International Core Journal of Engineering Volume 11 Issue 9, 2025
ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

A Review of YOLO Series Algorithms in Object Detection for UAV Aerial Images

Yuhan Yan, Lin Zhang*

School of Emergency Management and Safety Engineering, North China University of Science and Technology, Tangshan 063210, China

*Corresponding author: Lin Zhang

Abstract

With the rapid development of the low-altitude economy, the demand for object detection in UAV aerial images has become increasingly urgent in fields such as traffic monitoring, agricultural plant protection, and emergency rescue. The YOLO series algorithms have become the mainstream technology in this field due to their advantages of single-stage end-to-end detection. However, the characteristics of UAV aerial scenes, such as extremely small target scales, dense distribution, complex backgrounds, and variable attitudes, pose severe challenges to the accuracy and real-time performance of the algorithms. This paper systematically sorts out the evolution of YOLO series algorithms, from the pioneering exploration of YOLOv1 to the Transformer architecture innovation of YOLOv12, and analyzes the core improvements of each version in terms of anchor box mechanism, feature fusion, and detection head design. It focuses on discussing the difficulties faced by YOLO algorithms in UAV aerial scenes, such as insufficient detection accuracy of small targets, complex scale distribution, and the contradiction between real-time performance and accuracy, and summarizes the key technical solutions for feature extraction, network lightweighting, detection head optimization, and loss function improvement. Meanwhile, it introduces the characteristics of mainstream datasets such as StanfordDrone and VisDrone2019. The research aims to provide relevant scholars with the research status and development context in the field of object detection in UAV aerial images, and offer references for subsequent algorithm optimization and application implementation.

Keywords

Aerial Images; Object Detection; YOLO; Dataset.

1. Introduction

With the in-depth integration of the low-altitude economy and intelligent perception technologies, unmanned aerial vehicles (UAVs) have become a core tool for acquiring aerial images in fields such as traffic monitoring, agricultural plant protection, emergency rescue, and urban planning, thanks to their advantages of flexibility, low cost, and high resolution[1].UAV aerial images contain rich geospatial information, and object detection, as a core technology for extracting key information from images, directly determines the application effectiveness of UAVs in scene understanding and decision support. For example, in intelligent transportation, vehicle detection in UAV aerial images can monitor road congestion and violations in real-time; in agricultural monitoring, accurate identification of crops and pests can support precision fertilization and disaster early warning; in emergency rescue, rapid positioning of trapped people and rescue materials can improve rescue efficiency[2]. However, the particularity of UAV aerial scenes poses severe challenges to object

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

detection: high-altitude shooting leads to extremely small and densely distributed targets; complex backgrounds easily obscure target features; in addition, UAV attitude jitter can cause target rotation, blurring, and other issues, making it difficult for traditional object detection algorithms to balance accuracy and real-time performance[2].

Against this background, the YOLO (You Only Look Once) series algorithms, relying on the architectural advantage of "single-stage end-to-end detection", have demonstrated unique value in balancing speed and accuracy, becoming the mainstream choice for object detection in UAV aerial images[3]. Since the proposal of YOLOv1 in 2016, the series of algorithms have undergone years of iteration: from the first realization of real-time detection in v1, to the introduction of Feature Pyramid Network (FPN) in v3 to improve multi-scale capabilities, v5 achieving lightweight design through CSPNet, v8 optimizing positioning accuracy with anchor box adaptation and task-aligned head, and v10 breaking through the real-time bottleneck with a non-NMS architecture. The YOLO series has continuously optimized for the needs of small object detection, adaptability to complex scenes, and deployment on edge devices, gradually adapting to the technical pain points of UAV aerial photography[4].

This paper focuses on the research of YOLO series algorithms for object detection in UAV aerial images, analyzes the current difficulties faced by YOLO series algorithms in UAV object recognition, and summarizes the common optimization schemes for YOLO network structures, aiming to help other relevant researchers understand the current research status in the field of YOLO series algorithms for UAV detection in UAV aerial images.

2. Progress in the Development of YOLO Series Algorithms

2.1 YOLOv1: Pioneering Single-Stage Detection (2016)

In 2016, Joseph Redmon and Ali Farhadi [5]proposed YOLOv1, which completely transformed the paradigm of object detection. Previously, mainstream object detection algorithms (such as the R-CNN series) adopted a two-stage process of "first generating candidate regions and then performing classification", which was slow and difficult to meet real-time requirements.YOLOv1 was the first to convert object detection into a regression problem solved by a single forward propagation: a neural network directly outputs the coordinates of object bounding boxes and class probabilities without an additional region proposal step, achieving a real-time detection speed of 45 FPS, with a faster version even reaching 155 FPS.

Its network structure was designed based on GoogLeNet, consisting of 24 convolutional layers for feature extraction and 2 fully connected layers for result prediction. The input image was divided into a 7×7 grid, where each grid was responsible for predicting 2 bounding boxes and the corresponding class probabilities. Although this design broke through speed bottlenecks, it had obvious limitations: the coarse grid division led to poor small-object detection capabilities; each grid could only predict a small number of bounding boxes, making it difficult to handle dense object scenarios; and the localization accuracy needed improvement. These issues became the starting point for improvements in subsequent versions.

2.2 YOLOv2/YOLO9000: Optimization Toward Practicality (2017)

To address the shortcomings of YOLOv1, YOLOv2 (also known as YOLO9000[6]) was released one year later (2017) with the core goal of being "better, faster, and stronger." It introduced three key improvements: the anchor box mechanism replaced fully connected layers for bounding box prediction. By predefining prior boxes of different ratios, the model learned offsets instead of directly regressing coordinates, significantly improving localization stability; the Darknet-19 backbone network replaced GoogLeNet, using 19 convolutional layers combined with batch normalization to reduce computational load while accelerating training convergence; the multi-scale training strategy enhanced the model's adaptability to objects of different scales by dynamically adjusting the input image resolution.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

YOLO9000 also pioneered the "joint training" method, which fused ImageNet (classification data) and COCO (detection data) through a WordTree hierarchy, enabling the model to detect over 9,000 types of objects and achieving large-scale category detection for the first time. These improvements transformed the YOLO series from an academic prototype into a practical tool, especially demonstrating value in scenarios requiring rapid identification of a large number of categories.

2.3 YOLOv3: Maturity of Multi-Scale Feature Fusion (2018)

The core breakthrough of YOLOv3[7] was multi-scale feature fusion, aiming to address the inadequacy in detecting small objects. It adopted a new Darknet-53 backbone network with residual connections, avoiding gradient vanishing while deepening the network to 53 layers, thereby significantly enhancing feature extraction capabilities. More importantly, it detected large, medium, and small objects using three feature maps of different resolutions (13×13, 26×26, 52×52) respectively. Combined with upsampling and skip connections, shallow detail features and deep semantic features were fully fused, greatly improving the accuracy of small-object detection. In addition, YOLOv3 replaced SoftMax with independent logistic classifiers, supporting multi-label output (e.g., simultaneously identifying "person" and "person wearing a hat"), which better suited complex scenarios. These adjustments slightly reduced speed but brought accuracy close to that of two-stage algorithms at the time, making it a milestone in single-stage detection and laying the foundation for the "balance between accuracy and speed" design philosophy in subsequent versions.

2.4 YOLOv4: Integration of Systematic Optimization Strategies (2020)

YOLOv4[8] systematically integrated optimization strategies for object detection for the first time, proposing the concepts of "free bags" and "special offer bags": the former refers to training techniques that do not increase inference costs, such as Mosaic data augmentation and self-adversarial training; the latter refers to inference optimizations with low computational cost but high returns, such as attention mechanisms and improved loss functions. This approach allowed the model to maintain real-time performance while improving accuracy.

In terms of architecture, YOLOv4 clearly divided the network into three parts: backbone, neck, and head. The backbone used CSPDarknet53, which reduced redundant computations based on cross-stage partial connections; the neck introduced the SPP (Spatial Pyramid Pooling) module to expand the receptive field and the PANet (Path Aggregation Network) to enhance feature transmission; the head retained multi-scale prediction but adopted the CIoU loss function to improve bounding box regression accuracy. These improvements enabled YOLOv4 to be efficiently trained on a single GPU, making it the preferred model for industrial deployment and widely used in security, drones, and other scenarios.

2.5 YOLOv5: Engineering and Ecosystem Construction (2020)

YOLOv5 became one of the most popular versions due to its engineering design. Its most significant change was the shift to the PyTorch framework; previous versions were based on Darknet. Adopting PyTorch lowered the usage threshold, supporting dynamic computation graphs and flexible deployment. Additionally, it provided five model sizes (N/S/M/L/X), adaptable from edge devices to cloud servers, meeting the needs of different scenarios.

Technically, YOLOv5 introduced adaptive anchor boxes to automatically learn dataset-adapted prior boxes; the Focus module for lossless downsampling; the SPPF module for efficient feature aggregation; and a complete toolchain, including model export, automatic annotation, and visualization interfaces. This "algorithm + ecosystem" model attracted a large number of developers, promoting YOLO from a research tool to an engineering solution.

2.6 YOLOX: Innovation in Decoupling and Anchor-Free Mechanisms (2021)

Following YOLOv5, YOLOX[9], launched by the Meituan team in 2021, became a key turning point in the series' development with "dual breakthroughs in accuracy and speed." Its core innovations were the decoupled head design and anchor-free mechanism, which completely transformed the traditional

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

detection head architecture of YOLO. Traditional YOLOv1-v5 used a "coupled head," where a single convolutional layer simultaneously performed classification and bounding box regression. Although concise, this design easily caused task interference. YOLOX was the first to introduce a "decoupled head": splitting classification and regression into two independent convolutional branches, with the classification branch focusing on learning object category features and the regression branch focusing on optimizing bounding box coordinates, improving accuracy by 2-3 percentage points. Meanwhile, it abandoned the long-used anchor box mechanism and instead directly predicted the offset of the object center and the aspect ratio, not only reducing the cost of manual parameter tuning but also solving the "imbalanced positive and negative samples" problem caused by anchor boxes.

To further optimize sample utilization, YOLOX adopted the "SOTA positive-negative sample matching strategy": dynamically allocating samples based on object difficulty to ensure each ground truth box matched the most suitable prediction box, particularly effective in small-object and occluded scenarios. These improvements enabled YOLOX to surpass YOLOv5 in accuracy on the COCO dataset while maintaining comparable inference speed. More importantly, its decoupled head and anchor-free design were widely adopted by subsequent YOLOv6, v7, and v8, becoming the "standard configuration" for series architecture upgrades.

2.7 YOLOv6: Special Adaptation for Industrial Scenarios (2022)

YOLOv6[10], launched by the Meituan team in 2022, was optimized specifically for industrial applications, with the core goal of "achieving efficient deployment on commonly used hardware." It designed the EfficientRep backbone network, using reparameterized convolutions (multi-branch during training, merged into a single branch during inference) to balance accuracy and speed; the neck adopted the Rep-PAN (reparameterized path aggregation network) to enhance feature fusion efficiency; the head was replaced with an "efficient decoupled head" to separate classification and regression tasks for improved accuracy.

Addressing the actual needs of industrial scenarios, YOLOv6 optimized model compression, supporting INT8 quantization, reducing volume by 75%, accelerating inference, and achieving high FPS real-time detection on mainstream industrial GPUs such as Tesla T4. Its design philosophy demonstrated that industrial-grade detectors require not only excellent benchmark performance but also adaptation to hardware constraints and business scenarios.

2.8 YOLOv7: In-Depth Exploration of Training Strategies (2022)

YOLOv7[11] focused on "accuracy improvement without increasing inference costs," proposing the concept of "trainable free bags": exploring model potential by optimizing the training process rather than changing the inference architecture. Core innovations included the ELAN module (controlling gradient path length to enhance feature learning) and dynamic label assignment (dynamically allocating samples based on object difficulty to resolve inconsistencies between classification and regression tasks).

It also attempted "model differentiation" for the first time: launching "extended models" (YOLOv7-X) pursuing the highest accuracy and "optimized models" (YOLOv7-W6) adapted to high-resolution inputs, targeting different scenarios. This segmented design enabled YOLOv7 to surpass previous versions in accuracy on the COCO dataset while maintaining high detection speed, becoming a new benchmark for real-time detection, particularly excelling in scenarios sensitive to details such as satellite image analysis and precision manufacturing quality inspection.

2.9 YOLOv8: Modularity and Multi-Task Unification (2023)

YOLOv8 from the Ultralytics team adopted a fully modular architecture, unifying tasks such as detection, segmentation, and pose estimation into a single framework for the first time. It replaced YOLOv5's C3 module with the C2f module, adding residual connections to improve feature reuse; the detection head adopted an anchor-free design, directly predicting object centers and aspect ratios, reducing hyperparameter dependence and simplifying model tuning. In engineering, YOLOv8 supported automatic hyperparameter optimization, finding optimal configurations through

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

evolutionary algorithms, and with 11 deployment formats, it could even achieve real-time segmentation on mobile devices. This "one-stop" solution enabled its rapid application in multi-task scenarios such as AR/VR interaction and gesture control, further expanding YOLO's application boundaries.

2.10 YOLOv9: Programmable Optimization of Gradient Information (2024)

YOLOv9[12] fundamentally addressed the "information loss" problem in deep networks, proposing the Programmable Gradient Information (PGI) mechanism: preserving complete information of the input image through an auxiliary reversible branch to ensure gradient non-degeneration during backpropagation, allowing deep networks to learn efficiently. Combined with the Generalized Efficient Layer Aggregation Network (GELAN) (flexibly integrating multiple computational blocks to improve parameter utilization), it achieved accuracy improvements while reducing parameters by 49% and computational load by 43%.

This "lightweight + high-precision" design made YOLOv9 particularly suitable for scenarios with limited computing power, such as autonomous driving domain controllers, enabling efficient deployment on edge devices like Jetson AGX Orin, marking YOLO's shift from "pursuing performance limits" to "balancing efficiency and accuracy."

2.11 YOLOv10: Realization of End-to-End Detection (2024)

YOLOv10[13] from Tsinghua University achieved a major breakthrough by removing the non-maximum suppression (NMS) post-processing, realizing true end-to-end detection. Traditional NMS required manual threshold adjustment, easily leading to missed detections or duplicate annotations; YOLOv10 enabled the model to directly output unique bounding boxes through "dual label assignment" (using one-to-many enhanced supervision during training and one-to-one output during inference), significantly improving inference speed.

It also introduced spatial-channel decoupled downsampling (reducing information loss) and partial self-attention (enhancing global modeling), excelling in video stream detection scenarios such as drone tracking and sports event analysis. This "post-processing-free" design upgraded YOLO from a "fast detection algorithm" to an "efficient inference system."

2.12 YOLOv11: Extreme Lightweight Design (2024)

YOLOv11 continued the engineering route, focusing on "lightweight deployment." By replacing traditional modules with the C3k2 module (using two small convolutions instead of three in traditional modules), it reduced computational load by 15%; the C2PSA module fused spatial pyramid pooling with attention to enhance key region perception, reducing parameters by 20% at the same accuracy.

It supported 4-bit quantization for the first time, compressing the model size to below 8 MB, achieving 160 FPS inference on low-end devices like Jetson Nano, perfectly adapting to industrial IoT scenarios. Its design demonstrated that high-performance detection on resource-constrained devices can be achieved through architectural optimization rather than simply increasing parameters.

2.13 YOLOv12: Comprehensive Application of Attention Mechanisms (2025)

YOLOv12[14] broke the series' reliance on CNNs, adopting a pure Vision Transformer architecture and fully introducing attention mechanisms into single-stage detection. The core innovative Area-Attention (A^2) module divided feature maps into continuous regions, reducing computational complexity from $O(n^2d)$ to $O(n^2d/4)$ while maintaining a large receptive field, addressing the speed issue of Transformers. Combined with the Residual Efficient Layer Aggregation Network (R-ELAN) (optimizing gradient flow), YOLOv12 significantly improved detection robustness in complex scenarios (e.g., low light, dense objects). This "attention-first" design marked the YOLO series' transition from "CNN-dominated" to the "Transformer era," opening new directions for real-time detection.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

Table 1. Comparison of Different YOLO Versions

YOLO Version	Anchor Mechanism	Detection Head	Backbone	Input Size	Activation Function	Loss Function
YOLOv1	Anchor-Free	Coupled Head	GoogLeNet	448	LeakyReLU	IOU
YOLOv2	Anchor-Based	Coupled Head	Darknet-19	416	LeakyReLU	IOU
YOLOv3	Anchor-Based	Coupled Head	Darknet-53	608	LeakyReLU	IOU
YOLOv4	Anchor-Based	Coupled Head	CSPDarknet53	608	LeakyReLU	DIOU
YOLOv5	Anchor-Based	Coupled Head	CSPDarknet	608	LeakyReLU	CIOU
YOLOX	Anchor-Free	Decoupled Head	Darknet-53	640	SiLU	CIOU
YOLOv6	Anchor-Free	Decoupled Head	EfficientRep	640	SiLU	CIOU
YOLOv7	Anchor-Free	Decoupled Head	ELAN	640	SiLU	CIOU
YOLOv8	Anchor-Free	Decoupled Head	Darknet-53	640	SiLU	CIOU
YOLOv9	Anchor-Free	Decoupled Head	GELAN	640	SiLU	CIOU
YOLOv10	Anchor-Free	Decoupled Head	Darknet-53	640	SiLU	CIOU
YOLOv11	Anchor-Free	Decoupled Head	C3k2/C2PSA	640	SiLU	CIOU
YOLOv12	Anchor-Free	Decoupled Head	R-ELAN	640	SiLU	CIOU

Since the advent of YOLOv1 in 2016, the YOLO series algorithms have undergone continuous evolution from pioneering exploration to engineering maturity and paradigm breakthroughs, consistently innovating around the "balance between accuracy and speed." From the early YOLOv1 pioneering the single-stage detection paradigm by converting object detection into a single regression problem, to YOLOv2 introducing the anchor box mechanism and YOLOv3 achieving multi-scale feature fusion to gradually improve detection performance; to YOLOv4 systematically integrating optimization strategies and YOLOv5 promoting engineering and ecosystem construction to make algorithms more deployable; subsequent YOLOX brought architectural innovations with decoupled heads and anchor-free mechanisms, while YOLOv6 to v12 continued to break through in industrial adaptation, training strategies, multi-task unification, lightweight design, and application of attention mechanisms, transforming the YOLO series from an academic prototype into an efficient detection system covering multiple scenarios.

Table 1 clearly presents key changes in this evolution by comparing core features of each version: in terms of anchor mechanisms, from anchor-free in YOLOv1, to anchor-based in v2-v5, and back to anchor-free in YOLOX and later versions, reducing parameter tuning costs and sample imbalance issues; detection heads evolved from coupled heads in v1-v5 to decoupled heads in YOLOX and subsequent versions, improving classification and regression accuracy through independent branches; backbone networks advanced from GoogLeNet and Darknet series to CSPDarknet, ELAN, GELAN, and finally to the pure Vision Transformer-based R-ELAN in v12, achieving a leap in feature extraction capabilities; meanwhile, input sizes gradually stabilized at 640, activation functions transitioned from LeakyReLU to SiLU, and loss functions upgraded from IOU to DIOU and CIOU. These iterative mechanisms and components have collectively made the YOLO series a benchmark in the field of object detection.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

3. Difficulties and Challenges of YOLO Series Algorithms in UAV Aerial Image Object Detection

References are cited in the text just by square brackets

3.1 Challenges Arising from Inherent Scene Characteristics

3.1.1 Insufficient Accuracy in Small Object Detection

Due to the long shooting distance and wide coverage of UAV aerial images, small objects (such as pedestrians, small vehicles, etc.) in aerial images account for an extremely low proportion of pixels, which greatly increases the difficulty of detection[15]. Firstly, small objects have scarce feature information, with blurred edges and textures, making it difficult for the model to accurately capture their unique properties during feature extraction. Secondly, small objects occupy a minimal proportion in the image and are easily obscured by complex backgrounds. Finally, the arrangement and distribution of small objects in the image may be uneven, and this uncertainty makes it difficult for the model to adapt to all scenarios. These situations cause YOLO to struggle to effectively capture key information of small objects, resulting in low detection accuracy and frequent missed detections.

3.1.2 Extremely Complex Object Scales and Distributions

Affected by the UAV's flying height and shooting angle, the object scales in UAV images have a large span. For example, a single image may contain both close-range large vehicles and long-range small pedestrians; in dense scenarios such as crowded gatherings and parking lots, objects overlap severely. This causes YOLO's anchor box matching mechanism to be prone to false detections or duplicate annotations.

3.1.3 Complex Backgrounds and Interfering Factors

The background of UAV aerial images includes diverse elements such as sky, vegetation, roads, and buildings, which easily form "same-spectrum different-objects" interference (e.g., green vehicles being confused with vegetation)[16]. Changes in lighting during UAV shooting (backlighting, shadows) and weather interference (haze, rain and fog) further reduce object contrast, leading to a decline in YOLO's feature discrimination ability.

3.1.4 Diversity in Object Rotation and Posture

Different shooting angles and flight attitudes of UAVs result in vastly varying appearances of objects in images. Among them, jitter in the UAV's flight attitude can cause objects to exhibit arbitrary rotation angles. Traditional YOLO's horizontal bounding boxes struggle to accurately locate such objects, easily introducing background noise and reducing the IoU (Intersection over Union) matching degree between the detection box and the object.

3.2 Inherent Limitations of YOLO Series Algorithms

3.2.1 Conflict Between Real-Time Performance and Accuracy

In UAV embedded platforms and other onboard devices with limited computing power, improving object detection accuracy often requires deepening network layers or introducing complex multi-scale feature fusion modules (e.g., adding low-level feature branches). However, such optimization strategies lead to a sharp increase in model parameters (Params) and computational load (FLOPs), directly conflicting with real-time requirements, making the detection frame rate unable to meet practical application needs.

3.2.2 Insufficient Efficiency of Multi-Scale Feature Fusion

The multi-scale fusion mechanisms such as FPN/PAN adopted by YOLO series have significant performance bottlenecks when dealing with extreme scale differences in aerial images: the complementarity between high-level semantic features and low-level detail features is insufficient, easily causing the "feature imbalance phenomenon"-where large object features are redundantly accumulated while small objects lack sufficient effective feature supply[17]. This imbalance weakens the ability of shallow networks to learn small objects due to the lack of effective supervision signals,

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

ultimately reducing the quality of multi-scale feature fusion and making it difficult to balance the detection needs of objects of different scales.

3.2.3 Difficulty in Balancing Lightweight Design and Robustness

Lightweight YOLO variants designed to meet real-time requirements reduce computational complexity by compressing network width and reducing layers, but inevitably sacrifice feature representation capabilities. In complex aerial scenarios (e.g., dense small objects, drastic lighting changes, cluttered backgrounds), such models tend to exhibit significantly reduced anti-interference ability due to insufficient feature extraction, manifested as increased missed detection rates and intensified confidence fluctuations.

3.2.4 Insufficient Adaptability of Anchor Boxes

YOLO series rely on preset anchor boxes for object matching, but the sizes and ratios of these anchor boxes are mostly designed based on general datasets, making them difficult to adapt to objects with extreme scales in aerial scenarios. Significant deviations between anchor boxes and the actual sizes of objects lead to matching failures and cumulative positioning errors. Especially in scenes with a high proportion of small objects, the IoU between anchor boxes and objects is often below the threshold, directly resulting in missed detections or a sharp decline in bounding box regression accuracy.

4. Difficulties and Challenges of YOLO Series Algorithms in UAV Aerial Image Object Detection

4.1 Optimization of YOLO Feature Extraction

The performance of the YOLO object detection algorithm highly depends on the backbone network's ability to extract and fuse target features in images. In UAV aerial photography scenarios, small targets are characterized by scarce feature information, resulting in generally low recognition and localization accuracy of existing algorithms. To address this, researchers have conducted targeted explorations on backbone network optimization. By replacing inefficient network structures, improving feature extraction modules, and introducing attention mechanisms, they have significantly enhanced the model's capability to capture features of small UAV aerial targets.

As the core carrier of feature extraction, the structural design of the backbone network directly affects the expressive ability of target features. Introducing lightweight and computationally efficient network structures to replace the original backbone can reduce computational costs while enhancing feature extraction performance. Yuying Cao et al.[18] integrated EfficientViT into the backbone network of YOLOv11 and designed the C2PSA-CPCA module. This module not only enhances multi-scale feature perception but also improves the network's efficiency in extracting target features while effectively reducing computational complexity. Experimental results show that the detection mAP of the improved LightTassel-YOLO model on the dataset is 4 percentage points higher than that of the baseline model YOLOv11n. Jinpei Li et al. [19] replaced the backbone network of YOLOv8 with EfficientViT to enhance the ability to extract local image features by reducing redundant parameters; they also introduced the LSKNet selective kernel network, fused the attention mechanism of the C2f module, and added a bidirectional feature pyramid network and a small target P2 detection head. After multi-dimensional improvements, the model's recognition accuracy on the UAV-acquired dataset is 3.7% higher than that of YOLOv8, and mAP50 (mean average precision at 50) is increased by 3.9%, achieving accurate recognition and localization of small targets.

In response to the characteristics of variable target scales and strong background interference in UAV aerial photography, researchers have enhanced the model's ability to capture features of multi-scale targets in complex scenarios by optimizing core modules in the backbone network (such as C2f and C3K2). Tao Shi et al. [20] proposed the C2f-DCN module by combining the flexible sampling advantages of DCNv1-3, enhancing the model's adaptability in extracting features of targets of different scales; they also replaced the original detection head of YOLOv8 with Dynamic Head,

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

further improving detection performance by fusing scale, spatial, and task attention mechanisms. Experiments on the Mapsai dataset show that the mAP of the improved algorithm is increased by 6.2 percentage points, with a detection speed of 72.6 FPS, balancing accuracy improvement and real-time requirements, making it suitable for dynamic UAV detection scenarios. Yangyang Fan et al.[21] embedded a dilated residual (DWR) module into the C3K2 module and reconstructed the backbone network using a programmable gradient information strategy to strengthen the network's ability to extract features of infrared small targets. On the self-built underground infrared dataset, the detection mAP of the improved model is 3.8% higher than that of the baseline model, significantly enhancing the detection robustness in complex environments. Shen Chao et al.[22] designed a partial channel deformable convolution module (PCDConv) and improved the C2f module based on it; they also redesigned the neck network based on Bifpn to improve the efficiency of feature extraction and fusion for small targets; finally, they combined the DyHead with multi-attention mechanisms and a new bounding box loss function. The improved model's mAP on the VisDrone2019 dataset is 3.7% higher than that of the original model, effectively improving the detection accuracy of aerial small targets while meeting real-time detection requirements.

The attention mechanism dynamically adjusts feature weights, enabling the model to focus on key target regions (such as small targets) and suppress redundant background information, significantly improving the utilization of UAV target features in complex scenarios. Huizhi Xu et al.[23] addressed the problems of low detection accuracy, easy missed detection, and false detection of small targets in UAV aerial photography by introducing the GE attention mechanism into the backbone and neck networks of YOLOv7, optimizing the model by strengthening the use of contextual information and adding small target detection layers. Experiments on the VisDrone2019 dataset show that the accuracy of the new model is 2.5 percentage points higher than that of YOLOv7. Feng Wang et al. [24] aimed at the problem that small targets have a low imaging proportion and difficult feature extraction due to UAV flight height, embedded a small target detection structure (STC) in the network to enhance semantic information collection, and introduced global attention GAM in the bottom layer of the backbone to ensure the integrity of feature transmission. Experiments on the VisDrone2021 dataset show that the improved model's detection performance for small targets is significantly enhanced, with mAP 4.4% higher than that of the baseline model, outperforming mainstream algorithms such as SSD and YOLO series. Zhenhua Wang et al. [25] addressed the challenges of multi-scale, multi-orientation, and complex backgrounds of offshore UAV targets by improving the residual module in YOLOv3: enhancing the model's utilization of target features and global information through establishing dense connections and introducing the CBAM dual attention mechanism; they also replaced FPN with a two-level recursive feature pyramid network to weaken interference from complex environments. The improved YOLO-D model significantly outperforms the baseline model in detection accuracy for small-sized targets.

In summary, through efficient replacement of backbone networks, targeted improvement of feature modules, and integration of attention mechanisms, the YOLO series algorithms have significantly enhanced their ability to extract features of small UAV aerial targets, providing diverse technical paths for real-time detection in complex scenarios and laying a methodological foundation for subsequent research.

4.2 Optimization of YOLO Network Structure Size

UAV platforms are limited by computing power, storage, and power consumption, which put strict requirements on the lightweight and efficiency of target detection models. Although YOLO series algorithms perform well in detection accuracy, the original models have large parameters and high computational complexity, making them difficult to directly deploy on embedded devices. To this end, researchers have explored the optimization of network structure size, reducing model volume and computational costs while ensuring detection accuracy through strategies such as lightweight module replacement, network pruning, and structure simplification.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

By introducing efficient convolution modules and lightweight feature fusion structures to replace redundant components in the original network, the number of parameters and computational load can be reduced while maintaining or even enhancing feature extraction capabilities. Limei Song et al.[26] proposed the lightweight YOLO-VG algorithm based on YOLOv8s: using GSConv instead of traditional convolution in the backbone network to reduce redundant information; replacing ordinary convolution in the neck network with ODConv to enhance the model's adaptability and generalization ability; and replacing the original C2f module with the VoV-Ghost structure to further achieve model lightweighting and efficiency. Experimental results show that the improved model's computational efficiency on the dataset is increased by 24.6%, the model size is reduced by 20.4%, and it can more accurately capture key image information. Xiaoxiao Feng et al. [27] proposed the ESimB-YOLO algorithm to address the deployment limitations of UAV embedded devices: using ESNet fused with the ECA module as the backbone network, designing the C2f_SimAM module that integrates C2f and SimAM, and adopting a weighted bidirectional feature pyramid structure for feature fusion. Compared with the baseline model YOLOv8n, the new model's parameters are reduced by 57%, the model size is reduced by 53%, and the mAP value is increased by 1.8%, achieving a balance between lightweighting and detection accuracy. Hui Wang et al.[28] referred to the lightweight design concept of MobileNetV3, optimized the backbone and neck networks of YOLOv11 by combining GSConv and VoVGSCP modules, and enhanced multi-scale feature fusion capabilities; they also integrated Monte Carlo attention mechanism and partial convolution into the C3K2 module to improve the ability to extract features of small targets while reducing computational complexity. The improved YOLO-LiRa model has 24.4% fewer parameters and 12.5% lower computational complexity than the baseline model, achieving a good balance between detection accuracy and speed. Yifan Lyu et al.[29] addressed the difficulty of deploying deep learning models on UAV embedded devices by modifying the neck network structure based on YOLOv8 to enhance the detection ability of small targets; they designed an orthogonal feature enhancement module (OFEM) and a local attention module (LAM) to replace simple concatenation operations, effectively filtering irrelevant interference and optimizing feature representation. The improved LightUAV-YOLO algorithm maintains low parameters and computational complexity while increasing the mean average precision by 6.4%, making it more suitable for real-time UAV detection scenarios.

By removing redundant convolution kernels or channels in the network, the model size and computational load can be significantly reduced with minimal accuracy loss, which is an important means to achieve model lightweighting. Rui Shi et al.[30] proposed a generalized attribution pruning method, identifying convolution kernels closely related to target output through designing channel and spatial masks, and pruning irrelevant kernels layer by layer in the channel dimension; the model obtained after fine-tuning has scale and rotation invariance, with computational load reduced by 68.7% and accuracy increased by 0.4% compared with before pruning, making it more suitable for resource-constrained platforms such as UAVs. Zuxiang Situ et al.[31] addressed the deployment limitations of deep learning solutions on small mobile terminals, realizing model lightweighting based on YOLOv5 combined with transfer learning and channel pruning technology. The pruned new model has 81% fewer parameters and 48.8% fewer operations, and its superiority in embedded scenarios such as UAVs is fully verified through comparison with mainstream detection algorithms and lightweight backbone networks.

In summary, strategies such as lightweight module replacement, structure simplification, and network pruning have effectively reduced the number of parameters, computational complexity, and volume of YOLO series models, while balancing detection accuracy and real-time performance, providing feasible lightweight solutions for target detection tasks on resource-constrained platforms such as UAVs.

4.3 Optimization of YOLO Detection Head Structure

The YOLO algorithm uses multi-scale detection heads (FPN-PAN structure) to cope with target scale changes, but its original design still has limitations in UAV aerial photography scenarios - aerial data

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

has a high proportion of small targets and a large scale span, and the receptive field of existing detection heads does not match the features of small targets, which easily leads to missed detection or positioning deviation. To this end, researchers have optimized the detection head structure, improving the model's ability to detect multi-scale targets (especially small targets) by adding small target detection layers, introducing decoupled head design, and dynamically adjusting detection strategies.

To address the weak features of small targets in UAV aerial photography and the insufficient receptive field of existing detection layers, the ability to capture tiny targets is enhanced by adding dedicated small target detection layers or adjusting the distribution of detection levels. Jiahui Chen et al.[32] introduced a decoupled head to improve positioning accuracy to alleviate the prediction deviation of YOLOv5 in different tasks, while adding a small target detection layer and designing a multi-scale feature extraction module C3ResBlock. The improved model's mAP on the VisDrone dataset is 12.9 percentage points higher than that of the baseline model YOLOv5, outperforming mainstream target detection algorithms. Yuting Zhang [33] added a smaller-scale detection layer to the model to address the low detection accuracy of multi-scale targets, while appropriately cropping the large target detection layer to balance computational costs; combining the SPPCSPC pyramid pooling module, CARAFE upsampling operator, and EIoU loss function, she proposed the ECSH-YOLOv8 model. Experiments on self-collected datasets show that the new model's recall rate is increased by 4.1%, the mAP value is increased by 1.8%, and its sensitivity to small targets is significantly enhanced.

The original YOLO detection head couples classification and regression tasks, which easily causes task interference, especially affecting detection accuracy in complex scenarios. By separating tasks through decoupled head design and combining dynamic adjustment mechanisms, the model's task focus and adaptability can be improved. Si Wu et al. [34] adopted a new dynamic detection head strategy to address the small size and large span of targets from the UAV perspective, organically unifying the detection heads of different targets to reduce the impact of occlusion on detection in complex scenarios; they also introduced lightweight improved spatial depth convolution (SPDs-Conv) to improve the detection performance of small targets while reducing the number of parameters. The average precision of the improved model on the VisDrone2019 dataset is 15.1% higher than that of the baseline model, with an inference speed of 41fps, meeting real-time requirements. Zhijie Ren et al.[35] proposed a "scale-adaptive decoupled head" based on YOLOv5 to improve detection accuracy - automatically adjusting the number of channels according to the model size, enhancing network performance by separating classification and regression tasks; they also constructed a more efficient YOLO-SDH framework, using lightweight deformable convolution modules to reduce computational complexity. The mAP of the new model is 1.29% higher than that of the original YOLOv5, achieving a balance between accuracy and efficiency.

By designing a detection head with quality evaluation and combining loss function optimization, the positioning accuracy of detection boxes can be improved, reducing the impact of angle deviation and scale changes on small target detection in UAV aerial photography. Mengyang Li et al.[36] designed the LQEHead detection head, combined with a positioning quality estimator to evaluate the quality of detection boxes, and optimized the classification branch; introduced the MASRCNet feature extraction module for small target detection, and adopted a new NWD-InnerCIoU loss function to reduce the interference of aerial angles on small target positioning. Ablation experiments show that the model using the LQEHead detection head has an mAP@75 of 0.495 and an mAP@50:95 of 0.496, with significantly reduced model parameters and computational costs, making it easier to deploy on UAV platforms.

In summary, the scale expansion, decoupled design, and quality evaluation mechanism of the detection head have effectively improved the adaptability of YOLO series algorithms in detecting small UAV aerial targets. Combined with feature extraction and loss function optimization, the rate

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

of missed detection and false detection in complex scenarios is further reduced, providing a more accurate solution for multi-scale target detection.

4.4 Optimization of YOLO Loss Function

In UAV aerial images, a large number of small targets account for much less proportion than the background, and small targets have scarce feature information, which directly causes an imbalance between positive and negative samples in the dataset, seriously affecting detection accuracy. Although the new version of YOLO focuses on improving the detection ability of small targets and has become the main choice for target detection from the UAV perspective, the design of the loss function is crucial to solve the above problems. Therefore, loss function optimization has become an important improvement direction for YOLO target detection models in UAV aerial images.

Puyi Song et al.[37] replaced the GIoU Loss in YOLOv5 with CIoU Loss. This replacement not only accelerates the bounding box regression rate but also improves positioning accuracy. At the same time, they combined the squeeze-and-excitation module with a double frustum feature fusion structure to enhance feature extraction and fusion capabilities. The improved YOLOv5s model achieved an 86.3% mean average precision (mAP), which can effectively improve the target detection ability of UAV images even in complex backgrounds.

Mingjie Wu et al.[38] adopted the Focal-EIoU loss function to solve the problem of inaccurate regression results calculated by CIoU Loss in YOLOv5s. In addition, they further improved the model's detection accuracy for small targets by using double-layer routing attention and dynamic detection head DyHead. The final BD-YOLO model has an average precision index 0.062 higher than that of the YOLOv5s model on the VisDrone2019-DET dataset, and its detection effect on small targets is better than other mainstream models.

Xueqiang Zhao et al.[39] proposed to replace the original CIoU with WIoUv3 to improve the YOLOv8 model in response to the multi-scale problem of UAV images. At the same time, they introduced the large separable kernel attention LSKA mechanism of SPPF to weight multi-scale feature maps to better focus on feature information. Experimental results of the improved model on the dataset show that the mAP value is increased by 3.6% compared with the YOLOv8n model, and the frames per second are also increased from 78.7 to 83.3.

Li Tang et al.[40] proposed to replace the loss function in the original model with MFShaoe-IoU to address the problem that "black flying" UAV targets in complex backgrounds are difficult to identify due to variable scales and blurriness. This makes the model pay more attention to the shape and scale information of the bounding box itself, aggregate difficult samples, and improve target positioning accuracy. In addition, they fused the RepViTBlock structure and the efficient multi-scale attention mechanism EMA into the C2f module to improve the Bottleneck module, designing the C2f-RVB module, which enhances the model's feature extraction ability while reducing the number of parameters. Experiments on the public dataset CBD show that the average precision of the improved final model is 4.1 percentage points higher than that of the baseline model, and it meets the deployment requirements on mobile devices.

Jiayu Sun et al. [41] used the WIoU loss function to replace the original CIoU loss function, improving the adverse impact of low-quality small target data on gradients and accelerating network convergence. At the same time, combined with the idea of DCN, they designed a C2_DCf module to enhance the extraction of small target features, further improving the fusion of feature information for small targets. Experiments on the dataset show that the average precision of the improved new model is increased by 4.6 percentage points, fully demonstrating the effectiveness of the improved algorithm.

In summary, through the optimization of the loss function combined with other auxiliary improvement methods, the YOLO series algorithms have effectively improved the accuracy and efficiency of small target detection in UAV aerial images, providing strong support for solving target detection problems from the UAV perspective.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

4.5 Performance Evaluation of YOLO Detection Models

4.5.1 Accuracy Indicators

The confusion matrix, usually used in classification tasks, includes the following four values:

True Positive (TP): the number of samples that are actually positive and predicted as positive;

False Positive (FP): the number of samples that are actually negative but predicted as positive;

False Negative (FN): the number of samples that are actually positive but predicted as negative;

True Negative (TN): the number of samples that are actually negative and predicted as negative.

The above four values are used to evaluate the accuracy of classification tasks in target detection technology and subsequent indicators.

4.5.2 Model Efficiency and Complexity Indicators

Frame Per Second (FPS) indicates the number of images that the model can process per second, reflecting the running speed of the algorithm. A higher FPS value means a faster model running speed. The calculation method is as shown in formula (1), where t is the processing time required for a single image.

$$FPS = 1/t. \tag{1}$$

Parameters (Params) are the total number of trainable parameters in the model, directly affecting the model's storage requirements and memory usage.

Floating-Point Operations (FLOPs) are the number of floating-point operations performed by the model during a forward propagation, directly affecting device energy consumption and inference speed. A higher FLOP value means the model requires more support from computing hardware.

4.5.3 Detection Accuracy Indicators

Precision measures the reliability of the model's prediction results (Formula (2)). High precision means a higher proportion of samples predicted as positive by the model are actually positive. Recall measures the completeness of the model's defect detection (Formula (3)). High recall means the model can detect most of the actually existing defects.

$$Precision = TP/(TP+FP) \tag{2}$$

$$Recall = TP/(TP+FN) \tag{3}$$

Average Precision (AP) represents the precision of a single category, measuring the model's prediction accuracy in that category. A larger value indicates higher precision. The calculation method is as shown in formula (4), where R represents the recall rate under different intersection over unions.

$$AP = \int_0^1 P(R)dR \tag{4}$$

Mean Average Precision (mAP) represents the precision of multiple categories, measuring the overall performance of the model in all categories. It is obtained by averaging the APs of all categories. The calculation method is as shown in formula (5), where n is the total number of categories.

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{5}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

5. Datasets for Object Detection in UAV Aerial Images

5.1 StanfordDrone Dataset

The StanfordDrone dataset[42] is a large-scale dataset released by Stanford University's Computer Vision and Geometry Laboratory in 2016. It was captured by UAVs in an overhead view during crowded hours on campus, containing trajectory interaction information of 20k objects across 8 different scenarios. The objects in the dataset include multiple categories such as pedestrians, bicycles, scooters, and cars, with each object's trajectory labeled with a unique ID. This dataset aims to address tasks like object tracking and trajectory prediction, providing support for designing new algorithms, and can be widely used in fields such as intelligent transportation and smart security.

5.2 CARPK Dataset

The CARPK dataset[43], namely the CarParkingLot dataset, was proposed by Hsieh et al. in 2017. Captured by UAVs at an altitude of approximately 40 meters, it contains 1,448 images covering 4 unique parking lots, with information on nearly 90,000 cars. The dataset includes 989 training images and 459 test images. The vehicles in the captured images have high clarity, facilitating subsequent related research and playing a significant role in promoting the development and application of vehicle detection and counting technologies in the field of intelligent transportation systems.

5.3 UAVDT Dataset

The UAVDT dataset[44] is a large-scale UAV detection and tracking benchmark dataset released by a well-known domestic team. It covers various complex scenarios such as squares, highways, and intersections, containing approximately 80,000 representative frames extracted from 10 hours of original videos. Each frame is manually annotated with bounding boxes and useful attributes such as vehicle categories and occlusion status. Additionally, the dataset includes three target categories: cars, trucks, and buses. The UAVDT dataset can be used for multiple computer vision tasks such as object detection, single-object tracking, and multi-object tracking, providing high-quality and rich multi-task data for research on UAV applications in complex environments and promoting the development of fields like UAV video analysis.

5.4 VisDrone2019 Dataset

The VisDrone2019 dataset[45], collected and released by a team from Tianjin University, is used for research on UAV object detection and tracking. It contains 10,209 static images, involving 10 target categories such as people, cars, bicycles, vans, and trucks. These images cover various scenarios such as rural roads and urban streets, as well as different weather conditions (sunny, cloudy) and lighting environments, providing rich and diverse data support for UAV vision research.

5.5 AU-AIR Dataset

The AU-AIR dataset[46] is the first multimodal UAV dataset for object detection, specifically designed for low-altitude traffic monitoring. It contains over 2 hours of original videos, from which 32,823 labeled frames are extracted, covering 132,034 object instances across 8 target categories including humans, cars, buses, trucks, and bicycles. The unique feature of the AU-AIR dataset lies in its multimodal nature: it not only provides visual data and target annotations but also records rich flight data. This makes it a sufficient data resource for object detection and tracking tasks in static images and video streams, enabling the development of more advanced object detection and tracking algorithms and promoting the application of UAVs in autonomous surveillance in complex environments.

5.6 DroneCrowd Dataset

The DroneCrowd dataset[47] is a crowded crowd dataset proposed by the WenLongyin team in 2021, serving as a benchmark for algorithms in tasks such as object detection, crowd counting, and crowd density estimation in UAV-captured videos. It consists of 112 video clips, containing a total of 33,600 high-definition frames covering various scenarios such as campuses, streets, parks, parking lots, and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

playgrounds. The videos are recorded at 25 frames per second, and each image is meticulously annotated with head markers and trajectory labels. The dataset provides 20,800 human trajectories and over 4.8 million head annotation points, and also defines video-level attributes such as lighting conditions, target sizes, and crowd density. It offers multi-faceted considerations and data support for researchers to explore crowd behavior and crowd density estimation in UAV videos under the influence of different environmental factors.

5.7 FLAME3 Dataset

The FLAME3 dataset[48] was created by Clemson University and multiple collaborating institutions, aiming to promote the application of UAV thermal imaging technology in wildfire management. It contains synchronously collected visible spectrum and thermal imaging images, providing high-resolution RGB and thermal imaging TIFF files suitable for computer vision and wildfire modeling tasks. The dataset creation process involves using UAVs to collect data at designated wildfire sites and processing the data through an automated pipeline. The application of the FLAME3 dataset mainly focuses on wildfire detection, segmentation, and assessment, aiming to improve the efficiency and accuracy of wildfire management through high-precision thermal imaging data.

5.8 DroneVehicle Dataset

The DroneVehicle dataset[49] is a large-scale dataset collected and released by the Sun team from Tianjin University, focusing on RGB-infrared multimodal vehicle detection and counting tasks from a UAV perspective. It consists of 56,878 images captured by UAVs, with RGB images and infrared images each accounting for half. The research team has provided rich annotations with directional bounding boxes for five categories. Specifically, there are 389,779 annotations for cars in RGB images and 428,086 in infrared images; 22,123 annotations for trucks in RGB images and 25,960 in infrared images; 15,333 annotations for buses in RGB images and 16,590 in infrared images; 11,935 annotations for vans in RGB images and 12,708 in infrared images; and 13,400 annotations for freight cars in RGB images and 17,173 in infrared images.

5.9 UVSD Dataset

The UVSD dataset[50] is the first public UAV-based dataset for vehicle detection and segmentation, released by a team from Shandong University, aiming to address the challenges of vehicle detection and segmentation in UAV aerial images. It contains 5,874 high-resolution images and 98,600 vehicle instances, covering various scenarios such as urban and rural areas. The data was collected by UAVs at different heights and locations, with some images derived from the VisDrone dataset to ensure scenario diversity. The dataset is designed to target the particularities of UAV aerial photography, such as large differences in vehicle scales, diverse postures, and complex background interference. Through a combination of manual fine annotation and semi-automatic tools, the dataset ensures the accuracy and consistency of annotations. The UVSD dataset is of great value for researching vehicle detection and segmentation technologies under UAV shooting, providing an important benchmark for research in the field of UAV vision.

5.10 PDTDataset

The PDT dataset is a UAV target dataset jointly developed by the Shandong Computer Science Center (National Supercomputing Jinan Center) and Qilu University of Technology (Shandong Academy of Sciences), specifically for monitoring pest and disease data. It includes both high-resolution and low-resolution versions, with a total of 5,775 images covering healthy and pest-infected pine tree images. The dataset creation process involves field collection, data preprocessing, and manual annotation, aiming to provide high-precision target detection support for UAVs in precision spraying in agriculture. The application of the PDT dataset mainly focuses on agricultural UAV technology, aiming to improve the target recognition accuracy of UAVs in plant protection and address the shortcomings of traditional detection models in practical applications.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

Table 2 summarizes the above 10 mainstream datasets for object detection in UAV aerial images, covering a wide range of scenarios: they include general scenarios such as urban transportation and campus squares, specialized scenarios such as parking lot vehicles and dense crowds, and even extend to professional fields like wildfire management and forest pest monitoring, supporting applications in multiple fields such as intelligent transportation, public security, and ecological monitoring.

The data characteristics show significant diversity: resolutions range from 640×480 to 4000×3000, with 1 to 10 categories, and data volumes from 1,448 images to 80,000 frames, meeting the needs of different tasks. Technically, they highlight multimodal innovation (e.g., RGB with thermal imaging, infrared cross-modal data) and task expansion (from detection to segmentation, tracking, etc.).

Through scenario diversity, characteristic specificity, and task comprehensiveness, these datasets have constructed a comprehensive data support system for research on object detection in UAV aerial images. They not only provide a benchmark for basic algorithm innovation but also lay a data foundation for application deployment in specific fields, promoting the transformation of UAV vision technology from laboratory research to practical scenario applications.

Table 2. Common UAV Aerial Image Datasets

Name	Size	Categories	Quantity	Remarks
StanfordDrone	640×480	6	48,000 images	Multi-scale social interaction scenarios
CARPK	1280×720	1	1,448 images	Focus on vehicle and license plate detection
UAVDT	1080×540	1	80,000 frames	14 scenario attributes, urban-focused
VisDrone2019	1920×1080	10	10,209 images	Covers 14 cities, multi-task
AU-AIR	1280×720	8	32,283 images	Multimodal dataset, real-time scenario-focused
DroneCrowd	1920×1080	1	33,600 frames	Dense crowd scenarios
FLAME3	4000×3000	2	2,952 images	Includes synchronized RGB and thermal data
DroneVehicle	640×512	5	56,878 images	Cross-modal dataset, covers day-night scenes
UVSD	Multi-size	1	5,874 images	Extended from VisDrone
PDTDataset	640×640	1	5,775 images	Specialized for UAV forestry detection

6. Conclusion

This paper focuses on research on YOLO series algorithms for object detection in UAV aerial images. It systematically sorts out the evolutionary path of the YOLO series from v1 to v12, revealing its key breakthroughs in aspects such as anchor mechanisms (from anchor-free to anchor-based and back to anchor-free), detection head design (from coupled heads to decoupled heads), and feature fusion (from FPN to PANet and further to reparameterized networks).

Given the particularities of UAV scenarios, this paper summarizes core challenges such as insufficient accuracy in small object detection, complex object scales and distributions, and the conflict between real-time performance and accuracy. It also concludes optimization strategies from academia and industry: enhancing feature extraction capabilities by introducing attention mechanisms and deformable convolutions; achieving model lightweighting through network pruning and lightweight modules; and improving multi-scale adaptability by adding small object detection heads and optimizing loss functions.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

In addition, this paper introduces 10 mainstream UAV datasets, providing data support for algorithm verification. The above work presents the research status and technical context of the YOLO series in the field of UAV aerial object detection.

References

[1] Zhong S., Wang L. P.: A Review of Object Detection Technology in UAV Aerial Images, Laser & Optoelectronics Progress, Vol. 62 (2025) No. 10, p.71-89.

[2] Chen J. L., Wu Y. Q., Yuan Y. B.: Research Progress of YOLO Series Algorithms for Object Detection from UAV Perspective, Journal of Beijing University of Aeronautics and Astronautics, (2025) p.1-33. [Accessed 09 August 2025]. https://doi.org/10.13700/j.bh.1001-5965.2024.0420.

[3] Zhang L. L., Tong Q., Liu X. L.: A Review of Small Object Detection in UAV Images, Computer Simulation, Vol. 40 (2023) No. 12, p.1-7+32.

[4] Wang L. Y., Bai J., Li W. J., et al.: Research Progress of YOLO Series Object Detection Algorithms, Computer Engineering and Applications, Vol. 59 (2023) No. 14, p.15-29.

[5] Redmon J., Divvala S., Girshick R., et al.: You Only Look Once: Unified, Real-Time Object Detection, Proc. IEEE Conference on Computer Vision and Pattern Recognition (2016), p.779-788.

[6] Redmon J., Farhadi A.: YOLO9000: Better, Faster, Stronger, Proc. IEEE Conference on Computer Vision and Pattern Recognition (2017), p.7263-7271.

[7] Redmon J., Farhadi A.: YOLOv3: An Incremental Improvement, arXiv Preprint arXiv:1804.02767 (2018).

[8] Bochkovskiy A., Wang C. Y., Liao H. Y. M.: YOLOv4: Optimal Speed and Accuracy of Object Detection, arXiv Preprint arXiv:2004.10934 (2020).

[9] Ge Z., Liu S., Wang F., et al.: YOLOX: Exceeding YOLO Series in 2021, arXiv Preprint arXiv:2107.08430 (2021).

[10] Li C., Li L., Jiang H., et al.: YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications, arXiv Preprint arXiv:2209.02976 (2022).

[11] Wang C. Y., Bochkovskiy A., Liao H. Y. M.: YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023), p.7464-7475.

[12] Wang C. Y., Yeh I. H., Mark Liao H. Y.: YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information, European Conference on Computer Vision (Cham: Springer Nature Switzerland, 2024), p.1-21.

[13] Wang A., Chen H., Liu L., et al.: YOLOv10: Real-Time End-to-End Object Detection, Advances in Neural Information Processing Systems, Vol. 37 (2024), p.107984-108011.

[14] Tian Y., Ye Q., Doermann D.: YOLOv12: Attention-Centric Real-Time Object Detectors, arXiv Preprint arXiv:2502.12524 (2025).

[15] Yang Z. F., Yuan J. Z., Xu C., et al.: A Review of Research Progress on Object Detection Based on YOLO Series, Proc. 28th Annual Conference on Network New Technology and Application 2024 of China Computer Users Association Network Application Branch (Beijing Union University Beijing Key Laboratory of Information Service Engineering; Beijing Union University College of Robotics Brain and Cognitive Intelligence Beijing Laboratory; Beijing Open University, 2024), p.263-267. DOI:10.26914/c.cnkihy.2024.047826.

[16] Xing S. S., Zhao W. L.: A Review of UAV Object Detection in Complex Scenes Based on YOLO Series Algorithms, Application Research of Computers, Vol. 37 (2020) No. S2, p.28-30.

[17] Liu P., Zhang X. H., Zhang Z. L., et al.: A Review of Object Detection from RCNN to YOLO, Proc. 16th National Conference on Signal and Intelligent Information Processing and Application (Tianjin University of Technology and Education School of Information Technology Engineering; Tianjin Sino-German University of Applied Sciences School of Intelligent Manufacturing; Tianjin Huahong Technology Co., Ltd., 2022), p.16-23. DOI:10.26914/c.cnkihy.2022.053359.

[18] Cao Y. Y., Liu Y. C., Gao X. Y., et al.: LightTassel-YOLO: A Real-Time Detection Method for Maize Tassels Based on UAV Remote Sensing, Smart Agriculture (Chinese & English), (2025) p.1-15. [Accessed 09 August 2025]. https://link.cnki.net/urlid/10.1681.S.20250804.0915.002.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

[19] Li J. P., Meng X. L., Hu L. L., et al.: Small Target Crack Detection in Bridges Based on Improved YOLOv8, Journal of Tsinghua University (Natural Science Edition), Vol. 65 (2025) No. 07, p.1260-1271. DOI:10.16511/j.cnki.qhdxxb.2025.26.023.

[20] Shi T., Cui J., Li S.: Algorithm for Real-Time UAV Vehicle Detection by Optimizing and Improving YOLOv8, Computer Engineering and Applications, Vol. 60 (2024) No. 09, p.79-89.

[21] Fan Y. Y., Liu Y. S., Wang Q. S.: Infrared Small Target Personnel Detection Algorithm for Underground Unmanned Vehicles, Transducer and Microsystem Technologies, Vol. 44 (2025) No. 08, p.133-137+142. DOI:10.13873/J.1000-9787(2025)08-0133-05.

[22] Shen C., Zhao J.: Aerial Small Target Detection Algorithm Based on Multi-Feature Cross-Layer Fusion, Journal of Ordnance Equipment Engineering, Vol. 46 (2025) No. 04, p.243-251.

[23] Xu H. Z., Gu X. N.: Optimization Research on Traffic Small Target Image Detection Algorithm from UAV Perspective, Computer Engineering and Applications, Vol. 60 (2024) No. 21, p.194-204.

[24] Bao Z.: The UAV Target Detection Algorithm Based on Improved YOLO V8, Proc. International Conference on Image Processing, Machine Learning and Pattern Recognition (2024), p.264-269.

[25] Wang Z., Zhang X., Li J., et al.: A YOLO-Based Target Detection Model for Offshore Unmanned Aerial Vehicle Data, Sustainability, Vol. 13 (2021) No. 23, p.12980.

[26] Song L., Yu H., Yang Y., et al.: YOLO-VG: An Efficient Real-Time Recyclable Waste Detection Network, Journal of Real-Time Image Processing, Vol. 22 (2025) No. 2, p.79.

[27] Feng X. X., Ren A. H., Qi H.: Lightweight Highway Vehicle Detection Algorithm Based on YOLOv8n, Transducer and Microsystem Technologies, Vol. 44 (2025) No. 07, p.155-158+163. DOI:10.13873/J.1000-9787(2025)07-0155-04.

[28] Wang H., Liu Y.: YOLO-LiRa: Lightweight Detection Algorithm for Small Aerial Targets, Measurement Science and Technology, Vol. 36 (2025) No. 6, p.066009.

[29] Lyu Y., Zhang T., Li X., et al.: LightUAV-YOLO: A Lightweight Object Detection Model for Unmanned Aerial Vehicle Image, J. Supercomput., Vol. 81 (2025) No. 1, p.105.

[30] Shi R., Li T., Yamaguchi Y.: An Attribution-Based Pruning Method for Real-Time Mango Detection with YOLO Network, Computers and Electronics in Agriculture, Vol. 169 (2020), p.105214.

[31] Situ Z., Teng S., Liao X., et al.: Real-Time Sewer Defect Detection Based on YOLO Network, Transfer Learning, and Channel Pruning Algorithm, Journal of Civil Structural Health Monitoring, Vol. 14 (2024) No. 1, p.41-57.

[32] Chen J. H., Wang X. H.: Dense Small Target Detection Algorithm in UAV Aerial Images Based on Improved YOLOv5, Computer Engineering and Applications, Vol. 60 (2024) No. 03, p.100-108.

[33] Zhang Y. T.: Research on Nighttime Pedestrian and Vehicle Detection Based on Improved YOLO Algorithm (MS., Dalian Jiaotong University, China 2025). DOI:10.26990/d.cnki.gsltc.2025.000619.

[34] Wu S., Huang D. D., Liu Z., et al.: Improved YOLOv8n-Based Multi-Scale Object Detection Algorithm for UAV Ground Observation and Its Implementation, Computer Engineering and Applications, (2025) p.1-14. [Accessed 09 August 2025]. https://link.cnki.net/urlid/11.2127.tp.20250707.1324.022.

[35] Ren Z., Yao K., Sheng S., et al.: YOLO-SDH: Improved YOLOv5 Using Scaled Decoupled Head for Object Detection, International Journal of Machine Learning and Cybernetics, Vol. 16 (2025) No. 3, p.1643-1660.

[36] Li M., Yan N.: IPD-YOLO: Person Detection in Infrared Images from UAV Perspective Based on Improved YOLO11, Digital Signal Processing, (2025), p.105469.

[37] Song P. Y., Chen H., Gou H. B.: UAV Object Detection Algorithm Based on Improved YOLOv5s, Computer Engineering and Applications, Vol. 59 (2023) No. 01, p.108-116.

[38] Wu M. J., Yun L. J., Chen Z. Q., et al.: Small Target Detection Algorithm from UAV Perspective Based on Improved YOLOv5s, Computer Engineering and Applications, Vol. 60 (2024) No. 02, p.191-199.

[39] Zhao X., Chen Y.: YOLO-DroneMS: Multi-Scale Object Detection Network for Unmanned Aerial Vehicle (UAV) Images, Drones (2504-446X), Vol. 8 (2024) No. 11.

[40] Tang L., Jia Y., Zhang Y. N.: UAV Detection Algorithm with Bidirectional Multi-Scale Feature Fusion, Computer Engineering and Applications, Vol. 61 (2025) No. 10, p.267-278.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202509_11(9).0009

[41] Sun J. Y., Xu M. J., Zhang J. P., et al.: Optimized and Improved YOLOv8 Object Detection Algorithm from UAV Perspective, Computer Engineering and Applications, Vol. 61 (2025) No. 01, p.109-120.

[42] Robicquet A., Alahi A., Sadeghian A., et al.: Forecasting Social Navigation in Crowded Complex Scenes, arXiv Preprint arXiv:1601.00998 (2016).

[43] Hsieh M. R., Lin Y. L., Hsu W. H.: Drone-Based Object Counting by Spatially Regularized Regional Proposal Network, Proc. IEEE International Conference on Computer Vision (2017), p.4145-4153.

[44] Du D., Qi Y., Yu H., et al.: The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking, Proc. European Conference on Computer Vision (ECCV) (2018), p.370-386.

[45] Du D., Zhu P., Wen L., et al.: VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results, Proc. IEEE/CVF International Conference on Computer Vision Workshops (2019), p.0-0.

[46] Bozcan I., Kayacan E.: Au-Air: A Multi-Modal Unmanned Aerial Vehicle Dataset for Low Altitude Traffic Surveillance, Proc. 2020 IEEE International Conference on Robotics and Automation (ICRA) (IEEE, 2020), p.8504-8510.

[47] Wen L., Du D., Zhu P., et al.: Detection, Tracking, and Counting Meets Drones in Crowds: A Benchmark, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021), p.7812-7821.

[48] Shamsoshoara A., Afghah F., Razi A., et al.: Aerial Imagery Pile Burn Detection Using Deep Learning: The FLAME Dataset, Computer Networks, Vol. 193 (2021), p.108001.

[49] Sun Y., Cao B., Zhu P., et al.: Drone-Based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 32 (2022) No. 10, p.6700-6713.

[50] Zhang W., Liu C., Chang F., et al.: Multi-Scale and Occlusion Aware Network for Vehicle Detection and Segmentation on UAV Aerial Images, Remote Sensing, Vol. 12 (2020) No. 11, p.1760.