., Wu, J., Sharma, U., Eds.; Springer: Singapore, 2026; pp. 233–246.
84. Li, Z.; Qi, L. MSA-DETR: A multi-scale attention augmented model for small object detection in UAV imagery. *Remote Sens.* **2026**, *18*, 1179. [\[CrossRef\]](#)
85. Xu, Y.; Ji, H. Contextual-semantic Interactive Perception Network for Small Object Detection in UAV Aerial Images. *Remote Sens.* **2025**, *17*, 3581. [\[CrossRef\]](#)
86. Xue, H.; Gao, T.; Tang, Z.; Xia, Y.; Wang, L.; Li, L. A Configurable Global Context Reconstruction Hybrid Detector for Enhanced Small Object Detection in UAV Aerial Imagery. *Comput. Vis. Image Underst.* **2026**, *263*, 104598. [\[CrossRef\]](#)
87. Tao, W.; Wang, X.; Yan, T.; Wang, Y.; Yan, J. Small object detection in complex backgrounds with multi-scale attention and global relation modeling. *arXiv* **2026**, arXiv:2603.03788v1.
88. Wu, S.; Lu, X.; Guo, C.; Guo, H. MV-YOLO: An Efficient Small Object Detection Framework Based on Mamba. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5632814. [\[CrossRef\]](#)
89. Hu, L.; Yuan, J.; Cheng, B.; Xu, Q. CSFPR-RTDETR: Real-time Small Object Detection Network for UAV Images Based on Cross-Spatial-Frequency Domain and Position Relation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5638219. [\[CrossRef\]](#)
90. Tong, Y.; Ye, H.; Yang, J.; Yang, X. ACD-DETR: Adaptive Cross-Scale Detection Transformer for Small Object Detection in UAV Imagery. *Sensors* **2025**, *25*, 5556. [\[CrossRef\]](#) [\[PubMed\]](#)
91. Zhou, Y.; Wei, Y. UAV-DETR: An Enhanced RT-DETR Architecture for Efficient Small Object Detection in UAV Imagery. *Sensors* **2025**, *25*, 4582. [\[CrossRef\]](#) [\[PubMed\]](#)
92. Ahmad, H.M.; Rahimi, A. Dynamic focal loss for imbalanced learning. *SN Comput. Sci.* **2025**, *6*, 657. [\[CrossRef\]](#)
93. Liu, Y.; He, M.; Hui, B. ESO-DETR: An Improved Real-Time Detection Transformer Model for Enhanced Small Object Detection in UAV Imagery. *Drones* **2025**, *9*, 143. [\[CrossRef\]](#)
94. Chen, J.; Chen, D.; Xiong, F.; Qian, Y.; Xiao, L. Balanced hierarchical contrastive learning with decoupled queries for fine-grained object detection in remote sensing images. *arXiv* **2025**, arXiv:2512.24074v1.

95. Di, R.; Fan, H.; Ma, Y.; Wang, J.; Qian, R. GAME-YOLO: Global Attention and Multi-Scale Enhancement for Low-Visibility UAV Detection with Sub-Pixel Localization. *Entropy* **2025**, *27*, 1263. [\[CrossRef\]](#) [\[PubMed\]](#)
96. Wang, J.; Xu, C.; Yang, W.; Yu, L. A normalized gaussian wasserstein distance for tiny object detection. *arXiv* **2022**, arXiv:2110.13389v2.
97. Qu, J.; Tang, Z.; Zhang, L.; Zhang, Y.; Zhang, Z. Remote Sensing Small Object Detection Network Based on Attention Mechanism and Multi-Scale Feature Fusion. *Remote Sens.* **2023**, *15*, 2728. [\[CrossRef\]](#)
98. Guan, Z.; Fu, X.; Huang, P.; Zhang, H.; Du, H.; Liu, Y.; Wang, Y.; Ma, Q. Gaussian Combined Distance: A Generic Metric for Object Detection. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 8000905. [\[CrossRef\]](#)
99. Liao, D.; Zhang, J.; Tao, Y.; Jin, X. ATBHC-YOLO: Aggregate Transformer and Bidirectional Hybrid Convolution for Small Object Detection. *Complex Intell. Syst.* **2025**, *11*, 38. [\[CrossRef\]](#)
100. Han, Z.; Xu, Y.; Li, J.; Plaza, J.; Plaza, A.; Wei, Z.; Wu, Z. CAEM-DETR: Small Aerial Target Detection via Contrastive Attention-Enhanced Multidomain Fidelity Fusion. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5646614. [\[CrossRef\]](#)
101. Sun, H.; Li, Y.; Yang, L.; Cao, X.; Zhang, B. Uncertainty-Aware Gradient Stabilization for Small Object Detection. In Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–25 October 2025; pp. 8407–8417. [\[CrossRef\]](#)
102. Cheng, Y.; Wang, T.; Zhang, W. Real-time UAV Small Object Detection: An Efficient Approach Using SGFNet and Dynamic Loss Optimization. *Digit. Signal Process.* **2026**, *168*, 105543. [\[CrossRef\]](#)
103. Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.S. Detecting Tiny Objects in Aerial Images: A Normalized Wasserstein Distance and a New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 79–93. [\[CrossRef\]](#)
104. Shi, S.; Fang, Q.; Zhao, T.; Xu, X. Similarity Distance-Based Label Assignment for Tiny Object Detection. *arXiv* **2024**, arXiv:2407.02394v3.
105. Ge, L.; Wang, G.; Zhang, T.; Zhuang, Y.; Chen, H.; Dong, H.; Chen, L. Adaptive Dynamic Label Assignment for Tiny Object Detection in Aerial Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 6201–6214. [\[CrossRef\]](#)
106. Shi, S.; Fang, Q.; Xu, X. Dynamic Adaptive Label Assignment for Tiny Object Detection in Remote Sensing Images. *CAAI Trans. Intell. Technol.* **2026**, *11*, 428–446. [\[CrossRef\]](#)
107. Fan, P.; Wang, J.; Zhang, Y.; Liu, M.; Chen, Y.; Han, G. A Lightweight Small-Object Detection Algorithm for Drone Images with Optimized Feature Fusion and Knowledge Distillation. *J. Supercomput.* **2025**, *82*, 20. [\[CrossRef\]](#)
108. Belmouhcine, A.; Pham, M.T.; Hadjisoteriou, A.; Dimitriou, N. Upsampling and knowledge distillation improve small-object detection in satellite remote sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 16607–16622. [\[CrossRef\]](#)
109. Ma, Y.; Xi, C.; Ma, T.; Sun, H.; Lu, H.; Xu, X.; Xu, C. I-YOLOv11n: A Lightweight and Efficient Small Target Detection Framework for UAV Aerial Images. *Sensors* **2025**, *25*, 4857. [\[CrossRef\]](#) [\[PubMed\]](#)
110. Wang, H.; Li, Y.; Zhou, W.; Zhang, H. ADD-Net: Adaptive dynamic architecture with knowledge distillation for UAV small object detection. *Digit. Signal Process.* **2026**, *174*, 105990. [\[CrossRef\]](#)
111. Wang, C.; Liu, F. ISDC-YOLO: A Lightweight Small Object Detection Algorithm Based on Inverted Bottleneck and Semantic Detail Fusion Network for UAV Aerial Image. *IEEE Access* **2025**, *13*, 147874–147891. [\[CrossRef\]](#)
112. Tan, T.; Peng, X.; Zhang, J.; Zuo, H.; Zhang, Y.; Wu, Y.; Li, H. LDSNet: A lightweight detail-sensitive network for small object detection in low-altitude UAV scenarios. *J. Imaging* **2026**, *12*, 209. [\[CrossRef\]](#) [\[PubMed\]](#)
113. Xiao, Y.; Di, N. SOD-YOLO: A Lightweight Small Object Detection Framework. *Sci. Rep.* **2024**, *14*, 25624. [\[CrossRef\]](#) [\[PubMed\]](#)
114. Li, W.; Zhang, Y. Enhanced small object detection in UAV aerial imagery through attention gated backbone and context aware fusion. *Sci. Rep.* **2026**, *16*, 2366. [\[CrossRef\]](#) [\[PubMed\]](#)
115. Wu, Q.; Yang, Y.; Han, R.; Sun, W. LDA-YOLO: A Lightweight Deblurring-Aware Network for Real-Time Object Detection in Blurred Aerial Images. *J. Real-Time Image Process.* **2026**, *23*, 23. [\[CrossRef\]](#)
116. Lu, W.; Chen, S.B.; Li, H.D.; Shu, Q.L.; Ding, C.H.Q.; Tang, J.; Luo, B. LEGNet: A lightweight edge-gaussian network for low-quality remote sensing image object detection. *arXiv* **2025**, arXiv:2503.14012v3.
117. Wang, C.; Chang, Y.; Xie, S.; Yang, X.; Tian, Y.; Sun, W.; Hu, J. Dual-domain attentions for unmanned aerial vehicle small object detection. *Eng. Appl. Artif. Intell.* **2026**, *163*, 112849. [\[CrossRef\]](#)
118. Yu, J.; Wu, J.; Tang, Y. A Lightweight Adaptive Feature Selection Network for Enhanced Small Object Detection in UAV Imagery. *Vis. Comput.* **2025**, *41*, 10599–10615. [\[CrossRef\]](#)
119. Xia, Y.; Liu, C.; Xiang, T.; Tu, Z. EFSI-DETR: Efficient frequency-semantic integration for real-time small object detection in UAV imagery. *arXiv* **2026**, arXiv:2601.18597v1.
120. Guo, H.; Wu, Q.; Wang, Y. AUHF-DETR: A Lightweight Transformer with Spatial Attention and Wavelet Convolution for Embedded UAV Small Object Detection. *Remote Sens.* **2025**, *17*, 1920. [\[CrossRef\]](#)
121. Wu, W.; Liu, A.; Hu, J.; Mo, Y.; Xiang, S.; Duan, P.; Liang, Q. EUAVDet: An Efficient and Lightweight Object Detector for UAV Aerial Images with an Edge-Based Computing Platform. *Drones* **2024**, *8*, 261. [\[CrossRef\]](#)

122. Zhang, Q.; Jia, X.; Abdelmoniem, A.M.; Zhou, L.; An, J. Small Target Detection Based on Mask-Enhanced Attention Fusion of Visible and Infrared Remote Sensing Images. *arXiv* **2026**, arXiv:2603.06925v1.
123. Li, F.; Lv, X.; Zhao, M.; Wu, W. IFD-YOLO: A Lightweight Infrared Sensor-Based Detector for Small UAV Targets. *Sensors* **2025**, *25*, 7449. [[CrossRef](#)] [[PubMed](#)]
124. Li, J.; Liu, Q.; Fu, Y. Infrared small object detection based on attention mechanism. *Acta Aeronaut. Astronaut. Sin.* **2024**, *45*, 628959. [[CrossRef](#)]
125. Li, J.; Pan, Y. DNR-YOLO: An infrared small object detection algorithm for UAV-based imagery. *Discov. Appl. Sci.* **2026**, *8*, 423. [[CrossRef](#)]
126. Fang, H.; Guo, S.; Chen, Q.; Chang, Y.; Yan, L. Spatio-Temporal Context Learning with Temporal Difference Convolution for Moving Infrared Small Target Detection. In Proceedings of the 40th AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; Volume 40, pp. 3840–3848. [[CrossRef](#)]
127. Cui, Y.; Li, S.; Yang, X.; Wang, G.; Wang, Y. UAMF: A lightweight and powerful infrared small target detection model. *Appl. Soft Comput.* **2026**, *196*, 115121. [[CrossRef](#)]
128. Nan, S.; Bin, S.; Sun, G. CFPM-DETR: A lightweight model for UAV infrared small target detection. *Signal Image Video Process.* **2026**, *20*, 68. [[CrossRef](#)]
129. Ji, J.; Zhao, Y.; Li, A.; Ma, X.; Wang, C.; Lin, Z. Dense Small Object Detection Algorithm for Unmanned Aerial Vehicle Remote Sensing Images in Complex Backgrounds. *Digit. Signal Process.* **2025**, *158*, 104938. [[CrossRef](#)]
130. Zhou, H.; Cao, X.; Sun, K.; Wu, T.; Deng, B. Dense Small Object Detection via Multi-Scale Fusion and Context Information Enhancement. *J. Supercomput.* **2025**, *81*, 955. [[CrossRef](#)]
131. Teng, X.; Zhang, W.; Liu, T.; Yang, J.; Ma, M. Stff-rtdeetr: A Small Object Detection Algorithm Based on Drone Aerial Photography. *J. Supercomput.* **2025**, *81*, 928. [[CrossRef](#)]
132. Zhang, D.; Lin, C.; Pan, Y. A Lightweight Small Object Detection Network Inspired by the Visual Area V2. *Comput. Electr. Eng.* **2025**, *126*, 110471. [[CrossRef](#)]
133. Wan, Z.; Lan, Y.; Xu, Z.; Shang, K.; Zhang, F. DAU-YOLO: A Lightweight and Effective Method for Small Object Detection in UAV Images. *Remote Sens.* **2025**, *17*, 1768. [[CrossRef](#)]
134. Wang, X.; Fang, H.; Li, Q.; Wang, L.; Chang, Y.; Yan, L. Blur-Robust Detection via Feature Restoration: An End-to-End Framework for Prior-Guided Infrared UAV Target Detection. In Proceedings of the the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 20–27 January 2026; Volume 40, pp. 10181–10189. [[CrossRef](#)]
135. Park, S.; Kim, H.; Park, B.; Ro, Y.M. Language-guided Learning for Object Detection Tackling Multiple Variations in Aerial Images. *arXiv* **2025**. [[CrossRef](#)]
136. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 213–226. [[CrossRef](#)]
137. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking. *arXiv* **2018**, arXiv:1804.00518v1.
138. Hsieh, M.R.; Lin, Y.L.; Hsu, W.H. Drone-Based Object Counting by Spatially Regularized Regional Proposal Network. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4165–4173. [[CrossRef](#)]
139. Bozcan, I.; Kayacan, E. AU-AIR: A Multi-Modal Unmanned Aerial Vehicle Dataset for Low Altitude Traffic Surveillance. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 8504–8510. [[CrossRef](#)]
140. Jocher, G. Ultralytics YOLOv5. Version 7.0. 2020. Available online: <https://github.com/ultralytics/yolov5> (accessed on 13 September 2026).

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