YOLO11-LKAConv: Optimizing UAV Image Multi-Target Detection Based on Improved YOLO Architecture

Wu Xianyi
Belarusian State University
Minsk,Republic of Belarus
tigerv5872@gmail.com

Sergey Ablameyko
Belarusian State University
United Institute of Informatics Problems
National Academy of Sciences of Belarus
Minsk, Republic of Belarus
ablameyko@bsu.by

***Abstract*—This paper presents YOLO11-LKAConv, an improved model based on the YOLOv11 framework, to address the issue of low detection accuracy for small targets in UAV aerial images. By replacing standard convolutions with lightweight large-kernel attention convolution (LKAConv), the model integrates dynamic large-kernel receptive fields and channel-spatial attention mechanisms, enhancing its ability to capture long-range contextual features for small targets. Experiments on the VisDrone2019 dataset show that the improved model achieves an mAP50-95 of 0.173, a 1.2% increase over the baseline YOLOv11n, with improvements in both P and mAP50 while maintaining almost the same inference time. The results indicate that LKAConv effectively balances detection accuracy and computational efficiency through its lightweight large-kernel design, offering a better solution for real-time UAV target detection tasks.**

***Keywords*—Small target detection; YOLOv11n; UAV detection; LKAConv; Lightweightt**

I. Introduction

With the rapid development and popularisation of UAV technology, UAVs have been widely used in aerial photography, agriculture, security monitoring, disaster rescue, environmental monitoring and other fields. UAV aerial photography can not only provide high-definition images and videos, but also complete tasks such as monitoring, surveying and search and rescue of complex environments, which has become an indispensable technical means in modern society [1]. However, the flight safety and effective monitoring of UAVs have become an urgent problem, especially in complex scenarios, and the performance of target detection technology directly determines the intelligence level of UAVs [2], [3].

Traditional UAV target detection methods usually rely on hand-designed feature extractors and classifiers (e.g., HOG, SIFT, etc.), which perform reasonably well in specific scenarios, but their performance is often limited under complex backgrounds and variable target morphology [4], [5]. In recent years, the rapid development of deep learning technology provides new solutions for UAV target detection. Deep learning-based target detection methods significantly improve detection accuracy and robustness by training deep neural networks to automatically learn the feature representation of the target [6], [7]. Among them, the YOLO series of algorithms has become a research hotspot in the field of UAV target detection due to its fast speed and high accuracy [1], [4]. However, with the expansion of UAV application scenarios, the existing YOLO series network still has deficiencies in dealing with the problems of dense small targets and complex backgrounds [2], [3].

Aiming at the insufficient performance of the existing UAV target detection algorithms in the dense and complex background of small targets, this study aims to propose a lightweight kernel attention mechanism (LKAConv) in the YOLOv11 framework, which dynamically adjusts the kernel sensing field and the allocation of the attention weights, enhances the ability of the model to capture the features of the small targets, and improves the detection precision and the recall rate, so as to realise a lightweight design of the model and ensure that it can be used in the UAV application scenarios [2], [3]. The lightweight design ensures its real-time application on resource-constrained devices such as UAVs.

II. Method

A. YOLOv11

YOLOv11, a new generation of object detection algorithms introduced by Ultralytics in 2023, aims to further improve the accuracy and efficiency of object detection. It has made several improvements based on YOLOv8 [8] to adapt to a wider range of application scenarios and enhance model performance. YOLOv11 provides multiple versions of different scales, including YOLOv11n (ultra-light), YOLOv11s (small), YOLOv11m (medium), YOLOv11l (standard), and YOLOv11x (extra-large), to meet different needs. Compared with previous YOLO versions, YOLOv11 has made the following improvements:

1) **Backbone Network:** YOLOv11 introduces the `C3k2` module [9], replacing the `C2f` module in YOLOv8. The `C3k2` module uses smaller convolution kernels to improve computational efficiency while maintaining performance. It retains the spatial pyramid pooling fast (`SPPF`) module [10] and introduces the cross-stage partial and spatial attention (`C2PSA`) module [10], enhancing spatial attention in feature maps and improving detection accuracy.
2) **Neck Structure:** In the neck structure, YOLOv11 replaces the `C2f` module with the `C3k2` module, improving the speed and performance of feature aggregation. The `C2PSA` module enhances spatial attention, enabling the model to more effectively focus on key areas in the image and improving detection accuracy for small and partially occluded targets.
3) **Head Structure:** In the head structure, YOLOv11 uses multiple `C3k2` modules to process and optimize feature maps, improving the model's detection accuracy.

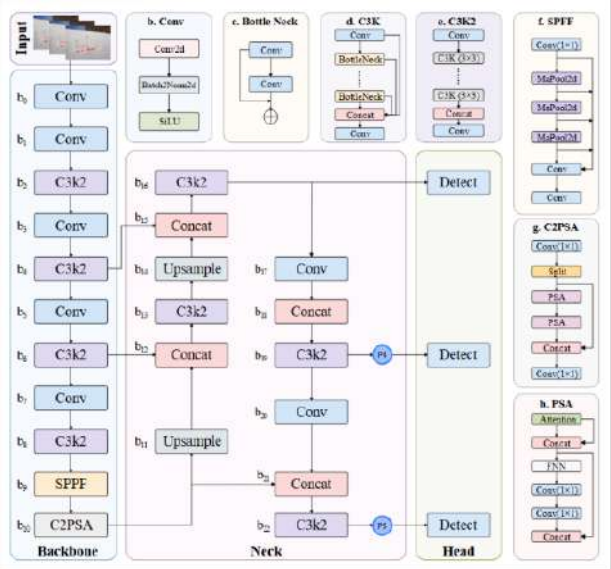

Figure 1. YOLOv11 Network Structure Diagram

Additionally, YOLOv11 adopts multi-scale training and data augmentation techniques during training to further improve the model's generalization and detection accuracy. Compared with previous generations, YOLOv11 shows significant improvements in inference speed and accuracy. In summary, YOLOv11 has made significant progress in the accuracy and efficiency of object detection by introducing innovative technologies such as the `C3k2` module and the `C2PSA` module. It performs well in models of different scales and demonstrates strong adaptability and practicality in various application scenarios. The network structure of YOLOv11 is shown in Figure 1.

B. LKAConv

`LKAConv` (Large Kernel Attention Convolution) is a key component of the LKA mechanism, used to implement the decomposition of large convolution kernels [11]. `LKAConv` captures long-range dependencies by decomposing a large convolution kernel into multiple small convolution kernels and dilated convolutions. This decomposition method not only retains local structural information but also effectively captures long-range dependencies while maintaining linear complexity. The principle of the large convolution kernel is shown in Figure 2:

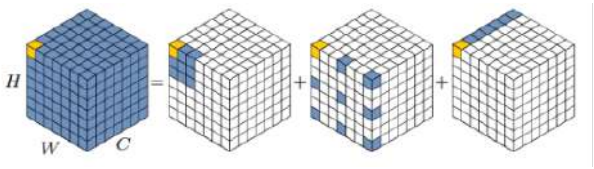

Figure 2. Decomposition Diagram of Large-Kernel Convolution

The core formula of LKA is as follows:

$$\text{Attention} = \text{Conv}_{1\times1}(DW - D - \text{Conv}(DW - \text{Conv}(F))), \quad (1)$$

$$\text{Output} = \text{Attention} \otimes F \quad (2)$$

where: $\mathbf{F}$ is the input feature map, $DW - Conv$ represents depthwise separable convolution, $DWD - Conv$ represents dilated depthwise separable convolution, $Conv_{1\times1}$ represents 1×1 convolution, and 1×1 represents element-wise multiplication.

In this study, we introduced the large kernel attention convolution to enhance the feature extraction capability of the YOLOv11 model, especially when processing UAV datasets. LKAConv is a new type of convolution module that combines the advantages of convolution and self-attention mechanisms, effectively capturing long-range dependencies and local structural information. In our model, the 5th layer adopted LKAConv with parameter settings of 512 input channels, a 3x3 convolution kernel, and a dilation rate of 2. In addition, the downsampling process was mainly completed by standard convolution and LKAConv. The features extracted by LKAConv were fused with the features of the previous layers through concatenation and convolution layers to form rich feature maps. This fusion method effectively combined multi-scale features, further enhancing the model's detection performance.

C. YOLO11-LKAConv

This paper takes YOLOv11n as the baseline and proposes an improved model, YOLO11-LKAConv, to address the problem of detecting small targets in UAV aerial

images. YOLOv11n still has high rates of missed and false detections in complex backgrounds, dense targets, and small target detection. The improved model maintains a lightweight design while significantly improving detection accuracy. The specific improvements are shown in Figure 3 below:

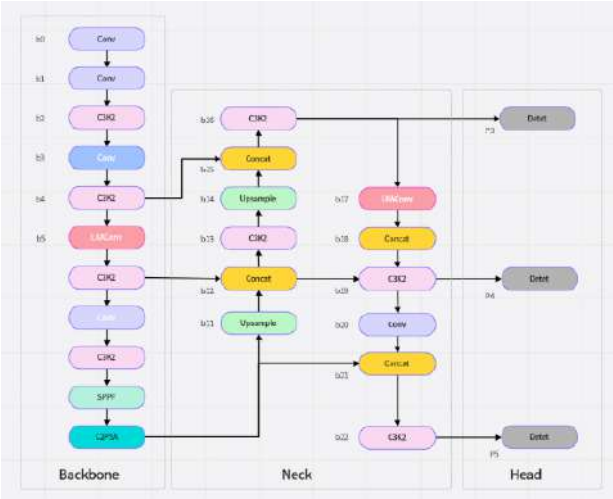

Figure 3. YOLO11-LKAConv Structure

The 5th layer of Backbone corresponds to the output of P4, which is in the middle-level feature stage. This layer is responsible for detecting medium-scale targets. The purpose of introducing LKAConv is to expand the receptive field of the middle-level features, suppress background noise interference, and balance computational efficiency. In drone aerial images, medium-sized targets are often mixed with ground or vegetation of similar colors. The channel attention mechanism of LKAConv can enhance target-related channels and suppress noise channels such as vegetation texture through feature recalibration. The 16th layer of Head corresponds to the feature fusion layer after P3 upsampling. This layer fuses deep semantic features and shallow detail features through Concat operation. The design of introducing LKAConv here is to strengthen cross-scale feature fusion and resist scale change interference. LKAConv can introduce multi-scale perception capabilities at this level. Large kernel convolution covers spatial patterns of different scales, and the attention mechanism focuses on the current dominant scale.

The two LKAConvs in the backbone and head form a "mid-level capture-high-level refinement" synergistic effect: capturing contextual association features of targets through large receptive fields and providing high-quality mid-level semantic information for subsequent FPN. Using attention mechanisms to screen cross-level features and reconstruct the complete spatial structure of small targets. This "coarse-fine" synergistic strategy retains the speed advantage of the YOLO series while breaking through the performance bottleneck of traditional methods in complex air-to-ground detection scenarios, providing a new technical paradigm for real-time UAV target detection.

III. Experiments

A. Dataset

The VisDrone2019 [12] dataset is a large-scale benchmark dataset created by the AISKYEYE team of the Machine Learning and Data Mining Laboratory of Tianjin University, China. The dataset was collected using various drone platforms (i.e., drones with different models) in various situations and under various weather and lighting conditions. It contains carefully annotated ground truth data for various computer vision tasks related to drone-based image and video analysis.

This experiment uses the static images in this dataset, which contains 10,209 images. This experiment divides the dataset into the following: 6,471 training sets, 548 validation sets, and 1,610 test sets. The current dataset has a total of 471,266 annotated targets, covering 11 categories: 0: pedestrian, 1: people, 2: bicycle, 3: car, 4: van, 5: truck, 6: tricycle, 7: awning-tricycle, 8: bus, 9: motor, 10: other. Others are non-valid target areas and are ignored in this project. The validation set contains 38,759 instances, of which small targets (pixel area <32×32) account for 68.2%, and dense scenes (>100 targets per image) account for 45%, which fully reflects the detection challenges from the perspective of drones.

The data annotation is shown in Figure 4 below:

Figure 4. Data annotation

B. Experimental Environment and Configuration

The experimental platform is a 64-bit Windows 11 operating system, with PyTorch 2.4.0 and CUDA 2.4.1 as the deep learning framework. An NVIDIA GeForce

RTX 4060 GPU is used for training, with Python 3.9 and CUDA version 12.41. Specific experimental environment configuration parameters are shown in Table I. Parameters not provided in this paper adopt the default parameters of YOLOv11n.

Table I
Table to test captions and labels

epochs	*batch*	*imgsz*	*device*	*optimizer*	*amp*
100	16	640	0	auto	true

C. Evaluation indicators

The experiment mainly uses mean average precision (mAP), accuracy (precision, P), and recall (recall, R) to evaluate the algorithm's target detection performance. At the same time, floating point operations (GFLOPs), parameter volume (params), model size (volume), and frames per second (FPS) are used to evaluate the model complexity and detection efficiency. The calculation formulas for precision, recall, and average accuracy are shown in equations (3) to (5).

$$P = \frac{TP}{TP + FP}, \tag{3}$$

$$R = \frac{TP}{TP + FN}, \tag{4}$$

$$\begin{cases} AP = \int_0^1 P(R)dR \\ mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \end{cases} \tag{5}$$

In the formula: P is precision; R is recall; TP is the number of samples predicted to be positive and actually positive; FP is the number of samples predicted to be positive but actually negative; FN is the number of samples predicted to be negative but actually positive; AP is the precision of each category in the dataset; mAP is the average accuracy of all categories in the dataset.

D. Cross-Model Comparison Results and Analysis

In order to verify the comprehensive performance of the YOLO11-LKAConv model proposed in this paper in the UAV target detection task, a comparative experiment was designed. We compared it with the mainstream lightweight versions of the YOLO series, including: YOLOv5n [13], YOLOv6n [14], YOLOv8n, YOLOv10n [15], and YOLOv11n. The experimental environment, configuration, and parameters are the same, and the experimental results are shown in Table II:

Through experiments, we found that on the VisDrone2019 dataset, each model performed differently in terms of accuracy and efficiency. The P values of v11n and v10n were both 0.402, and the R values were 0.308 and 0.299, respectively. They performed well in terms of

Table II
Comparison of Different Models on the VisDrone2019-DET Datase

	v11n	v10n	v8n	v6n	v5n	LKAConv
P	0.402	0.402	0.398	0.363	0.387	0.407
R	0.308	0.299	0.302	0.28	0.279	0.302
mAP50	0.297	0.291	0.291	0.271	0.273	0.301
mAP50-95	0.171	0.163	0.165	0.156	0.154	0.173
time	6.012	5.885	7.193	5.794	5.172	6.015
preprocess	0.1ms	0.2ms	0.2ms	0.2ms	0.2ms	0.2ms
inference	1.7ms	1.4ms	2ms	1.7ms	1.7ms	2ms
parameters	2.46M	2.39M	2.57M	3.96M	4.04M	2.51M
GFLOPs	6.3	7.1	8.2	11.5	11.8	6.7

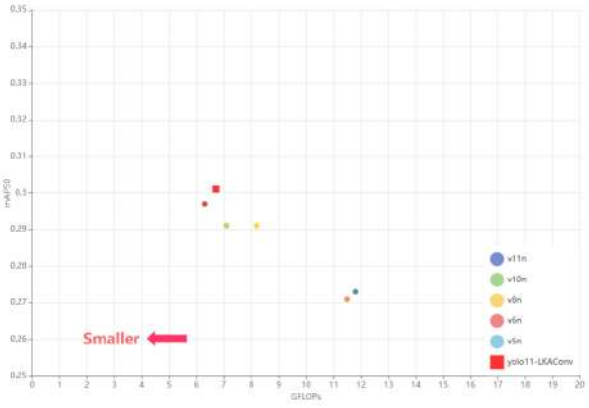

Figure 5. Comparison of GFLOPs of different models

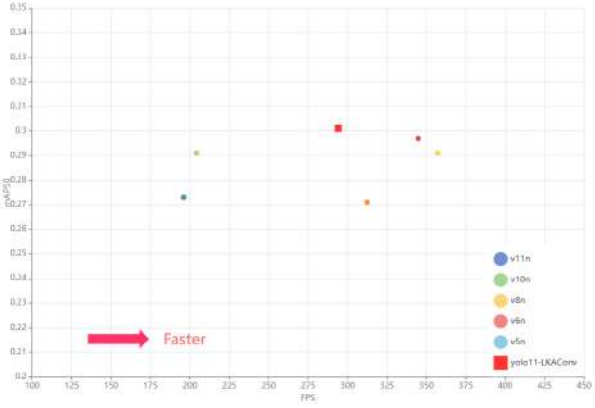

Figure 6. Comparison of FPS of different models

precision and recall, but their FPS were relatively not the highest, at 344.83 and 204.08, respectively, and a trade-off between accuracy and speed was required; v8n had the highest FPS, at 357.14, and its mAP50-95 was also relatively high, at 0.165. It achieved a high processing speed while ensuring a certain accuracy, and was suitable for scenarios with high real-time requirements; v5n and v6n had more parameters, and v5n had a longer post-processing time of 3.2 seconds. In terms of accuracy, the mAP50-95 value of yolo11-LKAConv reached 0.173, an increase of 1.2% over the baseline v11n, and significantly better than other models. This shows that it has better performance for complex scenes (such as occlusion and small targets). The value of yolo11-LKAConv on mAP50 reaches 0.301, indicating that the model is better at detecting medium-scale targets. The model maintains the same inference speed as the baseline v11n at the cost of a slight increase in the number of parameters and computation, as shown in Figures 5. and 6. The lightweight design effectively balances efficiency.

E. Comparison Results and Analysis of Different Convolution Layers

To verify the effectiveness of the improved YOLO11-LKAConv model, it is compared with different convolution layer variants of the baseline YOLOv11n, including RepViTblock, GSConv, and ADown. Experimental results are shown in Table 3. On the VisDrone2019 dataset,

Table III
Comparison of Different Convolution Models on the VisDrone2019-DET Dataset

	v11n	RepViTblock	GSConv	ADown	LKAConv
Box(P	0.402	0.395	0.39	0.401	0.407
R	0.308	0.308	0.303	0.304	0.302
mAP50	0.297	0.298	0.295	0.298	0.301
mAP50-95	0.171	0.169	0.166	0.171	0.173
time	6.012	5.982	6.067	6.086	6.015
preprocess	0.1ms	0.2ms	0.2ms	0.2ms	0.2ms
inference	1.7ms	1.9ms	1.7ms	1.7ms	2ms
parameters	2.46M	2.78M	2.45M	2.36M	2.51M
GFLOPs	6.3	6.3	6.1	6	6.7

RepViTblock has a P value of 0.395 and an R value of 0.308, which are close to v11n's P value of 0.402 and R value of 0.308, but its FPS is 312.50, which is lower than v11n's 344.83; GSConv has a P value of 0.39 and an R value of 0.303, which are slightly lower than v11n, and its time consumption is slightly higher than v11n, which is 6.067 seconds; ADown has a P value of 0.401 and an R value of 0.304, which are close to v11n, but it has the least number of parameters and may have certain advantages in model complexity;

In this experiment, YOLO11-LKAConv still achieves higher mAP50 and mAP50-95 values than other variant models. This indicates that YOLO11-LKAConv can effectively improve small target detection accuracy while maintaining inference efficiency. Although YOLO11-LKAConv's recall rate is slightly reduced due to the feature screening of large-kernel features by LKAConv, the model maintains a similar inference speed to the baseline YOLOv11n with a slight increase in parameters and computational cost, verifying the effectiveness of the lightweight large-kernel design.

IV. Summary

This study proposes an improved model, YOLO11-LKAConv, based on the YOLOv11 framework for UAV target detection tasks. By introducing the lightweight large-kernel attention module (LKAConv), the model significantly improves detection performance for small targets in complex scenes. Experiments on the VisDrone2019 dataset show that the improved model performs excellently in detection accuracy, computational efficiency, and model complexity. Specifically, YOLO11-LKAConv achieves an mAP50-95 of 0.173 and an mAP50 of 0.301, representing improvements of 1.2% and 1.3% over the baseline model YOLOv11n. The inference time (6.015 hours) is almost the same as the baseline model, and the increase in parameters and computational cost is kept within a small range, verifying the effectiveness of the lightweight design. Ablation experiments and visualization analysis further confirm the key role of the LKAConv module in improving model performance. This module dynamically adjusts the large-kernel receptive field and attention weight distribution, enhancing the model's ability to capture small target features and effectively suppressing interference from complex backgrounds. Additionally, comparative experiments with existing variant models (such as YOLO11-RepViTblock, YOLO11-GSConv, and YOLO11-ADown) show that YOLO11-LKAConv achieves better efficiency and lightweight levels while maintaining high detection accuracy. This study provides an efficient and accurate solution for UAV target detection tasks with significant practical application value. The real-time performance of the improved model on UAV edge computing devices makes it widely applicable in scenarios such as public interest litigation and environmental monitoring. Future research directions may include further optimizing the channel pruning strategy of the LKAConv module or introducing dynamic sparse computation to further improve recall rate and inference speed.

References

[1] Z. Zhang, "Drone-YOLO: An efficient neural network method for target detection in drone images," *Drones*, vol. 7, no. 8, p. 526, 2023, doi: 10.3390/drones7080526.

[2] K. Takeda, “Perception and sensing for autonomous vehicles under adverse weather conditions: A survey,” *Robotics and Autonomous Systems*, vol. 150, p. 103462, 2022, doi: 10.1016/j.robot.2022.103462.
[3] H. Chen, D. Liu, and X. Yan, “Infrared image UAV target detection algorithm based on IDOU-YOLO,” *Journal of Applied Optics*, vol. 45, no. 4, pp. 723–731, 2024, doi: 10.1234/jao.2024.723.
[4] S. Kumar et al., “A novel YOLOv3 algorithm-based deep learning approach for waste segregation: towards smart waste management,” *Electronics*, vol. 10, no. 1, p. 14, 2021, doi: 10.3390/electronics10010014.
[5] S. Lin, M. Garratt, and A. Lambert, “Unmanned aerial vehicles autonomous landing at nighttime using monocular vision,” *Sensors*, vol. 21, no. 18, p. 6226, 2021, doi: 10.3390/s21186226.
[6] W. Liu et al., “SSD: Single shot multibox detector,” in *Proc. European Conference on Computer Vision (ECCV)*, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.
[7] W. Liu et al., “SSD: Single shot multibox detector,” arXiv preprint arXiv:1512.02325, 2015.
[8] G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics YOLO* (Version 8.0.0) [Computer software], 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.
[9] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” arXiv preprint arXiv:1406.4729, 2014.
[10] R. Khanam and M. Hussain, “YOLOv11: An overview of the key architectural enhancements,” arXiv preprint arXiv:2410.17725, 2024. [Online]. Available: https://arxiv.org/pdf/2410.17725.
[11] M.-H. Guo et al., “Visual attention network,” *Journal of LaTeX Class Files*, vol. 14, no. 8, pp. 1–11, 2022. [Online]. Available: https://arxiv.org/pdf/2202.09741.
[12] P. Zhu et al., “Detection and tracking meet drones challenge,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.
[13] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal speed and accuracy of object detection,” arXiv preprint arXiv:2004.10934, 2020. [Online]. Available: https://arxiv.org/pdf/2004.10934.
[14] A. S. Geetha, “What is YOLOv6? A deep insight into the object detection model,” arXiv preprint arXiv:2412.13006, 2024. [Online]. Available: https://arxiv.org/abs/2412.13006.
[15] A. Wang et al., “YOLOv10: Real-time end-to-end object detection,” arXiv preprint arXiv:2405.14458, 2024. [Online]. Available: https://arxiv.org/pdf/2405.14458.

YOLO11-LKACONV: ОПТИМИЗАЦИЯ ОБНАРУЖЕНИЯ НЕСКОЛЬКИХ ЦЕЛЕЙ НА СНИМКАХ БПЛА НА ОСНОВЕ УЛУЧШЕННОЙ АРХИТЕКТУРЫ YOLO

By Сяньи, Абламейко С. В.

В данной статье представлен YOLO11-LKAConv – улучшенная модель, построенная на основе фреймворка YOLOv11, которая направлена на решение проблемы низкой точности обнаружения маленьких целей на аэрофотоизображениях БПЛА. Заменяя стандартные сверточные слои на легковесные сверточные слои с большим ядром и вниманием (LKA-Conv), модель интегрирует динамические крупные рецептивные поля и механизмы канально-пространственного внимания, что усиливает ее способность захватывать долгосрочные контекстные признаки для маленьких целей. Эксперименты на датасете VisDrone2019 показывают, что улучшенная модель достигает mAP50-95 в 0,173, что на 1,2% выше, чем у базовой YOLOv11n, причем показатели точности (P) и mAP50 также улучшены, а время вывода осталось почти неизменным. Результаты указывают на то, что LKAConv благодаря своему легкому дизайну с большим ядром эффективно балансирует точность обнаружения и вычислительную эффективность, предлагая лучшее решение для задач реального времени по обнаружению целей на БПЛА.

Received 25.03.2025