International Core Journal of Engineering Volume 11 Issue 7, 2025
ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0001

Enhancing Small Object Detection from UAV Perspectives via an Improved YOLOv11 Model

Yue Hong[1], Byung-Won Min[2], Shentao Wang[1], Yuxiao Hu[1]

[1] College of Yonyou Digital & Intelligence, Nantong Institute of Technology, Nantong 226000, China

[2] Division of Information and Communication Convergence Engineering, Mokwon University, Daejeon 35242, Korea

Abstract

Object detection in UAV aerial imagery faces significant challenges, including a high proportion of small objects, large viewpoint variations, severe occlusions in dense scenes, and difficulties in low-light or nighttime conditions. To address these issues, this paper proposes an improved version of YOLOv11 by incorporating a Partial Convolution (PConv) mechanism, which enhances the model's sensitivity to edge regions and sparse salient targets. Additionally, a Spatial and Channel Self-Attention (SCSA) module is introduced to improve feature focusing by jointly modeling spatial and channel-wise dependencies. Together, these enhancements form the YOLOv11-PCSCSA model, which achieves a lightweight structure and precise perception capabilities. Experiments conducted on two benchmark UAV datasets, HIT-UAV and VisDrone2019, demonstrate that the proposed model outperforms the original YOLOv11 in terms of both mAP@0.5 and mAP@0.95 metrics.

Keywords

YOLOv11; Object Detection; UAV Perspective; Aerial Imagery; Small Object Detection.

1. Introduction

UAV (Unmanned Aerial Vehicle) is an unmanned aerial vehicle. With the advancement of the "low-altitude economy" policy, the application of drones in real life is becoming more and more perfect [1]. With the continuous advancement of drone technology, it is increasingly being used in logistics distribution, emergency rescue, public safety, agricultural monitoring and other industries with its unique aerial vision and remote control capabilities. From the initial service of military reconnaissance to the current widespread use in civil scenarios such as urban management, disaster relief, crop monitoring and package delivery [2-4]. The application scope of drones continues to expand. Its rapid popularization is due to its ability to complete complex or dangerous aerial operations without the presence of personnel, which not only improves operational efficiency but also enhances safety, making it an important tool for coping with harsh environments and special space missions. The rapid development of artificial intelligence algorithms such as image processing and target detection has reduced the cost of obtaining image data and provided an important driving force for the application of target detection in drone systems. Image acquisition and target detection based on drone platforms have become a key technical path to promote the implementation of unmanned perception systems.

Unlike conventional ground-based surveillance systems, UAV-acquired imagery presents a range of unique challenges due to the dynamic nature of aerial data collection. First, because UAVs typically operate at higher altitudes, the objects captured in images are often small in size, making feature

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0001

extraction more difficult. This limitation significantly hampers the performance of current object detection algorithms in identifying small targets, ultimately affecting the overall effectiveness and operational efficiency of UAV-based applications. Second, variations in flight attitude and viewing angle lead to frequent changes in object appearance. Although UAV imagery often benefits from wide perspectives and rich scene information, this also introduces considerable background noise and irrelevant data, increasing computational complexity and reducing detection accuracy. The VisDrone2019 dataset used in this study exemplifies these challenges, containing images characterized by wide-angle views, occlusion, and clutter[5].

In addition, dense object distribution is common in UAV scenes, particularly in crowded environments or under adverse weather conditions, where overlapping targets, occlusion, rain, and fog contribute further to detection difficulty. Moreover, object detection using single-modality visible light images performs poorly under low-light or nighttime conditions, posing a significant barrier to achieving all-day, all-weather surveillance. To address this issue, infrared (IR) sensing has emerged as a vital complement. As a passive imaging modality, infrared can capture thermal radiation, reflecting the temperature distribution on object surfaces. Although IR images generally have lower contrast than RGB images, they exhibit strong penetration and interference resistance, maintaining robust detection performance under challenging conditions such as darkness, occlusion, or adverse weather[6].

Given these challenges-including small object detection difficulties, significant viewpoint variation, strong environmental interference, and poor nighttime detection capability-developing object detection algorithms that balance accuracy, speed, and model efficiency, while remaining deployable in real-world UAV applications, remains a pressing research problem. To this end, this paper utilizes the HIT-UAV, and VisDrone2019 datasets and proposes an all-weather UAV object detection framework based on dual-modality inputs. The system leverages visible light imagery to achieve high-precision detection in complex daytime environments and integrates infrared imagery to enhance target perception under low-light or nighttime conditions, ultimately enabling intelligent detection across full time periods and diverse scenarios.

Building upon the above analysis, this study presents an enhanced object detection model based on the YOLOv11s framework. The proposed network incorporates a Partial Convolution (PConv) mechanism and a Spatial-Channel Self-Attention (SCSA) module to improve detection accuracy, particularly for small and sparsely distributed objects. The primary contributions of this work are as follows:

A novel feature extraction module based on partial convolution (PConv) is introduced to strengthen the model's sensitivity to edge regions and isolated salient targets. To maximize its effectiveness, the original C3f2 module in YOLOv11 is replaced with the PConv structure.

The SCSA module, which jointly captures spatial and channel-wise dependencies, is integrated into the small-object detection layer of the network. This enhancement significantly improves the model's capability to detect small-scale targets in complex scenes.

2. Related Work

In 2015, R. Joseph et al. [7] introduced YOLO, a one-stage object detection algorithm that applies a single neural network to the entire image. It divides the image into a grid and simultaneously predicts both the bounding boxes and class probabilities for each cell. In the following years, the YOLO series evolved rapidly, with successive releases of YOLOv3 [8], YOLOv4 [9-10], and YOLOv5. In 2022, Meituan released YOLOv6 [11], a model designed for industrial applications, which introduced a multi-branch re-parameterization structure, self-distillation strategies, and quantization techniques. This model was further refined in 2023 with the release of YOLOv6 v3.0 [12].

Also in 2022, the original authors of YOLOv4 proposed YOLOv7[13], which focused on structural re-parameterization and dynamic label assignment to improve training efficiency and accuracy. In parallel, Xu et al. enhanced PP-YOLOv2 [14] and introduced PP-YOLOE [15], an anchor-free model

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0001

that featured the powerful CSPRepResStage module, an Efficient Transformer (ET) detection head, and the Task-aligned Assigner for dynamic label assignment.

In 2023, the Ultralytics team launched YOLOv8, which fused improvements from previous YOLO versions. It adopted an anchor-free detection paradigm, introduced the C2f module, and employed a decoupled detection head to enhance performance and flexibility.

Most recently, in 2024, a collaborative effort between Academia Sinica and National Taipei University of Technology led to the development of YOLOv9 [16]. This version addressed information degradation issues in deep networks, such as bottlenecks and the lack of invertibility, by proposing a novel auxiliary supervision method called Programmable Gradient Information (PGI). In addition, a lightweight network architecture named Generalized Efficient Layer Aggregation Network (GELAN) was designed based on gradient path planning. Experimental results demonstrated that PGI significantly improved detection performance, especially in lightweight models.

YOLOv11 represents a significant milestone in the evolution of the YOLO series, building upon the advancements of YOLOv5 and YOLOv8 by integrating state-of-the-art lightweight design principles, attention mechanisms, and object perception optimization strategies. These enhancements further improve the model's robustness in complex environments and its adaptability to multi-scale objects. YOLOv11 is particularly effective in addressing key challenges associated with aerial imagery, such as the difficulty of detecting small targets, severe occlusions, and complex background interference.The network structure of YOLOv11 is shown in Figure 1.

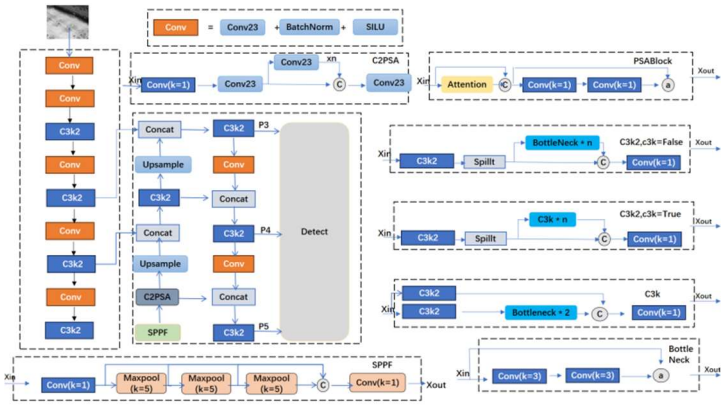

Figure 1. YOLOv11 Model Architecture

3. Method

In UAV aerial imagery, the challenges posed by long imaging distances, small object scales, complex backgrounds, and densely distributed targets often hinder the performance of conventional convolutional neural networks during the feature extraction stage. These models typically struggle with weak semantic representation of small objects and a tendency to lose fine-grained edge details. In particular, during multi-scale feature fusion, shallow-layer information is frequently overwhelmed by high-level semantic features, leading to a decline in small object detection performance.

Although YOLOv11, as an advanced framework within the YOLO family, demonstrates high detection speed and satisfactory overall accuracy in standard object detection tasks, it still faces notable performance bottlenecks when applied to UAV imagery. Such scenarios are characterized by a high proportion of small targets, significant viewpoint variation, heavy occlusion, and strong environmental interference. Specifically, the recall rate for small objects tends to be low, increasing the likelihood of missed detections. In the multi-scale fusion process, low-level features are often suppressed, weakening the representation of edge and boundary information. Moreover, under nighttime or low-light conditions, relying solely on RGB visible-spectrum imagery limits the model's perceptual capability. Due to architectural constraints, YOLOv11 also shows limitations in feature

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0001

focusing, region discrimination, and cross-scale information modeling in complex UAV environments.

3.1 PConv

Partial Convolution (PConv) was originally introduced by Liu et al. in 2018 [17] for the task of image inpainting, aimed at reconstructing irregularly occluded regions. This mechanism performs convolution operations only on the regions of the input feature map that are marked as "valid" by a binary mask, thereby excluding masked or irrelevant areas from the computation. Additionally, the output of each convolution is normalized based on the proportion of valid inputs, ensuring that the convolution result is influenced solely by meaningful information. By minimizing the interference from invalid regions, PConv enhances the model's ability to focus on locally salient features, improving its perception of structurally important areas.

The introduction of Partial Convolution (PConv) enhances the model's responsiveness to informative regions by selectively performing convolution only on valid input areas, while automatically disregarding invalid or missing regions. This core design principle enables the network to suppress background noise more effectively, which is particularly beneficial in small object detection scenarios. By incorporating a form of early-stage "selective attention," PConv guides the feature extraction process to focus on potential target regions from the outset. This is especially advantageous when dealing with small objects that exhibit blurry boundaries or low contrast with the background. In such cases, PConv improves the target's saliency within the feature map, making it easier for subsequent layers to recognize the presence of the object. As a result, detection accuracy and recall are significantly enhanced. An illustration of the PConv operation is shown in Figure 2.

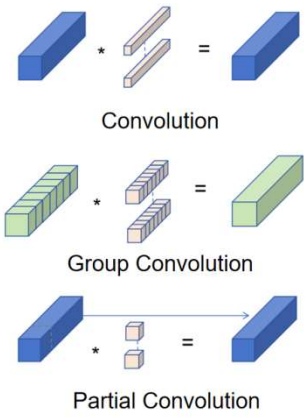

Figure 2. PConv structure

3.2 SCSA

To further enhance the representational capability of YOLOv11 in complex scenarios-particularly its ability to focus on densely distributed, boundary-located, and low-saliency targets-this paper incorporates a novel attention module, termed Spatial and Channel Synergistic Attention (SCSA)[18], into the detection framework, resulting in the improved model YOLOv11-SCSA. Originally proposed by Duan Jinshuo et al. from Nanjing University of Posts and Telecommunications in 2023, the SCSA module has demonstrated strong feature enhancement capabilities, especially within lightweight network architectures.

SCSA is designed to address the limitations of modeling spatial and channel attention independently by establishing a joint attention mechanism through the synergy of two submodules: Shared Multi-Semantic Spatial Attention (SMSA) and Progressive Channel Self-Attention (PCSA). The SMSA submodule captures both local and global spatial information using multi-scale receptive fields, enabling effective spatial response to salient regions. Guided by the spatial features produced by

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0001

SMSA, PCSA applies a refined intra-channel self-attention mechanism to model semantic similarity across channels. The outputs of both modules are multiplied and then added to the original feature map, completing the enhancement process.

This synergistic attention mechanism enables spatial enhancement across multiple scales and semantic focusing across channels, significantly improving the model's perception of complex and subtle targets. The architecture of the SCSA module is illustrated in Figure 3.

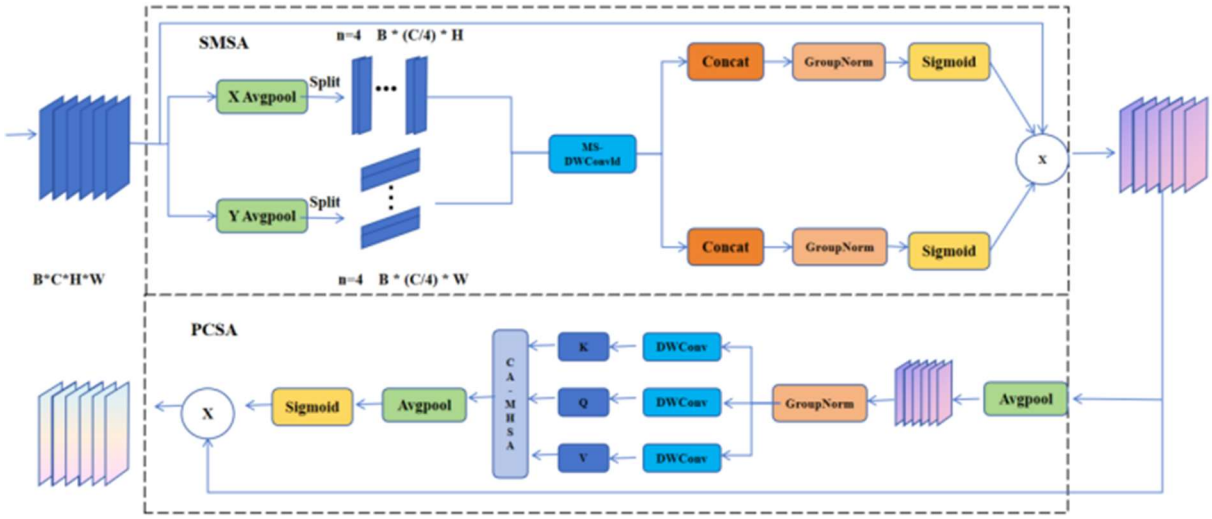

Figure 3. SCSA structure

4. Experiment

4.1 Dataset and Experimental Settings

The HIT-UAV dataset[19] is a high-resolution collection specifically designed for evaluating object detection and tracking algorithms in UAV-based imagery. It includes a wide range of aerial images captured under various environmental conditions, with annotations for multiple object categories such as vehicles, pedestrians, and small targets. This dataset is particularly valuable for testing models in real-world UAV applications, where objects may appear small, occluded, or in diverse orientations and lighting conditions. Its challenging scenarios make it a key resource for advancing object detection techniques, especially in surveillance and autonomous navigation tasks.

The VisDrone2019 dataset[20] is a large-scale collection intended for object detection and tracking tasks using UAV-captured videos and images. It covers a variety of scenes, including urban, rural, and industrial environments, with annotations for various object categories like pedestrians, vehicles, and small objects. Known for its challenging conditions-such as severe occlusions, small object detection, and complex backgrounds-VisDrone2019 provides an essential platform for the development and benchmarking of advanced object detection algorithms. Its diverse and difficult scenarios make it ideal for improving real-time UAV-based detection and tracking systems in dynamic environments.

To ensure the stability and fairness of ablation experiments between the baseline algorithm (YOLOv11) and the proposed improved models, all experiments were conducted under identical settings and fixed hyperparameters. Specifically, the training configuration included a batch size of 32, 400 epochs, and an input image resolution of 640. The Stochastic Gradient Descent (SGD) optimizer was employed throughout the training process. The hardware configuration of the server used for training is detailed in Table 1.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0001

Table 1. Experimental environment

Category	Configuration
Operating System	Ubuntu 22.04
Python Version	Python 3.9
CPU	Intel(R) Xeon(R) Platinum 8481C
GPU	NVIDIA GeForce RTX 4090
Deep Learning Framework	PyTorch 2.1.0

4.2 Experimental Results

The ablation experiments conducted on the HIT-UAV dataset are shown in Table 2. The combined use of the P-Conv and SCSA modules results in an improvement in mAP@0.5, reaching 90.2%, and mAP@0.5-0.95, which increases to 60.1%. At the same time, the model's parameter count is reduced to 8.18M, and the FLOPs are lowered to 20.1G, indicating a strong synergistic effect between these two modules. Compared to the baseline model, this joint optimization at both the structural and attention levels compensates for the original detection framework's shortcomings in terms of lightweight design, precise focusing, and feature representation. As a result, YOLOv11 achieves a better balance between accuracy, efficiency, and model complexity.

Table 2. Ablation experiment results on the HIT-UAV dataset

	mAP50/%	mAP50-95/%	GFLOPs/G	Parameters/M,
YOLOv11s	88.4	59.3	21.6	9.431
YOLOv11s+PConv	89.1	59.5	20.2	8.185
YOLOv11s+SCSA	89.5	59.9	21.6	9.433
Ours	**90.2**	**60.1**	**20.1**	**8.180**

The ablation experiment results on the VisDrone2019 dataset are shown in Table 3. With the simultaneous introduction of the P-Conv and SCSA modules, the model performance significantly improves. Specifically, mAP@0.5 reaches 33.3%, and mAP@0.5-0.95 increases to 18.8%, while the parameter count is reduced to 8.18M, and FLOPs decrease to 20.1G, achieving the optimal balance. This joint optimization strategy effectively compensates for the shortcomings of the baseline model in terms of receptive field, attention dispersion, and redundant computations, making the YOLOv11 model better suited for the efficient detection of multi-scale small objects in complex urban scenes and UAV top-down views, as seen in the VisDrone2019 dataset.

Table 3. Ablation experiment results on the VisDrone2019 dataset

	mAP50/%	mAP50-95/%	GFLOPs/G	Parameters/M,
YOLOv11s	31.6	17.9	21.6	9.431
YOLOv11s+PConv	31.9	18.2	20.2	8.185
YOLOv11s+SCSA	31.6	18.1	21.6	9.433
Ours	**33.3**	**18.8**	**20.1**	**8.180**

5. Conclusion

Based on the ablation experiments conducted on the HIT-UAV and VisDrone2019 datasets, the proposed improvements to the YOLOv11 framework have demonstrated significant performance

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0001

gains, particularly in small object detection within complex UAV imaging environments. The integration of the P-Conv and SCSA modules has proven to enhance both detection accuracy and model efficiency. Specifically, on the HIT-UAV dataset, the inclusion of both modules resulted in a mAP@0.5 of 90.2% and a mAP@0.5-0.95 of 60.1%, while reducing computational costs (FLOPs) and maintaining a lightweight model with 8.18M parameters. Similarly, on the VisDrone2019 dataset, the model achieved 33.3% mAP@0.5 and 18.8% mAP@0.5-0.95, further validating the effectiveness of these optimizations in handling complex urban scenarios and multi-scale small object detection.

These results confirm that the proposed P-Conv and SCSA modules significantly contribute to improving the model's ability to focus on relevant features, suppress background noise, and enhance the perception of small and occluded objects. The overall improvement in detection performance, coupled with a reduced computational footprint, demonstrates that the enhanced YOLOv11 model strikes an optimal balance between accuracy, efficiency, and model complexity. Therefore, this work provides an effective solution for real-time UAV-based object detection, especially in challenging conditions such as dense urban environments and low-light scenarios. Future work will focus on further optimizing the framework for multi-modal data and exploring its applicability in dynamic tracking and real-time applications.

References

[1] Xue Z, Xu R, Bai D, et al. YOLO-tea: A tea disease detection model improved by YOLOv5[J]. Forests, 2023, 14(2): 415.

[2] Zhou X, Xu X, Liang W, et al. Intelligent small object detection for digital twin in smart manufacturing with industrial cyber-physical systems[J]. IEEE Transactions on Industrial Informatics, 2021, 18(2): 1377-1386.

[3] Guo Y, Chen S, Zhan R, et al. LMSD-YOLO: A lightweight YOLO algorithm for multi-scale SAR ship detection[J]. Remote Sensing, 2022, 14(19): 4801.

[4] Yuan M, Zhou Y, Ren X, et al. YOLO-HMC: An improved method for PCB surface defect detection[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1-11.

[5] Bi H, Feng Y, Tong B, et al. RingMoE: Mixture-of-Modality-Experts Multi-Modal Foundation Models for Universal Remote Sensing Image Interpretation[J]. arXiv preprint arXiv:2504.03166, 2025.

[6] Yu C, Jiang X, Wu F, et al. Research on Vehicle Detection in Infrared Aerial Images in Complex Urban and Road Backgrounds[J]. Electronics, 2024, 13(2): 319.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C] Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[8] RedmonJ, Farhadi A.Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[9] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[10] Wang C Y, Bochkovskiy A, Liao H Y M. Scaled-yolov4: Scaling cross stage partial network[C]//Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. 2021: 13029-13038.

[11] Li C, Li L, Jiang H, et al. YOLOv6: A single-stage object detection framework for industrial applications[J]. arXiv preprint arXiv:2209.02976, 2022.

[12] Li C, Li L, Geng Y, et al. Yolov6 v3.0: A full-scale reloading[J].arXiv preprint arXiv:2301.05586, 2023.

[13] Wang CY, BochkovskiyA, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state- of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 7464-7475.

[14] Huang X, Wang X, Lv W, et al. PP-YOLOv2: A practical object detector[J]. arXiv preprint arXiv:2104.10419, 2021.

[15] Xu S, Wang X, Lv W, et al. PP-YOLOE: An evolved version of YOLO[J]. arXiv preprint arXiv:2203.16250, 2022.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0001

[16] Wang C Y, Yeh I H, Liao H Y M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information[J]. arXiv preprint arXiv:2402.13616, 2024.

[17] Chen, J.; Kao, S.h.; He, H.; Zhuo, W.; Wen, S.; Lee, C.H.; Chan, S.H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.

[18] Chen J, Kao S, He H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 12021-12031.

[19] Suo J, Wang T, Zhang X, et al. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection[J]. Scientific Data, 2023, 10(1): 227.

[20] Du D, Zhu P, Wen L, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results[C]//Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019: 0-0.