Теоретические основы информатики

Theoretical foundations of computer science

УДК 004.93

RLD-YOLO: НОВЫЙ МЕТОД ОБНАРУЖЕНИЯ ОБЪЕКТОВ НА ИЗОБРАЖЕНИЯХ БЕСПИЛОТНЫХ ЛЕТАТЕЛЬНЫХ АППАРАТОВ С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ НЕЙРОННОЙ СЕТИ YOLOv11

ВУ СЯНЬИ[1)]***, С. В. АБЛАМЕЙКО***[1), 2)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*
[2)]*Объединенный институт проблем информатики НАН Беларуси, ул. Сурганова, 6, 220012, г. Минск, Беларусь*

Аннотация. Изображения, получаемые с беспилотных летательных аппаратов, в настоящее время широко используются во многих приложениях. Однако эти изображения сталкиваются с рядом проблем (плотное распределение мелких объектов, переменные масштабы объектов и незаметные контурные особенности), которые приводят к пропускам объектов и ложному обнаружению объектов. Для решения этих проблем в данной статье предлагается улучшенный алгоритм обнаружения объектов RLD-YOLO, разработанный на основе версии YOLOv11n семейства

Образец цитирования:
Ву Сяньи, Абламейко СВ. RLD-YOLO: новый метод обнаружения объектов на изображениях беспилотных летательных аппаратов с использованием модели нейронной сети YOLOv11. *Журнал Белорусского государственного университета. Математика. Информатика.* 2025;2:105–117 (на англ.).
EDN: RBEXHB

For citation:
Wu Xianyi, Ablameyko SV. RLD-YOLO: new method for object detection in unmanned aerial vehicle images using YOLOv11 neural network. *Journal of the Belarusian State University. Mathematics and Informatics.* 2025;2:105–117.
EDN: RBEXHB

Авторы:
Ву Сяньи – аспирант кафедры веб-технологий и компьютерного моделирования механико-математического факультета. Научный руководитель – С. В. Абламейко.
Сергей Владимирович Абламейко – доктор технических наук, академик НАН Беларуси, профессор; профессор кафедры веб-технологий и компьютерного моделирования механико-математического факультета[1)], главный научный сотрудник отдела интеллектуальных информационных систем[2)].

Authors:
Wu Xianyi, postgraduate student at the department of web-technologies and computer simulation, faculty of mechanics and mathematics.
tigerv5872@gmail.com
https://orcid.org/0009-0003-6976-5386
Sergey V. Ablameyko, doctor of science (engineering), academician of the National Academy of Sciences of Belarus, full professor; professor at the department of web-technologies and computer simulation, faculty of mechanics and mathematics[a], and chief researcher at the department of intelligent information systems[b].
ablameyko@bsu.by
https://orcid.org/0000-0001-9404-1206

алгоритмов YOLO. Алгоритм RLD-YOLO включает технологию структурной репараметризации RepConv, сохраняющую способность к многоветочной экспрессии признаков во время обучения и автоматически преобразующуюся в эффективную одноветвевую структуру во время вывода. Этот алгоритм разрабатывает модуль большого ядерного внимания LKAConv для улучшения способности к захвату признаков мелких целей с помощью глубокой разделяемой свертки размера 7 × 7 и механизма пространственного внимания. Алгоритм RLD-YOLO вводит динамический адаптивный модуль слияния DASI для оптимизации многоуровневого взаимодействия признаков с помощью обучаемого распределения весов. Экспериментальные результаты показывают, что улучшенный алгоритм обнаружения объектов RLD-YOLO, который объединяет модули LKAConv, RepConv и DASI, увеличивает на наборе данных VisDrone2019-DET значения mAP50 и mAP50-95 на 2,02 и 1,17 % соответственно. Скорость постобработки оптимизирована на 9,09 %. Хотя время предварительной обработки увеличивается из-за операций по улучшению признаков, критический этап вывода все еще поддерживает реальное время выполнения 1,7 мс/кадр. Алгоритм RLD-YOLO, интегрированный с модулями LKAConv, RepConv и DASI, очень подходит для задачи обнаружения мелких объектов на изображениях беспилотных летательных объектов.

Ключевые слова: обнаружение мелких объектов; YOLOv11; изображения БПЛА; LKAConv.

RLD-YOLO: NEW METHOD FOR OBJECT DETECTION IN UNMANNED AERIAL VEHICLE IMAGES USING YOLOv11 NEURAL NETWORK

WU XIANYI [a], ***S. V. ABLAMEYKO*** [a, b]

[a] *Belarusian State University, 4 Niezaliezhnasci Avenue, Minsk 220030, Belarus*
[b] *United Institute of Informatics Problems, National Academy of Sciences of Belarus, 6 Surganava Street, Minsk 220012, Belarus*

Corresponding author: Wu Xianyi (tigerv5872@gmail.com)

Abstract. Unmanned aerial vehicle images are widely used now in many applications. However, these images face many challenges such as dense distribution of small objects, variable object scales, and inconspicuous edge features, leading to missed and false object detection. To address these issues, this paper proposes a lightweight enhanced solution based on the improved YOLOv11n model RLD-YOLO. The algorithm combines RepConv structural reparameterisation technology, retaining multi-branch feature expression capabilities during training and automatically converting to an efficient single-branch structure during inference. It designs the LKAConv large kernel attention module to enhance the feature capture ability of small targets through 7 × 7 depthwise separable convolution and spatial attention mechanism. It introduces the DASI dynamic adaptive fusion module to optimise multi-scale feature interaction through learnable weight allocation. Experimental results show that the improved RLD-YOLO object detection algorithm, which integrates LKAConv, RepConv, and DASI, increases mAP50 and mAP50-95 by 2.02 and 1.17 % respectively on the VisDrone2019-DET dataset. The post-processing speed is optimised by 9.09 %. Although the pre-processing time increases due to feature enhancement operations, the critical inference stage still maintains a real-time performance of 1.7 ms/frame. The RLD-YOLO model, fused with LKAConv, RepConv, and DASI, is very suitable for the task of small object target detection in unmanned aerial vehicle images.

Keywords: small object detection; YOLOv11; UAV images; LKAConv.

Introduction

With the continuous development of technologies, object detection technology has been widely applied in many practical fields, such as smart cities [1], precision agriculture [2], and disaster relief [3]. In recent years, with the rapid development of unmanned aerial vehicle (UAV) technology, its presence in people's lives has also increased. Nowadays, UAVs are widely used in traffic patrol, environmental monitoring, maritime search and rescue, and other fields [4]. However, object detection from a UAV perspective faces unique challenges: low-light environments such as night [5], fog or dawn, dusk [6], and complex background interference, such as building shadows, vegetation occlusion [7], which seriously affect the detection accuracy and efficiency of target objects. With the increasing real-time detection requirements of UAV platforms, UAV object detection algorithms need higher precision and speed, posing significant challenges to the design and optimisation of object detection algorithms in low-light scenes from a UAV perspective.

Currently, the YOLO series of models with outstanding performance have been widely used in object detection tasks in multiple fields [8] (for example, remote sensing target detection [9], intelligent parking [10], etc.).

However, when performing object detection from a UAV perspective, the maintenance of high detection accuracy while keeping the model lightweight has become a research challenge [7]. H. Chen, et al. [11], proposed an IDOU-YOLO (infrared detection of UAV-YOLO) algorithm model for UAV target detection based on thermal imaging, which improved the detection accuracy and convergence speed of the model by constructing a multi-scale fusion feature pyramid mechanism and introducing the bounding box loss function SIoU (smooth intersection over union). The researcher Z. Zhang [12] proposed a multi-scale UAV image target detection algorithm Drone-YOLO based on the YOLOv8 model. This method uses a three-layer path aggregation feature pyramid network (PAFPN) structure and combines large-scale feature maps with a detection head customised for small-sized objects, significantly enhancing the algorithm's ability to detect small-sized targets. Y. Huang, et al. [13], proposed a real-time detection algorithm for urban low-altitude multi-scale UAV images. It is used for UAV detection tasks with different image features during the day and night. During the day, the paper proposed a defogging detection structure to solve the detection problem in foggy environments. At night, the paper proposed a squeeze-and-excitation backbone (SE-backbone) [14] structure and SPD-PAFPN (spatial-to-depth PAFPN) [15] structure with a feature pyramid network, to obtain effective information from deeper feature maps for UAV detection in low-resolution images. A. He, et al. [16], proposed an ALSS-YOLO (adaptive lightweight channel splitting and shutting YOLO model) detection architecture based on the improved YOLOv8, which designed an ALSS module that adopts an adaptive channel segmentation strategy to optimise feature extraction and integrates a channel shuffle mechanism to enhance channel-wise information exchange. It improved the detection accuracy of blurred targets, especially when dealing with blurred and overlapping targets caused by jitter. S. Liu, et al. [17], proposed the LI-YOLO (low-illumination YOLO model) algorithm by improving YOLOv8, which proposed a feature enhancement module (FEB) and embedded the FEB into the C2f module at the end of the backbone network to enhance the algorithm's feature extraction ability. For the problems of low brightness, high noise, and blurry details in low-light images, the feature enhancement block and adaptive spatial feature fusion structure were used to improve the target detection performance in low-light scenes. X. Wu, et al. [18], proposed solutions from the perspective of deep learning models for three research directions: object detection in images, object detection in videos, and object tracking in videos. Currently, open datasets have been widely used for UAV target detection and tracking research, and performance evaluation has been carried out through four benchmark datasets.

The current mainstream of small target detection methods have the following problems. Firstly, the impact of dense small targets on object detection: small targets, due to their small size and low pixel ratio, have fewer grid points on the feature map, leading to inconspicuous recognisable features in the image, limiting the perception ability of the detection network for these targets, making it difficult for the model to learn enough discriminative information [19]. Secondly, the impact of complex backgrounds on object detection: images captured by UAVs usually contain complex backgrounds, such as trees, buildings, and changing terrain. These complex backgrounds may be similar to the features of small targets, increasing the probability of the model misjudging the background noise as the target. In addition, the high-speed movement of UAVs may cause image blurring, further reducing the detection accuracy of small targets [20]. Thirdly, UAV systems usually require detection algorithms to have high real-time performance to quickly respond and execute tasks [21]. However, detecting small targets usually requires more complex model structures or more computing resources, which conflicts with the requirement of real-time performance. To maintain real-time performance, some model complexity may need to be sacrificed, thereby affecting the detection accuracy.

In response to the above problems, this paper proposes an improved YOLOv11n aerial photography lightweight small target detection algorithm RLD-YOLO (RepConv-LKAConv-DASI-YOLOv11n). The model designs a reparameterisation convolution (RepConv) structure in the first layer of the backbone, which achieves performance separation in training and deployment stages through multi-branch convolution fusion and single-branch inference reconstruction technology, reducing the computational load of the first layer while retaining feature expression capabilities. To meet the demand for long-distance dependency modelling in complex scenes, the large kernel attention convolution (LKAConv) module is introduced in the deep feature extraction link, using a joint decomposition strategy of depthwise convolution and dilated convolution (5 × 5 depthwise convolution + 7 × 7 dilated rate 3 convolution) to build an equivalent 21 × 21 ultra-large receptive field, combined with spatial attention feature reweighting mechanism, to enhance the model's ability to capture features of dense small targets and occluded targets, and improve the detection accuracy of small targets. In complex environments, background noise often affects the detection of small targets. The DASI (dimension-aware selective integration) dynamic adaptive interaction unit is deployed in the head feature fusion stage, which achieves intelligent weight allocation and cross-level detail calibration of multi-scale features through the dual-path architecture of channel attention and spatial cross-correlation, enhancing the model's adaptability to dense targets and scale changes, and effectively suppressing the interference of background noise such as clouds and vegetation in aerial photography scenes. The RLD-YOLO model, through efficient inference of RepConv, global

perception of LKAConv, dynamic fusion of DASI, and collaborative optimisation, achieves dual improvements in precision and speed on the basis of maintaining the detection characteristics of YOLOv11n, providing an efficient solution for UAV target detection.

YOLOv11 model

YOLOv11 model (fig. 1) is a new generation of object detection algorithm, launched by company «Ultralytics» in 2023, aiming to further improve the precision and efficiency of the object detection. It has made many improvements on the basis of YOLOv8[1] to adapt to a wider range of application scenarios and improve model performance. YOLOv11 provides multiple versions of different scales, including YOLOv11n (ultra-lightweight), YOLOv11s (small), YOLOv11m (medium), YOLOv11l (standard), and YOLOv11x (extra-large) to meet different needs. Compared with previous versions of YOLO, YOLOv11 has made improvements in the following aspects:

1) backbone network. YOLOv11 introduced the C3k2 module [22], replacing the C2f module in YOLOv8. The C3k2 module uses smaller convolution kernels to improve computational efficiency while maintaining performance. It retained the spatial pyramid pooling fast (SPPF) [23] module and introduced the cross-stage partial and spatial attention (C2PSA) module [24], enhancing the spatial attention of feature maps and improving detection accuracy;

2) neck structure. In the neck structure, YOLOv11 replaced the C2f module with the C3k2 module, improving the speed and performance of feature aggregation. Through the C2PSA module, it enhanced spatial attention, enabling the model to focus more effectively key areas in the image and improve the detection accuracy of small targets and partially occluded targets;

3) head structure. In the head structure, YOLOv11 used multiple C3k2 modules to process and optimise feature maps, improving the model's detection accuracy.

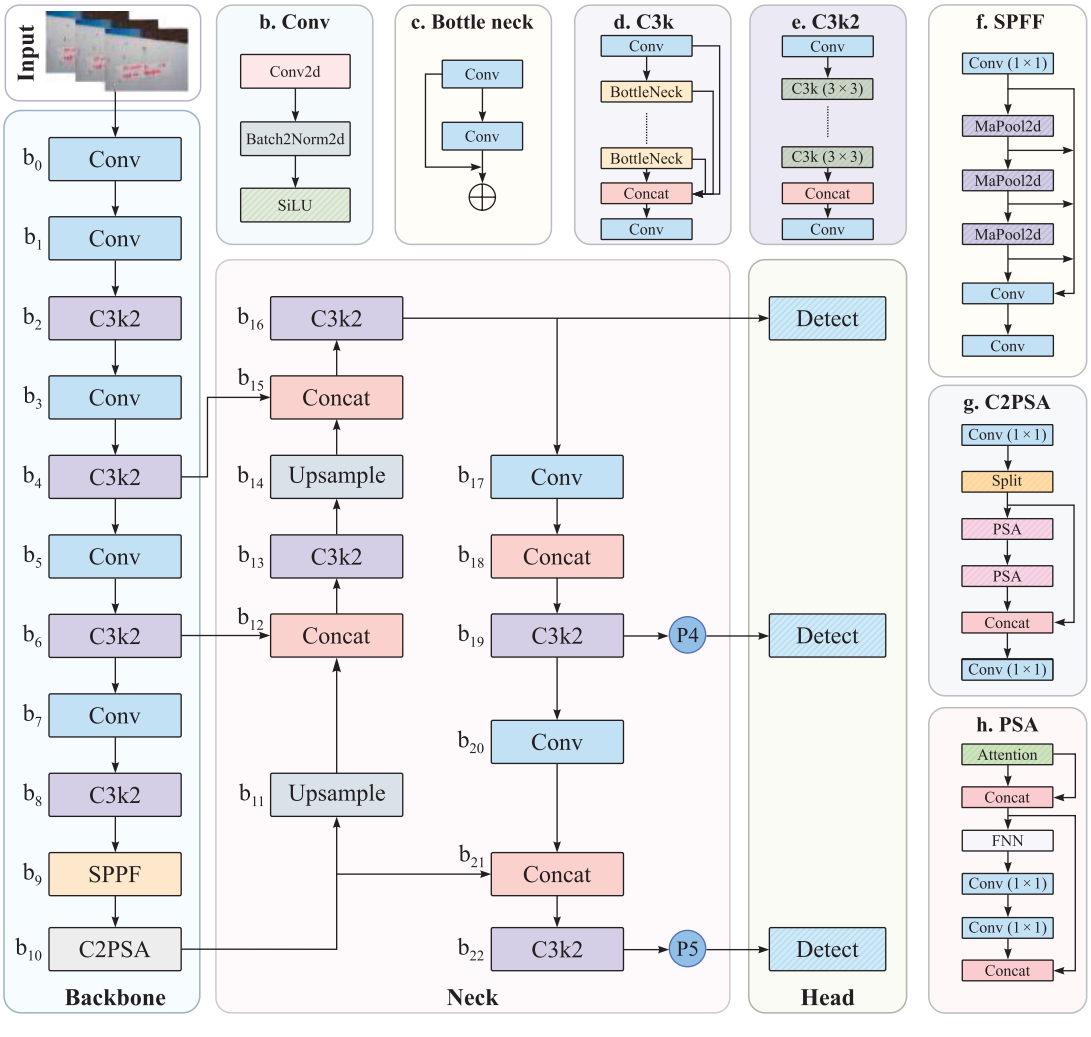

Fig. 1. YOLOv11 network structure diagram

[1] *Jocher G., Chaurasia A., Qiu J.* Ultralytics YOLO (version 8.0.0) [Electronic resource]. URL: https://github.com/ultralytics/ultralytics (date of access: 12.01.2025).

Meanwhile, YOLOv11 model adopted multi-scale training and data augmentation techniques during training, further improving the model's generalisation ability and detection accuracy. Compared with its predecessors, YOLOv11 model has made significant improvements in inference speed and accuracy. In summary, YOLOv11 model has made significant progress in the precision and efficiency of object detection through the introduction of innovative technologies such as the C3k2 module and the C2PSA module. It not only performs well in models of different scales but also demonstrates strong adaptability and practicality in various application scenarios.

RLD-YOLO model

This paper takes YOLOv11n as the baseline and targets the small target detection problem in UAV aerial photography images. YOLOv11n still has high rates of missed and false detections in complex backgrounds, dense targets, and small target detection. Therefore, an improved model RLD-YOLO based on YOLOv11n is proposed. The model significantly improves detection accuracy and robustness while maintaining lightweight characteristics. The specific improvements are as follows:

1) RepConv structural reparameterisation technology. The RepConv module is introduced in the initial layer and key paths of the backbone. RepConv uses a multi-branch structure during training to enhance feature expression capabilities and merges into a single 3 × 3 convolution through structural reparameterisation during inference, reducing computational load;

2) LKAConv large kernel attention module. The LKAConv module is introduced in the backbone and head parts, combining 7 × 7 depthwise separable convolution with a spatial attention mechanism to expand the receptive field to 120 × 120 pixels. LKAConv can effectively capture the context information of small targets, reducing the missed detection rate;

3) DASI dynamic adaptive fusion module. The DASI module is introduced in the neck part, dynamically adjusting the fusion ratio of multi-scale features through learnable weights. DASI optimises the static fusion defect of the traditional FPN + PAN structure and enhances the model's adaptability to dense targets and scale changes;

4) detection and feature enhancement. The head part retains three detection layers P3, P4, and P5, and dynamically fuses multi-scale features through the DASI module to enhance the model's ability to recognise targets of different scales;

5) lightweight design and efficiency optimisation. The C3k2 module of YOLOv11 is retained to reduce model parameters and computational load. A dynamic data augmentation strategy is adopted to improve the model's accuracy for scale changes and occlusions in UAV perspectives.

The improved model network structure is shown in fig. 2.

RepConv. The multi-branch structure of RepConv can capture richer feature patterns during training, enhancing the model's discriminative ability for complex backgrounds and small targets. In the RLD-YOLO model, RepConv is introduced in the initial layer, which is responsible for extracting basic image features such as edges and textures. Through convolution operations, the input image is convolved with the convolution kernel using a sliding window calculation to generate feature maps, which are crucial for the construction of subsequent high-level semantic information. The core idea of RepConv is to use a multi-branch structure during training to enhance feature expression capabilities and merge the branches into a single convolution through mathematical equivalence transformations during inference, reducing computational load. During the inference stage, RepConv merges the above branches into an equivalent 3 × 3 convolution kernel through structural reparameterisation. This process can be regarded as reparameterising the basic weights, enabling the new convolution kernel to learn more diverse representations. The computational load is the same as that of standard convolution, but the feature expression capability is stronger. The principle diagram of RepConv is shown in fig. 3.

The core idea of RepConv is to enhance the model's representation capability by establishing connections between convolution kernel parameters. For example, in the depthwise separable convolution, each convolution kernel channel focuses only on one channel of the input feature map. RepConv uses refocusing transformation to enable each convolution kernel channel to focus on the features of other channels, thereby learning richer representations. This process can be described by the following formula:

$$W_t = T\left(W_b, W_r\right),$$

where W_b is the basic weight; W_r is the trainable parameter of the refocusing transformation; T is the refocusing transformation function; W_t is the generated new convolution kernel parameter [25].

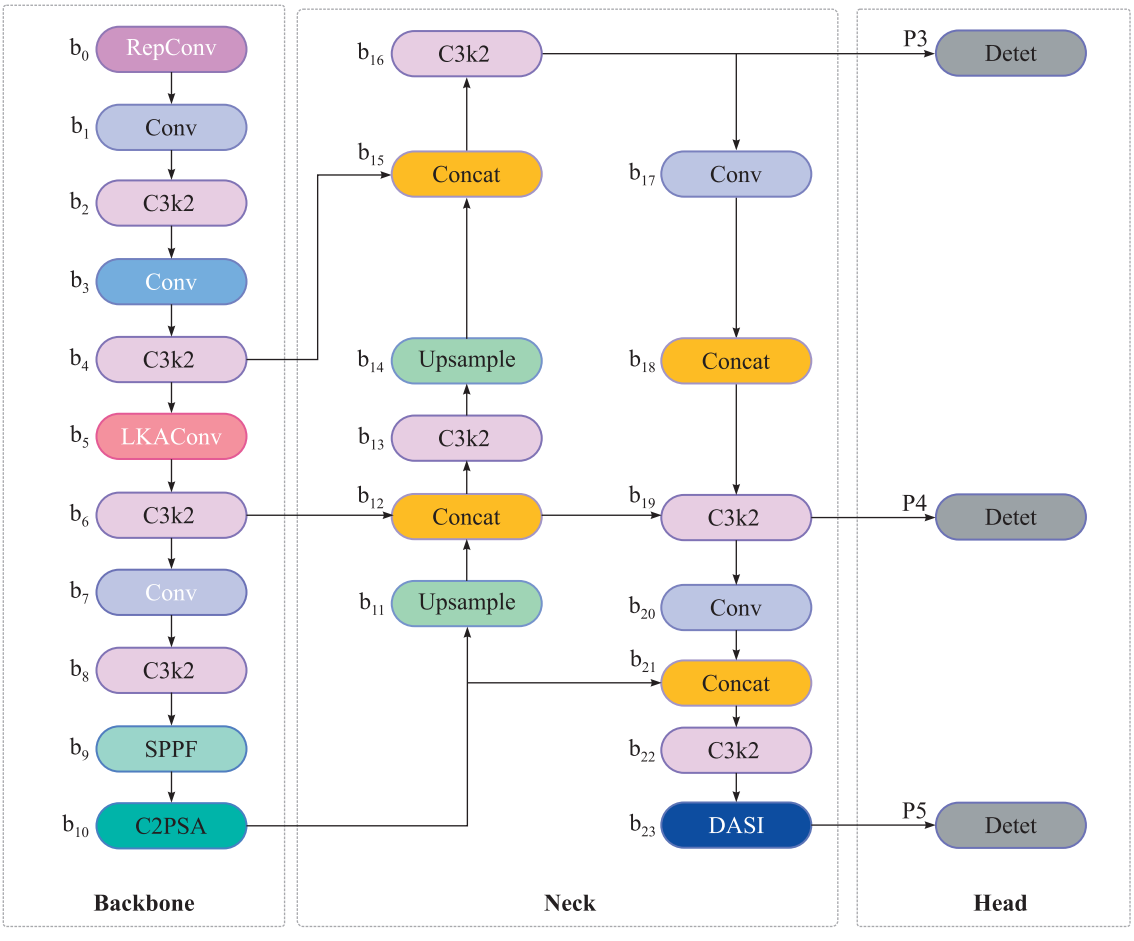

Fig. 2. RLD-YOLO network structure diagram

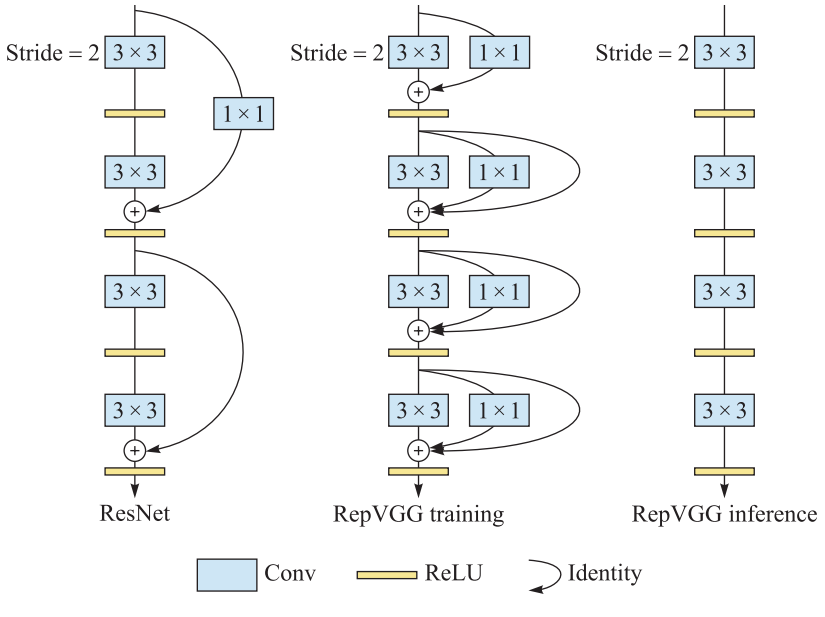

Fig. 3. RepConv principle diagram

LKAConv. LKAConv is a key component of the LKA mechanism, used to implement the decomposition of large convolution kernels. LKA 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 large convolution kernels is shown in fig. 4 [26].

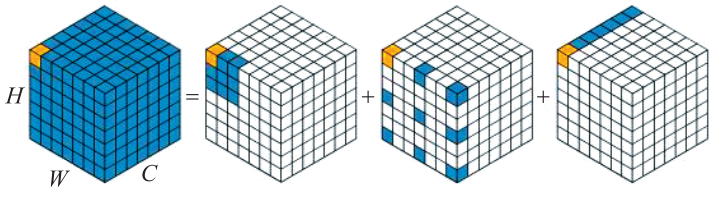

Fig. 4. Decomposition diagram of large kernel convolution

The core formula of LKA is as follows:

$$\text{Attention} = \text{Conv}_{1\times 1}\Big(\text{DW-D-Conv}\big(\text{DW-Conv}\big(F\big)\big)\Big),$$

$$\text{Output} = \text{Attention} \otimes F,$$

where F is the input feature map; DW-Conv represents depthwise separable convolution; DW-D-Conv represents dilated depthwise separable convolution; $\text{Conv}_{1\times 1}$ represents 1×1 convolution; $\otimes$ represents element-wise multiplication.

In this study, we introduced LKAConv 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 fifth layer adopted LKAConv with parameter settings of 512 input channels, a 3×3 convolution kernel, and a dilation rate of two. 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.

DASI. Since small targets occupy fewer pixels in the image and the background is complex, high-dimensional features may lose information about small targets during multiple downsampling processes, while low-dimensional features may not provide sufficient contextual information. The DASI module enhances the model's ability to capture features of different scales by adaptivel selecting and fusing features of different dimensions, increasing the saliency of small targets and thus improving detection performance. The main function of the DASI module is to perform selective fusion of features at different dimensions of the feature map. The principle diagram is shown in fig. 5.

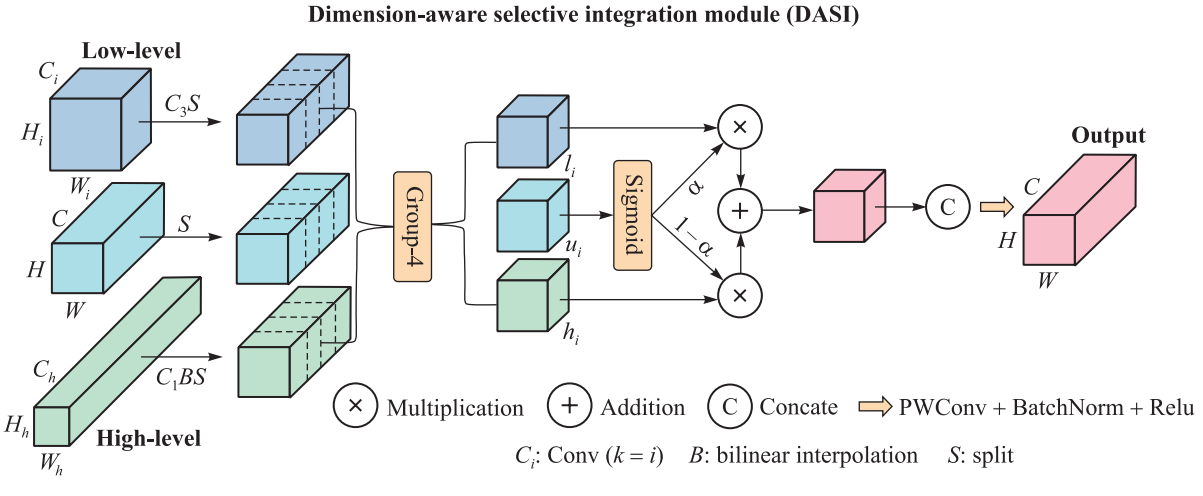

Fig. 5. Detail structure of DASI

Specifically, it aligns high-dimensional and low-dimensional feature maps and achieves feature adaptive fusion through channel splitting and selective aggregation [27]. This process can be described by the following formula:

$$\alpha = \sigma(u_i),\ u_i' = \alpha l_i + (1-\alpha)h_i,$$
$$F_u' = [u_1',\ u_2',\ u_3',\ u_4'],\ \hat{F}_u = \delta\left(B\left(\mathrm{Conv}\left(F_u'\right)\right)\right),$$

where σ represents the sigmoid activation function; u_i, l_i and h_i represent the i channel segment of the current layer, low-dimensional, and high-dimensional feature maps respectively; α is the weight calculated from the current layer's features, used to control the fusion ratio of low-dimensional and high-dimensional features; F_u' is the fused feature map; δ and B represent the ReLU (rectified linear unit) activation function and batch normalisation operation respectively; Conv represents the convolution operation.

Through this design, the DASI module can adaptively select appropriate features for the fusion based on the size and characteristics of the target, thereby enhancing the model's ability to detect small targets. In UAV dataset detection tasks, the DASI module effectively addresses the problems of small target loss and background interference, significantly improving the model's detection accuracy and maintaining stable functionality.

Dataset

The experiments are based on the VisDrone2019 [28] UAV object detection dataset, which is a large-scale UAV perspective dataset obtained by the AISKYEYE team from the machine learning and data mining laboratory at Tianjin University. The dataset contains 10 209 images with a total of 471 266 annotated targets covering 13 categories (such as pedestrians, cars, trucks, buses, etc.). The validation set contains 38 759 instances, with small targets (pixel area is less than 32 × 32) accounting for 68.2 %, and dense scenes (more than 100 targets per image) accounting for 45 %, fully reflecting the detection challenges of UAV perspectives.

Experimental results and analysis

Experimental environment and configuration. The experimental platform is a Windows (version 11) 64-bit operating system, and the experiment is based on the deep learning framework PyTorch (version 2.4.0) and the corresponding CUDA (version 2.4.1). We used an NVIDIA GeForce RTX 4060 GPU for training, with the development language Python (version 3.9) and the CUDA (version 12.41). The specific environmental configuration parameters are shown in table 1. Parameters not provided in this paper use the default parameters of the official YOLOv11n.

Table 1

Experimental configuration

Parameters	Values
epochs:	100
batch:	16
imgsz:	640
device:	0
optimizer:	auto
amp:	TRUE

Comparison of different attention mechanisms. In the experiments with different attention mechanisms, EfficientNetv2 is a lightweight attention mechanism [29] widely used in other object detection studies. However, in this experiment, the EfficientNetv2 model had the shortest post-processing time (0.8 ms), but its accuracy was significantly lower than other had. EMA is an efficient multi-scale attention module [30] that enhances the ability to fuse multi-scale features. In this experiment, the EMA module had a longer inference time (1.9 ms), but its accuracy was higher, suitable for scenarios with low real-time requirements. Squeeze-and-excitation (SE) channel attention mechanism [31] mainly improves the model's performance by compressing and exciting the input features. In this experiment, the SE module had the shortest inference time (1.6 ms), but its accuracy was low (mAP50 = 29.5). Simple attention module (SimAM) is an attention mechanism based on the local self-similarity of feature maps [32]. It dynamically adjusts the weight of each pixel by calculating the similarity between each pixel and its surrounding pixels in the feature map, thereby enhancing important features and suppressing irrelevant features. In this experiment, the overall performance of SimAM was not outstanding. StokenAttention is a method that improves the efficiency of capturing global dependencies by sampling super-tokens from

visual tokens through sparse associative learning [33], but the improvements were not significant enough. SwinTransformer is a transformer module based on the window self-attention [34], suitable for long-range dependency modelling. SwinTransformer had the highest computational load (16.6 GFLOPs), and although its accuracy was high, its computational complexity limited its deployment on edge devices. RLD-YOLO had moderate parameter quantity (3.21 mln) and computational load (6.9 GFLOPs), suitable for real-time application scenarios. RLD-YOLO maintained high accuracy (mAP50 = 30.3) while keeping the inference time the same as the baseline model (1.7 ms), performing excellently in terms of accuracy, inference time, and computational efficiency. In summary, the RLD-YOLO model proposed in this paper, through the collaborative design of large kernel attention, dynamic fusion, and structural reparameterisation, provides an efficient solution for UAV object detection.

Table 2

Comparison of different attention mechanisms on the VisDrone2019-DET dataset

Model	mAP50, %	mAP50-95, %	Inference, ms	Postprocess per image, ms	Parameters, mln	GFLOPs
YOLOv11n (baseline)	29.7	17.1	1.7	1.1	2.58	6.3
YOLOv11n-EfficientNetv2	18.8 ↓	10.3 ↓	1.9 ↓	0.8 ↑	2.1 ↑	3.0 ↑
YOLOv11n-EMA_attention	30.2 ↑	17.1	1.9 ↓	1.2 ↓	2.58	6.3
YOLOv11n-SE_attention	29.5	16.9	1.6 ↑	1.2 ↓	2.58	6.3
YOLOv11n-SimAM	29.4	16.8	1.8	1.1	2.58	6.3
YOLOv11n-StokenAttention	29.9	17.0	1.8	1.1	2.85	6.5
YOLOv11n-SwinTransformer	29.9	17.2 ↑	1.8	1.2 ↓	2.91	16.6 ↓
RLD-YOLO	30.3 ↑	17.2 ↑	1.7	1.0	3.21	6.9

Cross-model comparison experiment. To verify the comprehensive performance of the proposed RLD-YOLO model in UAV object detection tasks, we compared it with mainstream lightweight versions of the YOLO series, including YOLOv5 [35], YOLOv6 [36], YOLOv8, YOLOv10 [37], and YOLOv11. Through the experiments, we found that in terms of accuracy, RLD-YOLO achieved an mAP50 of 30.3 %, which is 0.6 % higher than the baseline YOLOv11n, and 1.2 % higher than YOLOv8n and YOLOv10n respectively. The mAP50-95 metric (17.2 %) was also the best, indicating stronger performance in complex scenes (such as occlusion, small targets). The inference time of RLD-YOLO was 1.7 ms, the same as YOLOv11n and YOLOv6n, but the post-processing time was optimised to 1.0 ms (a 9.1 % reduction compared to YOLOv11n). RLD-YOLO had a parameter quantity of 3.21 mln and a computational load of 6.9 GFLOPs, significantly lower than YOLOv5n (11.8 GFLOPs) and YOLOv6n (11.5 GFLOPs). RLD-YOLO achieved the best balance between precision and efficiency while maintaining real-time performance (1.7 ms/frame) through dynamic fusion and structural reparameterisation techniques (table 3).

Table 3

Comparison of different models on the VisDrone2019-DET dataset

Model	P, %	R, %	mAP50, %	mAP50-95, %	Speed preprocess, ms	Inference, ms	Postprocess per image, ms	Parameters, mln	GFLOPs
YOLOv5n	38.7	27.9	27.3	15.4	0.2	1.7	3.2	4.23	11.8
YOLOv6n	36.3	28	27.1	15.6	0.2	1.7	1.3	4.16	11.5
YOLOv8n	39.8	30.2	29.1	16.5	0.2	2.0	0.6	2.70	8.2
YOLOv10n	40.2	29.9	29.1	16.3	0.2	1.4	3.3	2.50	7.1
YOLOv11n	40.2	30.8	29.7	17.1	0.1	1.7	1.1	2.58	6.3
RLD-YOLO	41.5 ↑	31.1 ↑	30.3 ↑	17.2	0.2	1.7	1.0	3.21	6.9

YOLOv11n comparison experiment. UAVs have high real-time requirements for small target detection tasks, making lightweight model design very necessary. The goal in designing a lightweight model is to reduce the model size and computational load while maintaining or improving the model's detection accuracy as much as possible. To verify the effectiveness of the proposed modules (LKAConv, RepConv, DASI) for UAV object detection tasks, we introduced different modules step by step on the YOLOv11n baseline model and designed the following ablation experiment variants:

1) YOLOv11n. The baseline model is without any improvement modules;
2) YOLOv11n + LKAConv. The LKAConv large kernel attention module is integrated into the backbone;
3) YOLOv11n + RepConv. The RepConv structural reparameterisation technology is used in the initial layer;
4) YOLOv11n + DASI. The DASI dynamic adaptive fusion module is introduced in the neck part;
5) RLD-YOLO. The combined use of LKAConv, RepConv, and DASI modules.

Table 4

Experimental results for YOLOv11 for VisDrone2019-DET dataset

Model	mAP50, %	ΔmAP50, %	mAP50-95, %	Inference, ms	Parameters, mln	GFLOPs
YOLOv11n (baseline)	29.7	–	17.1	1.7	2.58	6.3
YOLOv11n + LKAConv	30.1	1.35 ↑	17.3 ↑	2.0	2.64	6.7
YOLOv11n + RepConv	29.8	0.34 ↑	17.1	1.8	2.58	6.3
YOLOv11n + DASI	29.9	0.67 ↑	17.0	1.9	3.15	6.5
RLD-YOLO	30.3	2.02 ↑	17.2 ↑	1.7	3.21	6.9

The LKAConv module improved the accuracy by 1.35 % (29.7 → 30.1), indicating that the large kernel attention mechanism effectively enhanced the feature extraction capability of small targets. However, the inference time increased up to 2.0 ms (+17.6 %), and the computational load increased up to 6.7 GFLOPs, mainly due to the additional overhead of the 7 × 7 depthwise separable convolution. The RepConv module had the same parameter quantity as the baseline, with an inference time of 1.8 ms (+5.9 %) and an mAP50 improvement of only 0.34 % (29.7 → 29.8), indicating that using RepConv alone had limited accuracy gains and required the synergy with other modules. The DASI module improved mAP50-95 by 0.6 % (17.1 → 17.2), indicating that dynamic weight allocation optimised multi-scale feature interaction. However, mAP50-95 decreased. After jointly using the three modules, the RLD-YOLO model significantly improved mAP50 by 2.02 % (29.7 → 30.3), and the inference time returned to the baseline level (1.7 ms), with the final computational load only increasing up to 6.9 GFLOPs (+9.5 %). The experiment showed that the synergistic design of LKAConv and RepConv achieved a 2.02 % improvement in mAP50 over the baseline model while maintaining real-time performance. It is worth noting that introducing the DASI module alone increased latency, but this negative impact could be completely offset by combining it with structural reparameterisation technology. In summary, the model proposed in this paper is better suited for small target detection tasks in UAV images.

Object detection experiments. To demonstrate more intuitively the detection effect of the proposed RLD-YOLO method, YOLOv11n and RLD-YOLO models were used to detect several different UAV aerial photography scenes in the VisDrone2019-DET dataset, and the detection effect comparison is shown in fig. 6. For small target dense distribution scenes, as it is shown in fig. 6, *a*, the improved algorithm can detect smaller targets further away compared to the YOLOv11n algorithm, such as the pedestrians in the upper right corner of the image. For small target detection in scenes with background interference, as it is shown in fig. 6, *b*, the RLD-YOLO algorithm can detect more targets. For dark light scenes, as it is shown in fig. 6, *c*, RLD-YOLO has a higher accuracy, such as the building on the left side of the road in the lower left corner of the image, which YOLOv11n detected as a car. For scenes with occluded targets, as it is shown in fig. 6, *d*, RLD-YOLO has a higher accuracy, and even if a part of an object is in the shadow or blocked by a large tree, RLD-YOLO can better detect the type of the object. In scenes with large scale changes, as it is shown in fig, 6, *e*, for small targets and very small targets such as bicycles, motorcycles, and pedestrians on both sides, the recognition accuracy has been improved to a certain extent. The YOLOv11n algorithm has problems of missed and false detections for small targets, while RLD-YOLO can make up for this problem.

Fig. 6. Detection effect comparison diagrams in various scenes:
small target-intensive scenarios (*a*), complex background scenes (*b*),
dark light scenes (*c*), target obscured scenes (*d*), multi-scale target scenarios (*e*).
The left side is the original image, the middle one is the algorithm detection effect diagram
of YOLOv11n, and the right side is the algorithm detection effect diagram of RLD-YOLO

Conclusions

This paper proposes an improved YOLOv11n lightweight small object detection algorithm for UAV images RLD-YOLO. The core innovation of RLD-YOLO lies in the collaborative design of three modules. Through RepConv structural reparameterisation technology, the training and inference stages are decoupled, enhancing feature expression capabilities while maintaining inference efficiency. Through the LKAConv large kernel attention module's depthwise separable convolution and spatial attention mechanism, the receptive field is expanded, significantly enhancing the feature capture ability of small targets. Finally, the DASI dynamic adaptive fusion module is introduced, with learnable weights dynamically adjusting the fusion ratio of multi-scale features, optimising the static fusion defect of traditional FPN, and improving the recall rate in dense scenes.

Compared with the original YOLOv11n model, RLD-YOLO has comprehensively outperformed YOLOv11n in performance. The YOLOv11n model, targeting small target detection in UAV aerial photography images, adaptability to dense scenes, and the demand for edge device deployment, has achieved collaborative optimisation of precision and efficiency through modular innovation design and technological integration. However, the model still faces some challenges and issues that need to be addressed. Specifically, there are the following problems, Firstly, the bottleneck of extremely small target detection, with a high missed detection rate for targets smaller than 16 × 16 pixels, which requires further optimisation through super-resolution preprocessing. Secondly, the adaptability to complex weather conditions, such as maintaining high detection accuracy in heavy rain and dense fog scenes. Future work will focus on improving the feature extraction capabilities of the network model to capture more subtle and distinctive features, in order to improve the classification performance for similar targets and increase the detection accuracy of extremely small targets, reducing missed detections.

References

1. Kumar S, Yadav D, Gupta H, Verma OP, Ansari IA, Ahn CW. A novel YOLOv3 algorithm-based deep learning approach for waste segregation: towards smart waste management. *Electronics*. 2021;10(1):1–20. DOI: 10.3390/electronics10010014.

2. Peng C, Vougioukas SG. Deterministic predictive dynamic scheduling for crop-transport co-robots acting as harvesting aids. *Computers and Electronics in Agriculture*. 2020;178:105742. DOI: 10.1016/j.compag.2020.105742.

3. Wu K, Wang X. Aligning pixel values of DMSP and VIIRS nighttime light images to evaluate urban dynamics. *Remote Sensing*. 2022;11(12):1463. DOI: 10.3390/rs11121463.

4. Klemas VV. Sensing from unmanned aerial vehicles: an overview. *Journal of Coastal Research.* 2015;31(5):1260–1267. DOI: 10.2112/JCOASTRES-D-15-00005.1.

5. Lin S, Jin L, Chen Z. Real-time monocular vision system for UAV autonomous landing in outdoor low-illumination environments. *Sensors.* 2021;21(18):6226. DOI: 10.3390/s21186226.

6. Zhang Y, Carballo A, Yang H, Takeda K. Perception and sensing for autonomous vehicles under adverse weather conditions: a survey. *Robotics and Autonomous Systems*. 2023;196:146–177. DOI: 10.1016/j.isprsjprs.2022.12.021.

7. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. SSD: single shot multibox detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. *Computer vision – ECCV2016. Proceedings of the 14th European conference on computer vision; 2016 October 11–14; Amsterdam, the Netherlands.* Cham: Springer; 2016. p. 21–37. DOI: 10.1007/978-3-319-46448-0_2.

8. Ouyang D, He S, Zhang G, Luo M, Guo H, Zhan J, et al. Efficient multi-scale attention module with cross-spatial learning. In: Institute of Electrical and Electronics Engineers. *Proceedings of the 2023 IEEE International conference on acoustics, speech and signal processing; 2023 June 4–10; Rhodes Island, Greece.* [S. l.]: Institute of Electrical and Electronics Engineers; 2023. p. 1–5. DOI: 10.1109/ICASSP49357.2023.10096516.

9. Wu X, Ablameyko SV. Efficient detection of building in remote sensing images using an improved YOLOv10 network. *Informatics*. 2025;22(2):33–47. DOI: 10.37661/1816-0301-2025-22-2-33-47.

10. Zhang S, Ma G, Yang W, Zuo F, Ablameyko SV. Car parking detection in images by using a semi-super-vised modified YOLOv5 model. *Journal of the Belarusian State University. Mathematics and Informatics.* 2023;3:72–81. EDN: XVDRSN.

11. Chen H, Liu D, Yan X. Infrared image UAV target detection algorithm based on IDOU-YOLO. *Journal of Applied Optics.* 2024; 45(4):723–731. DOI: 10.5768/JAO202445.0402001.

12. Zhang Z. Drone-YOLO: an efficient neural network method for target detection in drone images. *Drones*. 2023;7(8):526. DOI: 10.3390/drones7080526.

13. Huang Y, Qu J, Wang H, Yang J. An all-time detection algorithm for UAV images in urban low altitude. *Drones*. 2024;8(7):332. DOI: 10.3390/drones8070332.

14. Sunkara R, Luo T. No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects. In: Amini MR, Canu S, Fischer A, Guns T, Kralj Novak P, Tsoumakas G, editors. *Machine learning and knowledge discovery in databases (ECML PKDD 2022). Proceedings of the European conference; 2022 September 19–23; Grenoble, France. Part 3*. Cham: Springer; 2023. p. 443–459 (Goebel R, Wahlster W, Zhou Z-H, editors. Lecture notes in computer science; volume 13715). DOI: 10.1007/978-3-031-26409-2_27.

15. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Computer Vision Foundation. *Proceedings of the IEEE conference on computer vision and pattern recognition; 2018 June 18–22; Salt Lake City, USA.* Salt Lake City: Computer Vision Foundation; 2018. p. 7132–7141.

16. He A, Li X, Wu X, Su C, Chen J, Xu S, et al. ALSS-YOLO: an adaptive lightweight channel split and shuffling network for TIR wildlife detection in UAV imagery. arXiv:2409.06259 [Preprint]. 2024 [cited 2024 October 20]: [19 p.]. Available from: https://arxiv.org/abs/2409.06259.

17. Liu S, He H, Zhang Z, Zhou Y. LI-YOLO: an object detection algorithm for UAV aerial images in low-illumination scenes. *Drones*. 2024;8(11):653. DOI: 10.3390/drones8110653.

18. Wu X, Li W, Hong D, Tao R, Du Q. Deep learning for unmanned aerial vehicle-based object detection and tracking: a survey. *IEEE Geoscience and Remote Sensing Magazine.* 2022;10(1):91–124. DOI: 10.1109/MGRS.2021.3115137.

19. Lyu Y, Zhang T, Li X, Liu A, Shi G. LightUAV-YOLO: a lightweight object detection model for unmanned aerial vehicle image. *Journal of Supercomputing.* 2025;81:105. DOI: 10.1007/s11227-024-06611-x.

20. Chen N, Li Y, Yang Z, Lu Z, Wang S, Wang J. LODNU: lightweight object detection network in UAV vision. *Journal of Supercomputing.* 2023;79:10117–10138. DOI: 10.1007/s11227-023-05065-x.

21. Sun W, Dai L, Zhang X, Chang P, He X. RSOD: real-time small object detection algorithm in UAV-based traffic monitoring. *Applied Intelligence.* 2022;52:8448–8463. DOI: 10.1r007/s10489-021-02893-3.

22. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. *Computer vision – ECCV2014. Proceedings of the 13th European conference on computer vision; 2014 September 6–12; Zurich, Switzerland.* Cham: Springer; 2014. p. 346–361. DOI: 10.1007/978-3-319-10578-9_23.

23. Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements. arXiv:2410.17725 [Preprint]. 2024 [cited 2024 December 17]: [9 p.]. Available from: https://arxiv.org/pdf/2410.17725.

24. Ding X, Zhang X, Ma N, Han J, Ding G, Sun J. RepVGG: making VGG-style convnets great again. arXiv:2101.03697 [Preprint]. 2021 [cited 2024 November 14]: [10 p.]. Available from: https://arxiv.org/pdf/2101.03697.

25. Cai Z, Ding X, Shen Q, Cao X. RefCovn: re-parameterized refocusing convolution for powerful ConvNets. arXiv:2310.10563 [Preprint]. 2023 [cited 2024 December 21]: [18 p.]. Available from: https://arxiv.org/pdf/2310.105603.

26. Guo M-H, Lu CZ, Liu ZN, Cheng MM, Hu SM. Visual attention network. arXiv:2202.09741 [Preprint]. 2022 [cited 2024 October 25]: [12 p.]. Available from: https://arxiv.org/pdf/2202.09741.

27. Xu S, Zheng S, Xu W, Xu R, Wang C, Zhang J, et al. HCF-Net: hierarchical context fusion network for infrared small object detection. arXiv:2403.10778 [Preprint]. 2024 [cited 2025 January 20]: [6 p.]. Available from: https://arxiv.org/pdf/2403.10778.

28. Zhu P, Wen L, Du D, Bian X, Fan H, Hu Q, et al. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2021;44(11):7380–7399. DOI: 10.1109/TPAMI.2021.3119563.

29. Tan M, Le QV. EfficientNetv2: smaller models and faster training. arXiv:2104.00298 [Preprint]. 2021 [cited 2024 December 18]: [11 p.]. Available from: https://arxiv.org/pdf/2104.00298.

30. Ouyang D, He S, Zhang G, Luo M, Guo H, Zhan J, et al. Efficient multi-scale attention module with cross-spatial learning. In: Institute of Electrical and Electronics Engineers. *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing; 2023 June 4–10; Rhodes Island, Greece.* [S. l.]: Institute of Electrical and Electronics Engineers; 2023. p. 1–5. DOI: 10.1109/ICASSP49357.2023.10096516.

31. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. arXiv:1709.01507 [Preprint]. 2017 [cited 2024 November 13]: [13 p.]. Available from: https://arxiv.org/pdf/1709.01507.

32. Yang L, Zhang R, Li L, Xie X. SimAM: a simple, parameter-free attention module for convolutional neural networks. In: Meila M, Zhang T, editors. *Proceedings of the 38th International conference on machine learning; 2021 July 18–24* [Internet]. [S. l.]: [s. n.]; 2021 [cited 2024 December 27]. p. 11863–11874 (Proceedings of machine learning research; volume 139). Available from: https://proceedings.mlr.press/v139/yang21o/yang21o.pdf.

33. Huang H, Zhou X, Cao Ji, He R, Tan T. Vision transformer with super token sampling. arXiv:2211.11167 [Preprint]. 2022 [cited 2024 December 28]: [13 p.]. Available from: https://arxiv.org/pdf/2211.11167.

34. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. arXiv:2103.14030 [Preprint]. 2021 [cited 2025 January 10]: [14 p.]. Available from: https://arxiv.org/pdf/2103.14030.

35. Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: optimal speed and accuracy of object detection. arXiv:2004.10934 [Preprint]. 2020 [cited 2025 January 3]: [17 p.]. Available from: https://arxiv.org/pdf/2004.10934.

36. Geetha AS. What is YOLOv6? A deep insight into the object detection model. arXiv:2412.13006 [Preprint]. 2024 [cited 2024 November 30]: [16 p.]. Available from: https://arxiv.org/abs/2412.13006.

37. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. YOLOv10: real-time end-to-end object detection. arXiv:2405.14458 [Preprint]. 2024 [cited 2024 December 23]: [21 p.]. Available from: https://arxiv.org/pdf/2405.14458.

Received 04.03.2025 / revised 08.07.2025 / accepted 08.07.2025.