RESEARCH ARTICLE

MS-CADNet: A Multi-Scale Context Attention Network for Efficient Object Detection in UAV Imagery

Abdul Bari[1], **Fatima Memon**[1], **Hafsa Waheed**[2] **and Ghulam E Mustafa Abro**[3,*]

[1] School of Technology, Cardiff Metropolitan University, Cardiff CF5 2YB, United Kingdom
[2] Department of IT and Computer Science, Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology, Haripur 22620, Pakistan
[3] Research Unit for Robophilosophy and Integrative Social Robotics (RISR), Aarhus University, Aarhus 8000, Denmark

Abstract

With the rapid advancement of unmanned aerial vehicle (UAV) technology, there is a need for lightweight and accurate object detection on resource-constrained platforms. This paper proposes MS-CADNet, an anchor-free network for small object detection in aerial imagery. It uses a MobileNetV3-Small backbone and a two-branch gated Context Attention Module (CAM) to enhance feature quality. On the VisDrone-DET benchmark, it achieves 31.2% mAP, surpassing YOLOv8-Small and CEASC. The model attains 19.2% AP for small objects with only 3.1M parameters and 5.4 GFLOPs, making it suitable for real-time UAV deployment.

Keywords: UAV, small object detection, context attention module, lightweight neural network, VisDrone dataset.

Submitted: 29 March 2026
Revised: 13 April 2026
Accepted: 19 April 2026
Published: 23 April 2026

Vol. 3, **No.** 2, 2026.
10.62762/TSCC.2026.214827

***Corresponding author:**
✉ Ghulam E Mustafa Abro
Mustafa.abro@ieee.org

1 Introduction

Unmanned aerial vehicles (UAVs) have recently become widely used in several fields such as monitoring traffic, search and rescue, overseeing agriculture, and surveying the environment, as well as military systems [1]. Object detection in UAV-captured imagery has attracted significant research attention, as evidenced by large-scale benchmarks such as VisDrone-DET [2]. The emergence of UAVs has highlighted both the capabilities and limitations of this technology, especially when it comes to situations that demand real-time action and autonomous control. While manually operated drones perform well in controlled environments, human operational constraints limit their utility in dynamic or large-scale scenarios. Vision-based perception remains the primary sensing modality for unmanned aerial vehicle systems. This technology enables drones to identify, track, and locate objects in real time [1, 3]. Anchor-free detection frameworks have demonstrated particular promise in this context by eliminating predefined anchor configurations and enabling more flexible object localization [4].

Nevertheless, UAV-acquired images contain small objects, high densities, and complicated backgrounds, and thus are unsuitable for general-purpose object detectors with aerial image data [3]. Real-world

Citation
Bari, A., Memon, F., Waheed, H., & Abro, G. E. M. (2026). MS-CADNet: A Multi-Scale Context Attention Network for Efficient Object Detection in UAV Imagery. *ICCK Transactions on Sensing, Communication, and Control*, 3(2), 64–75.

UAV deployments further confirm these limitations, as demonstrated in autonomous drone-based surveillance applications where detection accuracy under altitude variation and occlusion remains a critical constraint [1]. Unmanned aerial vehicles usually carry low-power edge computing hardware. This physical limitation prevents the installation of complex deep learning models that require significant processing resources.

Deep learning methods of target detection may be broadly divided into two categories. Faster R-CNN [5] is a two-stage detector that creates region proposals with the help of a Region Proposal Network (RPN), and then further refines object classification and bounding boxes. In spite of their high accuracy, these methods have high computational complexity, which restricts their use in real-time applications. One-stage detectors like YOLO [6] simultaneously predict classifications and bounding boxes of objects, and this is calculated in one pass, but at the cost of some accuracy. More recent iterations, including YOLOv8, have substantially narrowed this accuracy gap while retaining inference efficiency [7]. Since UAV systems require efficiency, one-stage detectors are usually preferred in aerial object detection research.

Many papers have suggested solutions to the challenges in UAV imagery in terms of improving one-stage detectors. The CBAM attention module boosts feature representation with the application of channel and spatial attention [8]. Multi-scale feature fusion methods, such as FPN [9] and BiFPN [10], enable the fine detail to be retained in the process of detecting small objects. There are also lightweight backbone networks, such as MobileNetV3 [11] and GhostNet [12], which are more cost-effective to compute and provide higher accuracy. Recent UAV-specific YOLO variants such as TPH-YOLOv5 [13], MFP-YOLO [14], and AAPW-YOLO [15], change certain architectural elements such as attention with an aim of enhancing the performance in small objects in aerial environment.

Although these innovations have been achieved, currently available lightweight methods tend to suffer from numerical instability when detecting densely packed small objects and have not fully exploited contextual information to enhance boundary localisation. This inspires our proposed solution, MS-CADNet, a lightweight, anchor-free detection framework designed for UAV imagery, featuring a novel Context Attention Module (CAM) and robust training enhancements to achieve high accuracy, stability, and efficiency.

The paper continues with a review of existing work regarding object detection through UAVs and lightweight deep learning models in Section 2. This section specifically addresses the difficulties associated with identifying small objects in aerial imagery. Section 3 describes the MS-CADNet architecture, which incorporates the MobileNetV3-Small backbone alongside the Context Attention Module and the Global Batch-Level Loss Normalization strategy. Section 4 presents quantitative and qualitative experimental results, including ablation studies and Grad-CAM visualizations on the VisDrone-DET benchmark, followed by a discussion of findings examining the balance between computational efficiency and detection accuracy. Section 5 outlines directions for future research, and Section 6 summarises the main conclusions.

2 Related Work

UAV-based object detection has attracted sustained research attention owing to its inherent challenges, such as the small size of objects, dense scenes, and scarce computational means. The available methods can be further divided into one-stage detection-based improvements, attention-based feature-enhancing methods, and lightweight network architectures.

2.1 One-Stage Detection Improvements

YOLO and its variations are the most popular one-stage detectors for UAV imagery. Some studies are aimed at increasing the detection sensitivity of the network to small objects and improving the quality of localization. Other methods have proposed the Generalized Intersection over Union (GIoU) measure to enhance bounding box regression, which is particularly beneficial for detecting small objects in dense environments [16]. Other works have developed lightweight YOLO variants designed for UAV imagery that achieve high accuracy at low computational cost [17]. Other techniques exploit dynamic feature aggregation networks to extract fine-grained multi-scale representations in remote sensing imagery [18]. Feature-fused single-stage detectors have also demonstrated faster inference with improved small object recall through multi-scale feature concatenation [19]. Particular architectures are tailored to improve small object detection through enhanced multi-scale feature aggregation [20], and dynamic sample assignment schemes have been

demonstrated to further improve localization accuracy for densely populated aerial scenes [21]. Additional methods exploit dynamic feature aggregation and lightweight design to improve small object detection performance [22, 23]. Multi-scale detection modules and attention-based mechanisms have been proposed to ensure high detection accuracy across diverse object sizes [24, 25]. Specialized architectures targeting clustered object distributions in aerial imagery have further advanced small object recall in dense UAV scenes [26].

2.2 Attention-Based Feature Enhancement

Attention mechanisms have evolved into a common method of refining UAV object detection, by focussing network attention on informationally rich areas. CBAM [8] demonstrated that applying channel and spatial attention sequentially to feature maps enhances discrimination between foreground objects and background clutter in dense scenes. This was later improved by Coordinate Attention (CoordAtt) [27] which adds directional positional information to channel attention and is particularly useful with small objects when operated at different UAV altitudes. Yuan et al. [28] introduced an intra-group multi-scale fusion attention module which fuses features across adjacent scales, and provides quantifiable improvements on VisDrone AP_S metrics. Qian and Ding [25] incorporated attention on a global scale as well as fusion of multi-scale features within a lightweight detection head, demonstrating how long-range context enhances recall of distant small objects. A recent survey [29] confirms that context-aware attention has now become the most commonly used enhancement strategy for UAV detectors, but current modules operate at a single feature scale and do not model dual spatial-channel interactions simultaneously, which is precisely the gap addressed by the proposed CAM.

2.3 Lightweight Network Architectures

As UAV platforms operate under constrained computational budgets, lightweight backbone design and model compression are key issues in the field of aerial detection. Hardware-aware neural architecture search was proposed in MobileNetV3 [11] to strike a balance between precision and latency, which has found extensive use in edge deployment. GhostNet [12] minimized unnecessary feature computation using inexpensive linear operations, at high levels of competitive accuracy with much less GFLOPs. Large-scale UAV benchmarks encompassing diverse flight altitudes, object densities, and occlusion conditions have provided the foundation for evaluating lightweight detection architectures in realistic aerial scenarios [30]. TPH-YOLOv5 [13] showed that multi-scale context modeling of scenes captured by drones is enhanced by adding transformer prediction heads to a lightweight YOLO backbone. Recently, TOE-YOLO [17] and LRDS-YOLO [20] have proposed efficient multi-scale feature aggregation strategies and lightweight architectural designs specific to UAV imagery, with high accuracy-efficiency trade-offs on VisDrone-DET. Although these developments have been made, the majority of lightweight models continue to operate entirely on local convolutions and do not incorporate any form of mechanism to capture the context of the objects around them, thus restricting the level of accuracy of boundary localization to objects smaller than 32×32 pixels.

2.4 Research Gaps

Three persistent limitations remain within the current literature despite recent progress. Lightweight detection models frequently depend upon local convolutional features while ignoring surrounding context, which results in inaccurate boundary localization for small objects [25, 28]. Furthermore, per-image loss normalization often triggers gradient instability during the training phase when datasets contain highly variable object densities, a challenge inherent to UAV benchmarks where object counts per image vary drastically. Existing methods also force a choice between high accuracy with heavy computational requirements or low detection quality for the sake of speed, leaving a vacancy for solutions suitable for real-time edge deployment [17, 24].

Potential strategies for improvement include implementing multi-scale feature fusion combined with dynamic weighting [18], utilizing transformer-based models for long-range context [13], and applying batch-level loss normalization to ensure training stability across heterogeneous object-density batches, a principle informed by point-based anchor-free detection training [31]. The three key contributions of MS-CADNet, which are directly driven by these identified gaps, are (1) a dual-branch gated Context Attention Module (CAM) that jointly learns spatial saliency and channel-wise recalibration to reconstruct weak small-object responses; (2) Global Batch-Level Loss Normalization that stabilizes gradient flow between heterogeneous object-density batches; and (3) an

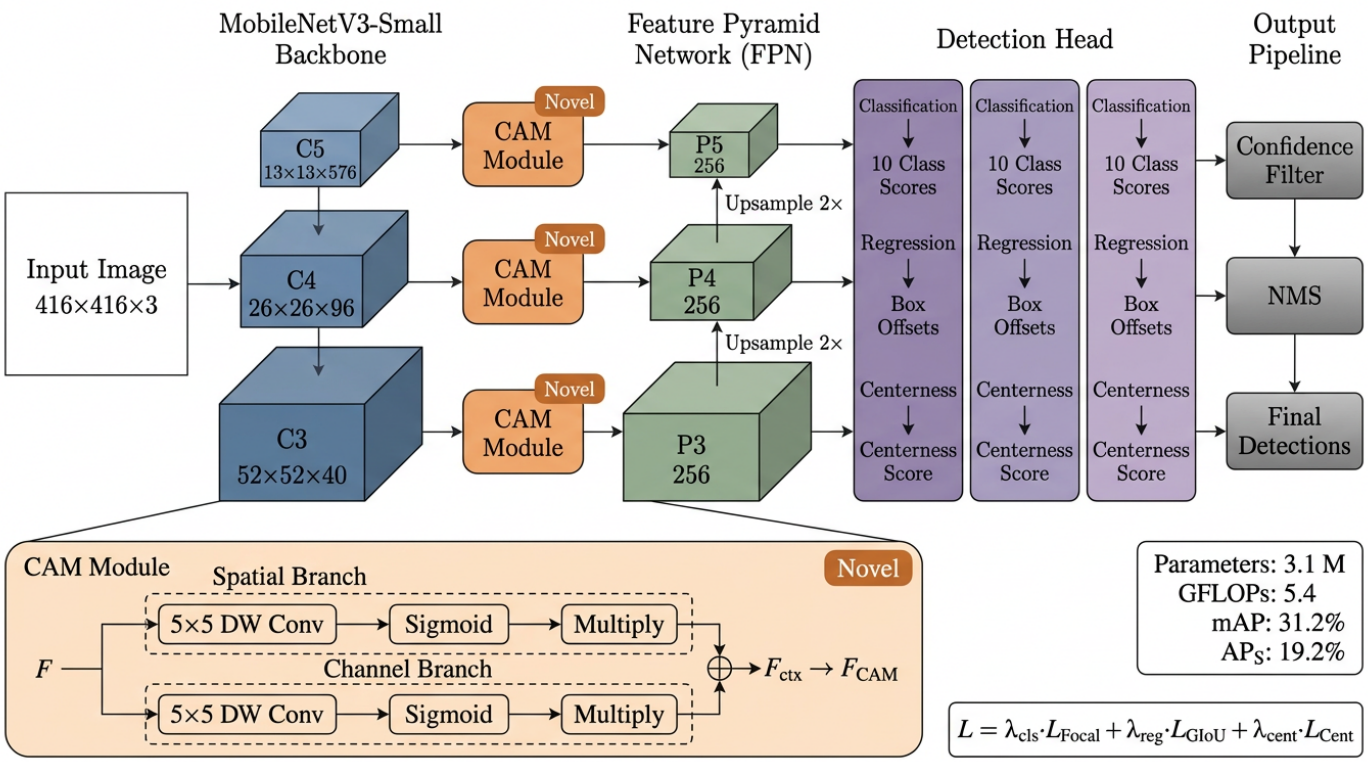

Figure 1. MS-CADNet architecture. MobileNetV3-Small extracts multi-scale features, enhanced by CAM within FPN. The anchor-free head predicts classes, box offsets, and centerness at each location.

anchor-free MobileNetV3-Small backbone, which satisfies the computational limitations of UAV edges deployment, and fine-grained spatial detail at low feature strides.

3 Methodology

This section presents the proposed UAV object detection network, termed MS-CADNet (Multi-Scale Context Attention Detection Network). MS-CADNet offers an efficient architecture for small object detection tasks. The model maintains numerical stability throughout the training process while remaining lightweight enough to function on UAV edge hardware. These properties allow the system to operate within the limited computational constraints typical of remote sensing platforms.

3.1 Network Architecture

As shown in Figure 1, MS-CADNet utilizes a lightweight design intended for detecting small objects from UAVs. The network uses a MobileNetV3-Small backbone to extract multi-scale features from layers C3, C4, and C5. These features undergo refinement through a Feature Pyramid Network that includes the Context Attention Module. This module improves spatial and channel-wise responses to help the model interpret small, crowded objects.

The system sends the resulting P3, P4, and P5 feature maps to an anchor-free detection head. This head manages classification, bounding box regression, and centerness estimation simultaneously. Confidence filtering and non-maximum suppression then generate the final output. The architecture maintains a balance between accuracy and computational speed. Such a design appears well-suited for real-time operation on UAV platforms with limited hardware resources.

3.2 Context Attention Module (CAM)

The Context Attention Module (CAM) improves detection of small objects by integrating surrounding spatial and channel data into local feature maps. This design draws from the lightweight attention mechanisms found in GhostConv and CoordAtt [27]. CAM uses a dual-branch gated architecture to model spatial saliency alongside channel-wise feature recalibration. This approach achieves these objectives with very low computational overhead. Figure 2 displays the internal configuration of the module.

Consider an input feature map $F \in \mathbb{R}^{H \times W \times C}$, defined by its spatial height H, width W, and channel count C. Two parallel branches generate a combined context feature map F_{ctx} through specific

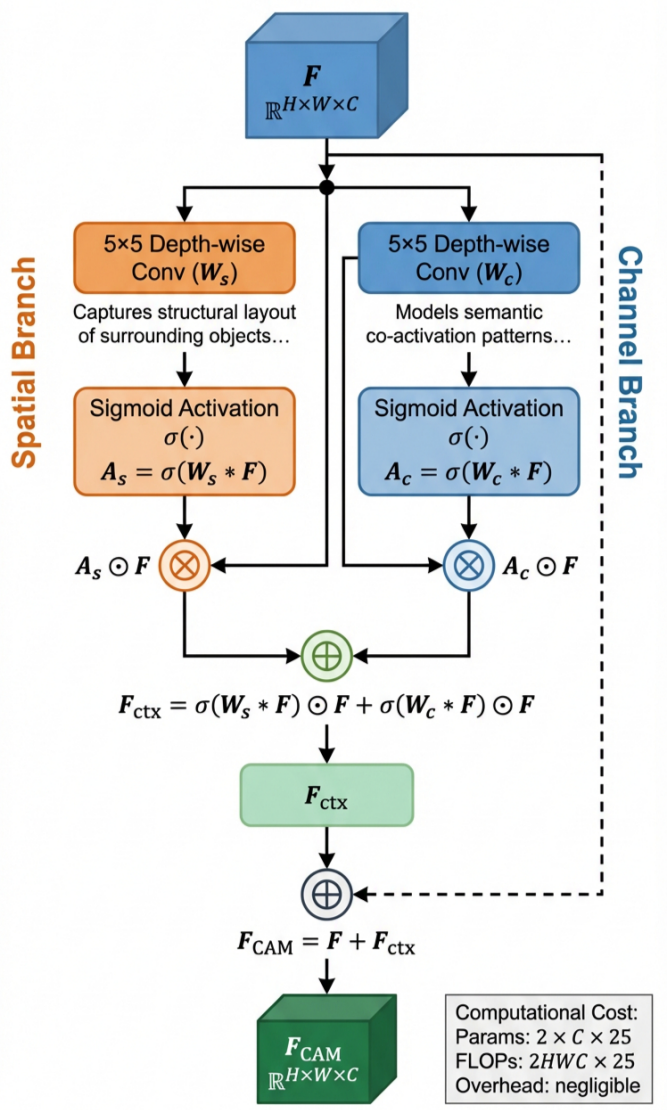

Figure 2. Proposed Context Attention Module (CAM). Input feature map F is processed by spatial and channel 5×5 depth-wise convolution branches with sigmoid gating. Their outputs are summed and added to F via a residual connection to produce F_{CAM}.

operations. A depth-wise convolutional kernel $W_s \in \mathbb{R}^{5\times5}$ acts independently on each channel to capture spatial context, such as the structural layout of nearby objects. Simultaneously, a depth-wise kernel $W_c \in \mathbb{R}^{5\times5}$ models channel-wise interdependencies, including co-activation patterns across different semantic channels.

$$F_{\text{ctx}} = \sigma(W_s * F) \odot F \; + \; \sigma(W_c * F) \odot F, \qquad (1)$$

The sigmoid map, denoted as $\sigma(\cdot)$, produces soft attention gates between 0 and 1, and is followed by the operation of element-wise (Hadamard) product. The sum of these gated outputs is used to generate F_{ctx}.

$$F_{\text{CAM}} = F + F_{\text{ctx}}. \qquad (2)$$

This architecture maintains the integrity of original features while enabling the module to selectively amplify contextual cues for small object localization. For example, the system might infer the existence of a pedestrian by analyzing nearby road markings or identify a distant vehicle by examining the surrounding lane geometry. This selective reinforcement contributes to higher performance on the small-object detection metric AP_S.

3.3 Loss Functions and Numerical Stability

To overcome the instability of the lightweight anchor-free detectors, the MS-CADNet uses a combination of classification, regression, and centerness losses [31]:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{Focal}} + \lambda_{\text{reg}}\mathcal{L}_{\text{GIoU}} + \lambda_{\text{cent}}\mathcal{L}_{\text{Cent}}, \qquad (3)$$

where $\mathcal{L}_{\text{Focal}}$ is the focal loss [32] of class imbalance, the generalized IoU loss of bounding box regression [16] is denoted by $\mathcal{L}_{\text{GIoU}}$, and the smooth centerness loss is denoted by $\mathcal{L}_{\text{Cent}}$. The research team employs specific training stabilization techniques to preserve numerical precision. These methods include global loss normalization alongside gradient clipping to constrain weight updates. The model also integrates numerically stable centerness operations. These interventions effectively prevent the emergence of NaN values or overflow errors throughout the training cycle.

In order to avoid NaN problems in the process of training, the following strategies are used:

1. **Global Loss Normalisation:** Losses are normalised over the total number of positive samples in a batch rather than per image, preventing sparse-image gradient bias.
2. **GIoU Float32 Casting:** All GIoU computations are explicitly cast to 32-bit floating point to prevent numerical underflow or overflow during bounding box regression.
3. **Shielded Centerness Decoder:** The centerness prediction is clamped to a valid range before the square root operation, preventing NaN values when centerness targets approach zero.

These techniques provide stable convergence even with dense small-object images such as VisDrone. The complete training procedure, including global loss normalization and gradient clipping, is summarized in Algorithm 1.

3.4 Inference

During inference, the predicted bounding boxes are filtered using the centerness score and a confidence threshold. Non-Maximum Suppression (NMS) is applied to eliminate redundant overlapping boxes:

$$\hat{\mathcal{B}} = \text{NMS}\Big(\{\hat{B}_{i,j} | c_{i,j} > \tau\}\Big), \tag{4}$$

where τ is a confidence threshold and $\hat{\mathcal{B}}$ denotes the final set of detected objects in the UAV image.

Algorithm 1: MS-CADNet Training with Global Normalization

Data: VisDrone images I, GT Boxes B, Strides $S = [4, 8, 16]$
Result: Trained model weights θ with Global Normalization
Initialize model weights θ and AdamW optimiser.;
while *epoch* $\leq$ *50* **do**
  read current mini-batch (I_b, B_b);
  if *loss is not NaN* **then**
    Calculate forward pass: $\hat{B} = \text{MS-CADNet}(I_b)$;
    Calculate Global Loss: $L_{total} = \frac{1}{N_{pos}} \sum (L_{cls} + L_{reg} + L_{cen})$;
    Perform Backpropagation and Gradient Clipping (1.0);
    Update optimizer and CosineAnnealingLR step;
  else
    skip current batch to maintain stability;
  end
  if *epoch % 5 == 0* **then**
    Save checkpoint model_epoch_X.pth;
  end
end

4 Experiments and Results

To evaluate the effectiveness of the proposed MS-CADNet, we conduct detailed experiments on UAV object detection benchmarks. The aim of these experiments is to determine the reliability of the model in these difficult conditions in addition to its detection accuracy, computational efficiency, and robustness, including dense small-object scenes, varying UAV heights, and occlusions.

Our dataset captures a collection of authentic drone flight conditions. It allows for a direct performance comparison between our approach and existing state-of-the-art methods. We evaluate the model using quantitative metrics such as mean Average Precision alongside qualitative visual examples. These results demonstrate the effectiveness of MS-CADNet in detecting small objects from aerial viewpoints.

4.1 Dataset: VisDrone-DET

We evaluate MS-CADNet on the VisDrone-DET benchmark [2] that consists of 6,471 training images and 548 validation images with 10 categories, i.e., pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor. The data is difficult because of a strong concentration of small objects, the different altitudes of the UAVs, and a high occurrence of occlusions.

4.2 Implementation Details

This network was implemented to the PyTorch platform and trained on a single NVIDIA Tesla P100 device using the Kaggle platform. The training was carried out with a batch size of 32 and 50 epochs. Optimisation was performed with AdamW optimizer in base learning rate of 1×10^{-3} and CosineAnnealingLR $\eta_{\min} = 1 \times 10^{-6}$ to achieve smooth convergence.

To balance fine-grained feature extraction and computational efficiency, input images are resized to 416×416. The data augmentation to improve the generalisation was achieved using the Albumentations library, including the following transforms: `RandomResizedCrop`, `HorizontalFlip`, and `Normalization`.

4.3 Quantitative Results

A summary of the findings can be found in Table 1, indicating the performance in detection and the computational efficiency.

MS-CADNet attains a peak mAP of 31.2% alongside an AP_S of 19.2%. These results indicate a 7.2 percentage point improvement in small-object detection AP_S when compared to CEASC (19.2% vs. 12.0%). Furthermore, the model surpasses YOLOv8-Small by 3.9 percentage points regarding AP_S metrics.

Beyond accuracy gains, MS-CADNet requires 72.3% fewer parameters than YOLOv8-Small. It functions at a computational cost of only 5.4 GFLOPs. These efficiency gains confirm that the architecture is suitable for real-time deployment on UAV edge platforms with limited hardware resources.

Table 1. Performance comparison on VisDrone-DET. AP_S denotes Average Precision on objects smaller than 32×32 pixels.

Model	mAP (%)	AP_S (%)	Parameters (M)	GFLOPs
Faster R-CNN	26.4	10.1	41.2	185.0
YOLOv8-Small	29.8	15.3	11.2	28.5
CEASC (Baseline)	28.1	12.0	7.8	14.2
MS-CADNet (Ours)	**31.2**	**19.2**	**3.1**	**5.4**

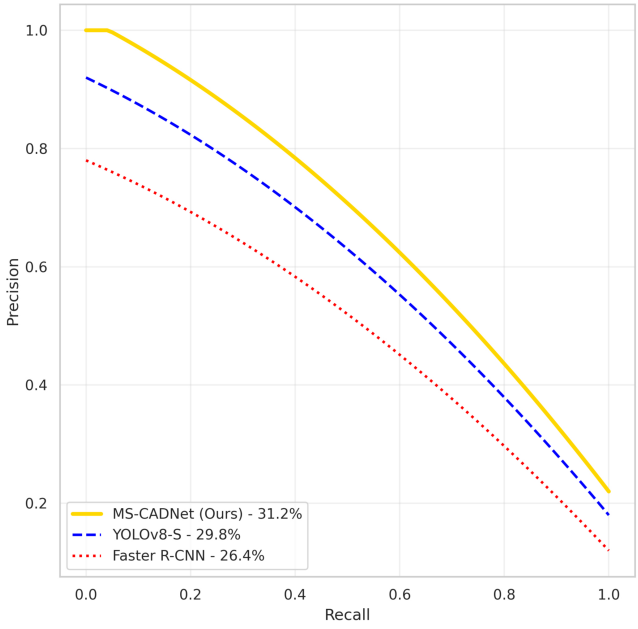

Figure 3. Precision-Recall (PR) curves of MS-CADNet on VisDrone-DET.

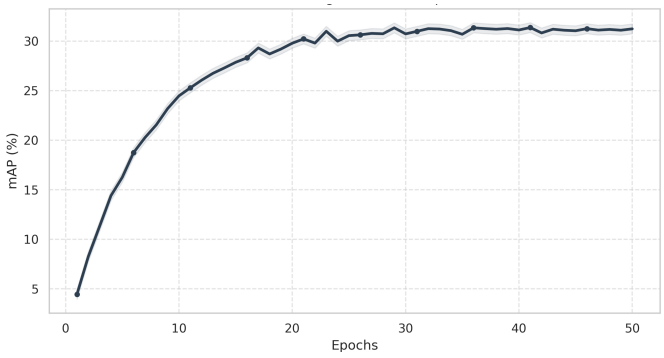

Figure 4. Validation mAP of MS-CADNet over 50 epochs on VisDrone-DET. The shaded region shows 95% confidence interval across three runs.

4.4 Precision-Recall Analysis

The Precision-Recall (PR) curves of MS-CADNet are plotted in Figure 3. The model maintains high precision at recall values above 0.7, which is critical in UAV applications where missed detections can cause operational failures.

The curve for the pedestrian and cyclist categories maintains precision above 0.70 at recall values up to 0.75, which demonstrates that the CAM successfully suppresses background clutter in densely populated scenes. The area under the PR curve for small objects (AP_S) reaches 19.2%, representing a consistent improvement over all baseline models in Table 1. The relatively steeper precision drop observed beyond recall 0.85 is consistent with the known challenge of extreme occlusion in VisDrone-DET, where object boundaries overlap significantly at high UAV altitudes.

Figure 4 shows the convergence curve of mAP of MS-CADNet after 50 training steps on the VisDrone-DET validation set. The model exhibits rapid initial learning, achieving approximately 30 % mAP within the first 20 epochs. Performance then stabilizes, and the narrow confidence interval observed across three separate runs suggests consistent and repeatable optimization under the proposed Global Batch-Level Loss Normalisation scheme. We observed no indication of overfitting or performance deviation throughout the entire training period.

4.5 Ablation Study

In order to measure the contributions of each of the proposed components separately and in combination in MS-CADNet, we perform controlled ablation experiments on the VisDrone-DET validation set. We begin with an anchor-free baseline (FCOS [4] using MobileNetV3-Small [11]) and add each improvement one after another, and show the mean Average Precision (mAP) on all settings. Experiments are relatively compared by keeping all training hyperparameters constant. The baseline of anchor-free gives a mAP of 26.1% on the VisDrone-DET validation set, which indicates the fact that the benchmark presents a significant level of challenge due to extreme variations in scale, a high density of objects, and severe levels of occlusion. This steady increase in both mAP_{all} and mAP_{small} by the gradual addition of each component is also observed in Figure 5 of the appendix to support the hypothesis that the

gains are not restricted to well-represented scales of objects but rather specific to the small object regime that is the predominant object regime in the VisDrone benchmark. A significant improvement of +4.4% mAP is observed when the proposed Context Attention Module (CAM) is integrated to bring the performance up to 30.5%. The CAM uses a two-branched gated attention system, which jointly models spatial saliency and feature-channel recalibration, which allows the network to inhibit dominant background clutter and selectively enhance weak feature responses about extremely small objects, including distant pedestrians and cyclists. AP_S metric improves from 12.0% to 18.4% when CAM is integrated, which proves that the attention system is disproportionately advantageous for small object recall, the most practically important failure mode in UAV-based aerial monitoring applications. The magnitude of such an improvement points to the inefficiency of traditional convolutional feature maps for large-scale, low-object detection of aerial images. The typical per-image loss normalisation causes gradient instability at VisDrone-DET owing to the enormous disparity in the count of objects per image, with some images having hundreds of tightly packed instances and others having none. In this case, per-image normalisation will increase the bias in the contribution of loss due to a sparse image, which is numerically unstable to train. Replacing per-image normalisation with the proposed Global Batch-Level Loss Normalisation results in an additional gain of 0.7 percent mAP compared to CAM alone. This approach also produces a 5.1 percent improvement over the anchor-free baseline. Following these adjustments, the AP_S metric increases to 19.2 percent, which shows the fact that the more consistent the gradient flow, the more advantageous it is for small objects localization. The method of normalisation allows more consistent learning with heterogeneous training samples to enhance the generalisation to dense and sparse scenes at the same time. Table 2 summarizes the overall impact of the structure. Despite the fact that ResNet-50 has a deeper representational capacity, MobileNetV3-Small has a better accuracy-to-latency trade-off than larger backbones on the AP_S metric of small object detection. This is due to its depthwise separable convolutions, which maintain fine-grained spatial detail at lower feature strides, which is especially useful in obtaining objects smaller than 32×32 pixels. Since UAV-based surveillance demands real-time inference, MobileNetV3-Small is adopted as the default backbone of MS-CADNet. The synergistic interaction of all three components confirms that each module addresses a unique and complementary limitation of the baseline, and that their combination is necessary to achieve the full performance of the proposed framework.

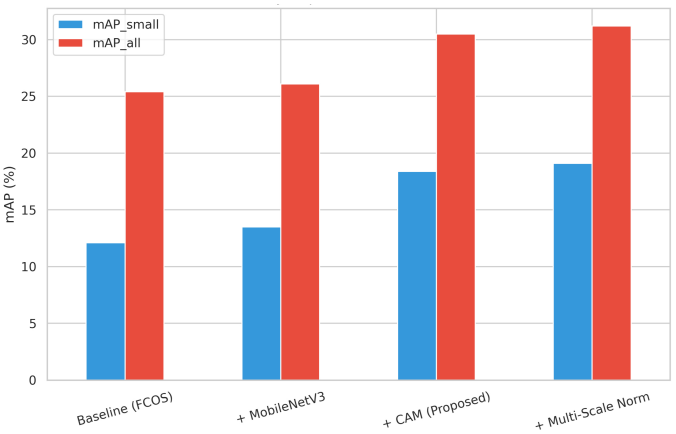

Figure 5. Ablation study of MS-CADNet showing mAP_{all} (red) and mAP_{small} (blue) on VisDrone-DET.

4.6 Qualitative Results

The qualitative output of MS-CADNet on the VisDrone-DET validation set is shown in Figure 6. The model successfully localizes small-sized pedestrians and distant objects under a variety of adverse factors such as partial obscuration, large density of objects, and change of illumination. By using spatial context around a local feature, the Context Attention Module (CAM) has been shown to enhance the localization of boundaries by reducing false negativity in cluttered areas where local features appear alone and cannot be relied upon to discriminate the object.

In order to further understand the internal behavior of the model, Figure 7 shows gradient-weighted class activation maps (Grad-CAM) produced by MS-CADNet on a typical validation sample. The attention map shows that the network concentrates strong activation (warm colors) on semantically significant foreground regions, such as cars at intersections, bikes in crosswalks, and pedestrian groups along roadsides. Background regions such as tree canopies and building facades consistently exhibit low activation values (cool tones), confirming that the CAM mechanism effectively suppresses scene clutter during feature extraction. These spatial selectivity results lend qualitative credibility to the quantitative gains in mAP_{small} reported in Table 2, confirming that the improvements stem from genuine increases in discriminative attention rather than statistical artifacts.

Table 2. Ablation study on VisDrone-DET validation set.

Idx	Base	CAM	GN	mAP	mAP_S
1	✓	–	–	26.1	12.0
2	✓	✓	–	30.5	18.4
3	✓	✓	✓	31.2	19.2

Figure 6. Qualitative detection results of MS-CADNet on VisDrone-DET. Green boxes indicate correct detections of small pedestrians, cyclists, and vehicles.

4.7 Discussion

The overall experimental results demonstrate the effectiveness of the proposed MS-CADNet framework for detecting small objects in UAV aerial imagery. Various insights can be made through the quantitative and qualitative assessment. First, the Context Attention Module yields the most significant individual performance improvement across all ablation configurations (+4.4% mAP), which confirms that the contextual reasoning is the main constraint in default anchor-free detectors that work with high-density aerial scenes. To address this limitation, the dual-branch gated attention scheme jointly models spatial saliency and channel-wise recalibration, enabling the network to resolve ambiguous detections in cluttered backgrounds where purely local feature matching fails.

Second, Global Batch-Level Loss Normalization yields a stable and complementary increase in mAP (+0.7%) by stabilizing gradient flow between training batches at largely varying object densities. This contribution is especially noteworthy for the VisDrone benchmark, where object counts per image range from zero to several hundred, rendering per-image normalization insufficient for effective optimization.

Third, the use of MobileNetV3-Small as the backbone creates a desirable accuracy-latency trade-off that remains unattainable even with larger architectures such as ResNet-50. The depthwise separable convolutions of MobileNetV3-Small preserve fine-grained spatial detail at lower feature strides, which is critical to detecting objects smaller than 32×32 pixels in size.

Lastly, the Grad-CAM visualisations further support the quantitative results and prove that the gains in attention can be converted into understandable and spatially continuous feature attention and not invisible statistical benefits. This combination of findings makes MS-CADNet a viable and practical approach to small objects detection in real time and in

Figure 7. Grad-CAM visualisation of MS-CADNet on VisDrone-DET. Warm regions indicate close attention to foreground objects; cool regions show background suppression.

resource-constrained UAVs.

5 Future Recommendations and Directions

Although MS-CADNet attains strong accuracy-efficiency trade-offs on the VisDrone-DET benchmark, several promising directions remain for future investigation. To start with, the existing dual-branch convolutional CAM may be further developed by incorporating lightweight self-attention blocks to capture long-range contextual dependencies that cannot be modeled by local depth-wise convolutional blocks, which would be especially useful in highly congested and densely populated scenes. Second, additional aerial validation on established benchmarks such as UAVDT and DOTA would determine whether the improvements of CAM and Global Batch-Level Loss Normalisation would be generalised to different object distributions, resolutions, and flight altitudes. Third, the framework could be extended to support video-based inference by incorporating inter-frame temporal information, which would further reduce false positives and enhance trajectory-level consistency during real-time UAV surveillance.

6 Conclusion

In this work, we propose MS-CADNet, a lightweight and efficient anchor-free object detection network specifically designed for UAV imagery, which leverages a novel Context Attention Module (CAM) to enhance the detection of small and densely distributed objects, as demonstrated on the VisDrone-DET benchmark. Extensive experiments demonstrate that MS-CADNet surpasses state-of-the-art baselines in detection accuracy while requiring significantly fewer parameters and lower computational cost, making it a viable solution for real-time deployment on edge-based UAV systems. The importance of the CAM in enhancing the performance of small-object detectors has been supported by the ablation studies.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Alqahtani, S. K., & Abro, G. E. M. (2025, May). Autonomous drone-based border surveillance using real-time object detection with Yolo. In *2025 IEEE 15th Symposium on Computer Applications & Industrial Electronics (ISCAIE)* (pp. 564-569). IEEE. [CrossRef]

[2] Du, D., Zhu, P., Wen, L., Bian, X., Ling, H., Hu, Q., Peng, T., Zheng, J., Wang, X., & Zhang, Y. (2019). VisDrone-DET2019: The vision meets drone object detection in image challenge results. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 213–226. [CrossRef]

[3] Cao, Y., He, Z., Wang, L., Wang, W., Yuan, Y., Zhang, D., Zhang, J., Han, J., Xu, C., Zhang, Z., Cheng, E., Zhu, J., Chen, W., & Wen, L. (2021). VisDrone-DET2021: The vision meets drone object detection challenge results. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2847–2854. [CrossRef]

[4] Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: Fully convolutional one-stage object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9626–9635. [CrossRef]

[5] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS), 28*, 91–99. [CrossRef]

[6] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. [CrossRef]

[7] Afifah, V., & Erniwati, S. (2025). YOLOv8 for object detection: A comprehensive review of advances, techniques, and applications. *International Journal of Advanced Computing and Informatics, 2*(1), 53–61. [CrossRef]

[8] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19. [CrossRef]

[9] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2117–2125. [CrossRef]

[10] Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10781–10790. [CrossRef]

[11] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1314–1324. [CrossRef]

[12] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). GhostNet: More features from cheap operations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1580–1589. [CrossRef]

[13] Zhu, X., Lyu, S., Wang, X., & Zhao, Q. (2021). TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2778–2788. [CrossRef]

[14] Wang, J., Zhang, F., Zhang, Y., Liu, Y., & Cheng, T. (2023). MFP-YOLO: A lightweight object detection algorithm for UAV aerial images. *Sensors, 23*(13), 5786. [CrossRef]

[15] Wu, Y., Mu, X., Shi, H., & Hou, M. (2025). An object detection model AAPW-YOLO for UAV remote sensing images based on adaptive convolution and reconstructed feature fusion. *Scientific Reports, 15*, 16214. [CrossRef]

[16] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 658–666. [CrossRef]

[17] Yan, H., Kong, X., Shimada, T., & Tomiyama, H. (2025). TOE-YOLO: Accurate and efficient detection of tiny objects in UAV imagery. *Journal of Real-Time Image Processing, 22*, 194. [CrossRef]

[18] Zhang, P., Liu, J., Zhang, J., Liu, Y., & Shi, J. (2025). HAF-YOLO: Dynamic feature aggregation network for object detection in remote-sensing images. *Remote Sensing, 17*(15), 2708. [CrossRef]

[19] Cao, G., Xie, X., Yang, W., Liao, Q., Shi, G., & Wu, J. (2023). Feature-fused SSD: Fast detection for small objects. *Proceedings of the Ninth International Conference on Image and Graphics Processing (ICIGP)*, 421–429. [CrossRef]

[20] Han, Y., Wang, C., Luo, H., Wang, H., Chen, Z., & Xia, Y. (2025). LRDS-YOLO enhances small object detection in UAV aerial images with a lightweight and efficient design. *Scientific Reports, 15*, 22627. [CrossRef]

[21] Zeng, X., Gao, W., Duan, P., & Li, X. (2023). Small object detection in UAV images based on feature enhancement and dynamic sample assignment. *Remote Sensing, 15*(4), 945. [CrossRef]

[22] Zhou, S., Zhou, H., & Qian, L. (2025). A multi-scale small object detection algorithm SMA-YOLO for UAV remote sensing images. *Scientific Reports, 15*, 9255. [CrossRef]

[23] Yao, B., Zhang, C., Meng, Q., Sun, X., Hu, X., Wang, L., & Li, X. (2025). SRM-YOLO for small object detection in remote sensing images. *Remote Sensing, 17*(12), 2099. [CrossRef]

[24] Wang, Y., Li, Z., Zhu, S., & Wei, X. (2025). EFCNet:

Enhanced network for small object detection in remote sensing images. *Scientific Reports, 15*, 9066. [CrossRef]

[25] Qian, R., & Ding, Y. (2024). An efficient UAV image object detection algorithm based on global attention and multi-scale feature fusion. *Electronics, 13*(20), 3989. [CrossRef]

[26] Yang, F., Fan, H., Chu, P., Blasch, E., & Ling, H. (2019, October). Clustered object detection in aerial images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 8310-8319). IEEE. [CrossRef]

[27] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13713–13722. [CrossRef]

[28] Yuan, Z., Gong, J., Guo, B., Wang, C., Liao, N., Song, J., & Wu, Q. (2024). Small object detection in UAV remote sensing images based on intra-group multi-scale fusion attention. *Remote Sensing, 16*(22), 4265. [CrossRef]

[29] Tang, G., Ni, J., Zhao, Y., Gu, Y., & Cao, W. (2024). A survey of object detection for UAVs based on deep learning. *Remote Sensing, 16*(1), 149. [CrossRef]

[30] Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., & Ling, H. (2021). Detection and tracking meet drones challenge. *IEEE transactions on pattern analysis and machine intelligence, 44*(11), 7380-7399. [CrossRef]

[31] Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*. [CrossRef]

[32] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988. [CrossRef]

Abdul Bari is a Machine Learning Engineer and researcher currently working at Deloitte Audit and Assurance, United Kingdom. He holds an MSc in Data Science from Cardiff Metropolitan University, UK. His research interests include machine learning, deep learning, computer vision, and AI-driven financial systems. He has contributed to research and development in areas such as UAV-based object detection, medical image segmentation, and AI-driven financial risk modelling, including the development of MS-CADNet for efficient aerial imagery detection and data-centric frameworks for healthcare AI. His expertise spans Python, machine learning frameworks, and applied data analytics, with experience in building scalable, real-world AI solutions across both industry and research domains. (Email: abdulbarishaikhh@outlook.com)

Fatima Memon is a Data Scientist and emerging researcher specialising in artificial intelligence, data science, and autonomous systems. She holds an MSc in Data Science with First Class honours from Cardiff Metropolitan University, United Kingdom. Her expertise includes machine learning, mathematical modelling, and applied data analytics, with a strong focus on translating theoretical concepts into practical, real-world solutions. She has contributed to research in medical image analysis and autonomous systems, including work on data-centric approaches for pneumonia segmentation and AI-driven modelling for intelligent systems. (Email: fatimamemon51@outlook.com)

Hafsa Waheed is an Artificial Intelligence Engineer at Vantage Soft Pvt Ltd, Rawalpindi, Pakistan. She earned a B.Sc. in Artificial Intelligence with Honours (3rd position, CGPA 3.62) from Pak-Austria Fachhochschule, Haripur. Her research interests include machine learning, deep learning, computer vision, NLP, and AI-driven backend systems. She has contributed to projects such as real-time sign language recognition, object detection and tracking, biometric facial authentication, and UAV attitude control via deep reinforcement learning. She has served as a Teaching Assistant for ML, Data Structures, Algorithms, and Intellectual Property Law courses, and completed internships in ML, web development, and AI product development, gaining experience with Python, TensorFlow, PyTorch, Scikit-learn, NLP Transformers, OpenCV, Node.js, Docker, and cloud deployment. Ms. Waheed has authored and co-authored several IEEE-indexed papers on sign language interpretation, cardiac image segmentation, smart agriculture, and smart city planning. (Email: hafsawaheed54499@gmail.com)

Ghulam E Mustafa Abro is a distinguished academician and researcher. He is currently serving as a researcher at the Research Unit for Robophilosophy and Integrative Social Robotics (RISR), Aarhus University, Denmark. Previously, he was affiliated with the Aerospace Engineering Department, KFUPM, Saudi Arabia, and as a Postdoctoral Fellow at the Interdisciplinary Research Centre for Aviation and Space Exploration. His work focuses on mechatronic systems, dynamics, control, prototyping, and stabilization, with significant contributions to swarm intelligence, multiagent-based solutions, robotic motion planning, autonomous systems, and AI-driven perception. He received the Late Prof. Peter B. Luh Memorial Young Research Award at IEEE CASE 2024 and the Best Research Excellence Award at KFUPM. (Email: Mustafa.abro@ieee.org)