

A hybrid YOLOv8-cornernet framework with sequential generative adversarial network augmentation for accurate unmanned aerial vehicle object detection on the UAVision-M3 dataset

Revathi Appavoo^{1*}, Madhuvanthi Thanikachalam¹

¹ Department of Computational Intelligence, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur Campus, Chengalpattu, India

* Corresponding author's e-mail: revathia1@srmist.edu.in

ABSTRACT

Accurate real-time object detection in unmanned aerial vehicle (UAV) is essential for surveillance, autonomous navigation, infrastructure inspection, and intelligent aerial monitoring. However, UAV imagery has significant challenges due to small object sizes, viewpoint variations, illumination changes, occlusions, and dynamic backgrounds. To address these limitations, this study proposes a hybrid UAV object detection framework that integrates sequential generative adversarial network (GAN)-based dataset augmentation with YOLO-CornerNet detection framework. A new multimodal benchmark dataset, UAVision-M3 (multimodal, multi-scenario, multi-object) was developed by combining UAV videos and high-resolution aerial imagery. To enhance the dataset diversity, a sequential CycleGAN-Conditional GAN (cGAN) augmentation pipeline was performed to adapt domain and condition-aware synthesis with annotation consistency. The proposed framework differs from traditional techniques, because a systematic comparative evaluation of YOLOv3-YOLOv9 was conducted under the same training environment to discover the most suitable backbone for hybrid integration. Experimental findings reveal that YOLOv8 achieves the optimal trade-off and is combined with CornerNet to improve localization of small and dense aerial objects through keypoint-based refinement. For UAVision-M3, the hybrid model achieves 95.1% Precision, 93.3% Recall and 97.8% mAP@0.5, which outperforms all YOLO versions alone. Extensive assessments on the Vis-Drone and UAVDT benchmarks further show persistent gains in detection accuracy and localization performance under tough airborne conditions while preserving real-time inference capacity. These findings suggest that sequential GAN-based dataset augmentation and hybrid object localization is a robust and successful approach for UAV object detection in demanding aerial scenarios.

Keywords: UAV object detection, synthetic data generation, CornerNet, YOLO, anchor-free detection, dynamic environments.

INTRODUCTION

Unmanned aerial vehicles (UAVs) have become an essential technology for civilian and military applications because of its versatility, cost-effectiveness and ability to operate in difficult and inaccessible situations. UAVs are widely used for surveillance, disaster management, infrastructure inspection, precision agriculture, environmental monitoring, search and rescue

operations, logistics, and autonomous navigation. However, with the continuous development of sensor technologies, onboard computers and artificial intelligence, their operational capabilities have been greatly improved, such as performing complicated perception and decision-making tasks with minimum human interaction. Therefore, one of the main needs for safe and autonomous UAV operation is to reliably and in real-time detect objects. More and more scientific

studies show the rising relevance of vision systems based on UAVs. Figure 1 describes the history of publications related to UAV object recognition and obstacle detection from 2005 to 2024. The trend shows that the research in this area is growing rapidly. The publication statistics were collected from the Scopus database using the keywords UAVs, object detection, obstacle detection, autonomous navigation, and aerial vision. The number of articles is increasing, reflecting a rising interest in the development of intelligent perception systems that can reliably recognize objects under different operating conditions.

Despite the recent advances in UAV vision technologies, accurate object detection in aerial imagery is still a challenging task. UAV images generally contain small and densely distributed objects with different viewpoints, illumination conditions, motion blurs, occlusions and complex backgrounds. These challenges significantly degrade the object localization and classification performance especially for real-time UAV applications, where reliability of detection directly affects navigation safety and mission accomplishment. These issues have been addressed by various sensing modalities and object identification methods. UAV perception systems consist of vision sensors, LiDAR, radar, infrared cameras, ultrasonic sensors and multimodal sensor fusion depending on the application needs. Among these techniques, deep learning-based vision systems have drawn much interest, since they have achieved an effective balance between detection accuracy and processing efficiency. Figure 2 shows a taxonomy of object recognition techniques and sensing modalities for UAV, including the main types of current methods and

their interactions. These approaches have gained great progress, but successfully recognizing tiny, heavily clustered, and partly occluded objects in complicated aerial scenes remains an outstanding research challenge.

Recent research has shown that the effectiveness of UAV object detectors relies not only on network design, but also on the quality and variety of training data. Generative adversarial networks (GANs) have proved effective in picture synthesis and domain adaptation, enhancing the resilience of deep learning models by synthesizing realistic training data under different environmental circumstances. Similarly, the YOLO family of models have significantly improved real-time object identification with increased feature extraction, multi-scale representation, and efficient detection heads. However, most existing studies are focused on optimizing a single detection architecture or applying a single augmentation strategy, while few studies systematically evaluate multiple detector architectures under the same experimental conditions, or combine complementary localization techniques to improve detection accuracy. The scientific novelty of this research is integration of four complementary contributions into a unified framework for UAV object detection.

- First, a new multimodal benchmark dataset, UAVision-M3 (multimodal, multi-scenario, multi-object) is generated based on UAV videos and high-resolution aerial pictures to reflect varied object categories and demanding environmental circumstances experienced in real-world UAV operations.
- Second, we present a sequential GAN-based augmentation framework with CycleGAN and

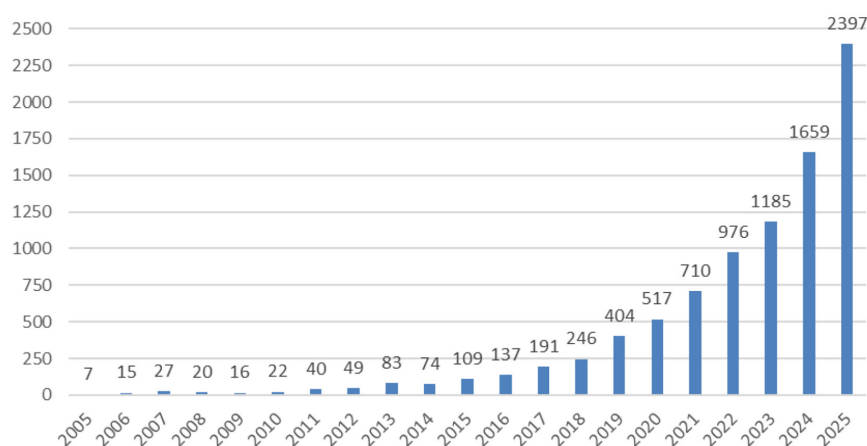


Figure 1. Trends in UAV object detection-related publications (2005–2025)

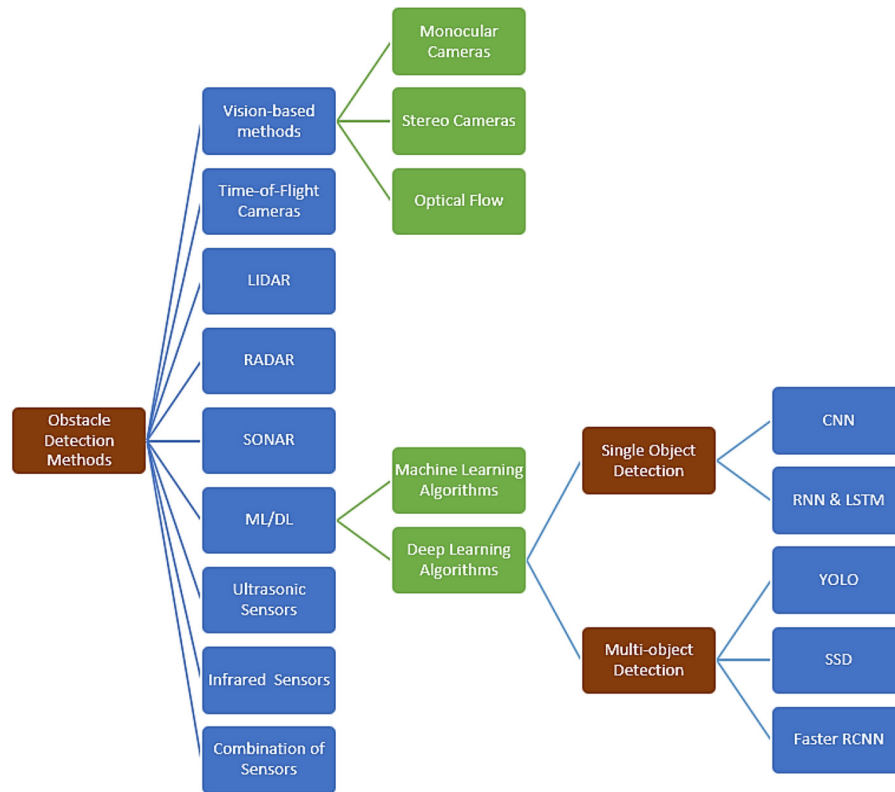


Figure 2. Classification of obstacle detection techniques and algorithms

then Conditional GAN (cGAN) for domain adaptation and condition-aware picture synthesis, significantly enhancing dataset variety while keeping annotation consistency.

- Third, this study provides a systematic comparative assessment of YOLOv3-YOLOv9 under the same experimental settings to choose the most appropriate backbone for UAV object identification, rather than directly using a single detector.
- Finally, the chosen YOLO backbone is included into CornerNet to combine the real-time object detection with the keypoint-based localization refinement to improve the recognition of tiny, densely dispersed and partly occluded aerial objects.

Different from previous works that mainly focus on detector performance or dataset enhancement separately, the proposed framework incorporates dataset diversity, backbone selection and localization refinement into a unified detection pipeline, providing a more comprehensive solution to UAV object detection in complex aerial environments. A novel UAVision-M3 dataset is created and augmented with CycleGAN and conditional GAN (cGAN) to

boost data variety and resilience across diverse environmental circumstances.

BACGROUND

This research [1] studied all the recent breakthroughs of DL-controlled autonomous IoT UAV systems. The article presented a review of DL techniques for autonomous obstacle detection and collision avoidance in the context of the challenges faced by UAV obstacle detection and avoidance. In [2], a monocular vision-based obstacle recognition and collision avoidance system using a MobileNet-SSD CNN is proposed for GPS-denied outdoor conditions using computationally low-range devices without GPU. The network is trained with four classes: autos, people, trees, and bicycles. This is important because in situations like rural regions where GPS connectivity is patchy at best. The proposed UAV object recognition system [3] with ResNet18 as the backbone achieves higher accuracy (IoU of 0.886), 19% faster speed and 18% lower power consumption than Darknet7. DroNet, while faster and more energy-efficient, sacrifices precision. The system reaches 28.5 FPS and 2.7 FPS/W on

the 2018 LPODC dataset, successfully addressing the issues of energy, accuracy and speed on embedded platforms, which makes it suitable for vision related UAVs' real-time applications. [4] this study provides different sensor-based integration strategies that make recognition more reliable. The escape path identified identifiable was divided into two regions, the free region and the obstacle-free region. The research proposes a CNN-Transformer hybrid model with a CSWin transformer backbone for effective recognition of objects in UAV pictures. The key advantages are: improved small object detection using a slicing-based inference method, multi-scale object detection, and hybrid patch embedding for low-level feature extraction. But it has performance gaps, expensive hardware requirements, and problems with pre-trained model effectiveness compared with the SOTA methods.

CIOFL [5] is a collaborative inference based online federated learning for real-time UAV swarm object recognition. CIOFL improves convergence rate and accuracy by 1.3x and 1.34x compared to conventional FL, by enabling UAVs to collaboratively train with high-quality, real-world data, making UAVs more effective in identifying tasks. It relies on continuous, high-quality data gathering and transmission by multiple UAVs, which may put network resources under strain and increase the delay. [6] focuses at providing a novel UAV planning and control framework to address dynamic and static environmental restrictions. The RE-PPO framework [7] can successfully address the wide state space and the partial observation issues in the search for random targets with continuous actions. RE-PPO outperformed the baseline in terms of faster convergence, greater success rate and lower time-out and collision rates. [8] proposed a method to detect a combination of static and moving objects in a low-level neighborhood area and to infer the state of unknown moving obstacles, employing a UAV SAA system sensor configuration design, including a visible light camera and LiDAR.

PP-YOLOE [9] is a refined and efficient object detector based on PP-YOLOv2. It features a novel anchor-free architecture, an improved backbone and an upgraded neck with CSPRepResStage, ET-head and TAL dynamic label assignment. The PP-YOLOE-l variant shows remarkable performance with 51.4 mAP and 78.1 FPS on the COCO test with a Tesla V100. It achieves a remarkable improvement of +1.9 AP

and +13.35% speed compared with PP-YOLOv2, and +1.3 AP and +24.96% speed over YOLOX. The inference speed is as high as 149.2 FPS by using the combination of TensorRT with FP16 accuracy, which further demonstrates the efficacy of this method in industrial applications. VAMYOLOX [10] is a revolutionary anchor-free object detector for UAV optical sensor imaging that is accurate, efficient, and single-stage, and balances accuracy and efficiency. A multi-positive sampling strategy and FL address sample imbalance, increase tiny object feature representation and alleviate occlusion problems, achieving a new record on the VisDrone-DET dataset. The model ODFC-YOLO [11] beats the previous best technique with an improvement of 9.07 mAP points, exhibiting remarkable advancements for detecting foggy objects. YOLOX-TBP technique improves aerial micro-object recognition by integrating the prediction head for multi-scale detection, transformers for global information and a weighted BiFPN for feature fusion [12]. It surpasses SOTA by 0.84% on VisDrone-DET 2021 with 40.27% mAP.

LODNU [13] is a lightweight YOLOv4 based object detection network for UAV, which maintains accuracy on resource-limited devices. LODNU reduces model parameters and enhances important feature extraction through depth-wise separable convolution and coordinate attention. It achieves 31.4% mAP on VisDrone2019 with just 13.1% parameters and 13.6% floating-point operations of YOLOv4. WildARe-YOLO [14] is a lightweight and optimized YOLOv5s for efficient wild animal identification in a resource-constrained distant ecological monitoring scenario. To reduce the computational effort without losing accuracy, the authors replaced the initial Focus layer with a StemBlock connected in parallel, added Depthwise Separable Convolutions (DWConv) and Mobile Bottleneck Blocks (MBBlock) to the backbone, added a BiFPN feature fusion neck, and used Focal-EIoU loss for bounding box regression. On the Wild Animal Facing Extinction dataset, the model achieves 94.4% mAP@0.5, with 28.55% less parameters (5.05M), 50.92% less FLOPs (8.0 GFLOPs), and 17.65% increased inference throughput (37.6 FPS) on low-power edge hardware such as Raspberry Pi 4B. The revised YOLOv5 model [15] trained on the integrated LSDUVD dataset (UCAS-AOD, VisDrone2019 and DOTA-V1.0) improves small object detection by 1.6%, 5.3%,

7.3%, 7.4% and 0.6% for planes, ships, storage tanks, small vehicles and pedestrians respectively. Gabor filter banks enhance the extraction, detection and regression of edge features. The idea is good for small dense items.

[16] presented an artificial intelligence-based system for UAV surveillance for person identification in security applications. The approach was based on thermal imaging using YOLOv5 and YOLOv8 object detection models on an NVIDIA Jetson Xavier NX edge computing platform and was evaluated on a custom thermal dataset including over 10,000 annotated frames captured by UAVs. The trials indicate strong detection performance, YOLOv5 achieved 90.1% accuracy and 93.1% recall, YOLOv8 achieved superior mAP@50–95 of 48.3%. However, the approach presented certain limitations, such as thermal-image generalization, detection confidence in complex forest environments and the absence of hybrid localization techniques. Using the CP dataset of 3,373 pictures, the research [17] developed YOLOv7 models with excellent detection accuracy for weed identification in chicory crops. The variants of YOLOv7 performed better than other detectors, with the highest mAP@0.5 for YOLOv7-x and the highest recall and accuracy for YOLOv7-w6 and YOLOv7-d6.

In [18], the authors proposed DDH-YOLO to tackle the dual issues of low target resolution and computational redundancy in UAV small-object detection. By replacing the standard three-head design with a dual-detection-head (DDH) structure and introducing a C2f-DD feature module and polymorphic feature fusion pyramid network (PFPN), they demonstrated that severe parameter reduction (down to 1.67M parameters) can be achieved without sacrificing performance. On the VisDrone2019 benchmark, their model achieved 40.4% mAP@0.5. While their design focuses primarily on parameter compression for edge devices, it highlights the importance of tailored feature fusion for small aerial targets – a challenge our proposed hybrid framework addresses by coupling anchor-free detection with keypoint localization refinement. A customized version of YOLOv8 to address the issues of tiny pineapple detection [19], significant occlusion by the canopy, and great visual resemblance to neighboring leaves in UAV aerial images. To solve these problems, proposed a dynamic interactive task alignment head (DITAH) to decouple and align the classification and localization tasks, a grouped

multi-scale convolution (GMSC) module with CBAM attention for background noise suppression, a spatial and channel synergistic attention (SCSA) mechanism for overlapping targets, and a Focaler_SIoU loss function to improve the bounding-box regression on small fruits. They achieved good performance on a high-resolution UAV dataset with 2.22M parameters for their model with 94.4% mAP@0.5, 94.6% Precision and 88.9% Recall. I think their task-alignment head and attention mechanisms work well to solve feature conflicts for small crop targets. The main limitations of the study are that the dataset was acquired at a single geographical site and for a single cultivar at an early fruiting stage, only 2D nadir RGB views were considered (which cannot capture pineapples that are completely hidden under leaves) and no multi-spectral or LiDAR sensor fusion was performed.

In [20] a better YOLOv8 model for UAV identification is presented. The model shows good performance on TIB-Net [21] and self-built datasets, but its recall loss in complex air space scenarios implies that it may not generalize to new backgrounds or habitats. WSNs deploy UAVs to help with data collection. A YOLOv8-based model [22] was trained to detect hindrances in the path of UAVs by using the captured aerial images and videos. The YOLOv8 F1 score is 96% after 200 epochs. [23] presented α S-YOLO, an improvement of YOLOv8, to keep a lightweight footprint while modeling long-range spatial interdependence in drone images. Their design adds the C2f_GE module that collects shared global context via efficient 1D channel attention, and a α -SIoU loss function that adaptively speeds up the convergence of directional box distance depending on angular alignment. Tested on the VisDrone-DET2019 benchmark, α S-YOLO obtained 30.4% mAP@0.5:0.95 and 17.2% mAP_s for small targets. However, the main shortcoming of their technique is the lower recall in complex aerial scenes and the ongoing difficulty in localizing very small or substantially obstructed items (e.g., bicycles and tricycles). Recently, the lightweight object identification method of UAV has been improved based on the modification of feature pyramid network and attention mechanism. [24] presented DAFPN-YOLO, which combines an adaptive feature pyramid (Drone-AFPN) and a 160×160 fine-grained detection head, together with Shape-IoU loss, reaching 44.1% mAP@0.5 on VisDrone2019,

and reducing the number of parameters to 9.2M. Similarly, [25] proposed YOLO-Air that substitutes big detecting heads with micro-heads (AeroFPN) and uses Squeeze-Excitation Coordinate Convolutions (SECAConv) together with adaptive scale fusion (ASFM) for 33.3% small-target accuracy (mAP_s). However, both of them have great limits. The inference throughput of DA-FPN-YOLO decreases (149.1 to 117.9 FPS) because to the processing of high-resolution feature maps. YOLO-Air increases GFLOPs more than twice (8.2 to 18.9 GFLOPs). Moreover, both approaches mainly use the typical bounding-box regression head, whereas our hybrid framework proposes to directly couple YOLO backbones with CornerNet keypoint localization refinement to overcome the boundary ambiguity of dense clusters.

NST-YOLO [26] is an improved version of YOLOv10 for detecting small and hard to find objects in complex aerial environments. To alleviate the feature degradation and localization inaccuracy, the authors proposed a Swin Transformer hierarchical attention mechanism in the backbone for global context perception, and a Neighborhood Weighted Distance Loss (NWD-Loss) to optimize the bounding-box regression geometry in dense small-object scenarios. NST-YOLO obtained 94.02% mAP@0.5 and 54.74% mAP@0.5:0.95 on the dataset of real targets and video sequences in 33.87 FPS, which shows its robustness in low-light conditions. Swin Transformer attention and neighborhood-weighted distance loss improve the global feature context and regression convergence. The extra processing cost of transformer blocks is around real-time threshold (~33.8 FPS). In the field of precision agriculture, [27] proposed a low altitude UAV framework based on YOLOv11 and YOLOv12 to identify tiny green tea buds (7–12 mm) and to create georeferenced spatial yield maps. Their work showed that YOLOv11-L obtained a mAP@0.5 of 83.95% on 5 m UAV images and that they found a good linear connection ($R^2 = 0.72$) between the projected number of buds and the harvested fresh-leaf weight. However, numerous drawbacks impede its widespread deployment: flying at ultra-low altitudes (~5 m) drastically limits field coverage rates and requires manual flight control, while extensive canopy occlusion and visual resemblance to young leaves lead to missed detections. [28] proposed MFR-YOLO, a modified architecture

based on YOLOv12, to tackle severe geometric distortions, small object density and background noise in UAV vision. The multi-scale feature extraction module (MSFEM) integrates the SPD-Conv with the Deformable Convolution (DCNv4) to adaptively change the size and direction of the receptive fields while preserving fine-grained information. A dual-branch global attention module (GAM) removes intricate background noise, and an upgraded C3K2-PPA bottleneck fuses local and global contextual data. MFR-YOLO achieved 40.1% and 54.6% mAP@0.5 on VisDrone2021 and UA-DETRAC at real-time frame rates (>128 FPS). MFR-YOLO combines DCNv4 with global attention to provide outstanding detection accuracy (40.1% mAP@0.5 on VisDrone2021). However, the primary drawbacks are structural complexity and a minor drop in inference throughput (from 140 FPS to 128 FPS compared to baseline YOLOv12). Bounding-box regression heads may help to better locate keypoints in the boundaries of thick or occluded regions.

Recent research proposed CS-YOLO (content-conscious and scale-sensitive YOLO) [29], an enhanced detection framework to enrich feature responses and background perception in UAV aerial photography with dense micro-objects and significant background interference. Three architectural innovations were proposed including a bottleneck attention module-cross-stage partial (BAM-CSP) to integrate channel attention to amplify small-target features, a multi-scale pooling attention fusion module (MPAFM) to suppress complex environmental clutter with the help of parallel pooling, and an FDFM to capture feature changes during the sampling. The gradient constraint weight of bounding-box regression was optimized via the GA-EIoU (Gradient Adaptive-Efficient IoU) loss function. CS-YOLO achieved 22.6% mAP@0.5:0.95 (+2.7% improvement over YOLO11n) on the VisDrone2019 benchmark and 53.8% on the HazyDet dataset, demonstrating high recall and robustness to environmental conditions in aerial inspection applications.

CycleGAN [30], a popular unsupervised image-to-image translation system capable of learning mappings between two picture domains without paired training data. The model utilizes two generators and two discriminators with a cycle consistency loss to retain the semantic contents of the pictures in the bidirectional translation.

CycleGAN has been extensively used in applications such as style transfer, object transfiguration, season translation and picture enhancement due to its capacity to conduct realistic domain adaptation [31]. Recent advancements in GANs [32] have significantly improved image-to-image translation techniques. Models such as Pix2Pix, CycleGAN, DualGAN, and cGAN have demonstrated effective domain adaptation and image generation capabilities for both paired and unpaired datasets. Despite notable improvements in image quality and translation accuracy, challenges including mode collapse, training instability, and limited diversity remain. These developments have established GAN-based frameworks as essential tools for applications in object detection, image enhancement, and autonomous vision systems. As shown in Table 1, the available UAV benchmark datasets are state-of-the-art with their unique limitations (e.g., fewer target classes, no environmental diversity, and single-sensor-based). This comparative review explicitly points out the study gap and emphasizes the

need of developing and enriching our innovative multimodal UAVision-M3 dataset.

METHOD

The proposed architecture has four consecutive stages: (1) generation of UAVision-M3 dataset, (2) data augmentation and domain adaption based on GAN, (3) assessment of different YOLO backbones, and (4) development of YOLO–CornerNet hybrid detector. Different from traditional techniques focused on a single object detector directly, the proposed methodology systematically investigates several YOLO topologies in the same experimental environment to discover the best backbone for object identification on UAVs. Then the chosen backbone is combined with CornerNet for better localization accuracy. The hybrid detector is evaluated on both the proposed UAVision-M3 dataset and public UAV benchmarks. The whole workflow of the proposed system is shown in Figure 3.

Table 1. Benchmark UAV datasets for object detection: features and applications

Dataset	Class	Size	Challenges	Applications
VisDrone	People; cars; bikes; trucks	10,000+ img; 96 vid	Urban/suburban scenes; dense traffic	Object detection; tracking; counting
UAVDT	Vehicles	23,000+ img	Weather; lighting variations	Vehicle detection and tracking
TinyPerson	People (tiny objects)	7,000+ img	Very small objects; crowd	Small/tiny object detection
AU-AIR	People; cars; bikes; trucks	32,000+ frames	Multiple viewpoints and altitudes	Autonomous driving; UAV-based perception
UAVDT-ODT	Vehicles	Subset of UAVDT	Low light; occlusion	Object detection for vehicles
DOTA	Planes; ships; vehicles	2,800 img	Various altitudes and backgrounds	General aerial object detection
Okutama-Action	People; vehicles; actions	77,000+ frames in 43 vid	High-resolution video; action detection	Action detection; multi-object tracking
DroneDeploy	Buildings; trees; vehicles	Commercially available	Agricultural/construction settings	Agriculture; construction; infrastructure
Stanford Drone	Pedestrians; vehicles	60,000+ frames	Complex urban scenes	Tracking; trajectory analysis
Wind-AID	Wind-powered drones; objects	Experimental	Windy conditions	Wind-powered UAV tracking
SeaDronesSee	Boats; humans	Maritime-focused img	Maritime environment; water reflections	Maritime surveillance; search and rescue
Aerial Maritime	Ships; boats	Maritime-focused img	Occlusion; water effects	Maritime surveillance
MODD	Marine vessels	Maritime img	Small object detection; waves	Search and rescue; marine monitoring
Manipal-UAV	People	13,462 img	unconstrained environment; camera altitude	crowd detection and counting; action detection; and person tracking
HIT-UAV [33]	Bicycle; car; person; other vehicle; DontCare	2,898 IR thermal img	Trade-off	Object detection; search and rescue mission at night
DroneVehicle [34]	Car; truck; bus; van; freight car	28,439 RGB – Infrared img	Occlusion and crowding	Traffic monitoring; research; aerial surveillance

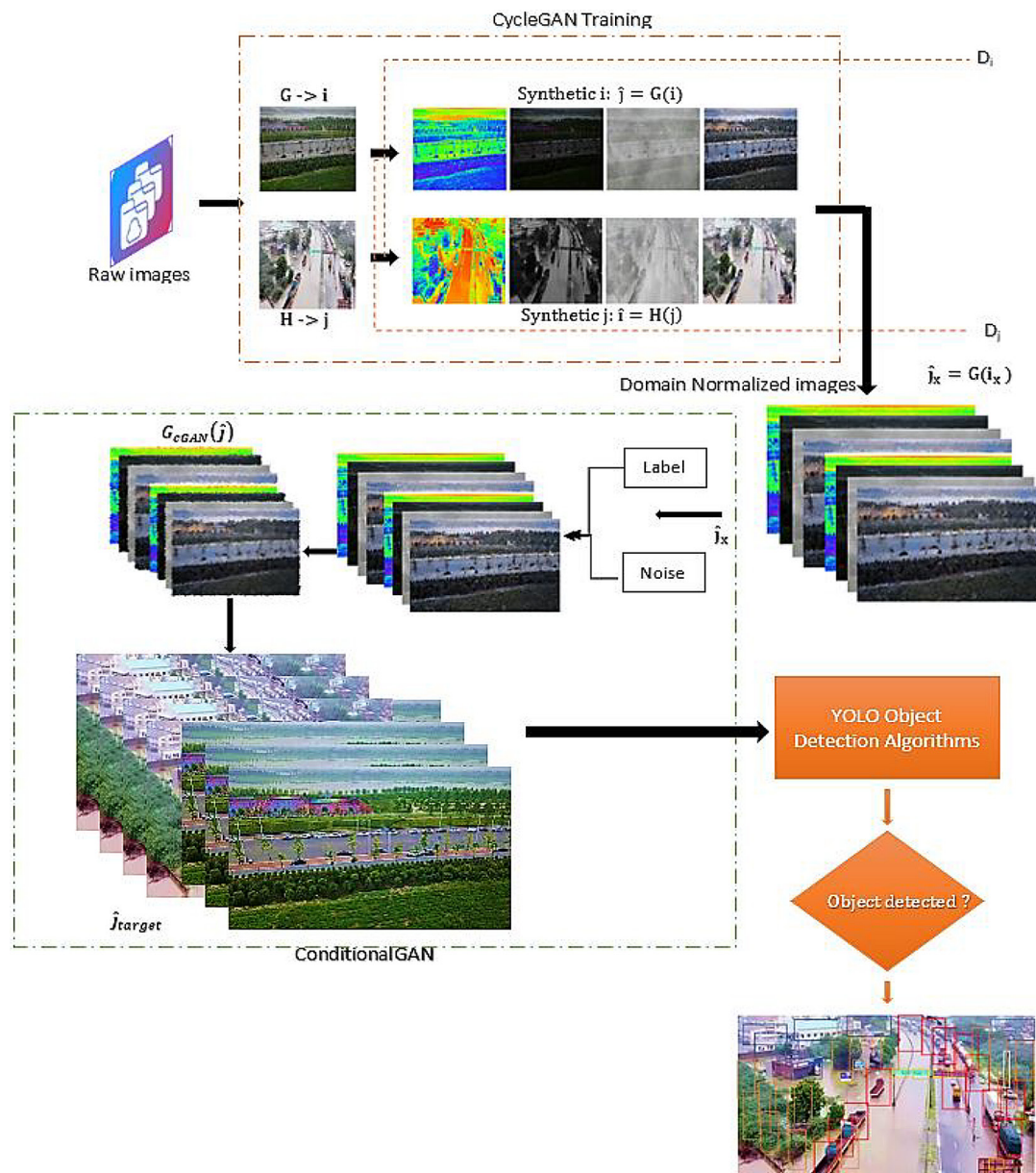


Figure 3. Integrated framework for domain adaptation and object detection using GANs and YOLO

Hardware setup

A machine running Windows 11 (64-bit) workstation with an Intel® Core™ i5-10210U processor (1.60 GHz base clock, up to 4.20 GHz Turbo Boost) and 16 GB of RAM were used for dataset gathering, annotation management, preprocessing, and visualization. All deep learning models were trained using Google Colab Pro cloud resources with NVIDIA Tesla T4 GPU (16 GB GDDR6 VRAM) due to the computational cost of training on the upgraded UAVision-M3 dataset. The software environment comprised of Python 3.10, PyTorch 2.1, CUDA 12.1 and cuDNN 8.9. So, it is called

as hybrid local–cloud computational pipeline. The UAV configuration is based on the AIR Drone model which is created from a combination of plastic, metal and advanced electronics. This drone operates on a 2.4 GHz receiving frequency and has four channels. The UAV is equipped with folding characteristics, making it portable and deployable in different operational settings. The AIR Drone is powered by a 3.7 V, 1800 mAh lithium battery and takes roughly 95 minutes to fully charge. It offers a dual charging (solar and USB) and a long flying endurance when fully charged. It has three settings, rapid, medium and slow for customers to customize their flying experience. The AIR

Drone is equipped with an upgraded 4K camera with facial recognition. The camera's lens angle may be manually changed to 90° for flexible framing, and Wi-Fi connection allows for the smooth sharing of information. Integrated camera cover protects against accidental impact.

Construction of the UAVision-M3 dataset

The UAVision-M3 dataset is provided to foster robust UAV-based computer vision research in object recognition, multi-object tracking and autonomous aerial navigation. Unlike traditional benchmarks, UAVision-M3 tackles the problems of aerial imaging, including tiny item representation, large perspective fluctuations, dynamic camera movements, altitude-dependent scale changes, cluttered backdrops, and different environmental conditions. The collection consists of 180 UAV video sequences, 145,000 sample frames obtained by systematic frame sampling, and 6,200 high-resolution aerial photographs. These illustrations show humans, automobiles, buildings, birds, drones, electrical infrastructure, flora and fake cutouts. The original dataset has 151,200 annotated samples and 1.45 million bounding box annotations for nine object categories: people, automobile, bird, building, drone, electrical line, pole, tree, and cuts. Also, 1,369 multi-modal sensor readings from UAV systems are employed in sensor fusion and autonomous decision-making research. To provide an unbiased performance assessment, the data set is separated into 70% for training, 15% for validation, and 15% for testing.

Augmentation of UAVision-M3 dataset using a sequential GAN-based image transformation pipeline

The dataset contains a variety of aerial perspectives; nevertheless, few samples of diverse meteorological situations may limit the generalization of deep learning models. To address this restriction, a sequential GAN-based data augmentation method is demonstrated in Figure 4, using CycleGAN and ConditionalGAN (cGAN). In the absence of UAV shots, aerial and satellite photos are employed as additional source domain data. CycleGAN learns bidirectional mappings between source and destination domains to translate aerial and UAV data in an unsupervised way. Adversarial learning and cycle-consistency constraints preserve structural information and provide realistic lighting and weather effects, shadows and backgrounds in translated photos. This domain normalization step aids the detector in learning domain-invariant visual representations.

cGAN synthesizes domain-normalized images of CycleGAN into condition-aware images. cGAN, unlike CycleGAN, combines object class and spatial information as conditional inputs to generate realistic object instances and image patches, while preserving semantic consistency and annotation integrity. So, the generated images show superior intra-class variations, object appearances, and contextual realism without explicit labelling. The generated samples are automatically linked to the original object annotations and added to the main UAVision-M3 dataset to enrich the training corpus.

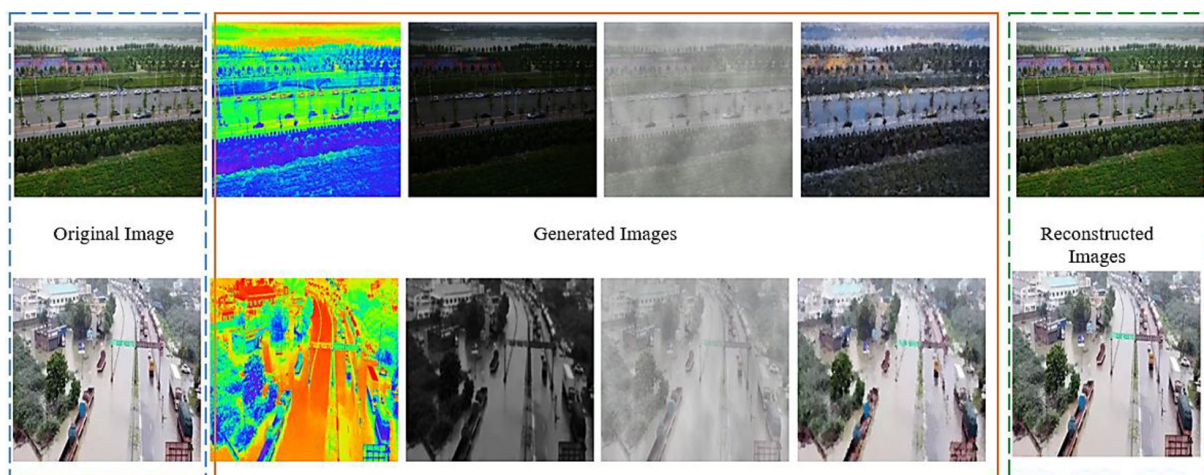


Figure 4. Sequential GAN based image transformation pipeline for UAVision-M3 dataset augmentation

Table 2 summarizes the quantitative impact of the suggested augmentation approach. The original UAVision-M3 dataset comprises 151,200 annotated samples with ~1.45 million bounding box annotations and takes 18 GB storage. The first augmentation step is domain adaptation using CycleGAN to transform the aerial and satellite images to the UAV-like images. The augmented dataset contains 241,500 images with over 2.30 million annotations and compressed size of 32 GB. Then the CycleGAN produced samples are sent to the conditional GAN (cGAN) which does the condition-aware picture synthesis to increase object look, contextual realism and environmental variety while keeping annotation consistency results in a final dataset with 322,800 samples and about 3.10 million annotations with a compressed size of 48 GB. The successive augmentation

method increases the dataset size on average by a factor of 2.14, while keeping the annotations consistent and considerably enhancing the diversity of the environment. The model was trained using Adam optimizer with learning rate 0.0002 and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. In the first half of training, the learning rate was kept constant, and then linearly decreased to zero in the second half to induce stable convergence. Training was done with batch size of 1 because of memory limits with high resolution frames. The quality and efficiency of image transformation was measured with a number of quantitative metrics by analyzing the CycleGAN and conditional GAN (cGAN) architecture at many stages during training. The enhanced dataset is then used as training corpus for the proposed YOLO–CornerNet hybrid detection method.

Algorithm 1

Start

Initialize CycleGAN:

- Initialize generators G and H
- Initialize discriminators D_i and D_j

CycleGAN Training Loop:

- For each epoch and for each batch of real images i and j
 - Forward Pass
 - Generate fake images: $\hat{j} = G(i), \hat{i} = H(j)$
 - Reconstruct images: $i_{recon} = H(\hat{j}), j_{recon} = G(\hat{i})$
 - Computing Losses
 - Adversarial losses: $L_{GAN}(G, D_j, i, j), L_{GAN}(H, D_i, j, x)$
 - Cycle-consistency losses: $L_{cy}(G, D_i, x, j)$
 - Overall loss: $L_{Total} = L_{GAN} + \alpha L_{cy}$
 - Backpropagation
 - Updating G and H using total loss
 - Updating D_i and D_j using adversarial losses
- Saving the trained CycleGAN models G and H

Image Generation using CycleGAN

- For each image i_x in the source domain i
 - Generating images correspondingly $\hat{j}_x = G(i_x)$

Initialize cGAN:

- Initialize cGAN generator G_{cGAN} and discriminator D_{cGAN}

cGAN Training loop:

- For each epoch and for each batch of CycleGAN-generated images $\hat{j}$ and real target images j_{target}
 - Forward pass
 - Generate fake target images: $\hat{j}_{target} = G_{cGAN}(\hat{j})$
 - Computing Losses
 - Adversarial loss: $L_{GAN}(G_{cGAN}, D_{cGAN}, \hat{j}, j_{target})$
 - Conditional loss: $L_{con}(G_{cGAN}, \hat{j}, j_{target})$
 - Overall loss: $L_{Total} = L_{GAN} + \beta L_{con}$
 - Backpropagation
 - Updating G_{cGAN} using total loss
 - Updating D_{cGAN} using adversarial loss
- Saving the trained cGAN model as G_{cGAN}

End

Algorithm 1.

Table 2. Growth of UAVision-M3 dataset via sequential GAN-based augmentation

Stage	Video frames	Static images	Total samples	Bounding boxes	Size
Base Dataset	145,000	6,200	151,200	1.45 M	18 GB
Base Dataset + CycleGAN	232,000	9,500	241,500	2.3 M	32 GB
Base Dataset + CycleGAN + cGAN (Final)	310,000	12,800	322,800	3.1 M	48 GB

The adversarial training and high-resolution aerial picture synthesis needed more processing power. Thus, cGAN took around 24–36 cumulative hours to complete 300 training epochs, whereas CycleGAN required approximately 48–60 cumulative hours to complete 300 training epochs. Model checkpoints were saved frequently and training was continued across numerous sessions until convergence to handle cloud session limits.

Multiple YOLO variants for backbone selection

Instead of using a single YOLO architecture based on earlier reported findings, the proposed framework conducts a full comparative assessment of several YOLO variations to discover the most suitable backbone for UAV-based object recognition. We separately trained and assessed eight sample models, i.e. YOLOv3, YOLOv4, YOLOv5, YOLOv6, YOLOv7, YOLOv7-tiny, YOLOv8, and YOLOv9 using the enhanced UAVision-M3 dataset with the same experimental parameters. By using a shared training setup, we verify that the observed differences in performance across the models are due only to architectural factors, and not because to differences in optimization tactics or training circumstances.

Before training, all input images are pre-processed with a uniform pipeline to guarantee all detection models are consistent. As given in Table 3, all YOLO variations are trained using the same optimization settings. We use the stochastic gradient descent (SGD) optimizer with momentum for its steady convergence characteristics for object identification problems. For all trials, the initial learning rate of 0.01, momentum of 0.937 and weight decay of 5×10^{-4} are used. All models are trained individually with a maximum of 300 epochs, a batch size of 16, and an input resolution of 640×640 pixels. Since Google Colab Pro has session constraints, we implemented an automatic checkpointing system

that allows us to continue extended experiments from the most recent stored checkpoint, spanning several sessions. The cosine annealing schedule is used to dynamically alter the learning rate, which stabilizes the convergence in the later stage of training. To fairly evaluate the computing cost, all tests are performed on the same hardware platform.

All models are part of the YOLO family yet each architecture brings specific advantages in feature extraction and object localization. YOLOv3 uses the Darknet-53 backbone and multi-scale prediction, giving a robust baseline for tiny object identification. YOLOv4 improves the feature representation by using CSPDarknet53, spatial pyramid pooling (SPP) and path aggregation network (PAN) that increases the feature fusion at many scales. YOLOv5 proposes a lightweight design based on efficient bottleneck modules, which reduces the computational complexity while retaining competitive detection accuracy. YOLOv6 considerably improves inference efficiency with an anchor-based detection head and an efficient decoupled prediction method for industrial deployment. YOLOv7 adopts Extended Efficient Layer Aggregation Networks (E-ELAN) to increase gradient propagation for better feature learning and faster convergence. YOLOv7-tiny is tailored for the embedded UAV systems with low compute resources, achieving fast inference speed with decreased model complexity. YOLOv8 employs an anchor-free detection approach with a decoupled detection head and an enhanced feature pyramid network, leading to a considerable boost in localization accuracy for tiny and densely clustered objects. Finally, YOLOv9 incorporates Programmable Gradient Information (PGI) and the generalized efficient layer aggregation network (GELAN) to enhance feature propagation and mitigate information loss during deep network optimization.

The backbone with the best overall performance on UAVision-M3 dataset is selected as the ideal feature extraction network of the

Table 3. Hyperparameters for training on All YOLO variants

Hyperparameter	Value
Input image size	640 × 640
Batch size	16
Number of epochs	300
Optimizer	SGD
Initial learning rate	0.01
Momentum	0.937
Weight decay	0.0005
Learning rate scheduler	Cosine annealing
Confidence threshold	0.25
IoU Threshold (NMS)	0.45
Data split	70% Train / 15% Validation / 15% Test

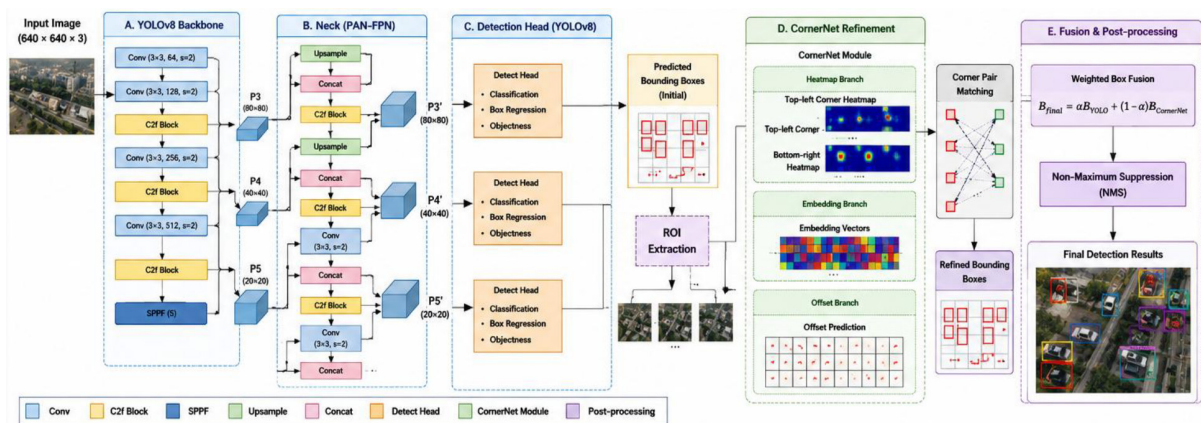
proposed hybrid detection framework. The decision is not based on the improvement of a single evaluation parameter, but rather on the entire trade-off among localization accuracy, computing efficiency and resilience for all the object categories. The anchor-free detection technique solved the localization ambiguity of tiny aerial objects, and the detached detection head separately enhanced classification and bounding-box regression. Furthermore, the upgraded feature pyramid network has the capability to provide more informative multi-scale feature representations, which made it more robust for detecting tiny, dense, and partially occluded objects in UAV images. YOLOv8 has better Precision, Recall, and mAP compared to previous YOLO versions, with competitive inference performance, and is the best choice for integration with the CornerNet localization refinement module.

Proposed YOLO-CornerNet hybrid detection framework

The complete architecture of the proposed YOLOv8–CornerNet hybrid detection system for precise and real-time UAV object identification is illustrated in Figure 5. The framework integrates YOLOv8's efficient feature extraction and multi-scale object identification capabilities with the accurate keypoint-based localization method of CornerNet. The overall architecture comprises of five main modules: (A) YOLOv8 backbone, (B) Neck (PAN-FPN), (C) Detection head, (D) CornerNet localization refinement and (E) Fusion and post-processing.

YOLOv8 backbone

The first level of the framework takes as input the preprocessed UAV picture with a defined spatial resolution (e.g., $640 \times 640 \times 3$). The input image is sent via YOLOv8 backbone, which extracts hierarchical visual characteristics at different semantic levels. The backbone consists of an initial convolutional layer, many C2f (Cross Stage Partial with feature fusion) blocks and a spatial pyramid pooling fast (SPPF) module. The convolution layers continuously decrease the spatial resolution while expanding the number of feature channels, allowing the network to learn both low-level textural information and high-level semantic information. The C2f blocks enhance the feature propagation with efficient gradient flow and feature reuse at relatively low computational cost. Finally, the SPPF module expands the receptive field by gathering contextual information from many pooling scales, which allows the detector to gather objects of different sizes,

**Figure 5.** Detailed YOLOv8–CornerNet hybrid architecture

which are often seen in UAV images. The backbone provides three multi-scale feature maps, P3, P4 and P5, which are high, middle and low-resolution feature representations respectively. These feature maps store fine spatial information for recognizing tiny items, while also include rich semantic information for bigger objects.

Neck (PAN-FPN)

Then, the recovered feature maps are sent to the neck network, where the path aggregation network (PAN) and feature pyramid network (FPN) are used for the enhancement of multi-scale feature fusion. The neck uses sequential up sampling, concatenation and C2f operations to fuse high-level semantic data with low-level spatial information. This two-way flow of information allows the detector to use contextual information from many feature resolutions concurrently. The neck produces a set of improved multi-scale features (P3', P4', and P5') which have better object representations across various object sizes. This feature aggregation is especially useful for UAV photography, when the target size varies dramatically due to changes in altitude and diverse camera viewpoints.

Detection head of YOLOv8

The revised feature maps are processed by three detecting heads at various scales separately. Each detection head forecasts the class, confidence score and bounding-box coordinates of the candidate objects. YOLOv8 is different to earlier YOLO versions since it uses an anchor-free detection technique, thereby doing away with the need for pre-defined anchor boxes. Instead, we directly regress object positions from feature maps, which reduces the hyperparameter adjustment linked to anchors and improves localization flexibility. The detector does simultaneously: object categorization, confidence estimate, bounding box regression. This one-stage prediction technique enables YOLOv8 to provide real-time inference with excellent detection accuracy for numerous UAV object types. The result of this step is the first set of bounding-box recommendations which are potential areas for the next localization refinement.

CornerNet localization refinement

YOLOv8 is capable of producing very precise object suggestions, but it may still

encounter localization issues when recognizing tiny, overlapping, or partly occluded objects. To overcome this constraint, we propose a CornerNet refinement module in the framework. First, regions of interest (ROIs) are acquired from the projected bounding boxes based on YOLOv8. These ROIs keep the candidate object areas with enough surrounding context. CornerNet then processes each ROI for object localization using keypoint-based representation instead of straight bounding-box regression. The CornerNet module is composed of three complementing branches:

- Heatmap branch: forecasts probability distributions of keypoints in top-left and bottom-right corners for each item.
- Embedding Branch: Learning associative embeddings to appropriately associate the expected corner keypoints with the same item.
- Offset Branch: predicts sub-pixel offsets to alleviate the spatial quantization caused by feature down sampling.

The paired corner predictions of CornerNet are used to rebuild the improved object boundaries with far greater localization accuracy than the standard regression-based detectors. This anchor-free refining method is especially applicable to UAV applications, because the aerial photos often include tiny objects with blurred outlines.

Fusion and postprocessing

In the last step, a weighted fusion method is used to merge the detection results from YOLOv8 and CornerNet. YOLOv8 can give reliable object recognition and confidence estimate, whereas CornerNet can provide more precise border localization. Thus, the fused bounding boxes may inherit the advantages of both detectors, which provide better localization accuracy without losing the inference time. After the fusion, the duplicate detections are eliminated using non-maximum suppression (NMS) based on the Intersection-over-Union (IoU) overlap criterion. The remaining detections are the final prediction results, which include optimized bounding boxes, confidence ratings, and item class labels. The system is able to recognize all the target categories specified in the UAVision-M3 dataset, including people, cars, birds, buildings, drones, electricity wires, poles, trees and cuts.

Algorithm 2

Start

Input Image Processing:

- Load input image I
- Resize I to match YOLO and CornerNet's input dimensions I'

Feature Extraction Using a Shared Backbone:

- Pass I' through a shared feature extractor F_θ : $F_\theta(I') = X$; X is the extracted feature map

Split feature maps into two branches:

- YOLO Branch: Processes grid-based detection
- CornerNet Branch: Processes keypoint-based detection

YOLO predicts object locations:

- Using bounding box regression equation, $\hat{B} = (x, y, w, h, c)$
 - Compute YOLO loss

$$L_{\text{YOLO}} = \lambda_{\text{coord}} \sum_i^N 1_{ij}^{\text{obj}} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] \\ + \lambda_{\text{size}} \sum_i^N 1_{ij}^{\text{obj}} \left[(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] \\ + \lambda_{\text{coord}} \sum_i^N 1_{ij}^{\text{obj}} (c_i - \hat{c}_i)^2$$

CornerNet Keypoint Detection:

- CornerNet detects object corners using top-left (p_{tl}) and bottom-right (p_{br}) keypoints: $B_{\text{corner}}(\hat{p}_{tl}, \hat{p}_{br}) = \text{CornerNet}(X)$

- Compute CornerNet loss using Heatmap Loss:

$$L_{\text{heatmap}} = -\frac{1}{N} \sum_i \begin{cases} (1 - \hat{p}_i)^\alpha \log(\hat{p}_i) & \text{if } y_i = 1, \\ (1 - y_i)^\beta (\hat{p}_i)^\alpha \log(1 - \hat{p}_i) & \text{otherwise} \end{cases}$$

- Compute CornerNet loss using Offset Loss:

$$L_{\text{offset}} = \sum \|p_{tl} - \hat{p}_{tl}\|^2 + \|p_{br} - \hat{p}_{br}\|^2$$

Bounding box fusion from YOLO & CornerNet:

- Combine YOLO bounding boxes $\hat{B}$ and CornerNet bounding boxes B_{corner} :

$$B_{\text{final}} = \alpha B_{\text{YOLO}} + (1 - \alpha) B_{\text{corner}}$$

Non-Maximum Suppression (NMS):

- Apply NMS using an IoU threshold T_{NMS}

Output Final detections:

- Return final object detections B_{final} with labels and confidence scores
 - The total loss function integrates both YOLO and CornerNet losses:

$$L_{\text{total}} = \lambda_{\text{YOLO}} L_{\text{YOLO}} + \lambda_{\text{CornerNet}} (L_{\text{heatmap}} + L_{\text{offset}})$$

End

Algorithm 2.

EXPERIMENTAL RESULTS AND ANALYSIS

All models were trained and assessed on the same experimental circumstances, including the same dataset splits, preprocessing techniques, hyperparameter settings and hardware resources to guarantee a fair comparison evaluation. The proposed methodology is assessed using the standard metrics of object detection, i.e., Precision, Recall, F1-score, mAP@0.5 and mAP@0.5:0.95 with respect to True Positives (TP), False Positives (FP) and False Negatives (FN). These metrics evaluate the performance of the detector

in terms of the capacity to reliably identify and locate items in complex UAV settings with targets such as people, cars, birds, buildings, poles, electrical cables, trees, and drones. Precision is the ratio of accurately identified items to the total number of detected objects; whereas Recall is a measure of how many of the ground truth items the detector successfully detected and it is calculated by Equation 1 and Equation 2:

$$\text{Precision} = TP / (FP + TP) \quad (1)$$

And

$$\text{Recall} = TP / (TP + FN) \quad (2)$$

High Precision means that the detector has less false alarms which is critical especially for UAV navigation applications where not misclassifying the backdrop (e.g. shadows or landscape) as obstacles is crucial to prevent unnecessary avoidance movements, while high Recall is critical for UAV safety, since missed detections (e.g. birds, utility poles, or electrical lines) may create dangerous circumstances during autonomous flying operations. The F1-score is the harmonic mean of Precision and Recall, which can be assessed by Equation 3:

$$F1 - score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (3)$$

Equation 3 offers a fair evaluation of detection performance since it takes into account false positive and false negative at the same time. The mean average precision at IoU threshold of 0.5 (mAP@0.5) measures the overall detection performance under a relatively loosened localization constraint. A detection is considered to be accurate if the predicted bounding box has the score of at least 0.5 with the ground-truth annotation correspondingly. mAP@0.5:0.95 is calculated based on COCO assessment methodology, which computes average precision (AP) averaged across IoU criteria from 0.50 to 0.95 with a step size of 0.05. This measure offers a more rigorous evaluation of localization quality and is especially useful for UAV photography since tiny objects usually comprise only a few pixels.

Finally, the inference efficiency and the real-time deployment of each model are quantified by frames per second (FPS). The higher the FPS, the shorter the inference latency which is important for time sensitive UAV applications. Collectively, these indicators provide a thorough assessment of the accuracy and real-time performance of the proposed framework.

Performance analysis on UAVision-M3 dataset

Evaluation of backbone selection for the UAVision-M3

This experiment found the best appropriate YOLO backbone for UAV-based object identification before the introduction of the CornerNet localization refinement module. The testing circumstances are kept the same, therefore the variations in performance seen can be attributed only to architectural advancements from one generation of YOLO to the next. The experimental findings are presented in Table 4. YOLOv8 provides the greatest overall detection performance among the tested detectors, being the most balanced in terms of localization accuracy, recall, inference speed and computational complexity. Therefore, the proposed hybrid detection framework adopts YOLOv8 as the backbone.

With the evolution of YOLO generations, there are progressive advances in the backbone architecture, the multi-scale feature extraction and detection head design. We find that YOLOv3 and YOLOv4 provide competitive baseline performance but are hampered in their capacity to identify extremely tiny airborne objects. YOLOv5 and YOLOv6 enhance the localization accuracy by efficient feature aggregation and improved detecting heads. YOLOv7 uses the E-ELAN architecture to boost gradient propagation, whereas YOLOv7-tiny has the greatest inference speed but suffers from the loss of detection accuracy. For YOLOv9, such additional computing complexity only provides modest accuracy gain while achieving same performance. The superiority of because

Table 4. Evaluation of standalone YOLO models on UAVision-M3

Model	Precision (%)	Recall (%)	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameters (M)	FPS	Training time (h)
YOLOv3	82.6	80.8	81.7	84.2	58.9	61.5	54	~24
YOLOv4	84.9	83.2	84.0	86.5	61.4	64.4	58	~28
YOLOv5	88.3	87.6	87.9	90.6	67.2	7.2	74	~12
YOLOv6	89.5	88.8	89.1	91.7	69.5	18.5	81	~16
YOLOv7	90.6	90.1	90.3	92.8	71.8	36.9	83	~20
YOLOv7-tiny	87.8	86.5	87.1	89.6	64.8	6.2	148	~8
YOLOv8	92.4	91.8	92.1	94.1	75.8	4.5	91	~10
YOLOv9	91.9	91.0	91.4	93.6	74.7	7.3	86	~14

to its efficient architecture design. Concretely, the standard C3 blocks are replaced by a C2f (cross-stage partial with two convolutions) module, which concatenates intermediate bottleneck outputs to boost the multi-path gradient propagation and feature reuse with a compact parameter footprint. Moreover, the anchor-free detection technique directly predicts object center coordinates without depending on pre-defined anchor priors. The solution solves the anchor mismatch problem for aerial targets with variable scales and minimizes the number of candidate predictions processed in the post-processing procedure, which helps to maintain high inference speed. Furthermore, the detached detection head performs classification and bounding-box regression with separate parallel branches, which alleviates the task interference during optimization. Hence, YOLOv8 offers a good compromise between detection accuracy, computational economy and inference speed, achieving 94.1% mAP@0.5 at 91 FPS with just 4.5 million parameters.

Evaluation of YOLO–CornerNet for the UAVision-M3

The suggested localization refinement technique is verified by combining CornerNet with each form of YOLO after picking the best YOLO backbone. The goal is to measure the contribution of CornerNet without dependence on the backbone architecture. All hybrid models are trained using the same dataset and experimental parameters as the solo detectors.

Table 5 shows the effect of adding the CornerNet localization refinement module to different YOLO backbones. The solo YOLO models are the same as those provided in Table 4 so that the comparison is fair under the same experimental

parameters. Adding CornerNet consistently increases Precision, Recall, F1-score and mAP by around 2.5–3.3% across all architectures. This shows that keypoint-based boundary refinement complements the coarse bounding-box regression of YOLO. The highest improvements are found for older YOLO versions, where CornerNet considerably improves the localization accuracy for tiny and densely packed airborne objects. The increase is especially noticeable in Precision and mAP, suggesting an improvement in bounding-box estimate and a reduction in false positives. YOLOv8–CornerNet obtains the best performance among all hybrid configurations with 95.1% precision, 93.3% recall, and 97.8% mAP@0.5 and keeps the real-time inference capabilities. In Table 5, Δ mAP (%) is the simple increase of mAP after adding CornerNet to the relevant YOLO model.

Figure 6 presents the comparative evaluation of standalone YOLO architectures along with their corresponding YOLO–CornerNet hybrid models on the proposed UAVision-M3 dataset. The results reveal that the performance advantages are not only due to the better feature extraction, but also the complementing localization capabilities brought by CornerNet, which makes the hybrid architecture suitable for the hard task of UAV object recognition.

Performance analysis on VisDrone and UAVDT dataset

Table 6 presents the performance of the proposed YOLO–CornerNet hybrid framework on two popular UAV object identification benchmarks, namely VisDrone and UAVDT. The experimental findings demonstrate that all hybrid models have superior detection

Table 5. Evaluation of YOLO and YOLO–CornerNet hybrid models on UAVision-M3

Model	Precision (%)	Recall (%)	F1-score (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	FPS	Δ mAP (%)	Training time (h)
YOLOv3 + CornerNet	85.2	83.4	84.3	87.5	62.8	50	+3.3	~29
YOLOv4 + CornerNet	87.4	85.8	86.6	89.7	65.3	54	+3.2	~34
YOLOv5 + CornerNet	90.9	89.8	90.3	93.5	71.0	69	+2.9	~15
YOLOv6 + CornerNet	92.0	91.0	91.5	94.6	73.4	76	+2.9	~20
YOLOv7 + CornerNet	93.1	92.0	92.5	95.6	75.5	78	+2.8	~24
YOLOv7-tiny + CornerNet	90.0	88.9	89.4	92.5	68.9	136	+2.9	~10
YOLOv8 + CornerNet	95.1	93.3	94.2	97.8	80.4	84	+2.9	~14
YOLOv9 + CornerNet	94.3	92.7	93.5	96.2	78.8	80	+2.6	~18

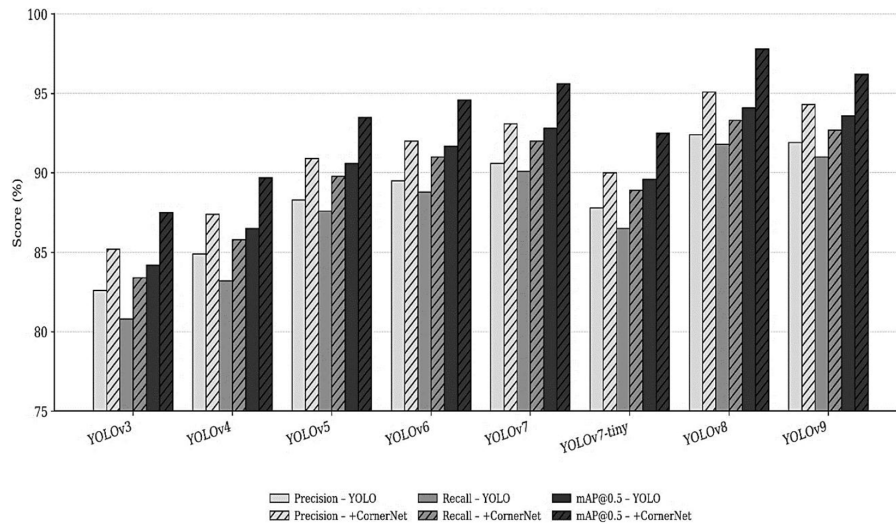


Figure 6. Standalone YOLO and YOLO–CornerNet hybrid models on the UAVision-M3 dataset – comparison of precision, recall, and mAP@0.5

performance on VisDrone compared to UAVDT, mostly because UAVDT exhibits more demanding scenarios like motion blur, wide perspective fluctuations, scale changes and intense traffic settings. YOLOv9-CornerNet also achieves competitive results but the improvement over the suggested method is modest, while it introduces more computational complexity. Previous hybrid designs like YOLOv3–CornerNet and YOLOv4–CornerNet had somewhat worse performance due to less effective feature extraction for tiny aerial objects. The lightweight models YOLOv7-tiny–CornerNet are faster to infer, but with less capacity, they lose detection accuracy. The proposed YOLOv8-CornerNet obtains the best performance among all examined models on both datasets. It gets 93.3% Precision, 92.4% Recall and 95.1% mAP@0.5 on VisDrone, while 92.1% Precision, 91.0% Recall and 94.0% mAP@0.5 on UAVDT.

Comparative discussion with existing methods

Some drawbacks of current UAV object detection motivated the proposed YOLOv8-CornerNet system. Existing methods, e.g., α S-YOLO, MFR-YOLO, DAFPN-YOLO, and CS-YOLO, mainly use the standard bounding-box regression process, which may cause localization error of tiny, dense, and partly occluded aerial objects. Besides, transformer-based modules, deformable convolutions, and multi-scale feature fusion may also enhance the feature representation. However, they often bring extra processing cost, which limits the applicability in real-time UAV scenarios.

To solve these issues, we provide a framework that blends the lightweight anchor-free design of YOLOv8 with the keypoint-based localization refinement of CornerNet. Experimental assessment indicates that the solo YOLOv8 obtained

Table 6. Performance evaluation of YOLO–CornerNet hybrid models using VisDrone and UAVDT datasets

Hybrid model	VisDrone precision (%)	VisDrone recall (%)	VisDrone mAP@0.5 (%)	UAVDT precision (%)	UAVDT recall (%)	UAVDT mAP@0.5 (%)
YOLOv3 + CornerNet	82.7	80.9	84.3	80.9	79.1	82.8
YOLOv4 + CornerNet	85.5	83.7	87.1	83.6	82.4	85.5
YOLOv5 + CornerNet	89.6	88.2	91.3	87.8	86.9	90.2
YOLOv6 + CornerNet	90.5	89.4	92.1	88.9	87.8	91.4
YOLOv7 + CornerNet	91.4	90.6	93.0	89.8	89.0	92.5
YOLOv7-tiny + CornerNet	87.4	85.6	89.5	85.3	83.5	87.8
YOLOv8 + CornerNet	93.3	92.4	95.1	92.1	91.0	94.0
YOLOv9 + CornerNet	92.6	91.8	94.2	91.3	90.4	93.2

94.1% mAP@0.5 with 91 FPS, while the suggested YOLOv8–CornerNet model increased detection performance to 97.8% mAP@0.5 and 80.4% mAP@0.5:0.95 with a real-time throughput of 84 FPS. Furthermore, the suggested technique has performed better than previous hybrid variations such as YOLOv7–CornerNet (95.6% mAP@0.5, 78 FPS) and YOLOv9–CornerNet (96.2% mAP@0.5, 80 FPS). The results show a 3.7% mAP gain over the YOLOv8 baseline with a slight decrease in inference speed, demonstrating the effectiveness of incorporating CornerNet in addressing the localization issues reported in prior UAV object detection works while maintaining computational efficiency.

CONCLUSIONS

In this research, a hybrid YOLO–CornerNet architecture is proposed for robust UAV-based object detection in complicated aerial situations. A new dataset, UAVision-M3 was constructed and augmented using a sequential CycleGAN–Conditional GAN (cGAN) augmentation to represent challenging dynamic environmental conditions and diverse categories of object, which commonly encountered in UAV operations. The augmentation technique boosted the dataset from 151,200 to 322,800 samples, with almost 3.1 million high-quality item annotations and improved variety of training samples.

To find the most appropriate backbone for UAV object detection, a comprehensive experiment was carried out under identical experimental settings to evaluate the YOLOv3–YOLOv9. The comparison study indicated the superiority of YOLOv8 in terms of the balance of the detection accuracy, computing efficiency and real-time inference and was chosen as the backbone of the proposed hybrid framework. Based on the experimental findings, YOLOv8 was integrated with CornerNet for keypoint-based localization refinement, which greatly decreased the bounding-box localization errors while maintaining the fast detection performance of YOLOv8. The proposed framework achieved 95.1% Precision, 93.3% Recall and 97.8% mAP@0.5 on UAVision-M3 benchmark, outperforming the evaluated standalone YOLO models. The experimental results on UAVision-M3, VisDrone and UAVDT datasets verified consistent improvements in Precision, Recall and mAP. The results demonstrates that the combination of GAN-based dataset improvement with

hybrid object localization detection mechanism is a realistic and scalable approach for UAV perception.

Although the proposed framework shows promising performance, several drawbacks are still shown. Since the experiments were performed with a fixed training setup, no multi-seed statistical validation, and the framework was assessed mostly on RGB imagery. Future work will focus on analyzing statistical significance by repeating experiments with different random seeds, extending the framework to multimodal sensing data such as thermal and LiDAR imagery, exploring lightweight transformer-based detection architectures for resource-limited UAV platforms, and validating the proposed approach in real-world autonomous flight scenarios.

REFERENCES

1. Fraga-Lamas P., Ramos L., Mondéjar-Guerra V., Fernández-Caramés T.M. A review on IoT deep learning UAV systems for autonomous obstacle detection and collision avoidance. *Remote Sensing*. 2019;11(18): 2144. <https://doi.org/10.3390/rs11182144>
2. Levkovits-Scherer, D.S., Cruz-Vega, I., Martinez-Carranza, J. (2019). Real-Time Monocular Vision-Based UAV Obstacle Detection and Collision Avoidance in GPS-Denied Outdoor Environments Using CNN MobileNet-SSD. In: Martínez-Villaseñor, L., Batyrshin, I., Marín-Hernández, A. (eds) *Advances in Soft Computing. MICAI 2019. Lecture Notes in Computer Science()*, vol 11835. Springer, Cham. https://doi.org/10.1007/978-3-030-33749-0_49
3. Deng J., Shi Z., Zhuo C. Energy-efficient real-time UAV object detection on embedded platforms. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*. 2019; 39(10): 3123-7. <https://doi.org/10.1109/TCAD.2019.2957724>
4. Ramli M.F., Shamsudin S.S. Obstacle detection technique to solve poor texture appearance of the obstacle by categorising image's region using cues from expansion of feature points for small UAV. *International Journal of Computational Vision and Robotics*. 2023; 13(1): 91-115. <https://doi.org/10.1504/IJCVR.2023.127300>
5. Wu F., Dong C., Qu Y., Sun H., Zhang L., Wu Q. CIOFL: Collaborative inference-based online federated learning for UAV object detection. In: *2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*. Denver, CO, USA, 2022. 258-259. <https://doi.org/10.1109/MASS56207.2022.00043>
6. Tu G.T., Juang J.G. UAV path planning and obstacle avoidance based on reinforcement learning in 3d

- environments. *Actuators*. 2023; 12(2): 57. <https://doi.org/10.3390/act12020057>
7. Pethő M., Zsedrovits T. UAV obstacle detection with bio-motivated computer vision. In: 2021 17th International Workshop on Cellular Nanoscale Networks and their Applications. Catania, Italy, 1-4. <https://doi.org/10.1109/CNNA49188.2021.9610823>
8. Huang X., Wang W., Ji Z., Cheng B. Representation enhancement-based proximal policy optimization for UAV path planning and obstacle avoidance. *International Journal of Aerospace Engineering*. 2023; 6654130. <https://doi.org/10.1155/2023/6654130>
9. Xu S., Wang X., Lv W., Chang Q., Cui C., Deng K., Wang G., Dang Q., Wei S., Du Y., Lai B. PP-YOLOE: An evolved version of YOLO. *arXiv preprint arXiv:2203.16250*. 2022; 16250. <https://doi.org/10.48550/arXiv.2203.16250>
10. Yang Y., Gao X., Wang Y., Song S. VAMYOLOX: An accurate and efficient object detection algorithm based on visual attention mechanism for UAV optical sensors. *IEEE Sensors Journal*. 2022; 23(11): 11139-55. <https://doi.org/10.1109/JSEN.2022.3219199>
11. Fang W., Zhang G., Zheng Y., Chen Y. Multi-task learning for uav aerial object detection in foggy weather condition. *Remote Sensing*. 2023; 15(18): 4617. <https://doi.org/10.3390/rs15184617>
12. Wang S., Xu G., Song Q., Yang Z., Xue Y. A uav object detection algorithm based on improved yolox. In: 5th International Conference on Intelligent Robotics and Control Engineering (IRCE). 2022; 98-102. <https://doi.org/10.1109/IRCE55557.2022.9963076>
13. Chen N., Li Y., Yang Z., Lu Z., Wang S., Wang J. LODNU: lightweight object detection network in UAV vision. *The Journal of Supercomputing*. 2023; 79(9): 10117-38. <https://doi.org/10.1007/s11227-023-05065-x>
14. Bakana S.R., Zhang Y., Twala B. WildARe-YOLO: A lightweight and efficient wild animal recognition model. *Ecological Informatics*. 2024; 80:102541. <https://doi.org/10.1016/j.ecoinf.2024.102541>
15. Zhang H., Shao F., He X., Zhang Z., Cai Y., Bi S. Research on object detection and recognition method for UAV aerial images based on improved YOLOv5. *Drones*. 2023; 7(6): 402. <https://doi.org/10.3390/drones7060402>
16. Gutt J., Kaczor M., Paleta G., Polec K., Stiborski K., Strzpek F., Czyba R., Czekalski P., Domin J., and Gebeyehu, N. Artificial intelligence you only look once – based unmanned aerial system for remote sensing in security surveillance. *Advances in Science and Technology Research Journal*. 2024; 18(7): 364-378. <https://doi.org/10.12913/22998624/193368>
17. Gallo I., Rehman A.U., Dehkordi R.H., Landro N., La Grassa R., Boschetti M. Deep object detection of crop weeds: Performance of YOLOv7 on a real case dataset from UAV images. *Remote Sensing*. 2023; 15(2): 539. <https://doi.org/10.3390/rs15020539>
18. Liu Y., Yang Z. DDH-YOLO: A dual-head YOLOv8 model for small object detection in UAV aerial images. *Signal, Image and Video Processing*. 2025; 816. <https://doi.org/10.1007/s11760-025-04429-5>
19. Xue Z., Liu Y., Chen Y., Dong M., Hao X., Shen W., Zhang Z., Sun H., Li H., Wang Y. YOLO-pineapple: enhanced pineapple detection in UAV images using an optimized YOLOv8 model. *Expert Systems with Applications*. 2026; 323: 132486. <https://doi.org/10.1016/j.eswa.2026.132486>
20. Zhai X., Huang Z., Li T., Liu H., Wang S. YOLO-Drone: an optimized YOLOv8 network for tiny UAV object detection. *Electronics*. 2023; 12: 3664. <https://doi.org/10.3390/electronics12173664>
21. Sun H., Yang J., Shen J., Liang D., Ning-Zhong L., Zhou H. TIB-Net: Drone detection network with tiny iterative backbone. *Ieee Access*. 2020; 15(8): 130697–707. <https://doi.org/10.1109/ACCESS.2020.3009518>
22. Rahman S., Rony J.H, Uddin J., Samad M.A. Real-time obstacle detection with YOLOv8 in a WSN using UAV aerial photography. *Journal of Imaging*. 2023; 9(10): 216. <https://doi.org/10.3390/jimaging9100216>
23. Hou W., Wu H., Wu D., Shen Y., Liu Z., Zhang L., Li J. Small object detection method for UAV remote sensing images based on α s-YOLO. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2025; 18: 8984-8994. <https://doi.org/10.1109/JSTARS.2025.3539873>
24. Wang H., Zhang Y., Zhu C. DAFPN-YOLO: An improved UAV-based object detection algorithm based on YOLOv8s. *Computers, Materials and Continua*. 2025; 83(2): 1929-49. <https://doi.org/10.32604/cmc.2025.061363>
25. Qiu J., Cai F, Fu N., Y. Yao. “YOLO-Air: An Efficient deep learning network for small object detection in drone-based imagery. *IEEE Access*. 2025; 13: 79718–79735. <https://doi.org/10.1109/ACCESS.2025.3565560>
26. He J., Wang W. NST-YOLO: Improved YOLOv10 model for small target UAV detection. *Ain Shams Engineering Journal*. 2025; 16(12): 103787. <https://doi.org/10.1016/j.asej.2025.103787>
27. Prasangini N., Jang G., Hyeon-Seok P., Sub A.H., Kim H.J. Detection of green tea buds and spatial mapping using low-altitude UAV imagery and object detection models. *Smart Agricultural Technology*. 2026; 14: 102068. <https://doi.org/10.1016/j.atech.2026.102068>
28. Ge J., Lv H., Guo Y., Chen Y., Xu Y., Yang X. MFR-YOLO: advancing UAV object detection

- with multi-scale feature refinement via deformable convolution and global attention. *Scientific Reports*. 2026; 16: 15587. <https://doi.org/10.1038/s41598-026-45641-8>
29. Fan R., Jiao R., Nan W., Meng H., Jiang A., Yang X., Zhao Z., Dang J., Wang Z., Tian Y., Dong B. CS-YOLO: A small object detection model based on YOLO for UAV aerial photography. *Signal Processing: Image Communication*. 2025; 142:117460. <https://doi.org/10.1016/j.image.2025.117460>
30. Alotaibi A. Deep generative adversarial networks for image-to-image translation: A review. *Symmetry*. 2020;12(10):1705.<https://doi.org/10.3390/sym12101705>
31. Zhu, J.Y., Park, T., Isola P., Efros A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017; 2223–2232. <https://doi.org/10.48550/arXiv.1703.10593>
32. Shamsolmoali P, Zareapoor M, Granger E, Zhou H, Wang R, Celebi ME, Yang J. Image synthesis with adversarial networks: A comprehensive survey and case studies. *Information Fusion*. 2021; 72:126-46. <https://doi.org/10.1016/j.inffus.2021.02.014>
33. Suo J., Wang T., Zhang X., Chen H., Zhou W., Shi W. HIT-UAV: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection. *Scientific Data*. 2023; 10(1): 227. <https://doi.org/10.1038/s41597-023-02066-6>.
34. Sun Y., Cao B., Zhu P., Hu Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*. 2022; 32(10): 6700–13. <https://doi.org/10.1109/TCSVT.2022.3168279>