International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

SPNC-YOLO: Improved YOLOv11 for Small Object Detection in UAV Aerial Images

Yuhang Zhang[1]
School of Computer and Engineering, Anhui University of Finance and Economics, Bengbu, Anhui, China-233030
Email address: 3195738339@qq.com

Abstract*—To address challenges such as small object size, ambiguous feature information, and high missed/false detection rates in unmanned aerial vehicle (UAV) aerial images, this study proposes SPNC-YOLO, a novel object detection method based on the improved YOLOv11. Firstly, an enhanced lightweight Space-to-Depth Convolution (SPD-Conv) is introduced to strengthen the extraction and retention of small object features, achieving a balance between accuracy improvement and parameter efficiency. Secondly, a dedicated small object detection layer (P2 layer) is integrated into the detection head, expanding the receptive field for distant small targets to reduce detection errors. Finally, based on the original C3K2 module, a Selective Boundary Aware (SBA) module is incorporated, while the Conv layers in the Bottleneck block are replaced with lightweight Partial Convolution (PConv) layers. The optimized C3K2 module selectively aggregates shallow boundary information and deep semantic features, ensuring precise maintenance of object boundaries and accurate calibration of object positions in images. Experimental results on the VisDrone2019 dataset demonstrate that compared with the YOLOv11 model, the proposed Ours-P2345 achieves mAP@0.5 and mAP@0.5:0.95 values of 38.0% and 23.0%, reflecting improvements of 6.4% and 4.5% respectively. This approach effectively adapts to object detection tasks in complex UAV aerial scenarios.*

***Keywords*—** *Object detection, small targets, YOLOV11*

I. INTRODUCTION

Nowadays, UAV technology is rapidly developing and being widely applied across various fields[1]. UAVs, characterized by their fast flight speed and adaptability to complex terrain environments, serve as efficient data collection tools in traffic monitoring, power line inspection, crop analysis, disaster rescue, and other domains. To enhance the intelligence of UAVs, target detection technology has been integrated into their d ata processing workflows. Current feasible methods for combining YOLO with UAV systems typically require the integration of embedded processors. The YOLO algorithm, when paired with embedded processors, enables real-time target detection on captured image data.

The primary challenges faced by UAV platform-based target detection include the following. First, due to variations in UAV shooting angles and flight altitudes, targets at different distances may coexist within the same image, leading to scale imbalance. Second, UAV-captured images often contain small targets in clustered states, which are prone to mutual occlusion. Third, diverse weather conditions and lighting variations cause changes in target appearance, demanding strong model generalization. Additionally, class imbalance in UAV images—where some categories have significantly fewer samples—further complicates model training. Finally, high-resolution UAV images, necessary for capturing detailed information, impose substantial computational overhead on target detection algorithms. Given the limited computational power of UAV platforms and the requirement for real-time performance, deploying complex models is challenging, necessitating model lightweighting.

Current UAV small-target detection algorithms are primarily based on deep learning. These algorithms are generally categorized into single-stage and two-stage methods. Single-stage algorithms directly predict target locations and categories from images, with representative examples including the YOLO series[2-6] and SSD[10]. Two-stage algorithms, such as R-CNN[12] and Fast R-CNN[11], first generate region proposals and then classify and localize them. While two-stage methods achieve higher accuracy, their slow detection speed fails to meet real-time requirements, making them unsuitable for UAV deployment. Single-stage algorithms, particularly the YOLO series, are well-suited for UAV platforms due to their balance of accuracy, speed, and lightweight design. Moreover, YOLO-based algorithms can be extended to process multispectral data (e.g., infrared or hyperspectral images) by increasing input channels or adopting dual-stream network architectures, enhancing their adaptability to new UAV sensor data. In summary, UAV small-target detection primarily faces two challenges: improving model accuracy and reducing computational complexity to ensure lightweight deployment.

Addressing the first challenge, this paper comprehensively redesigns the YOLOv11 model, introducing improvements to its backbone, neck, and detection head. First, the conventional Conv layers in the backbone are replaced with SPD

Conv layers[7], which are more efficient and precise for small-target detection, thereby enhancing feature extraction capabilities and processing speed. Second, the neck is enhanced by introducing a small-target detection layer and a Selective Boundary Aggregation (SBA) module[8]. The SBA module selectively aggregates shallow boundary details and deep semantic information, improving boundary precision and spatial calibration while preserving small-target details. Finally, the detection head is optimized to boost detection and localization accuracy, collectively reducing missed detections. For the second challenge, this work optimizes model lightweighting. Building on the four-head detection architecture of the improved model, all C3K2 modules in the network are refined. Specifically, the Conv layers in the Bottleneck blocks of C3K2 modules are replaced with

Yuhang Zhang, "SPNC-YOLO: Improved YOLOv11 for Small Object Detection in UAV Aerial Images," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 12, pp. 426-429, 2026.

lightweight PCConv layers, which maintain high detection accuracy while significantly reducing model parameters and computational costs. This modification accelerates training speed without compromising overall accuracy[14-15].

II. METHOD

2.1 SPD-Conv Module

The YOLOv11 model exhibits significant advantages in general object detection, but it struggles with specific challenges. For instance, extremely small targets (e.g., those with resolutions below 32×32 pixels), such as UAV-captured distant objects or road cracks, occupy minimal pixel areas in images. Their subtle features make accurate recognition and localization difficult, often resulting in false positives or missed detections. Additionally, UAV-captured images frequently contain cluttered backgrounds (e.g., skies, trees, transmission towers) where background elements share color or texture similarities with targets. The model's limited ability to suppress such complex backgrounds further degrades the accuracy of feature extraction and image segmentation [16–18].

The SPD-Conv module is designed to enhance the processing capability for low resolution images and small objects. It consists of two components: a Space-to-Depth (SPD) layer followed by a non-strided convolutional layer. The core idea of this design is to replace traditional strided convolutions and pooling operations, thereby retaining more fine-grained information to improve performance in handling small sized objects and low-resolution images.

Specifically, the SPD layer generates a series of sub-feature maps by slicing intermediate feature maps. These sub-feature maps are derived by partitioning specific regions of the original feature map, with each sub-map down sampled using a predefined scale factor. This process reduces spatial dimensions while increasing channel dimensions. Subsequently, a non-strided convolutional layer processes the transformed features from the SPD layer. This layer employs learnable parameters to reduce channel numbers, mitigating information redundancy caused by the expanded channels while preserving discriminative feature information.

2.2 Improved C3K2 Module

C3K2 Module is an important feature extraction component in the YOLOv11 model, designed as an improved version of the traditional C3 module. By combining variable convolutional kernels (e.g., 3×3, 5×5) and channel separation strategies, it provides stronger feature extraction capabilities, particularly suited for complex scenarios and deep-level feature extraction tasks.

To improve the training speed and reduce the parameter count of the enhanced model, we replaced the convolutional layers in the original C3K2 module's Bottleneck block with partial convolutional layers (PConv). Unlike traditional convolutions that process all input channels, PConv only performs convolution on a subset of input channels while leaving the remaining channels unprocessed. This modification significantly increases the model's training speed while ensuring the parameter count remains manageable and preserving the model's accuracy.

2.3 Improvements to the Neck Network

In multi-scale feature fusion, shallow features typically contain more distinct boundary information and richer details. However, due to the lack of high-level semantics, they often struggle to accurately represent the holistic semantic meaning of targets. Conversely, deep features possess more complete and abstract semantic information but tend to lose object details, resulting in ambiguous or neglected local contours. Therefore, directly fusing low-level features with high-level features may introduce redundancy, confusion, or inconsistency. To address this challenge, this paper proposes integrating the SBA (Selective Boundary-Aware) module into the original C3K2 model. This method selectively aggregates shallow boundary information and deep semantic features, maintaining the precision of target boundaries while more accurately calibrating object positions in images.

III. EXPERIENCE

3.1 Dataset and Experimental Setup

The dataset used in this study is the publicly available UAV small-target dataset VisDrone2019. The VisDrone2019 dataset contains 10,209 high-resolution images with a resolution of 2000×1500, covering ten categories: pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor. Among these, 6,471 images are used for training, 548 for validation. Since the test-challenge subset is reserved for competitions and lacks annotations, the test-dev subset (1,610 images) is adopted as the test set.

The hardware configuration includes an Intel i9-13900HX CPU, Windows 11 22H2, and an RTX 4060 16G GPU. All experiments are implemented with Python 3.10, CUDA 11.8, and PyTorch 2.1.1. Training spans 200 epochs with an input image resolution of 640×640. The optimizer is AdamW, no pre-trained weights are used.

3.2 Comparative Experiments

TABLE 1. performance comparison on VisDrone2019.

Methods	Param	mAP@0.5%	mAP@0.5-0.95%	R%
DMNet	-	30.3	-	-
RetinaNet	-	28.3	-	-
CenterNet	-	32.3	-	-
YOLOv5n	1.8	32.4	10.8	35.1
YOLOv7-t	6.0	31.7	16.1	35.5
YOLOv8n	3.1	31.5	18.1	32.1
YOLOv9-t	2.0	30.4	17.6	33.1
YOLOv11n	2.58	31.6	18.5	31.5
YOLOv12n	2.6	32.5	19.1	32.7
Ours-P234	3.7	37.6	22.8	36.8
Ours-P2345	4.6	38.0	23.0	37.1

The proposed models Ours-P234 and Ours-P2345 significantly outperform baseline models. Ours-P2345 achieves 38.0% mAP@0.5 and 23.0% mAP@0.5:0.95, which are 5.5% and 3.9% higher than the best baseline YOLOv12n.

Yuhang Zhang, "SPNC-YOLO: Improved YOLOv11 for Small Object Detection in UAV Aerial Images," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 12, pp. 426-429, 2026.

3.3 Ablation Experiments

TABLE 2. presents the ablation study results.

Models	mAP @0.5%	mAP@0.5-0.95%	R%	P%
YOLOv11n	31.6	18.5	25.8	31.5
YOLOv11n+p234	35.8	21.6	35.6	42.6
YOLOv11n+p234 +SPD-Conv	35.9	21.6	35.9	44.2
YOLOv11n+p234 +SPD-Conv+PC-Conv	36.2	21.9	36.4	44.3
YOLOv11n+p234 +SPD-Conv+PC-Conv+SBA	37.6	22.8	36.8	46.8
YOLOv11n+p2345 +SPD-Conv+PC-Conv+SBA	38.0	23.0	37.2	47.0

To validate the effectiveness of the proposed modules, we conducted experiments by progressively integrating SPD-conv, PC-conv, SBA modules, and two distinct detection heads (P234/P2345) into thebaselineYOLOv11nmodel. The experimental results are summarized inTable2.

When introducing the P234 detection head to the baseline YOLOv11n (mAP@0.5=31.6%), the model achieved an improved mAP@0.5 of 35.8%(+4.2%) with recall surging from 31.5% to 42.6%, demonstrating that multi-level feature fusion significantly enhances object localization capabilities. Subsequent integration of SPD-conv increased mAP@0.5:0.95 from 21.6% to 21.9% and precision to 44.3%, validating the effectiveness of the small-object detection head for multi-scale detection. The addition of PC-conv further improved training speed with minimal performance trade-offs, raising precision to 44.2%while slightly reducing mAP@0.5(36.2%→35.9%). The SBA module delivered comprehensive performance gains: mAP@0.5 reached 37.6%(+6.0%overbaseline) with precision at 46.8%, confirming SBA's ability to suppre-ss interference and balance precision-recall trade-offs.

Extending the detection head from P234 to P2345 moderately increased parameters but elevated mAP@0.5:0.95 to23.0% (+1.4%overP234), with recall and precision reaching 37.2% and 47.0%, respectively. This indicates deeper paths capture richer semantic features, particularly effective for de-nse small objects, though computational costs require deployment considerations.

The final composite model (YOLOv11n+P2345+SPD-conv+PC-conv) achieved optimal mAP@0.5(38.0%, +20.3% over baseline) and precision (47.0%,+49.2%). Notably, the synergistic use of SPD-conv and PC-convstabilized mAP@0.5:0.95at 23.0%, proving their complementary nature, while SBA alleviated false detections in complex scenes through dynamic feature selection.

IV. Conclusion

To address high miss rates and low precision in small-object detection, we propose SPNC-YOLOv11– an enhanced YOLOv11 algorithm. First, we redesign the backbone by introducing C3K2-PConv modules to optimize Bottleneck structures with PC-conv. SPD-Conv integration substantially improves computational accuracy, while dual detection head strategies (adding P2 layers or removing P5 structures) significantly enhance precision while reducing parameters. The SBA module in the Neck structure amplifies feature fusion efficacy. Evaluations on VisDrone2019 demonstrate SPNC-YOLOv11 surpasses baseline models in detection accuracy and generalization capability, showing robust performance across diverse scenarios. This work adapts YOLOv11 for UAV platforms with limited computational resources. Future directions include model pruning, knowledge distillation, parallel computing, and hardware acceleration to further optimize real-time detection performance.

References

[1] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded sparse query for accelerating high-resolution small object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13668–13677.

[2] T.-Y. Lin, P. Dollár, R. Girshick, K. He, and B. Hariharan, "Feature pyramid networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117–2125.

[3] P. Zhu et al., "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2019, pp. 213–226.

[4] D. Du et al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 370–386.

[5] J. Wang et al., "AI-TOD: Aerial image tiny object detection dataset," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4390–4399.

[6] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7464–7475.

[7] A. Wang, H. Chen, L. Liu, and K. Chen, "YOLOv10: Real-time end-to-end object detection," arXiv preprint arXiv:2405.14458, 2024.

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

[9] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," in Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Springer, 2022, pp. 443–459.

[10] J. Chen et al., "Run, don't walk: Chasing higher FLOPS for faster neural networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12021–12031.

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

[12] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7464–7475.

[13] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," *arXiv preprint arXiv:2402.13616*, 2024.

[14] A. Wang, H. Chen, L. Liu, and K. Chen, "YOLOv10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[15] P. Zhu *et al.*, "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019, pp. 213–226.

[16] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proceedings of the IEEE/CVF*

International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 2778–2788.
[17] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "ClusDet: Clustering detection for high-resolution aerial imagery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9015–9024.
[18] C. Li *et al.*, "Density map guided object detection in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 190–191.
[19] J. Wang *et al.*, "A normalized Gaussian Wasserstein distance for tiny object detection," *arXiv preprint arXiv:2110.13389*, 2021.
[20] J. Wang *et al.*, "AI-TOD: Aerial image tiny object detection dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4390–4399.
[21] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," in *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, Springer, 2022, pp. 443–459.
[22] J. Chen *et al.*, "Run, don't walk: Chasing higher FLOPS for faster neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 12021–12031.
[23] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.
[24] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13713–13722.
[25] D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 370–386.
[26] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3974–3983.
[27] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13668–13677.
[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.
[20] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.
[29] S. Deng et al., "A Global-Local Self-Adaptive Network for Drone-View Object Detection," in IEEE Transactions on Image Processing, vol. 30, pp. 1556-1569, 2021, doi: 10.1109/TIP.2020.3045636.
[30] M. Zhao and H. Cui, "Faster-CEASC+: A Lightweight Object Detection Network for Drone Images," 2024 10th International Conference on Computer and Communications (ICCC), Chengdu, China, 2024, pp.876880, doi:10.1109/ICCC62609.2024.10942251.