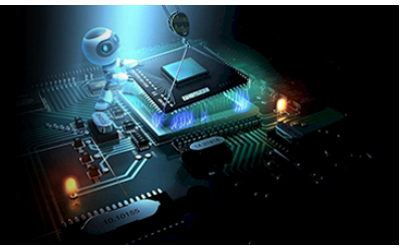
International Journal of Engineering in Computer Science



ISSN Print: 2663-3582
ISSN Online: 2663-3590
IJECS 2026; 8(4): 01-09
www.computersciencejournals.com/ijecs
Received: 14-08-2026
Accepted: 19-09-2026
Published: 16-10-2026

Narender Kumar
Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Engineering and Technology, Hisar, Haryana, India

Abhishek Kajal
Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Engineering and Technology, Hisar, Haryana, India

Sunil Kumar
Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Engineering and Technology, Hisar, Haryana, India

Sunil Kumar
Department of Artificial Intelligence and Data Science, Guru Jambheshwar University of Science and Engineering and Technology, Hisar, Haryana, India

Corresponding Author:
Narender Kumar
Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Engineering and Technology, Hisar, Haryana, India

From CNNs to state space models: A review of small and tiny object detection in aerial imagery

Narender Kumar[1], Abhishek Kajal[2], Sunil Kumar[3], Sunil Kumar[4]

DOI: https://www.doi.org/10.33545/26633582.2026.v8.i4a.316

Abstract
Object detection using aerial images captured from unmanned aerial vehicle (UAV) have a variety of applications such as traffic monitoring, surveillance, search-and-rescue, and remote sensing. Usually targeted scenes are small but are densely distributed, frequently occluded, and embedded in cluttered backgrounds. These adverse conditions make them suffer from severe recall loss, largely because aggressive early-stage downsampling erases the fine-grained features on which these small objects depend. The problem of the sever loss during recall have been addressed by many architectures starting from basic convolutional neural network (CNN), transformer, attention based architectures and nowadays State Space Models (SSMs) like Mamba architecture. This review article examines the deep learning methods based on the mentioned architectures for finding recent evolutions, open challenges and future avenues in the field of small and tiny object detection. It has been advocated that some mechanism like linear time sequence modelling, local convolution and quadratic attenuation changes over the feature preservation mechanism while design techniques for small object, detection head and contextual design usually do not. Multi-scale feature fusion, detection head reallocation and common steps like classification and boundary box regression also remained unaltered during small object detection. A unified taxonomy is presented in the review so that comparative analysis may be performed over the standard datasets viz. VisDrone, UAVDT, DroneVehicle, AI-TOD and DOTA. Discussion for a very recent and prominent research area i.e. SSM-based detectors, is included in this paper as a significant contribution as the area is not yet covered by existing review articles. The review is concluded by outlining open challenges, including extremely small (< 3 pixels) recall, the GPU-parallelism limits of SSMs, edge deployment, and multimodal low-light perception.

Keywords: Small object detection, Tiny object detection, Aerial imagery, UAV, State space models, Mamba, Feature pyramid, Real-time detection

Introduction

In recent past, object detection for aerial images has become a much demanding field in research community thanks to the development in deep learning and unmanned aerial vehicle (UAV) imagery. This area has great demand because of its prevalent uses in industry, automated driving, smart farming, warfare, security, surveillance, intelligent traffic monitoring (Kong *et al.*, 2024) [1], environmental monitoring, emergency management and many more. Reliable localization and classification from an aerial viewpoint are key factors to ensure the usefulness of the object detection especially in aerial imagery captured by unmanned aerial vehicles (UAVs) has consequently received continuous interest in both Computer Vision and Remote Sensing fields.

Still, detection in aerial imagery remains considerably harder as compared to the ground-level, static-camera images for which most general-purpose detectors were employed. The defining characteristic of aerial scenes is the overwhelming prevalence of small and tiny objects. As per COCO convention is widely used, defining objects smaller than 32 × 32 pixels are considered as small (Lin *et al.*, 2014) [2]. Tiny objects are typically less than 16 × 16 pixels and in the extreme case just a few pixels may be considered (Nautiyal *et al.*, 2026) [3].

In benchmark aerial datasets, these targets are so predominant that they make up about 60% of the total dataset on VisDrone, for instance, while large objects make up just about 5% of objects in the dataset. Such a large scale difference is exacerbated by dense distributions, occlusion and truncation. Moreover variable viewpoints, illumination and cluttered backgrounds are easy to mistake for real targets. Small objects have low discriminative texture to begin with, and the low-level texture continues to get lost as it goes through the layers of a deep network: Each layer of downsampling using strides eliminates fine-grained cues critical to small object recognition. Common detectors therefore have a high miss rate and/or false alarm rate on aerial data, thus inspiring a large and continually growing body of literature on specialized architectural remedies.

26633590, 2026, 8(4), Downloaded from https://www.doi.org/10.33545/26633582.2026.v8.i4a.316 by International Journal of Engineering in Computer Science on [16/10/2026]. See the Terms and Conditions (https://www.computersciencejournals.com/ijecs) for rules of use; OA articles are governed by the applicable Creative Commons License

Previous approaches to these difficulties have been along the line of the general development of visual representation learning, which has been used as the organizing principle for this review (Fig. 1). The first generation of solutions were developed with convolutional neural networks (CNNs). Nevertheless, two-stage detectors like Faster *R*-CNN (Ren *et al.*, 2017) [4] and one-stage detectors like YOLO family of detectors (Terven *et al.*, 2023) [5] and RTMDet (Lyu *et al.*, 2022) [6] are still the backbone of aerial detection due to their efficiency and robust local feature extraction abilities, with the addition of feature-pyramid designs like FPN (Lin *et al.*, 2017) [7], PANet (Liu *et al.*, 2018) [8] and BiFPN (Tan *et al.*, 2020) [9]. The local nature of convolution, however, poses a constraint on its ability to account for long-range spatial structure for distinguishing objects from background clutter and separating densely packed objects.

This locality constraint was overcome with a second wave that introduced Transformer and hybrid structures. Global context is captured by self-attention in Vision Transformer (ViT) (Dosovitskiy *et al.*, 2021) [10], (Carion *et al.*, 2020) [11] and Detection Transformer (DETR) family (RT-DETR (Zhao *et al.*, 2024) [12] and DINO (Zhang *et al.*, 2022) [13]), and hybrids like TPH-YOLOv5 (Zhu *et al.*, 2021) [14] or Swin-Transformer-based detectors (Liu *et al.*, 2021) [15] use convolutional locality and attention-based global reasoning. Deformable DETR is special detector which mainly focuses on a few point rather than the entire image, enhances the small object detection performance as a result thereby (Zhu *et al.*, 2021) [16]. VitDet is a specialized ViT that uses plain Vision Transformer as backbone network for detection without the usual feature pyramid network [17]. These approaches all gained accuracy in scenes with clutter but were very expensive. The quadratic complexity of self-attention with respect to the number of tokens becomes intractable for the high resolution feature maps. This become a problem since a resolution feature map (Chen *et al.*, 2024) [18] is often required for the small objects of interest in small-object detection. One of important issue in pure Transformers is that they often lose ground on small objects in cluttered scenes.

Recently, a third paradigm has taken hold and begun to redefine the space: State Space Models (SSMs) and in particular the selective SSM architecture Mamba (Gu *et al.*, 2024) [19] . However, in order to provide a global receptive field with only linear complexity and limited memory overhead, the vision adaptations in the aforementioned literature have been proposed: Vision Mamba (Zhu *et al.*, 2024) [20] and VMamba (Liu *et al.*, 2024) [21]. These two Mamba models (Vision Mamba and VMamba) introduce selective, direction-aware scanning, which works for the 2D feature maps more effectively. HD Mamba-YOLO is a hybrid architecture representing third paradigm introduced recently to utilize SSD-CNN hybridization to exploit global context (Mamba as Backbone) and local details (Conv. Modules) specifically designed for UAV based small object detection (Wei *et al.*, 2026) [22]. This is seen in early detection scenarios like Mamba-YOLO (Wang *et al.*, 2025) [23], (Cao *et al.*, 2025) [24] which demonstrate that a Transformer can achieve global-modeling capabilities compared to SSMs while maintaining real-time processing speeds, thereby highlighting its potential as an attractive alternative for aerial small-object detection, which is the main theme of this review.

What is surprising about these three generations is that the global-context (Du *et al.*, 2023) [25] modelling mechanism is the only real change: local-convolution, quadratic-attention, or linear-complexity state-space scanning, while the small-object-specific machinery is surprisingly stable. Whatever the backbone family may be, the following strategies (i) to (iv) are approximately common in effective aerial detectors: (i) scale reallocation of the detection-head, such as removing the large-object head and adding a high-resolution tiny-object head; (ii) fine-grained feature preservation that alleviates downsampling loss; (iii) multi-scale feature fusion to balance shallow detail and deep semantics; and (iv) attention- or context-guided enhancement to suppress background interference.

Several comprehensive surveys of small and tiny object detection have appeared in 2024–2025 (Nautiyal *et al.*, 2026) [3], (Nikouei *et al.*, 2025) [26], along with dedicated reviews of aerial small-object detection (Hua *et al.*, 2025) [27] and of visual Mamba more broadly (Zhang *et al.*, 2024) [28]. These provide valuable taxonomies, but to our knowledge no existing survey centers the state-space paradigm within small-object detection, nor consolidates the CNN–Transformer–SSM progression under a unified small-object design taxonomy. General vision-Mamba surveys, conversely, treat detection only in passing and do not address the specific difficulties of tiny targets. This review fills that gap. The major contributions of the review paper are:

- The review article presents, to the best of our knowledge, the first review that organizes small and tiny object detection in aerial imagery around the architectural evolution from CNNs through Transformers to State Space Models, emphasizing the emerging and previously unsurveyed class of Mamba-based detectors.
- We propose a unified taxonomy of small-object design strategies and show that these strategies recur across paradigms, clarifying which advances are attributable to global-context modeling and which are shared machinery.

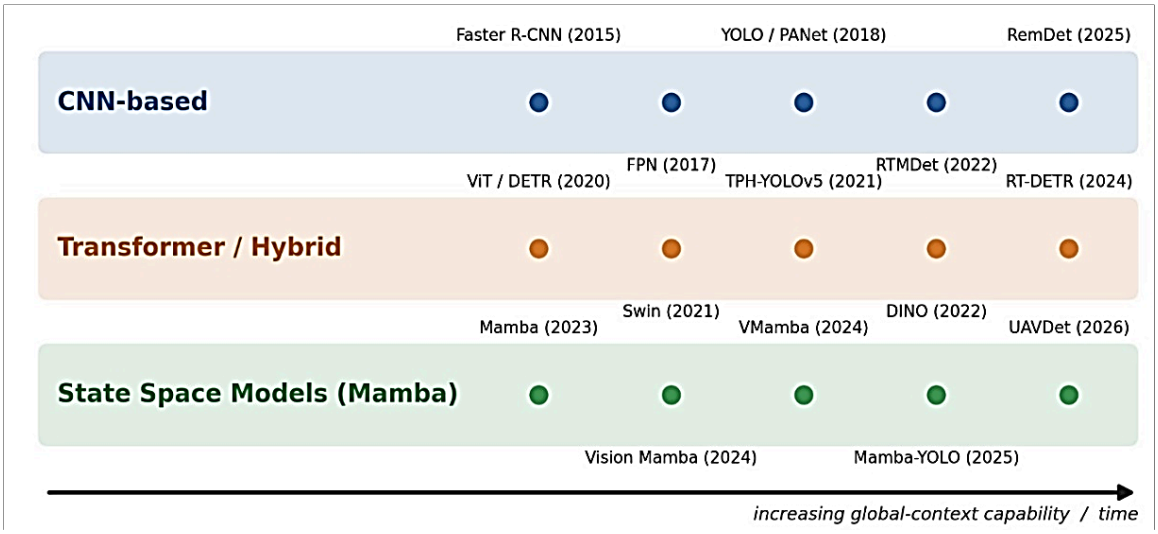


Figure 1. Evolution of aerial object-detection paradigms and representative models.

26633590, 2026, 8(4), Downloaded from https://www.doi.org/10.33545/26633582.2026.v8.i4a.316 by International Journal of Engineering in Computer Science on [16/10/2026]. See the Terms and Conditions (https://www.computersciencejournals.com/ijecs) for rules of use; OA articles are governed by the applicable Creative Commons License

- We review the standard aerial benchmarks and evaluation protocols and provide a normalized performance comparison that highlights and helps reconcile reporting inconsistencies in the current literature.
- We identify open challenges and directions, including recall of extremely small targets (< 3 px), the GPU-parallelism limits of SSMs, edge-oriented compression, and multimodal low-light perception.

Remaining part this paper is organized in the following manner: Section 2 provides a set of preliminaries, Section 3 reviews the evolution of the architectural aspects over the CNN, Transformer, and SSM paradigms, Section 4 discusses design approaches that are commonly used across the whole field of small object detection, Section 5 summarizes the benchmarks experimental studies and Section 6 presents conclusions and future directions in this domain.

Preliminaries

Problem definition and scale conventions

Object detection is a task in which all detected instances of any predefined categories are located and classified, at the same time. Aerial small object detection is a special case in which the number of instances is dominated by instances with low number of pixels. The meaning of scale depends on different communities: Table 1 outlines some of the more common definitions. The MS-COCO convention is the de facto benchmark for natural images, and the aerial benchmarks, like AI-TOD and DOTA, have stricter thresholds due to the systematic reduction in their target pixels. The distribution of object sizes is shown in Fig. 2 for the VisDrone2019 test set, where it is obvious that it is mainly composed of small targets, which calls for detector designs that focus on low-level, high resolution features.

Table 1. Common scale conventions for small and tiny objects.

Convention	Small / tiny threshold	Typical use
MS-COCO	small < 32^2; medium 32^2–96^2; large > 96^2 px	Natural-image detection
Tiny (extended)	tiny < 16 × 16 px	UAV ablation metrics (AP_T)
AI-TOD	≈ 72% of objects in 8–16 px; mean ≈ 12.8 px	Aerial tiny-object benchmark
DOTA	small ≈ 10–50 px	Remote-sensing detection
Extremely small	< 3 px area	Sub-pixel UAV targets

Evaluation metrics

Detection quality is measured through precision P and recall R, derived from true positives (TP), false positives (FP), and false negatives (FN):

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN} \tag{1}$$

Average Precision (AP) integrates precision over recall, and mean AP (mAP) averages over K categories:

$$AP = \int_0^1 P(R)\, dR, \quad mAP = \frac{1}{K} \sum_{i=1}^{K} AP_i \tag{2}$$

A prediction is matched to a ground-truth box when their Intersection-over-Union (IoU) exceeds a threshold:

$$IoU = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \tag{3}$$

Where B_p and B_g are predicted boundary box and actual (ground truth) boundary box. Following COCO, AP is averaged over ten IoU thresholds from 0.50 to 0.95. Scale-aware variants AP_S, AP_M, and AP_L report accuracy on small, medium, and large objects, and many UAV studies add a tiny-object metric AP_T (< 16 × 16 px). Efficiency is reported through the parameter count, floating-point operations (GFLOPs), and frames per second (FPS). Because AP_S and AP_T isolate the small-object regime that review study concerns, we treat them as the primary comparison metrics in Section 5.

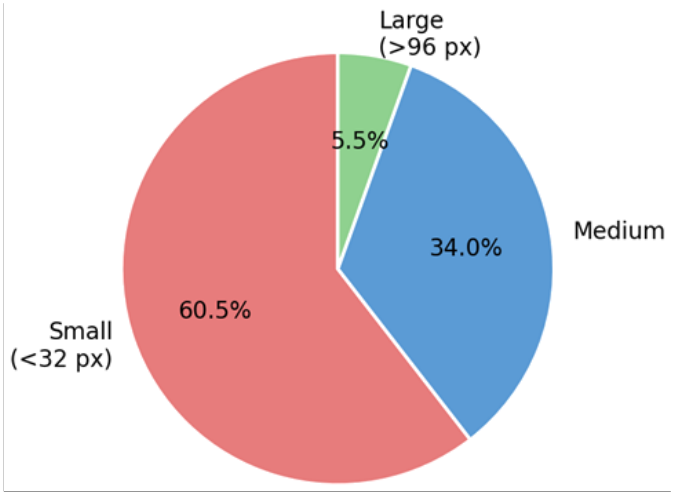


Figure 2. Object-size distribution on VisDrone2019 (small targets dominate) (Yang *et al.*, 2026) [29].

Benchmark datasets

Table 2 summarizes the datasets most frequently used to evaluate aerial small-object detectors. VisDrone2019 is the dominant benchmark and appears in essentially all recent studies; UAVDT and DroneVehicle test cross-scenario and multimodal generalization, while AI-TOD and DOTA push toward the tiny and remote-sensing regimes.

Architectural evolution

This section maps the three paradigms on the x-axis of Fig. 3. The key distinction is that convolution does aggregation in a local kernel, self-attention does a quadratic cost operation of pairing up all tokens, and state-space scanning does a linear cost operation on the feature map in a direction.

26633590, 2026, 8(4), Downloaded from https://www.doi.org/10.33545/26633582.2026.v8.i4a.316 by International Journal of Engineering in Computer Science on [16/10/2026]. See the Terms and Conditions (https://www.computersciencejournals.com/ijecs) for rules of use; OA articles are governed by the applicable Creative Commons License

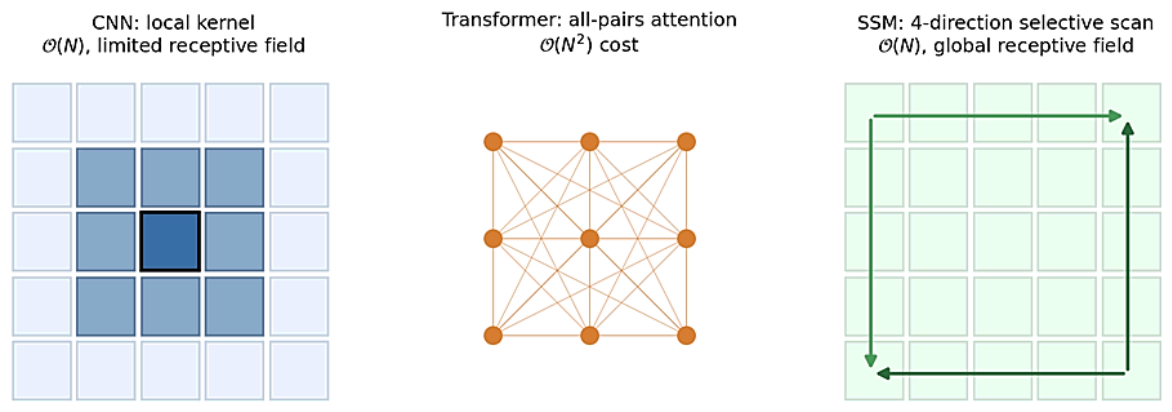


Figure 3. Global-context modeling in the three paradigms.

CNN-based detectors

CNN detectors are still the main elements of aerial detection. There are two kinds of network architectures: two-stage frameworks (Faster *R*-CNN (Ren *et al.*, 2017) [4] and Cascade *R*-CNN (Cai *et al.*, 2018) [30]) for accurate localization with region proposals and one-stage detectors (YOLO family (Terven *et al.*, 2023) [5] and anchor-free RTMDet (Lyu *et al.*, 2022) [6]) for fast predictions that are dominant in real-time deployments. Feature Pyramids like FPN (Lin *et al.*, 2017) [7], PANet/PAFPN (Liu *et al.*, 2018) [8] and BiFPN (Tan *et al.*, 2020) [9] combine shallow spatial information with deep semantic information for recovery of multi-scale information lost when downsampling. The non-uniform CNN design is enhanced by these UAV-specific CNN designs: RSOD and HR-FPN utilize shallow feature reuse, UAV-YOLO boosts the strength of early-layer convolutions, and the efficient design RemDet (Li *et al.*, 2025) [31] employs multiplicative operations instead of feed-forward blocks. An important commonality is that they have a fixed, local receptive field that hinders global reasoning and makes them susceptible to background noise and clutter and dense object interference.

Table 2. Standard benchmarks for aerial small-object detection.

Dataset	Images	Categories	Notes
VisDrone2019	≈ 8,629	10	UAV; ~60% small objects; primary benchmark
UAVDT	≈ 38k	3	Cross-scene (urban/highway/rural), varied altitude
DroneVehicle	≈ 28,439	5	RGB–Infrared paired; multimodal, day/ night
AI-TOD	≈ 28,036	8	Tiny objects; mean size ≈ 12.8 px
DOTA	≈ 2,806	15/18	Remote sensing; oriented boxes

Transformer and hybrid detectors

Self-attention is the mechanism used by transformers to deal with locality. Images were set by ViT (Dosovitskiy *et al.*, 2021) [10] and set prediction was redefined by DINO (Zhang *et al.*, 2022) [13] and the real-time RT-DETR (Zhao *et al.*, 2024) [12]. Due to the high cost of pure attention and its poor ability to capture small objects, most of the aerial works have taken hybrids, such as TPH-YOLOv5 (Zhu *et al.*, 2021) [14] which introduces a transformer prediction head to YOLOv5, introducing hierarchical attention in CNN detectors, such as Swin-based and CSWin-based models (Liu *et al.*, 2021) [15], and some designs that fuse the convolution and windowed attention together for preserving the local detail. The quadratic scaling of attention with resolution of feature map as well as structural redundancy makes the deployment of these hybrids difficult on resource-constrained UAV hardware.

State space models (Mamba)

Aerial imagery is an obvious application of State Space Models, because it requires global modeling at linear cost. A linear SSM is a continuous operation that transforms input sequence $x^{(t)}$ to output $y^{(t)}$ via a latent state $h^{(t)}$ as given in the Eq. 4:

$$h'(t) = Ah(t) + Bx(t), \quad y(t) = Ch(t) \tag{4}$$

A, B and C are the state, input and output matrices respectively whereas $h'(t)$ is known as state derivative that represents the change in time t. For discrete inputs the system is discretized, commonly by zero-order hold with step size Δ, giving discrete matrices $\bar{A}$ and $\bar{B}$:

$$\bar{A} = exp(\Delta A), \quad \bar{B} = (\Delta A)^{-1}(exp(\Delta A) - I)\,\Delta B \tag{5}$$

Which yields an efficient linear recurrence (equivalently a global convolution):

$$h_k = \bar{A}\,h_{k-1} + \bar{B}\,x_k, \quad y_k = C\,h_k \tag{6}$$

In Mamba (Zhu *et al.*, 2021) [16], the SSM parameters are content adaptive (selective scan), which retain linear complexity but enable propagation based on content. This is the critical distinction over attention given by Eq. 7:

$$\Omega_{attn} = \mathcal{O}(N^2 d) \quad vs. \quad \Omega_{SSM} = \mathcal{O}(Nd) \tag{7}$$

In the Vision Mamba (Zhu *et al.*, 2024) [20] method, the bidirectional scanning is applied to 2-D images, while VMamba (Liu *et al.*, 2024) [21] is proposed by us to sweep the feature map in four directions

26633590, 2026, 8(4), Downloaded from https://www.doi.org/10.33545/26633582.2026.v8.i4a.316 by International Journal of Engineering in Computer Science on [16/10/2026]. See the Terms and Conditions (https://www.computersciencejournals.com/ijecs) for rules of use; OA articles are governed by the applicable Creative Commons License

(top-down, bottom-up, left-right and right-left). Detection-oriented instances follow quickly: Mamba-YOLO (Wang *et al.*, 2025) [23] provides a simple SSM baseline, and hybrid CNN–Mamba detectors integrate a state-space branch alongside convolutional branches to combine local detail with global reasoning. A representative recent example couples a Cross-Stage-Partial structure with a visual SSM block, using the convolutional path for stable gradient flow and fine detail and the SSM path for linear-complexity global context, reporting notable small-object gains over CNN and Mamba-only baselines while cutting parameters. MART-DETR (Cai *et al.*, 2027) [32] integrates Mamba and DETR, with a channel-split gating Mamba module (CGMamba) to efficiently model features for small object detection. Because this line of work is only two to three years old, it is the least surveyed and, we argue, the most consequential frontier for aerial small-object detection.

Cross-cutting design strategies

Independently of the backbone paradigm, effective aerial detectors converge on the four strategy families as depicted in Fig. 4. This convergence is the empirical basis for our research that the small-object mechanism is largely paradigm-invariant.

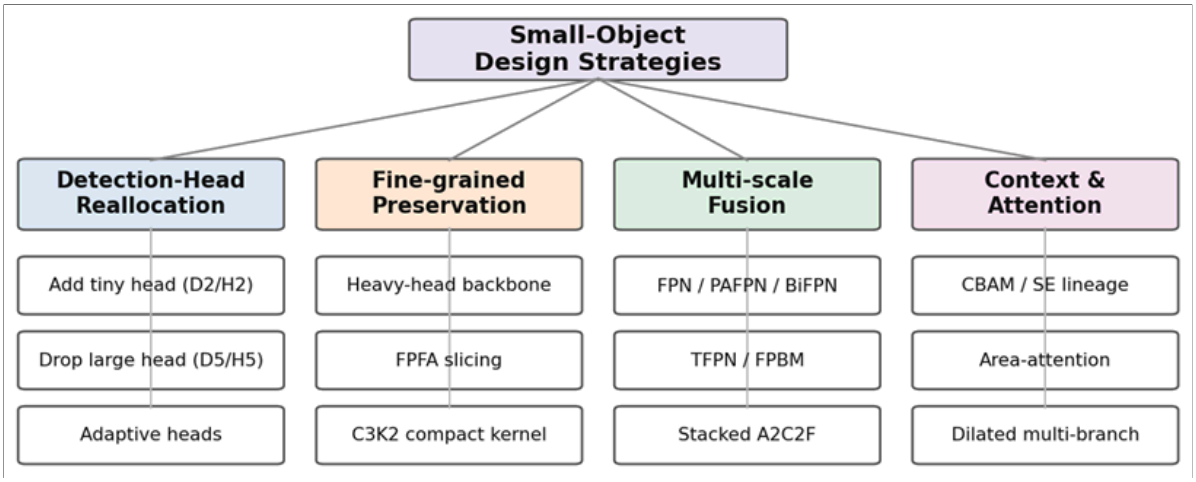


Figure 4. Taxonomy of recurring small-object design strategies across paradigms.

Detection-head reallocation

The single most consistent design choice across recent detectors is to reallocate detection heads toward higher resolution. Because aerial datasets contain very few large objects, the conventional large-object head is nearly pure overhead, whereas a high-resolution head is essential for tiny targets. Multiple recent CNN and YOLO (Sun *et al.*, 2022) [33], (Khanam *et al.*, 2024) [34] detectors independently remove the large head and add a high-resolution head, reporting simultaneous gains in small-object AP and reductions in parameter count. The reductions confirm that the high-resolution head drives most of the tiny-object improvement while the deepest head contributes little. It is an important finding that recurs across otherwise unrelated architectures.

Fine-grained feature preservation

Since downsampling is the primary cause of small-object information loss, a second family of methods preserves fine detail during feature extraction. Backbone re-balancing allocates more capacity to early, high-resolution stages (a 'heavy-head, light-bottom' configuration). Loss-less or near-loss-less downsampling modules replace strided convolution with slicing plus average pooling to retain all input elements, while compact-kernel blocks reduce parameters without discarding spatial resolution. These modules are backbone-agnostic and appear in CNN, hybrid, and SSM detectors alike. A hybrid architecture has been proposed in (Yang *et al.*, 2026) [29] to combat the problems small object in complex background.

Multi-scale feature fusion

Feature fusion has evolved from FPN (Lin *et al.*, 2017) [7] through PANet/PAFPN (Liu *et al.*, 2018) [8] to weighted BiFPN (Tan *et al.*, 2020) [9], and more recently to task-specific necks that add higher-resolution pathways, prune redundant top-down propagation, and insert global-context modules (Transformer- or SSM-based) at the mid-level fusion stages to suppress background noise. An approach utilizing deformable convolution and optimized feature fusion adopted to maintain the small object details in real time is proposed i.e. DCEF 2-YOLO (MDPI, 2026) [35]. The unifying goal of multi-scale feature fusion is to enhance the shallow, high-resolution features that carry small-object detail while limiting the propagation of background clutter that inflates false positives. A prominent study by L *et al* 2025 proposed multimodal fusion (RGB and infrared) in hybrid CNN-Mamba neck architecture for efficient small object detection (Li *et al.*, 2025) [36].

Context and attention enhancement

A tiny object has almost negligible pixels, while small object has very few samples. Therefore context enhancement mechanism is highly demanded so as to sharpen the identity of the weak small object. CIE-YOLO introduced joint attention to enhance the context and in parallel dimmed the background while up sampling (Ma *et al.*, 2025) [37]. Attention and context modules are high-pass filters for weak Small Objects. Most commonly used for channel/spatial attention is the Convolution Block Attention Module (CBAM) (Wang *et al.*, 2024) [38]/Squeeze-and-Excitation (SE) family of attention modules, which emphasize specific smaller regions is an area-based attention, and dilated multi-branch convolutions that increase the receptive field without additional stride. Class imbalance (background is usually overwhelmingly large compared to foreground which is small) is a frequent problem in aerial data, and is often solved by focal loss (Eq. 8), where easy negative examples are down-weighted. It is worth mentioning here that focal loss is the updated cross-entropy specifically formulated for difficult and misclassifies classes.

$$L_{local} = -\alpha_t(1 - p_t)^{\gamma} log(p_t) \quad (8)$$

Where α_t and γ are the balancing-weight and the focusing parameters respectively; p_t is the probability of the true class. It is to note that the modulating factor $(1 - p_t)^{\gamma}$ is an indicator that shrinks when p_t is high.

26633590, 2026, 8(4), Downloaded from https://www.doi.org/10.33545/26633582.2026.v8.i4a.316 by International Journal of Engineering in Computer Science on [16/10/2026]. See the Terms and Conditions (https://www.computersciencejournals.com/ijecs) for rules of use; OA articles are governed by the applicable Creative Commons License

Label assignment and tiny-object losses

There is also a more nuanced problem with IoU matching: When the boxes are small, one pixel's movement can cause a significant change in IoU. This is mitigated by a similarity measures (Eq. 9) based on distribution: The Normalized Gaussian Wasserstein Distance (NWD) models each box as a two dimensional Gaussian and compares distributions, giving a smooth, scale-invariant similarity for label assignment and loss.

$$NWD\left(N_p, N_g\right) = exp\left(- \frac{\sqrt{W_2^2\left(N_p, N_g\right)}}{C} \right) \quad (9)$$

Where N_p, *and* N_g are the predicted box and ground truth box respectively especially utilized for tiny object localization; W is the Wasserstein Distance and C is the normalization constant. It is worth remembering that the squared quantity $\sqrt{W_2^2\left(N_p, N_g\right)}$ is the second order Wasserstein Distance that calculates the cost while moving from one distribution to other. Similar approaches to the allocation of positive samples, such as receptive-field-based assignment (RFLA), are also effective for tiny targets. It is a relatively unexplored region in previous review articles, but is also central to the performance of tiny objects, and we emphasize this as a possible pathway.

Benchmark synthesis and discussion

The normalized comparison is given on the test set of VisDrone2019 in a recent study by Yang *et al.* (2026) [39] for all the representative detectors in the three paradigms. They report AP, small-object AP_S results, parameters, GFLOPs and FPS at a common 640 × 640 input for results to be comparable The accuracy–efficiency trade-off is shown in Fig. 5 which shows clearly shows UAVDet, a hybrid CNN-Mamba architecture is the most efficient for small object detection on *VisDrone2019* (Yang *et al.*, 2026) [39].

Method characterization

Over the time many detectors were developed towards the detection of the small objects listed in Table 3 which a very concise view of all the studies especially conducted in the last decade. The table presented the paradigm followed, which is of course the pivotal point the current paper article, along with the method's name, year and primary strategies utilized for detection. The next column of the table represents the idea and /or the methodologies followed to achieve the target of small object detection.

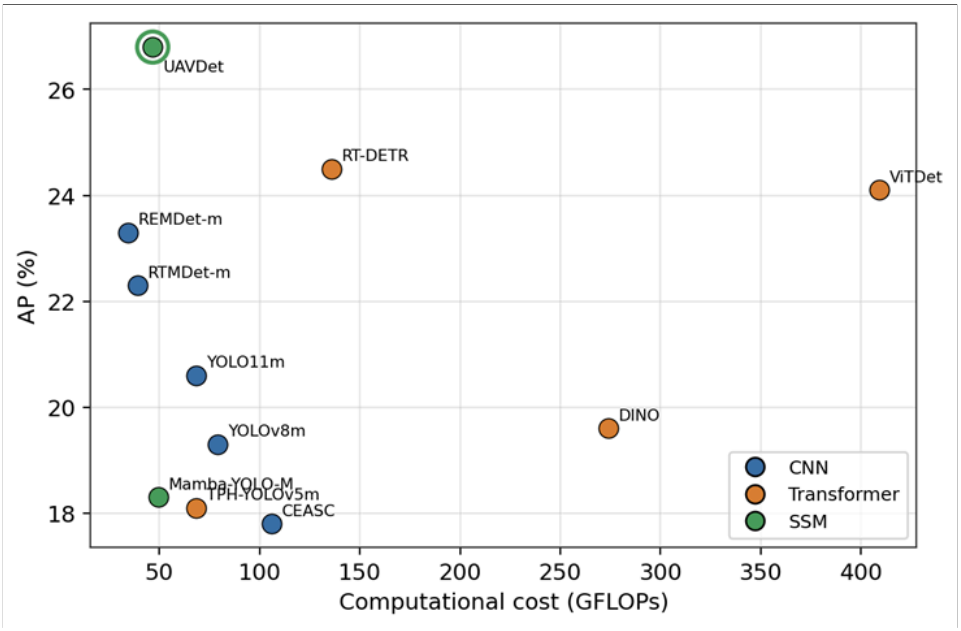


Figure 5. Accuracy–efficiency trade-off on VisDrone2019 test set (Yang *et al.*, 2026) [29].

Many complexity are involved while detecting small object especially in real time scenario like edge deployment. An important study SOD-YOLO introduced such method which cuts the computations to make it light-weight and edge deployable in UAVs for efficient small object detection (Xiao *et al.*, 2024) [40], (Liu *et al.*, 2025) [41].

Table 3. Method characterization and description of the SO detectors in last decade.

Method	Year	Paradigm	Primary small-object strategy (→ §4)	Key module / idea	Ref.
Faster *R*-CNN	2015	CNN (two-stage)	Baseline	RPN + ROI head	[4]
FPN	2017	CNN	Multi-scale fusion	Top-down pyramid	[7]
PANet / PAFPN	2018	CNN	Multi-scale fusion	Bottom-up path augmentation	[8]
Cascade *R*-CNN	2018	CNN (two-stage)	Label-quality refinement	Multi-stage IoU heads	[30]
RSOD	2022	CNN	Fine-grained preservation + fusion	Shallow-feature adaptive FPN	[33]
RTMDet	2022	CNN	Efficient one-stage baseline	CSPNeXt + PAFPN	[6]
CEASC	2023	CNN	Efficiency (edge)	Sparse conv + adaptive masking	[25]

(continued on next page)

26633590, 2026, 8(4), Downloaded from https://www.doi.org/10.33545/26633582.2026.v8.i4a.316 by International Journal of Engineering in Computer Science on [16/10/2026]. See the Terms and Conditions (https://www.computersciencejournals.com/ijecs) for rules of use; OA articles are governed by the applicable Creative Commons License

HR-FPN	2023	CNN	Fine-grained preservation	High-resolution pyramid	[18]
DCEF2-YOLO	2024	CNN	Multi-scale fusion	Deformable conv + efficient fusion	[35]]
CRL-YOLOv5	2024	CNN	Fine-grained / receptive field	Extra shallow detection layer	[38]
SOD-YOLO	2024	CNN	Efficiency (edge)	Deeper-narrower progressive concat	[40]
CIE-YOLO	2025	CNN	Context enhancement	Contextual cue integration	[37]
RemDet	2025	CNN	Efficiency (edge)	Multiplicative feature blocks	[31]
YOLO11	2024	CNN	Efficient one-stage	Anchor-free C3k2 / C2PSA	[34]
DETR	2020	Transformer	Global modeling	Set prediction, no NMS	[11]
Deformable DETR	2021	Transformer	Global + efficiency	Deformable attention	[16]
Swin Transformer	2021	Transformer	Global backbone	Shifted-window attention	[15]
TPH-YOLOv5	2021	Transformer hybrid	Context / head	Transformer prediction head	[14]
DINO	2022	Transformer	Label assignment	Denoising anchor queries	[13]
ViTDet	2022	Transformer	Global backbone	Plain-ViT detection backbone	[17]
RT-DETR	2024	Transformer	Real-time global modeling	Efficient hybrid encoder	[12]
ESO-DETR	2025	Transformer	Context enhancement	Multi-scale self-attention (AIFI)	[41]
Mamba	2023	SSM	Foundation	Selective state-space scan	[19]
Vision Mamba (Vim)	2024	SSM	Global backbone (linear)	Bidirectional scan	[20]
VMamba	2024	SSM	Global backbone (linear)	4-direction cross-scan	[21]
Mamba-YOLO	2025	SSM	Global modeling (linear)	ODSS detection block	[23]
HDMamba-YOLO	2025 †	SSM-CNN hybrid	Fine-grained + global	Stage-wise heterogeneous SSM/CNN	[22]
MambaRefine-YOLO	2025 †	SSM (multimodal)	Fusion + refinement	HFAN cross-modal refine	[24]
UAVD-Mamba	2025 †	SSM (multimodal)	Fusion	Deformable token fusion	[36]
MART-DETR	2025 †	SSM + Transformer	Global modeling	CGMamba channel-split gating	[32]
UAVDet	2026	SSM-CNN hybrid	Head realloc. + fine-grained + global	CSPMB + TFPN	[39]

Three observations followed, discussed are presented here as; First, CNN–Mamba has the least number of parameters as compared to Transformer detectors for reaching the best average-precision (AP) and average_precision_for small object (AP_S), which proves that global modeling with linear complexity is suitable for aerial small object detection. Second, there is a mismatch between quadratic attention and high resolution UAV images, as pure Transformer detectors (ViTDet, DINO) achieve the highest FLOPs but are not most accurate on small objects. Third and crucially from a reproducibility point of view, the reviewed papers have a wide range of 'state of the art' baselines, and sometimes internally conflicting metrics; a rigorous paper review should therefore standardize input resolution, test split and metric definitions instead of regurgitating headline numbers. This discrepancy in reporting is noted and it is suggested that future benchmarks include per scale AP in addition to code and configuration files.

Conclusion and future directions

The current study has covered an exhaustive review on detection of small and tiny objects (STO) in aerial images from the perspective of architectural evolution, starting from Convolutional networks, moving to Transformer and hybrid networks, and ending in the novel class of State Space Models. In all of these paradigms, the core mechanism that is changing is the global-context mechanism, which can be either local convolution or quadratic attention or linear-complexity scanning, while a consistent set of small-object mechanisms (head reallocation, fine-grained preservation, multi-scale fusion, and context enhancement) reoccurs. State-space detectors, like CNN–Mamba hybrids, are a promising, yet underexplored approach in the field, as they form a hybrid model with linear cost global modeling and impressive small-object accuracy. We hope the unified taxonomy, normalized benchmark, and reporting cautions presented here provide a useful foundation for future work on efficient, reliable aerial perception.

Based on the review carried out, here we highlight the critical challenges for future research in this area:

- **Extremely small recall:** Major challenges arise for the extremely small (< 3 px) recall, even high-resolution heads networks miss sub-pixel targets. Super-resolution obtained by upsampling of tiny object's pixels assisted detection, and frequency-domain features are also promising but still immature and leave wide scope of improvement.
- **GPU-parallelism of SSMs**: Although Mamba is linear in theory, its sequential state update underutilizes GPU parallelism relative to attention. Most of the time GPU sits idle as SSMs scan stays bounded with memory and tensor cores remains underutilized. Small object detection worsen it by expanding feature maps and thereby increasing memory scans in different directions and increasing sequence length as well. So hardware-aware scan kernels and hybrid parallel formulations are open problems.
- **Edge deployment:** The detection of the small object require sound computations as results of super-resolution feature maps, large backbone networks, feature fusion needed at multiple scales, and multidirectional scans. Real UAV hardware demands quantization, pruning, and distillation that fits into limited budget available. Therefore systematic study of SSM detectors under INT8 and on embedded accelerators is largely absent.
- **Multimodal and low-light perception**: Low illumination makes the small object detection more challenging in adverse weather conditions. Cross modality SSMs and RGB-feature fusion along with thermal image inclusion provide a solution to some extent but stills lacks in identification of small and tiny objects because of the fusion of different sources of diverse nature. The RGB–Infrared fusion (e.g., DroneVehicle) and cross-modality SSMs address night and adverse conditions but lack consolidated benchmarks and fusion theory.

26633590, 2026, 8(4), Downloaded from https://www.doi.org/10.33545/26633582.2026.v8.i4a.316 by International Journal of Engineering in Computer Science on [16/10/2026]. See the Terms and Conditions (https://www.computersciencejournals.com/ijecs) for rules of use; OA articles are governed by the applicable Creative Commons License

- **Label assignment for tiny objects**: During the training phase it essential to know that whether current pixel is background or the object. Even one pixel shift disturb the stability and that is most likely in case of small/tiny object detection. Distribution-based assignment (NWD, RFLA) remains under-integrated with modern one-stage and SSM detectors.
- **Standardized reporting**: The field needs shared evaluation protocols and reproducible configurations to make cross-paper comparison meaningful

References

1. Kong X, Ni C, Duan G, Shen G, Yang Y, Das SK. Energy Consumption Optimization of UAV-Assisted Traffic Monitoring Scheme With Tiny Reinforcement Learning. IEEE Internet of Things Journal. 2024;11(12):21135–21145.
2. Lin TY, *et al.* Microsoft COCO: Common Objects in Context. Computer Vision – ECCV 2014. Cham: Springer International Publishing; 2014:740–755.
3. Nautiyal R, Deshmukh M. Tiny object detection: An in-depth survey of techniques, challenges, and future directions. Digital Signal Processing. 2026;174:105995.
4. Ren S, He K, Girshick R, Sun J. Faster *R*-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;39(6):1137–1149.
5. Terven J, Córdova-Esparza DM, Romero-González JA. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. MAKE. 2023;5(4):1680–1716.
6. Lyu C, *et al.* RTMDet: An Empirical Study of Designing Real-Time Object Detectors. arXiv. 2022:arXiv:2212.07784.
7. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature Pyramid Networks for Object Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:936–944.
8. Liu S, Qi L, Qin H, Shi J, Jia J. Path Aggregation Network for Instance Segmentation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018:8759–8768.
9. Tan M, Pang R, Le QV. EfficientDet: Scalable and Efficient Object Detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020:10778–10787.
10. Dosovitskiy A, *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv. 2021:arXiv:2010.11929.
11. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End Object Detection with Transformers. Computer Vision – ECCV 2020. Cham: Springer International Publishing; 2020:213–229.
12. Zhao Y, *et al.* DETRs Beat YOLOs on Real-time Object Detection. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024:16965–16974.
13. Zhang H, *et al.* DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. arXiv. 2022:arXiv:2203.03605.
14. Zhu X, Lyu S, Wang X, Zhao Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). 2021:2778–2788.
15. Liu Z, *et al.* Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021:9992–10002.
16. 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:2010.04159.
17. Exploring Plain Vision Transformer Backbones for Object Detection | Springer Nature Link. 2026.
18. Chen Z, Ji H, Zhang Y, Zhu Z, Li Y. High-Resolution Feature Pyramid Network for Small Object Detection on Drone View. IEEE Transactions on Circuits and Systems for Video Technology. 2024;34(1):475-489.
19. Gu A, Dao T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv. 2024;arXiv:2312.00752.
20. Zhu L, Liao B, Zhang Q, Wang X, Liu W, Wang X. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. arXiv. 2024;arXiv:2401.09417.
21. Liu Y *et al.* VMamba: Visual State Space Model. Advances in Neural Information Processing Systems. 2024:103031-103063.
22. Wei L, Fan J, Mai Y, Qi Y. HDMamba-YOLO: Efficient State-Space Perception and Local Spatial Reconstruction for UAV Small Object. arXiv. 2026;arXiv:2609.23061.
23. Wang Z, Li C, Xu H, Zhu X, Li H. Mamba YOLO: A Simple Baseline for Object Detection with State Space Model. Proceedings of the AAAI Conference on Artificial Intelligence. 2025;39(8):8205-8213.
24. Cao S, Chen M, Song Y, Chen Z, Zhang X. MambaRefine-YOLO: A Dual-Modality Small Object Detector for UAV Imagery. arXiv. 2025;arXiv:2511.19134.
25. Du B, Huang Y, Chen J, Huang D. Adaptive Sparse Convolutional Networks with Global Context Enhancement for Faster Object Detection on Drone Images. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023:13435-13444.
26. Nikouei M *et al.* Small object detection: A comprehensive survey on challenges, techniques and real-world applications. Intelligent Systems with Applications. 2025;27:200561.
27. Hua W, Chen Q. A survey of small object detection based on deep learning in aerial images. Artif Intell Rev. 2025;58(6):162.
28. Zhang H *et al.* A Survey on Visual Mamba. Applied Sciences. 2024;14(13):5683.
29. Yang Y, Guo F, Niu P. UAVDet: A CNN–Mamba hybrid network for efficient small object detection in UAV imagery. Computer Vision and Image Understanding. 2026;264:104637.
30. Cai Z, Vasconcelos N. Cascade *R*-CNN: Delving Into High Quality Object Detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018:6154-6162.
31. Li C, Zhao R, Wang Z, Xu H, Zhu X. RemDet: Rethinking Efficient Model Design for UAV Object Detection. Proceedings of the AAAI Conference on Artificial Intelligence. 2025;39(5):4643-4651.
32. Cai C, Liu H, Shen Y. MART-DETR: Mamba-Enhanced Real-Time DETR for UAV Small Object Detection. Singapore: Springer Nature. 2027:374-385.
33. Sun W, Dai L, Zhang X, Chang P, He X. RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring. Appl Intell. 2022;52(8):8448-8463.
34. Khanam R, Hussain M. YOLOv11: An Overview of the Key Architectural Enhancements. arXiv. 2024:arXiv:2410.17725.
35. DCEF2-YOLO: Aerial Detection YOLO with Deformable Convolution–Efficient Feature Fusion for Small Target Detection. MDPI. 2026.
36. Li W, Tang J, Li Y, Xia B, Tan L, Qin H. UAVD-Mamba: Deformable Token Fusion Vision Mamba for Multimodal UAV Detection. 2025 IEEE Intelligent Vehicles Symposium (IV). 2025:561-568.
37. Ma Z, Zhou L, Wu D, Zhang X. A small object detection method with context information for high altitude images. Pattern Recognition Letters. 2025;188:22-28.
38. Wang Z *et al.* Improved Small Object Detection Algorithm CRL-YOLOv5. Sensors. 2024;24(19):6437.

26633590, 2026, 8(4), Downloaded from https://www.doi.org/10.33545/26633582.2026.v8.i4a.316 by International Journal of Engineering in Computer Science on [16/10/2026]. See the Terms and Conditions (https://www.computersciencejournals.com/ijecs) for rules of use; OA articles are governed by the applicable Creative Commons License

39. Yang Y, Guo F, Niu P. UAVDet: A CNN–Mamba hybrid network for efficient small object detection in UAV imagery. Computer Vision and Image Understanding. 2026;264:104637.
40. Xiao Y, Di N. SOD-YOLO: A lightweight small object detection framework. Sci Rep. 2024;14(1):25624.
41. 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(2):143.

How to cite this article: Kumar N, Kajal A, Kumar S, Kumar S. From CNNs to state space models: A review of small and tiny object detection in aerial imagery. *International Journal of Engineering in Computer Science*. 2026;8(4):01-09. doi: https://www.doi.org/10.33545/26633582.2026.v8.i4a.316

26633590, 2026, 8(4), Downloaded from https://www.doi.org/10.33545/26633582.2026.v8.i4a.316 by International Journal of Engineering in Computer Science on [16/10/2026]. See the Terms and Conditions (https://www.computersciencejournals.com/ijecs) for rules of use; OA articles are governed by the applicable Creative Commons License