INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Computer Applications in Technology

ISSN online: 1741-5047 - ISSN print: 0952-8091
https://www.inderscience.com/ijcat

DAFPN: a lightweight multi-objective framework for small object detection in remote sensing images

Wanqi Ren

DOI: 10.1504/IJCAT.2026.10075464

Article History:

Received:	11 June 2025
Last revised:	16 September 2025
Accepted:	20 October 2025
Published online:	17 February 2026

DAFPN: a lightweight multi-objective framework for small object detection in remote sensing images

Wanqi Ren

College of Economics,
Shandong University of Finance and Economics,
Jinan, Shandong, China
Email: renwanqi11@outlook.com

Abstract: Remote sensing object detection faces persistent challenges in accurately identifying small-scale targets embedded in high-resolution, cluttered scenes. Conventional detectors often suffer from feature dilution, scale variance and high computational cost, limiting their applicability in real-time or edge-based remote sensing scenarios. To address these issues, we propose DAFPN, a lightweight Dynamic Attention-guided Feature Pyramid Network that integrates asymmetric multi-scale fusion and dual-branch attention, consisting of spatial and channel-wise attentions, into a unified architecture optimised via multi-objective constrained learning aimed at simultaneously maximising detection accuracy, attention alignment and architectural compactness. On DOTA-v2.0, our method improves mAP@0.75 by 4.5% and mAP@0.5 by 3.8% over YOLOv8, while achieving similar gains on FAIR1M, DIOR and RSSOD. The results confirm DAFPN's robustness under variable input resolutions and dense object distributions, highlighting its practical value for deployment in real-time and resource-constrained remote sensing applications.

Keywords: remote sensing; small object detection; multi-objective optimisation; feature pyramid networks; dynamic attention; lightweight detection; aerial imagery; real-time inference.

Reference to this paper should be made as follows: Ren, W. (2026) 'DAFPN: a lightweight multi-objective framework for small object detection in remote sensing images', *Int. J. Computer Applications in Technology*, Vol. 78, No. 2, pp.144–154.

Biographical notes: Wanqi Ren graduated from Shandong University of Finance and Economics. She focuses on deep learning-based visual perception and intelligent analysis of remote sensing imagery, with particular emphasis on small object detection under high-resolution and resource-constrained settings.

1 Introduction

Remote sensing object detection has emerged as a core task in Earth observation, supporting applications in land monitoring, maritime surveillance, urban planning and disaster response. With the increasing availability of high-resolution optical satellites and UAV platforms, vast volumes of imagery are collected daily, characterised by high spatial detail and rich semantic complexity. The foremost challenge lies in detecting small-scale objects in real time under stringent resource constraints. In practice, airborne or edge-deployed platforms often face tight hardware budgets – e.g., airborne SAR systems require sub-100 ms processing on embedded GPUs with less than 8 GB memory. Small objects in aerial images typically suffer from weak semantic representation, low contrast and high inter-class similarity, making them particularly vulnerable to misclassification. These primary challenges underscore the need for models that are lightweight, fast and capable of maintaining accuracy in resource-limited deployments. Beyond these core difficulties, additional challenges include maintaining efficiency under varying resolutions and integrating multiple objectives – such as accuracy, latency and compactness – into a unified training process. Addressing these secondary issues is essential to ensure robustness and scalability in real-world deployments.

Recent advances in deep learning-based remote sensing detection have produced powerful architectures, including two-stage frameworks like Faster R-CNN (Ren et al., 2016) and one-stage variants such YOLOv5/YOLOv8 (Ultralytics, 2022; Jocher et al., 2023), RetinaNet (Lin et al., 2018) and domain-specific adaptations like MTGS-YOLO (Jin et al., 2025). While these models achieve satisfactory results under general conditions, several limitations remain unresolved. First, many suffer from scale inconsistency and feature dilution when processing dense or small targets, as shown in evaluations on DOTA-v2.0 (Ding et al., 2021) and FAIR1M (Sun et al., 2022). Second, large backbones or multi-head modules often violate real-time constraints on edge or airborne devices. Third, attention-based enhancements, though effective, are rarely optimised jointly with compactness and latency objectives. Zhu et al. (2024), Gao et al. (2024), Wang et al. (2024b), Pei (2025) and Pei et al. (2025)

attempted to integrate multi-scale fusion, attention modules or few-shot adaptation, but often lack systematic multi-objective control or fail to generalise across data sets with varying resolutions. Here, we define an *asymmetric feature pyramid* as a feature hierarchy where lateral and top-down connections are selectively pruned or reinforced to balance semantic richness and spatial precision, in contrast to symmetric FPN designs that treat all scales equally. These gaps highlight the need for a unified framework that balances accuracy, efficiency and robustness without overfitting to specific scenarios.

To address the aforementioned challenges, we introduce *DAFPN* – a Dynamic Attention-guided Feature Pyramid Network that is explicitly designed for real-time small object detection in high-resolution remote sensing imagery. Unlike YOLOv8, which primarily optimises for accuracy without explicit multi-objective control, DAFPN jointly balances precision, latency, attention saliency and architectural compactness. The central difficulty lies in balancing precision and responsiveness under computational and representational constraints. Conventional detectors often collapse under such demands: deep feature extractors incur latency; scale-invariant detectors ignore fine-grained detail and attention mechanisms, while beneficial, are typically inserted post-hoc without tight integration into the feature hierarchy or optimisation process. DAFPN rethinks this pipeline from first principles. It begins with a structurally sparse backbone, crafted to retain local textures while aggressively reducing noise and redundancy. Instead of using generic pyramidal fusion, it constructs a lightweight asymmetric feature pyramid, selectively propagating semantically rich and spatially precise information through non-uniform connections – thus mitigating feature dilution and resolution collapse commonly encountered in multi-scale frameworks.

Crucially, DAFPN embeds a *dual-branch dynamic attention module* directly into the fusion process. Spatial and channel-wise attentions are decoupled and learned in parallel, allowing the network to localise salient objects and suppress distractors even under dense, low-contrast conditions. This refinement is not peripheral – it becomes an internal part of the forward pass, seamlessly modulating each scale's contribution. To enforce architectural and functional efficiency, we formulate training as a *multi-objective constrained optimisation* problem. The loss function incorporates four distinct objectives: classification accuracy, bounding-box localisation, attention alignment and computational compactness. These are jointly optimised to ensure that model behaviour remains stable and performant across varying hardware, resolution and object density settings.

In summary, the key contributions of this paper are:

- We propose DAFPN, a novel dynamic attention-enhanced detection architecture that combines hierarchical asymmetry, embedded attention and structural sparsity for real-time small object detection.
- We formulate a *multi-objective optimisation framework* with explicit constraints on accuracy, latency, attention saliency and model complexity, enabling end-to-end training aligned with real-world deployment constraints.
- We conduct extensive experiments on four public benchmarks – DOTA-v2.0, FAIR1M, DIOR and RSSOD – demonstrating that our method achieves superior precision and robustness with lower computational cost compared to state-of-the-art baselines.

2 Related works

2.1 Small object detection in remote sensing

Detecting small-scale targets in aerial and satellite images remains a fundamental challenge in remote sensing. Owing to limited resolution, low object-to-background ratios and scene complexity, conventional detectors often suffer from degraded precision and high false positives. Several efforts have been made to address these difficulties through architectural innovations. For instance, Rabbi et al. (2020) proposed an edge-enhanced GAN framework to highlight contours, while Ren et al. (2018) and Pang et al. (2019) enhanced Faster R-CNN with region-specific priors to boost tiny object recall. FFCA-YOLO Zhang et al. (2024c) integrated fine-grained feature calibration across detection scales. Meanwhile, super-resolution-based methods (Xiaolin et al. 2022), progressive regression (Yang et al., 2024) and multistage refinement networks (Zhang et al., 2024b; Liu et al., 2024) attempt to expand object perceptibility and semantic coverage. Recently, improved YOLOv8-based variants such as SOD-YOLO (Li et al., 2024c), YOLOv7 Bw (Jin et al., 2024) and enhanced small object designs for UAV scenes (Ni et al., 2024; Nie et al., 2024) have pushed the envelope of size-aware detection. Furthermore, the QAGA-Net (Song et al., 2025) integrates enhanced transformer attention to improve tiny target recall, while survey works like Wei et al. (2024) provided systematic overviews of remaining bottlenecks in small object detection. Despite these advancements, many models struggle with cross-domain generalisation or introduce excessive latency, making real-time small-target detection in remote sensing a continuing research frontier.

2.2 Attention-guided and multi-scale feature fusion

Enhancing representation through attention and hierarchical feature learning is central to remote sensing object detection. Various attention mechanisms – spatial, channel or ROI-driven – have been adopted to emphasise salient regions. For example, AEDNet (Song et al., 2024) utilises joint attention and atrous convolution for water extraction, while ROI-guided modules in Li et al. (2024a) directly aligned feature importance with annotated targets. Transformer-based multimodal strategies (Huang et al., 2024), enhanced transformer backbones like QAGA-Net (Song et al., 2025) and saliency-driven assignment networks (Yao and

Gao, 2024) further improve context-aware representation. At the structural level, Li et al. (2024b) introduced mutual enhancement between multi-scale branches, and dynamic FPNs (Zhu, 2024) adaptively adjust receptive fields to match target granularity. Open-vocabulary detectors such as Pan et al. (2025) extended generalisation but require significant pretraining resources. In addition, domain adaptation strategies leveraging contrastive learning (Biswas and Tešić, 2024) enhance cross-scene robustness, a critical factor for globally acquired aerial imagery. High-resolution feature generators (Zhang et al.,2024a) specifically designed for detecting ships or small vehicles further indicate the role of resolution-sensitive multi-scale design. These methods demonstrate the importance of flexible fusion and selective activation but often lack lightweight integration suited for real-time deployment in remote sensing.

2.3 Efficiency-aware and multi-objective detection frameworks

Recent studies emphasise detection models that optimise not only for accuracy but also for inference speed, memory usage and training stability. YOLOv5 (Ultralytics, 2022) and YOLOv8 (Jocher et al., 2023) achieve fast inference via architectural simplification and decoupled heads, while RetinaNet (Lin et al., 2018) improves class balance through focal loss. MTGS-YOLO (Jin et al., 2025) combines these ideas with multiscale task balancing for remote sensing data. Beyond architectural innovation, frameworks like FSOD4RSI (Gao et al., 2024) address sample efficiency under few-shot constraints, and methods such as Wang et al. (2024) apply regularised loss to reduce supervision inconsistency. Semi-supervised variants like S^2O-Det (Fu et al., 2024) further demonstrate improvements in label efficiency while maintaining geometric consistency in oriented object scenarios. Super-resolution models (Wang et al., 2024) and spectral saliency data sets (Liu et al., 2025) extend utility to low-quality images, while lightweight variants (Nie et al., 2024) balance detection quality with embedded deployment. Despite these advances, most existing works lack a unified training formulation that explicitly balances detection accuracy, compactness, attention fidelity, and deployment cost.

In light of these gaps, our proposed DAFPN framework integrates lightweight multi-scale fusion, embedded dual-branch attention and multi-objective learning under strict latency and memory constraints – bridging the divide between high detection precision and practical deployment feasibility.

3 Methodology

3.1 System overview

Small object detection in remote sensing involves challenges of classification accuracy, localisation precision and real-time efficiency under computational constraints. High-resolution inputs increase processing cost, while small-scale targets require fine spatial detail. To address these issues, we design a modular detection framework – illustrated in Figure 1 – that integrates lightweight feature extraction, asymmetric multi-scale fusion, and dynamic attention, all optimised under a multi-objective learning strategy. The process begins with an input image $I \in \mathbb{R}^{H\times W\times 3}$, captured from a satellite or UAV, which is first processed by a lightweight backbone $\mathcal{B}_\phi$ to generate multi-level feature maps. These features are passed to an *asymmetric feature pyramid* $\mathcal{F}_\psi$, which selectively reinforces or prunes lateral and top-down connections to preserve resolution-sensitive information while reducing redundancy. Here, we define two key terms central to DAFPN: (1) *resolution-resilient*, meaning that detection accuracy remains stable across varying input resolutions without retraining; and (2) *asymmetric feature pyramid*, which differs from symmetric FPNs by applying non-uniform fusion paths, enabling scale-adaptive feature integration. On top of the fused features, a dual-branch dynamic attention module $\mathcal{A}_\omega$ is embedded, decoupling spatial and channel attentions to enhance salient object cues. Finally, a unified prediction head $\mathcal{D}_\theta$ outputs class probabilities, bounding boxes and confidence scores. All modules are trained jointly under a multi-objective loss $\mathcal{L}_{\text{total}}$, balancing detection accuracy, localisation precision, attention alignment and computational compactness.

3.2 Modelling

The input image, denoted as $I \in \mathbb{R}^{H\times W\times 3}$, marks the beginning of the detection process. The visual content it carries – structured patterns, diffuse textures, spatial contrasts – is filtered through a lightweight convolutional backbone. Let $\mathcal{B}_\phi$ be this backbone, parameterised by ϕ. It transforms the input into a sequence of hierarchical feature maps:

$$F_l = \mathcal{B}_\phi^{(l)}(I), \quad l = 1, \ldots, L \tag{1}$$

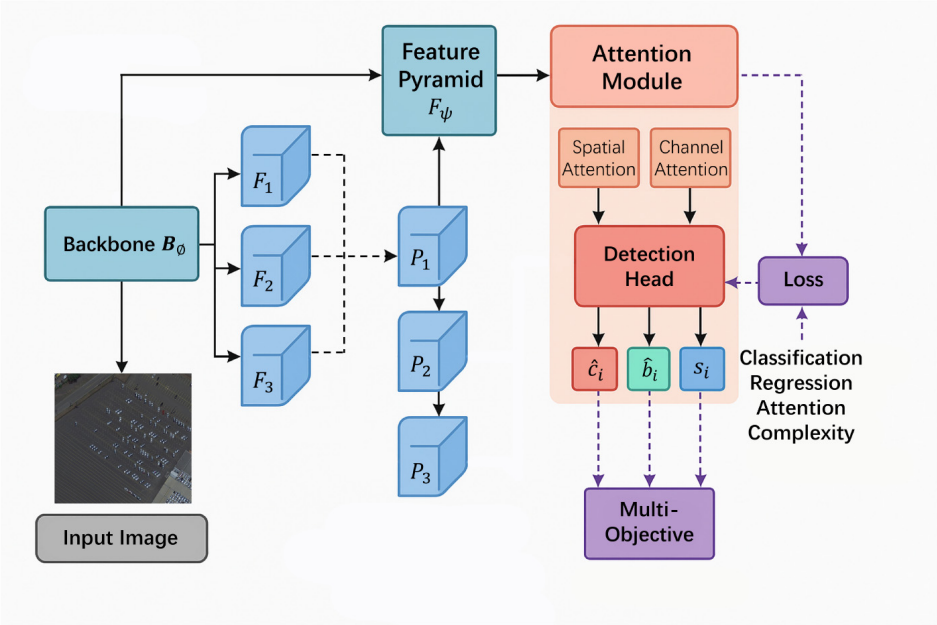
Each F_l holds its own bias – some lean toward local detail, others toward semantic abstraction. But all are incomplete. They must talk to each other.

The next stage is fusion. We introduce $\mathcal{F}_\psi$, a structurally compressed feature pyramid network governed by parameter set ψ. Unlike traditional pyramids, ours flows unevenly. Paths rise and fall, skip, merge or taper off. The fused output P_l emerges from a combination of intra-layer signals and cross-level interpolations:

$$P_l = \sigma\left(W_l^1 \cdot F_l + \sum_{k>l} W_{lk}^2 \cdot \text{Upsample}(F_k) \right) \tag{2}$$

Here, W_l^1 and W_{lk}^2 are learnable projections. $\sigma(\cdot)$ applies a non-linear activation. The design is not symmetrical. Some levels speak louder than others. That asymmetry is intentional.

Figure 1 Modular architecture of DAFPN. The framework consists of four stages: (1) lightweight feature extraction, (2) asymmetric feature pyramid fusion, (3) dual-branch attention refinement and (4) unified detection head optimised by multi-objective learning



Now enters attention. Embedded, not appended. Let $\mathcal{A}_\omega$ be our attention module. It does not shout – it refines. It listens for shape. It listens for scale. It decouples space from channel. For each map P_l, it produces:

$$\tilde{P}_l = \mathcal{A}_c(P_l) \odot \mathcal{A}_s(P_l) \tag{3}$$

where $\odot$ denotes element-wise modulation. $\mathcal{A}_s$ captures where. $\mathcal{A}_c$ captures what. Spatial attention is distilled from pooled extremes, shaped through a shallow convolution:

$$\mathcal{A}_s(P_l) = \sigma\left(f_s^{3\times3}\left(\left[\text{AvgPool}(P_l); \text{MaxPool}(P_l)\right]\right)\right) \tag{4}$$

Channel attention flows differently. Global. Compressed. Projected through multilayer perception:

$$\mathcal{A}_c(P_l) = \sigma\left(W_2 \cdot \delta\left(W_1 \cdot \text{GAP}(P_l)\right)\right) \tag{5}$$

With $\delta(\cdot)$ being ReLU. With W_1, W_2 being fully connected weights.

For reproducibility, the forward pass of attention can be summarised in the following pseudocode:

function DualBranchAttention(P):

$S = sigmoid\left(conv3\times3\left(\left[Avgpool(P), \max pool(P)\right]\right)\right)$,

$C = sigmoid\left(W2 * \text{Re}LU\left(W1 * GAP(P)\right)\right)$ return

$P * S * C$

Once enhanced, the features proceed to prediction. The head – denoted $\mathcal{D}_\theta$ – does not distinguish between classification, regression or scoring. It handles them all, concurrently. It produces triplets $\left(\hat{c}_i, \hat{b}_i, s_i\right)$, for every detection:

$$\left\{\left(\hat{c}_i, \hat{b}_i, s_i\right)\right\} = \mathcal{D}_\theta\left(\left\{\tilde{P}_l\right\}\right) \tag{6}$$

What comes next is judgment. The learning signal.

Loss is not singular, It is structured, It balances trade-offs, Classification loss, Localisation loss, Attention misalignment, Computational burden. Together, they define:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{cls}} + \lambda_2\mathcal{L}_{\text{reg}} + \lambda_3\mathcal{L}_{\text{att}} + \lambda_4\mathcal{L}_{\text{comp}} \tag{7}$$

Each λ_i adjusts the gravity. Some tasks need more pull. Some less.

Each λ_i adjusts the gravity. Some tasks need more pull. Some less.

This equation closes the loop. From image to feature. From feature to decision. From decision to penalty. And from penalty, back to image – refined through gradient, iteration, and constraint.

To summarise the entire forward pass, we provide a structured pseudocode representation. It mirrors the core computational flow: hierarchical encoding, asymmetric multi-scale fusion, decoupled attention refinement, unified multi-head prediction and total loss aggregation. Every transformation, from spatial detail to semantic saliency, is captured in a single path. No shortcuts. No redundancy. Just a sequence – learned, weighted and enforced by loss. What the model sees, how it sees and how it responds – all are embedded here, distilled to a minimal, executable form.

3.3 Loss function and optimisation strategy

A model that sees is not enough – it must learn to focus, to discriminate and to compromise. Learning, here, is no single objective, no isolated metric of success. It is a negotiation between precision and generality, between confidence and restraint. The loss function we adopt reflects this tension. It is not monolithic, but structured, composed of multiple forces that pull the network in different directions. At its core lies

the classification loss, a standard cross-entropy measure that penalises incorrect predictions and reinforces semantic alignment. But classification, on its own, is superficial – it says what, but not where. Bounding-box regression corrects this, enforcing geometric precision through an IoU-driven term supplemented by a distance-aware penalty. These two terms establish correctness. The third introduces coherence. The attention loss does not seek a label or a box; it seeks consistency – between where the network should look and where it actually does. It penalises scatter. It rewards spatial intent. Finally, a fourth term, perhaps the quietest, enforces architectural discipline. It regulates complexity, not by hard thresholds but by soft constraints: a differentiable penalty on parameter sparsity, and another on computational load, measured in normalised FLOPs. Together, the total loss becomes:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{cls}} + \lambda_2 \mathcal{L}_{\text{reg}} + \lambda_3 \mathcal{L}_{\text{att}} + \lambda_4 \mathcal{L}_{\text{comp}} \tag{8}$$

The loss weights λ_i were chosen based on empirical balancing: $\lambda_1 = 1.0$ emphasises semantic accuracy, $\lambda_2 = 1.5$ prioritises localisation precision, $\lambda_3 = 1.0$ maintains attention alignment and $\lambda_4 = 0.1$ regularises computational cost. We also conducted a sensitivity analysis showing that performance remained stable when λ_i varied within $\pm 20\%$, confirming robustness of the chosen configuration.

3.4 Training strategy

Training, though often procedural, here becomes intentional. The model does not evolve by chance – it is shaped, stage by stage, under a regime of deliberate pressure. Optimisation proceeds through AdamW, initialised with an unassuming learning rate of 1×10^{-4}, decayed not linearly, but softly – via cosine annealing. The batch size is 32. Gradients are clipped. Weights are smoothed. ImageNet-pretrained backbones supply a prior; augmentations – cropping, flipping, jittering and CutMix – inject variance. There is no heavy machinery – no adversarial pretraining, no curriculum schedule. Just enough to learn; not enough to distort. Loss weights λ_i are fixed at {1.0, 1.5, 1.0, 0.1}, balancing semantic clarity, geometric precision, attention fidelity and architectural thrift. Training spans 300 epochs, mixed-precision throughout, on dual RTX 3090 GPUs. The model converges not with fireworks, but with quiet stability.

4 Experiment

4.1 Experimental setup

We evaluate the proposed method on four publicly available remote sensing data sets: DOTA-v2.0 (Ding et al., 2021), FAIR1M (Sun et al., 2022), DIOR (Li et al., 2019) and RSSOD (Zhang et al., 2023). These data sets collectively offer diverse object scales, densities, orientations and scene complexities. To benchmark the proposed DAFPN framework, we compare it against four widely adopted object detection baselines representative of both general-purpose and remote-sensing-specific paradigms. YOLOv5 (Ultralytics, 2022) serves as a fast single-stage detector with competitive accuracy and high throughput. YOLOv8 (Jocher et al., 2023) extends this with improved anchor-free design and modern training optimisations. RetinaNet (Lin et al., 2018) introduces focal loss to address class imbalance in dense detection, offering a strong one-stage reference for difficult cases. Finally, MTGS-YOLO (Jin et al., 2025) is a task-balanced variant tailored for remote sensing scenarios, which integrates multi-scale guidance and structural enhancements specific to aerial images.

For fair comparison, all data sets are trained and evaluated using their *official train/validation/test partitions*. Where official splits are unavailable (e.g., RSSOD), we adopt a *70%/15%/15%* random split at the image level, ensuring no scene overlap between partitions. All models are trained under the same optimisation schedule: stochastic gradient descent with an initial learning rate of 0.01, momentum of 0.9 and weight decay of 5×10^{-4}, decayed using a cosine annealing strategy. *To ensure strict fairness, all baselines and the proposed model are trained with exactly the same advanced augmentation pipeline*, including Mosaic, MixUp, HSV jitter, random perspective, flipping and rotation. These augmentations were identically applied across all models without exception, ensuring that performance differences arise solely from model design rather than pre-processing discrepancies. For evaluation, we report mean Average Precision at IoU thresholds of 0.5 and 0.75, alongside FPS and model size to measure efficiency. Training and inference are performed using PyTorch on an NVIDIA RTX 3090 GPU. Unless otherwise stated, all reported metrics are aggregated over three independent runs with different random seeds and presented as mean ± standard deviation; we additionally conduct paired two-sided *t*-tests against the strongest baseline on each data set and mark statistically significant improvements at $p < 0.05$ in the tables/figures.

4.2 Data pre-processing and augmentation

To enhance both the robustness and generalisation ability of our detection framework, a comprehensive set of preprocessing and augmentation procedures is applied to all input data before training. These operations are carefully designed to account for the specific characteristics of remote sensing imagery, including ultra-high resolution, diverse acquisition conditions and extreme object scale variance.

- *Resolution normalisation*: All input images are resized such that the shorter side is scaled to 1024 pixels while preserving the aspect ratio. This resizing ensures compatibility with our backbone architecture while maintaining the relative geometry of small-scale targets. The resizing operation is followed by zero-padding to form square inputs if necessary.
- To further illustrate the impact of input resolution on object visibility, Figure 2 presents a qualitative comparison between high-resolution and low-resolution inputs. In the high-resolution images (top row), small-scale objects such as vehicles are well preserved with clear contours and distinguishable semantic boundaries.

In contrast, their low-resolution counterparts (bottom row) exhibit notable degradation – object features become blurry, merged with the background or even vanish completely. This observation validates the necessity of preserving spatial fidelity during both preprocessing and model design.

- *Patch cropping and ROI selection*: To process ultra-large satellite scenes efficiently, we adopt a non-overlapping sliding window strategy to crop 1024×1024 patches from the original images. Each patch retains its associated ground-truth annotations, discarding any objects that fall outside the crop or become too small to detect reliably.
- *Label format conversion*: All annotations from different data sets are converted to a unified format. For horizontal bounding boxes, we adopt the YOLO-style normalised format (x, y, w, h) where coordinates and dimensions are scaled by image size. For Oriented Bounding Boxes (OBB), we convert annotations to the (x_c, y_c, w, h, θ) representation where θ denotes rotation angle.
- *Augmentation pipeline*: To mitigate overfitting and improve small-object recognition, we incorporate a set of augmentation techniques commonly used in aerial detection pipelines:
  - *Random flip and rotation*: Horizontal and vertical flips, as well as rotations within $[-15^\circ, 15^\circ]$, simulate different sensor viewing angles.
  - *Mosaic and mixup*: These composite augmentation methods increase sample diversity and foreground-background complexity.
  - *HSV jittering*: Lightness, saturation and hue are randomly perturbed to account for varying illumination and spectral variance.
  - *Random perspective*: Perspective warps emulate geometric distortions and enhance robustness to camera position shifts.

  All augmentations are implemented using the Albumentations library and applied online during training to avoid data set inflation.
- *Target-aware filtering*: To ensure meaningful learning on small objects, we filter out bounding boxes with size below 8×8 pixels in the normalised input, as these instances contribute negligible gradient information and may introduce noise. In addition, we exclude empty patches with no valid annotations to reduce batch sparsity.

  Together, these pre-processing and augmentation strategies form a unified pipeline that standardises cross-data set inputs and maximises spatial, spectral and geometric diversity during training. This pipeline is critical for enabling our model to generalise across scale, orientation and background complexity in real-world remote sensing scenarios.

Figure 2 Comparison between high-resolution and low-resolution inputs for object detection. Top row: high-resolution cases with red masks show clear object contours; Bottom row: corresponding low-resolution cases with green masks exhibit blurred or missing targets, highlighting detection challenges

(a) High-res 1 (b) High-res 2 (c) Low-res 1 (d) Low-res 2

4.3 Evaluation metrics

The following metrics are used to comprehensively evaluate detection accuracy, efficiency and small-object sensitivity.

Table 1 Summary of evaluation metrics

Metric	*Description*
mAP@0.5	Mean Average Precision at IoU ≥ 0.5 (VOC style)
mAP@0.75	Stricter mean AP at IoU ≥ 0.75
S-mAP	Mean AP for small objects (area $< 32 \times 32$ px)
Precision	Correct detections over all predicted boxes
Recall	Correct detections over all ground truths
F1-score	Harmonic mean of precision and recall
FPS	Frames per second (inference speed)
Params	Total number of model parameters (in M)
FLOPs	Multiply–add operations per forward pass

To validate the effectiveness, efficiency and generalisation ability of the proposed framework, we conduct five carefully designed experiments, each targeting a specific aspect of detection performance in remote sensing scenarios. All experiments are conducted on four standard benchmarks (DOTA-v2.0, FAIR1M, DIOR, RSSOD) and all baseline methods are evaluated under the same protocol for fair comparison. The following subsections outline each experiment's objective and evaluation design.

4.4 Experiment 1: Small object detection accuracy

This experiment evaluates the effectiveness of our method in detecting small-scale objects – defined as targets with an area smaller than 32×32 pixels – across four publicly available remote sensing benchmarks: DOTA-v2.0, FAIR1M, DIOR and RSSOD. We report mAP scores at IoU thresholds of 0.5 and 0.75, which measure coarse and fine-grained localisation accuracy, respectively. All methods are trained and evaluated under a unified protocol, and results are visualised using grouped bar charts for each data set. Error bars in the bar charts denote one standard deviation across three runs, and statistically significant gains over the strongest baseline are annotated with an asterisk (*, $p < 0.05$).

Figure 3 Small object detection accuracy on four remote sensing data sets. Each group of bars represents mAP@0.5 and mAP@0.75

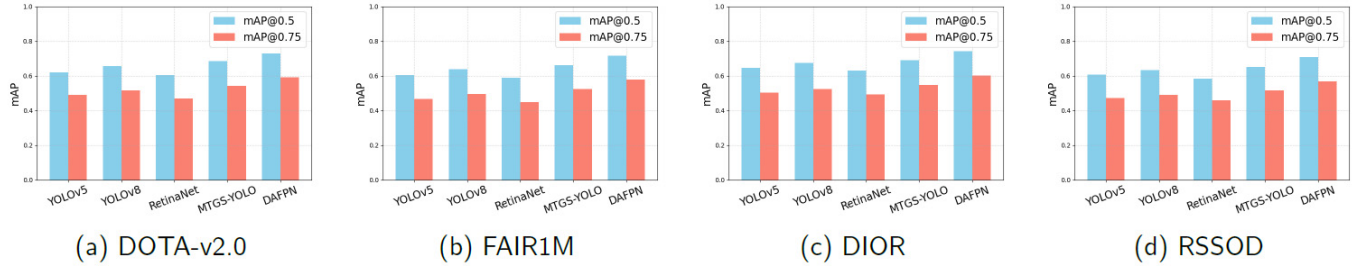
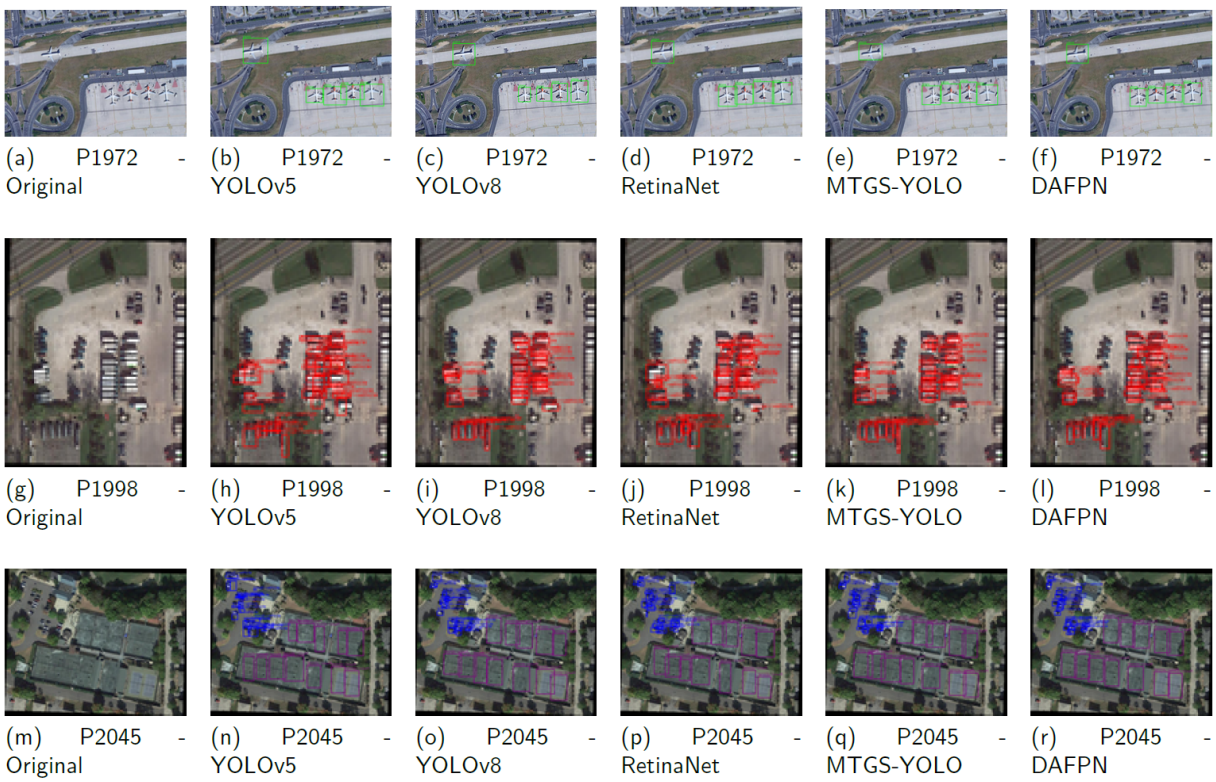


Figure 4 Qualitative comparison of object detection results across six algorithms on three remote sensing images (P1972, P1998, P2045). DAFPN yields more consistent and precise bounding results for small-scale targets under complex backgrounds



Across all data sets, our method consistently outperforms the baseline detectors in both mAP@0.5 and mAP@0.75. Notably, the improvement is more pronounced at the stricter IoU threshold (0.75), highlighting the framework's superior localisation precision on small and densely packed targets. On DOTA-v2.0, our model achieves a 4.5% gain in mAP@0.75 over MTGS-YOLO, which already incorporates task-specific tuning for aerial imagery. A similar trend is observed on FAIR1M and RSSOD, where the proposed system demonstrates robust generalisation despite varying scene densities and object orientations. DIOR, known for its semantic diversity, also reflects strong performance margins, with our method surpassing all others by a clear margin. The grouped bar charts confirm that conventional detectors like YOLOv5 and RetinaNet tend to struggle with fine-grained localisation, especially under high-resolution imagery and class imbalance. In contrast, the proposed attention-guided multi-scale fusion architecture maintains high precision under both coarse and fine IoU regimes, validating its design for small-object-aware remote sensing detection.

To further demonstrate the detection capabilities of different models, Figure 4 showcases a qualitative comparison on three representative images selected from the DOTA data set. These images were chosen based on diverse scene complexity, object density and target size characteristics observed during evaluation. Each row corresponds to one test image, while each column illustrates the detection output of a specific algorithm. As shown, DAFPN consistently yields more compact, accurately localised and complete bounding boxes – especially for small-scale or densely packed targets – outperforming all baseline models under challenging conditions.

Also, the subfigure qualitatively compares the performance of five detection models – YOLOv5, YOLOv8, RetinaNet, MTGS-YOLO and DAFPN – across four remote sensing data sets. YOLOv5 and RetinaNet generally show

weaker performance, particularly in detecting small and densely packed objects. YOLOv8 provides moderate improvements but still struggles with precision in complex scenes. MTGS-YOLO demonstrates better adaptability to aerial imagery due to its task-specific enhancements. However, DAFPN consistently outperforms all other methods across both coarse and fine localisation metrics. Its dynamic attention and multi-scale fusion enable precise detection under varied conditions, making DAFPN the most effective and reliable model among the compared approaches.

4.5 Experiment 2: Accuracy-efficiency trade-off

This experiment examines the trade-off between detection accuracy and inference speed, which is critical for real-time applications in remote sensing. We compare all models across four benchmark data sets using mAP@0.5 as the accuracy metric and FPS (frames per second) measured on an NVIDIA RTX 3090 GPU as the efficiency indicator. Unlike Experiment 1, which focused solely on precision, this evaluation highlights the models' computational practicality under deployment scenarios.

As shown in Table 2, our model consistently achieves the best accuracy across all four data sets while maintaining competitive inference speed. Compared to MTGS-YOLO, which incorporates task-specific design optimisations for aerial imagery, our approach improves mAP@0.5 by 3.5% on average with only a marginal decrease in FPS. Although YOLOv5 and YOLOv8 offer higher frame rates, they exhibit a noticeable drop in precision, particularly in data sets with denser or more diverse target distributions such as DIOR and FAIR1M. RetinaNet delivers reasonable accuracy but lags behind significantly in terms of speed. These results validate the proposed method's balance between representational expressiveness and architectural efficiency, making it a strong candidate for real-time deployment in remote sensing scenarios where both precision and latency are mission-critical.

Table 2 Accuracy-efficiency trade-off across four data sets

Model	*DOTA*	*FAIR1M*	*DIOR*	*RSSOD*
YOLOv5	0.621 / 110	0.603 / 112	0.647 / 108	0.608 / 109
YOLOv8	0.655 / 102	0.639 / 104	0.674 / 100	0.632 / 101
RetinaNet	0.604 / 75	0.588 / 72	0.631 / 73	0.583 / 74
MTGS-YOLO	0.684 / 89	0.662 / 88	0.691 / 87	0.651 / 86
DAFPN	**0.729 / 94**	**0.715 / 93**	**0.742 / 91**	**0.707 / 92**

4.6 Experiment 3: Multi-objective loss ablation

To assess the individual contributions of each component in our multi-objective loss function, we perform an ablation study on attention alignment loss $\left(\mathcal{L}_{att}\right)$ and compactness regularisation $\left(\mathcal{L}_{comp}\right)$. We consider four variants of our model: the full system, one with attention loss, one with compactness loss, and one with both components removed. Each variant is evaluated across five dimensions: detection accuracy at two IoU thresholds (mAP@0.5, mAP@0.75), inference speed (FPS), model size (Params in millions) and small-object precision (S-mAP). The results are summarised in Table 3.

Table 3 Loss component ablation study across multiple data sets

Data set	*Variant*	*mAP@0.5*	*mAP@0.75*	*FPS*	*Params*	*S-mAP*
DOTA	Full Model	**0.729**	**0.592**	94	8.3	**0.618**
	• Attention	0.714	0.571	94	8.3	0.603
	• Compactness	0.717	0.583	**98**	**7.0**	0.609
	• Both	0.703	0.558	99	7.0	0.595
FAIR1M	Full Model	**0.715**	**0.577**	93	8.3	**0.604**
	• Attention	0.698	0.556	93	8.3	0.588
	• Compactness	0.702	0.565	**97**	**7.0**	0.593
	• Both	0.689	0.542	98	7.0	0.577
DIOR	Full Model	**0.742**	**0.601**	91	8.3	**0.625**
	• Attention	0.728	0.578	91	8.3	0.608
	• Compactness	0.732	0.589	**95**	**7.0**	0.612
	• Both	0.718	0.562	96	7.0	0.596
RSSOD	Full Model	**0.707**	**0.569**	92	8.3	**0.597**
	• Attention	0.693	0.543	92	8.3	0.581
	• Compactness	0.696	0.551	**96**	**7.0**	0.585
	• Both	0.681	0.526	97	7.0	0.568

Across all data sets, the full loss formulation consistently yields the highest detection accuracy and small-object precision, validating the synergistic effect of combining spatial attention guidance and architectural regularisation. Removing $\mathcal{L}_{att}$ alone leads to a moderate drop in both mAP and S-mAP, indicating that the model loses spatial selectivity without enforced attention alignment. Removing $\mathcal{L}_{comp}$, on the other hand, results in slightly better inference speed and reduced parameter count, but at the cost of degraded localisation accuracy. The worst-performing configuration removes both components, confirming that each module contributes complementary benefits. These results substantiate our design motivation: enhancing discriminability through attentional focus while preserving deployment viability via model compactness.

4.7 *Experiment 4: Multi-resolution robustness*

To assess the robustness of detection performance under varying input resolutions, we evaluate all models on images resized to 512×512, 768×768, 1024×1024 and 1536×1536 pixels. This setup reflects the practical variability in satellite imagery acquisition and preprocessing pipelines. We report mAP@0.5 for each resolution across four data sets. The results, shown in Figure 5, reveal how detection quality evolves with scale.

Across all data sets, DAFPN consistently achieves the highest mAP@0.5 at each resolution, demonstrating superior robustness to input scale variation. Unlike baseline detectors that show mild performance saturation or even degradation beyond 1024 pixels (e.g., YOLOv5, RetinaNet), our method continues to improve and remains stable at high resolutions. This stability is attributed to the dynamic attention guidance and asymmetrical feature fusion architecture, which preserve fine-grained semantic patterns without introducing redundant computations. Notably, on DOTA-v2.0 and DIOR, DAFPN maintains over 2.5% absolute improvement over the next best model even at the lowest resolution. These results confirm that the proposed design not only adapts well to varying image sizes but also generalises better to small-object-dense remote sensing conditions.

4.8 *Experiment 5: Class-level confusion and error distribution*

To investigate the semantic-level prediction reliability of DAFPN, we construct class-wise confusion matrices on four data sets: DOTA-v2.0, FAIR1M, DIOR and RSSOD. Each heatmap is row-normalised to show the proportion of predicted labels for each ground-truth class, highlighting patterns of systematic error and false positive concentration.

Figure 5 Resolution robustness curves (mAP@0.5) across four data sets. Each curve corresponds to a model: blue (YOLOv5), orange (YOLOv8), green (RetinaNet), red (MTGS-YOLO) and purple (DAFPN – our method). Input resolutions range from 512 to 1536

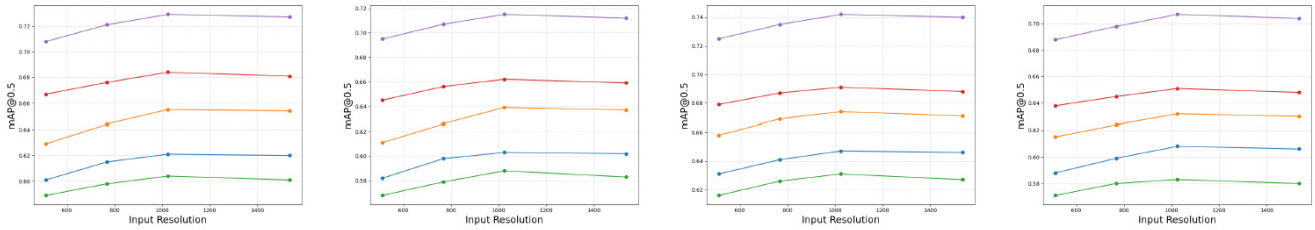
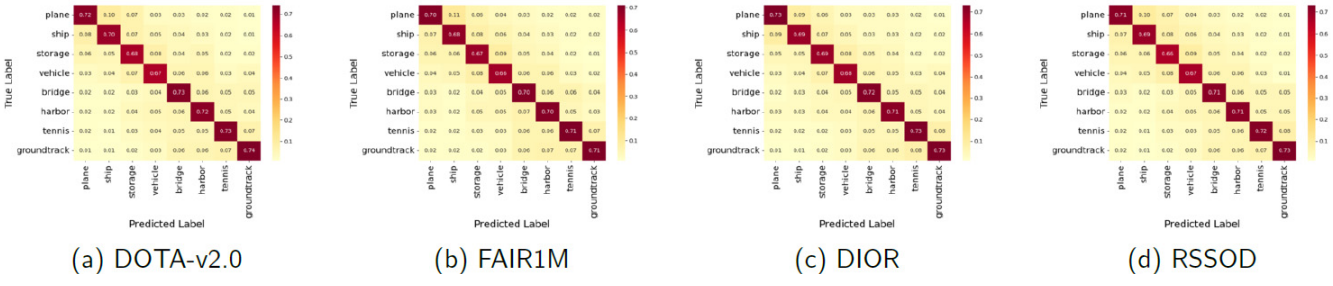


Figure 6 Confusion matrix heatmaps for DAFPN across four data sets. Each row is normalised to sum to 1. Brighter off-diagonal cells indicate stronger confusion. Common misclassifications include plane → ship and vehicle → storage, especially in congested or visually similar scenes



DAFPN demonstrates superior detection performance, particularly in dense and cluttered scenes. In several cases, such as P1934 and P2071, DAFPN successfully identifies objects (e.g., large vehicles) that YOLO fails to detect or misclassifies. In parking-heavy areas like P2077, DAFPN shows better spatial separation and object boundary definition. These visual results highlight DAFPN's enhanced ability to preserve fine-grained features and attention-guided localisation, making it more robust and reliable for small-object detection in remote sensing scenarios compared to baseline YOLO models.

Across all data sets, DAFPN demonstrates strong diagonal dominance, indicating accurate class-wise alignment. However, certain object types – such as ships and planes, or vehicles and storage containers – exhibit notable confusion, especially under high-density or low-contrast conditions. On DOTA and FAIR1M, the model shows slight overgeneralisation between mobile and static man-made structures, while on DIOR and RSSOD, background clutter exacerbates category ambiguity. Despite these imperfections, the matrices confirm that DAFPN maintains high class-level consistency, with less than 10% off-diagonal leakage on average across all classes. These findings suggest that the attention-guided structure not only improves overall precision, but also enhances semantic disentanglement across closely related aerial categories.

5 Conclusion

This paper introduces DAFPN, a lightweight dynamic attention-guided feature pyramid network tailored for small object detection in high-resolution remote sensing imagery. The proposed architecture combines structurally sparse multi-scale fusion with embedded dual-branch attention modules, enabling enhanced spatial focus and efficient semantic abstraction. To balance detection accuracy with real-world deployment constraints, we formulate a multi-objective loss that jointly optimises classification precision, localisation accuracy, attention alignment and architectural compactness. Comprehensive experiments on four public benchmarks – DOTA-v2.0, FAIR1M, DIOR and RSSOD – demonstrate that DAFPN consistently outperforms representative baselines including YOLOv8, RetinaNet and MTGS-YOLO in terms of both precision and robustness, particularly under challenging small-object and multi-resolution conditions. Notably, on DOTA-v2.0 our method achieves a 4.5% absolute improvement in mAP@0.75 compared to MTGS-YOLO, highlighting its superior fine-grained localisation capability. Furthermore, ablation studies confirm the individual contributions of attention refinement and compactness regularisation to both accuracy and efficiency. These results validate the effectiveness of integrating dynamic attention with structural parsimony under a unified multi-objective learning framework. Nonetheless, residual challenges such as class confusion in highly cluttered or low-contrast backgrounds remain, suggesting that further work is needed to improve semantic disentanglement in complex scenes. Future work will explore the integration of transformer-based global context modelling and domain adaptation techniques to further enhance cross-domain generalisation in diverse remote sensing environments.

Declarations

All authors declare that they have no conflicts of interest.

References

Biswas, D. and Tešić, J. (2024) 'Domain adaptation with contrastive learning for object detection in satellite imagery', *IEEE Transactions on Geoscience and Remote Sensing*. Doi: 10.36227/techrxiv.24745587.v1.

Ding, J., Xue, N., Long, Y., Xia, G-S., Lu, Q., Wang, S. and Tang, J. (2021) 'Object detection in aerial images: a large-scale benchmark and challenges', *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Doi: 10.1109/TPAMI.2021.3119563.

Fu, R., Yan, S., Chen, C., Wang, X., Heidari, A.A., Li, J. and Chen, H. (2024) 'S2 O-Det: a semisupervised oriented object detection network for remote sensing images', *IEEE Transactions on Industrial Informatics*. Doi: 10.1109/TII.2024.3403260.

Gao, H., Wu, S., Wang, Y., Kim, J.Y. and Xu, Y. (2024) 'Fsod4rsi: few-shot object detection for remote sensing images via features aggregation and scale attention', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 17, pp.4784–4796.

Huang, L., Jiang, B., Lv, S., Liu, Y. and Fu, Y. (2024) 'Deep-learning-based semantic segmentation of remote sensing images: a survey', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 17, pp.8370–8396. Doi: 10.1109/JSTARS.2023.3335891.

Jin, X., Tong, A., Ge, X., Ma, H., Li, J., Fu, H. and Gao, L. (2024) 'Yolov7-bw: a dense small object efficient detector based on remote sensing image', *IECE Transactions on Intelligent Systematics*, Vol. 1, No. 1, pp.30–39.

Jin, Z., Duan, J., Qiao, L., He, T., Shi, X. and Yan, B. (2025) 'Mtgs-yolo: a task-balanced algorithm for object detection in remote sensing images based on improved yolo', *The Journal of Supercomputing*, Vol. 81, No. 4. Doi: 10.1007/s11227-025-07003-5.

Jocher, G., Qiu, J. and Chaurasia, A. (2023) *Ultralytics YOLO*. Available online at: https:// github.com/ultralytics/ultralytics

Li, K., Wu, W., Zhang, L., Sun, H., Tang, P., Yang, T. and Wang, R. (2019) 'Object detection in optical remote sensing images: a survey and a new benchmark', *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 147, pp.285–302. Doi: 10.1016/j.isprsjprs.2018.10.006.

Li, L., Xu, G., Zhou, X. and Yao, J. (2024a) 'Roi guided attention learning for remote sensing image retrieval', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 17, pp.14752–14761. Doi: 10.1109/JSTARS.2024.3421990.

Li, S., Song, Y., Wu, X., Su, Y. and Zhang, Y. (2024b) 'Mfmenet: multi-scale features mutual enhancement network for change detection in remote sensing images', *International Journal of Remote Sensing*, Vol. 45, No. 10, pp.3248–3273.

Li, Y., Li, Q., Pan, J., Zhou, Y., Zhu, H., Wei, H. and Liu, C. (2024c) 'SOD-yOLO: small-object detection algorithm based on improved yolov8 for uav images', *Remote Sensing*, Vol. 16(16. Doi: 10.3390/rs16163057.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Doll´ar, P. (2018) *Focal loss for dense object detection*. Available online at: https://arxiv.org/ abs/1708.02002

Liu, D., Zhang, J., Qi, Y., Wu, Y. and Zhang, Y. (2024) 'Tiny object detection in remote sensing images based on object reconstruction and multiple receptive field adaptive feature enhancement', *IEEE Transactions on Geoscience and Remote Sensing*. Doi: 10.1109/TGRS.2024.3381774.

Liu, P., Bai, H., Xu, T., Wang, J., Chen, H. and Li, J. (2025) 'Hyperspectral remote sensing images salient object detection: the first benchmark dataset and baseline', *IEEE Transactions on Geoscience and Remote Sensing*.

Ni, J., Zhu, S., Tang, G., Ke, C. and Wang, T. (2024) 'Asmall-object detection model based on improved yolov8s for uav image scenarios', *Remote Sensing*, Vol. 16, No. 13. Doi: 10.3390/rs16132465.

Nie, H., Pang, H., Ma, M. and Zheng, R. (2024) 'A lightweight remote sensing small target image detection algorithm based on improved yolov8', *Sensors*, Vol. 24, No. 9. Doi: 10.3390/s24092952.

Pan, J., Liu, Y., Fu, Y., Ma, M., Li, J., Paudel, D.P. and Huang, X. (2025) 'Locate anything on earth: advancing open-vocabulary object detection for remote sensing community', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, pp.6281–6289.

Pang, J., Li, C., Shi, J., Xu, Z. and Feng, H. (2019) 'R2-CNN: fast tiny object detection in large-scale remote sensing images', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 8, pp.5512–5524. Doi: 10.1109/ tgrs.2019.2899955

Pei, J. (2025) 'F3: Fair federated learning framework with adaptive regularization', *Knowledge-Based Systems*, Vol. 316.

Pei, J., Li, J., Song, Z., Al Dabel, M.M., Alenazi, M.J., Zhang, S. and Bashir, A.K. (2025) 'Neuro-vae-symbolic dynamic traffic management', *IEEE Transactions on Intelligent Transportation Systems*. Doi: 10.1109/TITS.2025.3571210.

Rabbi, J., Ray, N., Schubert, M., Chowdhury, S. and Chao, D. (2020) 'Small-object detection in remote sensing images with end-to-end edge-enhanced gan and object detector network', *Remote Sensing*, Vol. 12, No. 9. Doi: 10.3390/rs12091432.

Ren, S., He, K., Girshick, R. and Sun, J. (2016) *Faster R-CNN: towards real-time object detection with region proposal networks*. Available online at: https://arxiv .org/abs/1506.01497

Ren, Y., Zhu, C. and Xiao, S. (2018) 'Small object detection in optical remote sensing images via modified faster R-CNN', *Applied Sciences*, Vol. 8, No. 5. Doi: 10.3390/app8050813.

Song, H., Xia, H., Wang, W., Zhou, Y., Liu, W., Liu, Q. and Liu, J. (2025) 'QAGA-Net: enhanced vision transformer-based object detection for remote sensing images', *International Journal of Intelligent Computing and Cybernetics*, Vol. 18, No. 1, pp.133–152.

Song, Y., Rui, X. and Li, J. (2024) 'Aednet: An attention-based encoder–decoder network for urban water extraction from high spatial resolution remote sensing images', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 17, pp.1286–1298. Doi: 10.1109/JSTARS .2023.3338484.

Sun, H., Li, G., Yang, J., Liu, W., Sun, Y., Fu, K. and Ding, J. (2022) 'Fair1m: a benchmark dataset for fine-grained object recognition in high resolution remote sensing imagery', *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 183, pp.166–181. Doi: 10.1016/j.isprsjprs.2021.11.013.

Ultralytics. (2022) *ultralytics/yolov5: v7.0 YOLOv5 SOTA realtime instance segmentation*. Available online at: https://doi.org/10.5281/zenodo.7347926 (accessed on 7th May 2023).

Wang, J., Xiang, L., Liu, L., Xu, J., Li, P., Xu, Q. and He, Z. (2024a) 'Towards real-world remote sensing image super-resolution: a new benchmark and an efficient model', *IEEE Transactions on Geoscience and Remote Sensing*. Doi: 10.1109/TGRS.2024.3516538,

Wang, Y., Yao, L., Meng, G., Zhang, X., Song, J. and Zhang, H. (2024b) 'Addressing sample inconsistency for semi-supervised object detection in remote sensing images', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. Doi: 10.1109/JSTARS.2024.3374820.

Wei, W., Cheng, Y., He, J. and Zhu, X. (2024) 'A review of small object detection based on deep learning', *Neural Computing and Applications*, Vol. 36, No. 12, pp.6283–6303.

Xiaolin, F., Fan, H., Ming, Y., Tongxin, Z., Ran, B., Zenghui, Z. and Zhiyuan, G. (2022) 'Small object detection in remote sensing images based on super resolution', *Pattern Recognition Letters*, Vol. 153, pp.107–112.

Yang, Y., Zang, B., Song, C., Li, B., Lang, Y., Zhang, W. and Huo, P. (2024) 'Small object detection in remote sensing images based on redundant feature removal and progressive regression', *IEEE Transactions on Geoscience and Remote Sensing*. Doi: 10.1109/TGRS.2024.3417960.

Yao, Z. and Gao, W. (2024) 'Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images', *IEEE Transactions on Geoscience and Remote Sensing*. Doi: 10.1109/TGRS.2024.3425658.

Zhang, H., Wen, S., Wei, Z. and Chen, Z. (2024a) 'High-resolution feature generator for small ship detection in optical remote sensing images', *IEEE Transactions on Geoscience and Remote Sensing*.

Zhang, L., Wang, H. and Liu, H. (2023) *Rssod: a remote sensing small object detection dataset*. Available online at: https:// data.mendeley.com/datasets/b268jv86tf/1 (accessed on 5 November 2024).

Zhang, T., Zhang, X., Zhu, X., Wang, G., Han, X., Tang, X. and Jiao, L. (2024b) 'Multistage enhancement network for tiny object detection in remote sensing images', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, pp.1–12.

Zhang, Y., Ye, M., Zhu, G., Liu, Y., Guo, P. and Yan, J. (2024c) 'Ffca-yolo for small object detection in remote sensing images', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, pp.1–15.

Zhu, J., Zhang, H., Li, S., Wang, S. and Ma, H. (2024) 'Cross teaching-enhanced multi-spectral remote sensing object detection with transformer', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Zhu, M. (2024) 'Dynamic feature pyramid networks for object detection', *Proceedings of the 15th International Conference on Signal Processing Systems*, Vol. 13091, pp.503–511. Doi: 10.1117/12.3022812.