Rethinking Detection Heads: Enhancing YOLO for Drone Image Object Detection

Rutvik Patel
Indian Institute of Technology Delhi
India
rutvik2900@gmail.com

Anureet Chhabra
Indian Institute of Technology Delhi
India
siy257576@iitd.ac.in

Ankita Raj
Indian Institute of Technology Delhi
India
ankita.raj@cse.iitd.ac.in

Chetan Arora
Indian Institute of Technology Delhi
India
chetan@cse.iitd.ac.in

Abstract

This work addresses the problem of object detection in drone imagery, which presents distinct challenges compared to ground-based detection due to factors such as variable flight altitudes, weather conditions, and lighting variations. Existing state-of-the-art detectors, including YOLO-based architectures, often struggle with detecting ultra-small objects, which frequently appear in aerial perspectives due to the higher vantage point of drones. Our analysis attributes this limitation to the absence of a dedicated detection head tailored for ultra-small object detection. As our first contribution, we introduce an additional ultra-small-object detection head into the YOLOv11 architecture. Furthermore, empirical analysis reveals that the large-object detection head offers negligible benefit in aerial scenarios, where targets are typically captured from altitudes exceeding 10 meters. Consequently, our second contribution involves removing the large-object detection head, thereby simplifying the architecture without sacrificing accuracy. On the VisDrone benchmark, our approach attains a mean Average Precision (mAP) of 40.62%, surpassing the current state of the art. Moreover, the proposed modifications are modular and transferable, enabling improved small-object detection across other YOLO-based detectors. Full source-code and pretrained models will be made publicly available upon acceptance.

CCS Concepts

• **Computing methodologies** → **Object detection**; *Vision for robotics.*

Keywords

Drone Image Object Detection, 2D Object Detection, YOLO

ACM Reference Format:
Rutvik Patel, Anureet Chhabra, Ankita Raj, and Chetan Arora. 2025. Rethinking Detection Heads: Enhancing YOLO for Drone Image Object Detection. In *Indian Conference on Computer Vision, Graphics, and Image Processing (ICVGIP 2025), December 17–20, 2025, Mandi, India.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3774521.3774563

ICVGIP 2025, Mandi, India

ACM ISBN 979-8-4007-1930-1/2025/12
https://doi.org/10.1145/3774521.3774563

1 Introduction

Drone-image Object Detection. Unmanned Aerial Vehicles (UAVs), commonly known as drones, are increasingly being deployed across a wide range of applications, including environmental monitoring, disaster response, precision agriculture, infrastructure inspection, and defense [1, 9, 14]. These drones heavily depend on accurate and efficient object detection to operate and interact with their surroundings. Robust object detection allows drones to recognize obstacles, track targets, and make informed navigation decisions in real time, directly impacting their safety, efficiency, and task effectiveness [5]. Although, deep learning techniques have significantly advanced object detection for natural images, many state-of-the-art detectors [2, 27, 34, 40] developed for such settings often struggle to maintain high performance when applied to aerial images captured from drones. Unlike fixed or ground-based cameras, operating drones at varying heights and encounter diverse weather and lighting conditions (c.f. Figure 1, introduces substantial variability in object scale and appearance. Additionally, due to payload and power limitations, the deployed models must be lightweight enough to run on edge devices without having to sacrifice accuracy. Therefore, balancing detection performance and real-time inference speed is critical for practical deployment in UAV-based applications.

Current State-of-the-Art. Drone-image object detection systems are broadly categorized based on where the processing takes place: (i) **ground-based systems** and (ii) **onboard (drone-mounted) systems**. The choice depends on application requirements like latency, accuracy, and communication bandwidth. Ground-based processing is viable in civilian scenarios when bandwidth is sufficient and high detection accuracy is desired. In such scenarios, transformer-based detectors such like DQ-DETR [11] and DNTR [20] perform well in dense scenes using dynamic queries, deformable attention, and denoising features. However, their significant computational demands and inference latency make them impractical for deployment on lightweight drone hardware. Onboard detection is critical in real-time or long-range UAV operations where transmitting high-resolution images to a remote server is slow, unreliable, or simply infeasible, e.g., in most defense applications. Lightweight architectures like YOLC [19] and YOLOv11 [13] are better suited

for these scenarios, but often yield lower detection performance. Thus, the trade-off between computational efficiency and detection accuracy plays a central role in selecting the appropriate system architecture for drone-based object detection. In this work, we focus on lightweight models like YOLOv11 [13], aiming to improve their performance while maintaining real-time suitability for onboard deployment.

Key Insights. Our ablation studies reveal a key limitation of YOLOv11 in detecting ultra-small objects (smaller than 16×16 pixels), which are particularly common in aerial imagery. A closer analysis suggests that this shortcoming stems from the absence of a detection head specifically designed for such small object scales. The default YOLOv11 architecture employs three detection heads to handle large, medium, and small objects, a configuration that performs well in ground-level image scenarios where object sizes are more evenly distributed. However, this design is less suited to drone-based settings, where the vast majority of objects are small or ultra-small, and large objects are rare (see Figure 2). As a result, the existing design fails to provide sufficient resolution or attention to the smallest object scales, leading to a noticeable drop in detection performance for such cases.

Our Proposal. We propose a lightweight architectural modification to YOLOv11, aimed at improving detection performance in drone imagery, particularly for ultra-small objects. Our method introduces a dedicated detection head optimized for ultra-small object scales, which are common in aerial scenes, while removing the large-object detection head, which is typically redundant in drone views. This design improves detection performance with a limited computational overhead, making it suitable for real-time, onboard deployment. The proposed detection head is modular and can be integrated into other YOLO-based architectures to enhance small object detection in drone images.

Contributions Main contributions of this work are:

(1) We add a detection head tailored for ultra-small objects and remove the large-object head, aligning the architecture with the object size distribution typically found in aerial imagery.
(2) The proposed design is modular and can be seamlessly integrated into existing YOLO-based detectors to improve small object detection.
(3) When integrated with YOLOv11, our method achieves state-of-the-art performance on the VisDrone dataset, with a mAP of 40.62.
(4) Our approach consistently improves over YOLOv11 and SPAR-YOLO on both VisDrone and AI-TODv2 datasets. The proposed YOLOv11 model gains 1.65 mAP points overall, and 1.82 mAP points on small objects and the proposed architecture with SPAR gains 1.5mAP on the VisDrone dataset .

2 Related Work

2.1 Datasets for Drone-image Object Detection

Effective drone-image object detection requires datasets that capture the unique challenges of aerial imaging, including small object sizes, high-density scenes, large scale variations, and dynamic environmental conditions. Among the most widely used datasets is VisDrone[36], which provides over 10,000 high-resolution images and 2.6 million annotated objects across 10 categories such as pedestrians and vehicles. The dataset captures diverse scenarios involving varying altitudes, weather, and lighting conditions, making it a benchmark for real-world UAV applications. AI-TOD [31], is another popular dataset that focuses specifically on tiny object detection, where the average object size is just 12.8 pixels, pushing the limits of detection algorithms for extremely small and sparsely distributed targets. Other datasets such as DOTA [32] and DroneVehicle[28] are also considered for drone imaging. While DOTA [32] is primarily used in satellite and military imagery analysis, DroneVehicle [28] provides a cross-modal benchmark with RGB and infrared imagery captured from drones. Together, these datasets serve as foundational resources for training and evaluating object detectors in drone-image contexts.

2.2 General-Purpose Object Detection

Object detection has traditionally been dominated by convolutional neural network (CNN) based models. Two-stage detectors like Faster R-CNN [27] achieved strong accuracy by generating region proposals followed by classification and bounding box regression, while one-stage detectors like YOLO [26] and SSD [22] offered faster inference with direct dense predictions. These models have evolved significantly over time, with newer versions of YOLO[12, 29] and RetinaNet [17] achieving a better trade-off between speed and accuracy, making them widely adopted in both academic and industrial settings.

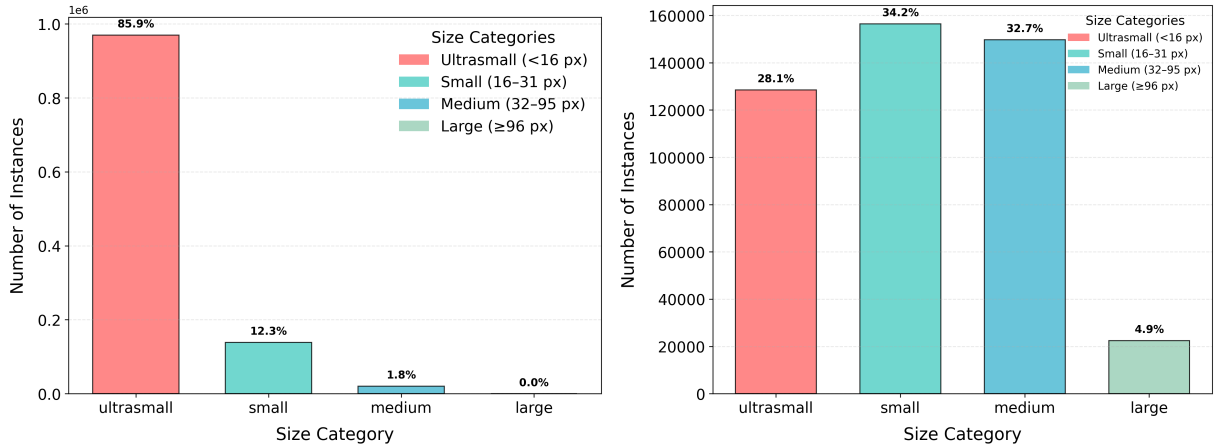
In recent years, there has been a shift towards transformer-based models, particularly those based on the DETR framework [4]. Models like DINO [34], Group DETR v2 [6], and Co-DETR [40] represent state-of-the-art solutions on datasets like COCO [18]. DINO improves query quality and training speed through contrastive denoising and dynamic anchor boxes, while Group DETR v2 scales performance further with group-wise training and large-scale pretraining. Co-DETR pushes the frontier by introducing collaborative hybrid assignments and auxiliary query heads to address the sparsity problem in positive sample generation. Other notable contributions include Deformable DETR [39], which improves multi-scale attention, and Conditional DETR [25], which improves convergence using conditional queries. Despite their accuracy, the large parameter counts, high latency, and computational demands of these models make them impractical for onboard UAV deployment, especially under real-time constraints and limited edge resources.

2.3 Drone-Specific Object Detection

Recent advances in drone-image object detection can be broadly categorized into transformer-based and CNN-based approaches, each targeting the challenges of tiny object detection, real-time processing, and resource efficiency. Transformer-based models like DQ-DETR [11] and DNTR [20] focus on improving detection in dense aerial scenes. DQ-DETR introduces dynamic query scaling, density-guided feature enhancement, and learnable positional encodings to improve detection of small and clustered objects. DNTR addresses feature noise in multi-scale fusion using denoising-based contrastive learning, improving feature clarity for tiny targets.

(a) (b) (c) (d)

Figure 1: (a) Variation in object appearance and scale due to changing drone altitude. (b) Changes in the drone's camera angle lead to a change in perspective and features .(c) Small and densely packed objects in drone images. (d) Lighting changes caused by time of day.



(a) AI-ToDv2 dataset. **(b) VisDrone dataset.**

Figure 2: Size-wise distribution of objects in (a) AI-ToDv2 and (b) VisDrone datasets.

In contrast, CNN-based architectures like YOLOv10 [30] and YOLC [19] emphasize lightweight inference for onboard deployment. YOLC improves detection through a Local Scale Module that zooms into dense regions, Gaussian Wasserstein Distance for more accurate regression of tiny objects, and high-resolution heatmaps for improved localization. YOLOv10 takes a different approach by removing Non-Maximum Suppression (NMS) and introducing architectural efficiencies that balance accuracy, inference speed, and memory usage, making it particularly well-suited for onboard drone deployment. Building upon this trend, YOLOv11 [13] incorporates advanced components such as the C3k2 (Cross Stage Partial with kernel size 2) block, SPPF (Spatial Pyramid Pooling – Fast), and C2PSA (Convolutional block with Parallel Spatial Attention), collectively enhancing detection speed, robustness, and precision. These models reflect a shift toward designs that prioritize real-time performance and edge compatibility without compromising detection quality for small and densely packed objects.

A third category combines transformer elements with YOLO backbones. TPH-YOLOv5 [37] augments YOLOv5 with Transformer Prediction Heads and scale-aware modules, yielding improvements in accuracy and robustness to scale variation and motion blur common in aerial footage. SPAR-YOLO [16] fuses multi-scale features from YOLOv8 with CLIP embeddings to construct graph representations, enabling graph convolution for enhanced contextual

reasoning. While these hybrid approaches achieve competitive accuracy, their high computational overhead can hinder deployment on resource-limited UAV platforms.

3 Proposed Methodology

3.1 Baseline Architecture

YOLOv11 [13] is a recent addition to the YOLO series, optimized for real-time object detection with high accuracy and low latency. The core architecture of YOLOv11, like other YOLO models, comprises three fundamental components: the backbone, responsible for feature extraction; the neck, which handles feature aggregation and enhancement; and the head, which generates the final predictions. The backbone of YOLOv11 involves a series of convolutional layers and custom blocks that generate features at multiple resolutions. It introduces efficient convolutional blocks (C3K2) and an attention mechanism (C2PSA) that helps focus on important regions in the image. The SPPF (Spatial Pyramid Pooling Fast) block helps capture multi-scale contextual information efficiently by pooling features at different spatial scales without significantly increasing computation. The neck, which also uses C3K2 blocks for speed and C2PSA blocks for spatial attention, aggregates feature maps from different resolutions and passes them to the detection head. The head generates the final predictions, *i.e.* bounding boxes, class probabilities, and confidence scores.

YOLOv11 employs three detection heads to handle objects at different scales. Assuming an input feature map of size $H \times W = 640 \times 640$, the neck produces three feature maps at different resolutions:

$$\text{Head}_{\text{small}} : \frac{H}{8} \times \frac{W}{8} = 80 \times 80 \tag{1}$$

$$\text{Head}_{\text{medium}} : \frac{H}{16} \times \frac{W}{16} = 40 \times 40 \tag{2}$$

$$\text{Head}_{\text{large}} : \frac{H}{32} \times \frac{W}{32} = 20 \times 20 \tag{3}$$

Each detection head takes its respective feature map and predicts bounding boxes and class probabilities at every spatial location. The 80×80 feature map provides finer spatial resolution, making it more effective for detecting small objects. The 40×40 and 20×20 feature maps target medium and large objects, respectively. This multi-scale detection mechanism allows YOLOv11 to robustly detect objects of varying sizes across the image.

3.2 Proposed Ultra-Small Object Detection Head

A key limitation identified in our baseline analysis is that the feature map associated with $\text{Head}_{\text{small}}$ lacks the spatial granularity necessary for accurately localizing ultra-small objects (i.e., instances occupying less than 16 pixels on the image plane). This deficiency is particularly problematic in UAV-based imagery, where objects are often captured from significant altitudes, resulting in severe scale reduction. Dataset statistics reinforce this challenge: ultra-small objects constitute approximately 28.1% of all annotated instances in the VisDrone dataset and an overwhelming 85.9% in the AI-TODv2 dataset, as shown in Figure 2b and Figure 2a.

To address this issue, we enhance the YOLOv11 neck by incorporating an additional high-resolution feature map at 160×160 resolution. This resolution corresponds to $\frac{H}{4} \times \frac{W}{4}$ of the input dimensions and is specifically optimized for retaining fine spatial detail. We then introduce a dedicated detection head, denoted as $\text{Head}_{\text{ultra-small}}$, to process this feature map:

$$\text{Head}_{\text{ultra-small}} : \frac{H}{4} \times \frac{W}{4} = 160 \times 160 \tag{4}$$

Architecturally, the classification sub-network of $\text{Head}_{\text{ultra-small}}$ employs depthwise separable convolutions (3×3 kernels) followed by 1×1 pointwise convolutions to maintain computational efficiency without sacrificing representational capacity. The regression sub-network comprises three fully convolutional layers with 3×3 kernels, enabling precise bounding box estimation for densely packed, minute targets.

Table 1: Performance comparison of the proposed model on the VisDrone validation set against existing methods. Rows marked with $*$ denote results reported in prior literature, while the remaining results are obtained from our own experimental evaluations.

Model	Venue	Year	mAP	AP50	AP75
$*$ Faster R-CNN [27]	NeurIPS	2015	21.7	40.7	19.9
$*$ RetinaNet [17]	CVPR	2017	13.9	32.6	14.8
$*$ UFPMP [10]	AAAI	2022	36.6	62.4	36.7
$*$ CEASC [7]	CVPR	2023	28.7	50.7	24.7
$*$ SDP [23]	TGRS	2023	30.2	52.5	28.4
$*$ CZDet [24]	CVPR	2023	33.2	58.3	33.2
$*$ DNTR [20]	TGRS	2024	33.1	53.8	34.8
$*$ DQ-DETR [11]	ECCV	2024	37.0	60.9	37.9
TPH-YOLOv5 [38]	ICCVW	2021	33.8	54.2	35.2
TPH-YOLOv5++ [35]	ICCVW	2023	32.2	51.8	33.5
YOLOv10x [30]	NeurIPS	2024	31.1	49.6	32.23
YOLC [19]	T-IT	2024	37.9	62.0	39.9
SPAR-YOLO [16]	AAAI	2025	37.90	59.02	40.00
SPAR-YOLO + Ours [16]	-	-	39.4	61.6	41.61
YOLOv11x (Baseline) [13]	-	2024	38.97	59.93	41.63
YOLOv11x + ours	-	-	**40.62**	**62.05**	**43.17**

3.3 Removal of Large Object Detection Head

Owing to the imaging characteristics of UAV platforms, where aerial footage is captured from altitudes exceeding 10 meters, most objects appear small on the image plane, while large objects are rare. For instance, objects larger than 96×96 pixels account for only 4.9% of all instances in the VisDrone dataset, while AITODv2 has no large objects (see Figure 2). Under such conditions, the large-object detection head, which processes coarse $H/32 \times W/32$ feature maps, carries limited utility. Our assessment of YOLOv11's large-object detection head confirms that it contributes negligibly to the detector's performance on the VisDrone validation set (see Table 4).

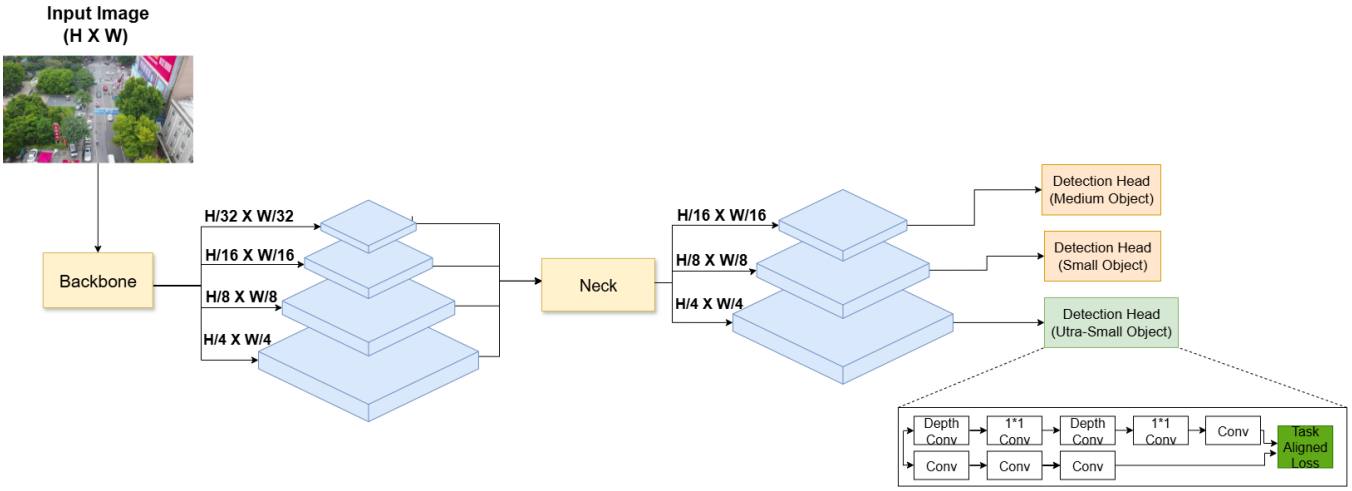


Figure 3: Proposed modifications to the YOLOv11 architecture. We remove the large detection head, and add a new detection head for ultra-small objects.

Consequently, we remove the large objects detection head to streamline the architecture. This modification reduces the parameter count by 21.34% (from 56.89M to 44.75M) while preserving competitive detection performance, aligning the model more closely with the spatial scale distribution inherent to UAV imagery. The

An overview of the proposed modifications to the YOLOv11 architecture is illustrated in Figure 3. The comparison in Figure 4 shows the trade-off between parameter count and mAP. The YOLO model with our proposed modifications achieves higher accuracy with fewer parameters as compared to the YOLO baseline. The same trend is observed in modified SPAR-YOLO as compared to the baseline.

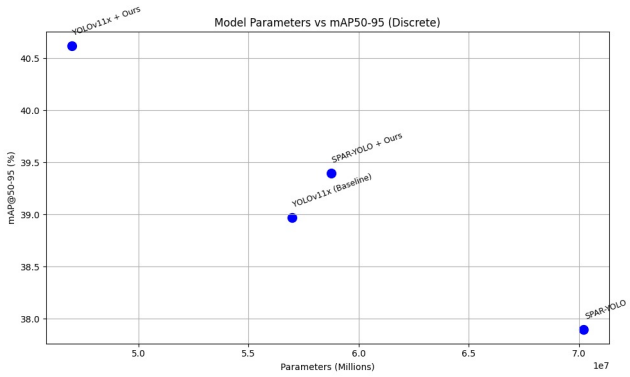


Figure 4: Comparison between number of Parameters and mAP for Baseline and the proposed model

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets. We evaluate our approach on two challenging aerial detection benchmarks: VisDrone [36] and AI-TODv2 [31]. VisDrone consists of 10,209 high-resolution static images, with 6,471 for training, 548 for validation, and two test sets (1,610 in Test-Dev and 1,580 in Test-Challenge). We perform the evalution on the validations split of the Visdrone dataset. It contains 10 object categories, characterized by dense scenes and a wide range of object sizes. The mean bounding box size is 35.8 pixels with a standard deviation of 32.8, highlighting the scale diversity and prevalence of small and ultra-small objects. AI-TODv2 is tailored for tiny object detection in aerial views, containing 28,036 images (800×800 resolution) and 700,621 annotated instances. Notably, 85.9% of objects in AI-TODv2 are smaller than 16×16 pixels, making it an ideal benchmark for ultra-small target detection in aerial imagery.

4.1.2 Implementation Details. We train our models on an NVIDIA V100 GPU with 32GB memory capacity for 150 epochs using a batch size of 4. Training uses an initial learning rate of 4×10^{-2}, weight decay of 5×10^{-4} and dropout rate of 5×10^{-2}. We initialize the weights from the baseline model and keep all the layers trainable. To enhance generalization, we incorporated several data augmentation strategies, such as Mosaic [2], MixUp [33], and Copy-Paste [8]. These augmentations are particularly beneficial for improving robustness in cluttered and small-object-heavy aerial imagery.

The VisDrone [36] dataset comprises images of varying resolutions, with most exceeding 1024×1024 pixels. To align with this distribution and preserve detail, we set the input size to 1024×1024. For the AI-TOD dataset, we follow its standard setting, training and validating on 800×800 images.

4.1.3 Evaluation Metrics. We evaluate our model using a comprehensive set of metrics to assess detection accuracy, robustness, and practical deployment performance. The primary metrics reported are mean Average Precision (mAP), AP_{50}, and AP_{75}. mAP represents the average precision across multiple IoU thresholds (from 0.5 to 0.95 with a step size of 0.05), providing a balanced view of overall performance. AP_{50} and AP_{75} are AP scores at fixed IoU thresholds of 0.5 and 0.75, respectively, highlighting performance at both lenient and strict localization levels. We follow the official

Table 2: Category-wise performance of the proposed architecture on the AI-TODv2 dataset. Results are shown for SPAR-YOLO and YOLOv11 architectures, with the improvement in AP from the proposed method reported for each category. The average object size (in pixels) for each category is also provided.

Model Avg Size (px)	mAP	Airplane 28.24	Bridge 18.38	Storage-tank 16.52	Ship 16.23	Swimming-pool 20.72	Vehicle 15.49	Person 13.38	Wind-mill 12.80
SPAR-YOLO [15]	28.73	45.436	20.298	43.94	40.204	21.888	32.434	15.859	9.8307
SPAR-YOLO + Ours	**29.32**	44.054	**21.452**	**45.259**	**42.859**	20.536	**32.67**	**16.783**	**10.973**
YOLOv11x (Baseline) [13]	27.41	41.133	22.608	42.164	39.249	19.986	32.386	14.103	7.6961
YOLOv11 + Ours	**29.18**	40.895	21.574	**44.6**	**43.412**	**21.826**	**34.043**	**17.943**	**9.2263**

Table 3: Category-wise performance of the proposed architecture on the VisDrone dataset. 'Ped.' and 'Awn.' are short for Pedestrian and Awning-tricycle, respectively. Results are shown for SPAR-YOLO and YOLOv11 architectures, with the improvement in AP from the proposed method reported for each category. The average object size (in pixels) for each category is also provided.

Model Avg Size (px)	mAP	Ped. 23.19	People 21.40	Bicycle 26.61	Car 49.96	Van 49.89	Truck 70.36	Tricycle 38.20	Awn. 39.45	Bus 68.16	Motor 26.16
YOLOv10 [30]	31.1	27.189	19.193	11.453	63.76	39.297	32.688	23.1	14.177	51.524	28.694
SPAR-YOLO [15]	37.9	35.531	25.927	20.874	67.575	46.278	40.926	30.807	17.368	58.161	35.565
SPAR-YOLO + Ours	**39.40**	**39.305**	**28.909**	**22.973**	**69.399**	**47.187**	40.736	**31.741**	**18.629**	57.085	**38.049**
YOLOv11x (Baseline) [13]	38.97	37.178	25.301	21.75	68.405	45.756	42.544	32.526	19.966	59.714	36.643
YOLOv11 + Ours	**40.62**	**40.158**	**29.472**	**23.641**	**69.651**	**47.131**	**43.685**	**33.346**	19.659	**60.422**	**39.072**

Table 4: Ablation study on VisDrone for 1024 × 1024 images

Model	mAP	AP50	AP75	Avg. Time (ms)
YOLOv11	38.97	59.93	41.63	34.08
- Large Head	38.54	59.46	40.62	30.61
- Large + Ultra-small Head (Proposed)	**40.62**	**62.05**	**43.17**	48.86
SPAR-YOLO	37.9	59.02	40.00	33.56
- Large Head	37.6	59	39.74	31.12
- Large + Ultra-small Head (Proposed)	**39.4**	**61.6**	**41.61**	55.92

VisDrone-DET-toolkit for the VisDrone [36] dataset and the COCO evaluation protocol [18] for AI-TODv2 [31] dataset.

4.2 Comparisons with State-of-the-art

4.2.1 Compared Methods. We compare with general object detection methods, such as Faster-RCNN [27], RetinaNet [17], Deformable DETR [39], etc., as well as methods specifically optimized for drone object detection, such as TPH-YOLO [37], and SPAR-YOLO [16].

4.2.2 Quantitative Comparison. Tables 1 and 5 present the quantitative results of the proposed architecture on the two datasets, VisDrone and AI-TODv2, respectively. On VisDrone, our method achieves a mean Average Precision (mAP) improvement of 1.65% over the baseline YOLOv11, while on AI-TODv2 the improvement is 1.71%. These gains, although seemingly modest in absolute value, are significant given the extremely challenging nature of these datasets, which contain large numbers of ultra-small objects. To further assess the generality of our approach, we integrate the proposed module into the SPAR-YOLO baseline. This integration yields an additional 1.5% mAP gain on VisDrone and a 1.77% gain on AI-TODv2, highlighting that our architectural modifications complement other advanced detection pipelines and are not tied to a specific base model. These consistent improvements across different architectures and datasets confirm the robustness and adaptability of our design.

Table 5: Performance comparison of the proposed model on the AI-TODv2 dataset against existing methods. Rows marked with ∗ are from prior literature; others are from our experiments.

Model	mAP	AP50	AP75
∗ Faster-RCNN [27]	12.8	29.9	9.4
∗ Cascade R-CNN [3]	15.1	34.2	11.2
∗ Deformable DETR [39]	18.9	50.0	10.5
∗ DAB-DETR [21]	22.4	55.6	14.3
TPH-YOLOv5 [38]	20.1	46.3	14.2
TPH-YOLOv5++ [35]	25.4	60.5	17.5
SPAR-YOLO [16]	28.7	64.00	22.28
SPAR-YOLO + Ours	**29.32**	**65.14**	22.31
YOLOv11x (Baseline) [13]	27.41	60.25	21.18
YOLOv11x + Ours	29.18	62.71	**22.95**

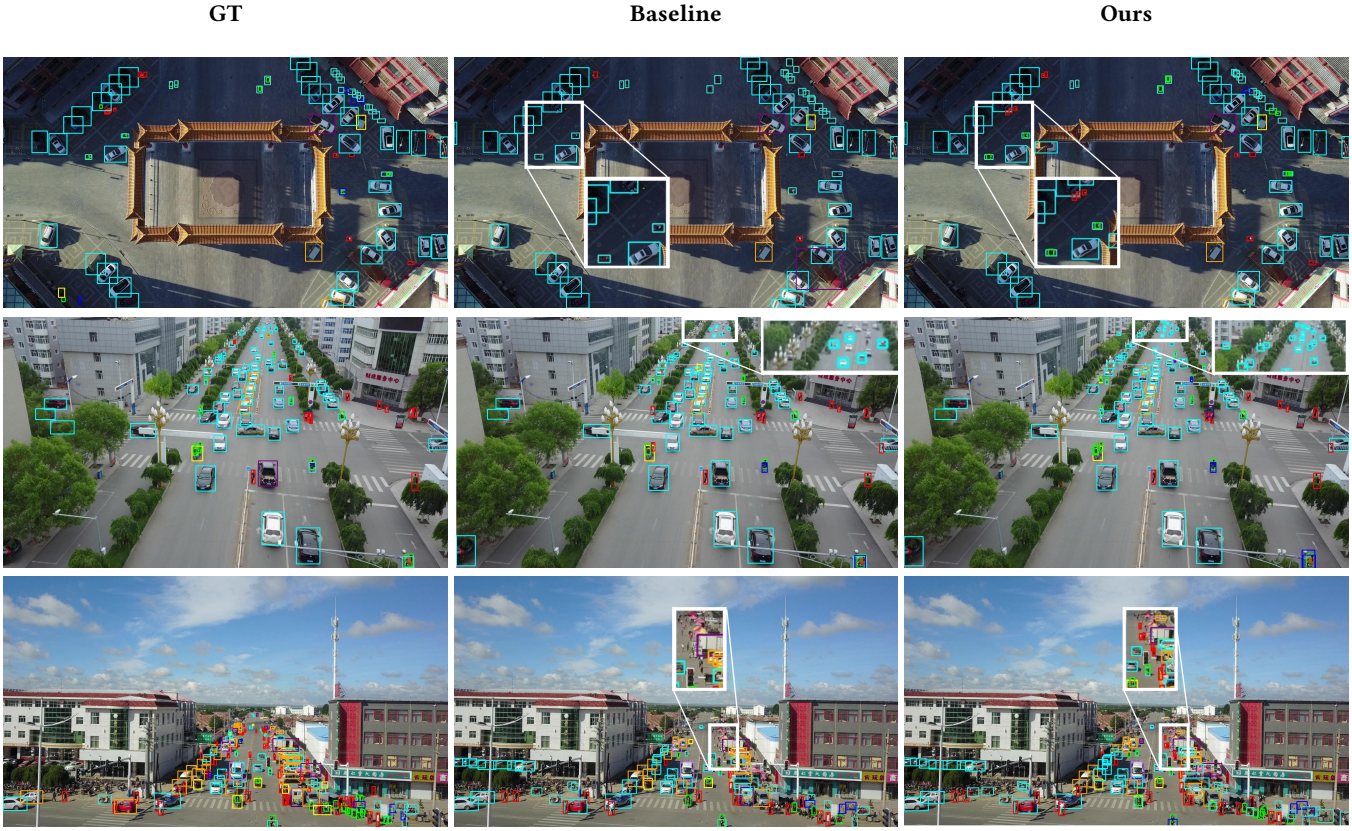


Figure 5: Comparison of Ground Truth (GT), Baseline, and Our proposed method on the VisDrone dataset. Object classes are color-coded as: pedestrian (red), people (green), bicycle (blue), car (cyan), van (orange), truck (purple), tricycle (yellow), awning-tricycle (indigo), bus (pink), and motor (teal). In the top row, the baseline misclassifies people as cars, while our method labels them correctly. In the second row, it misses far-field cars that our method detects. In the last row, it overlooks pedestrians which our method successfully detects.

Table 6: Comparison of Performance Across Object Sizes on the VisDrone Dataset

Model	mAP Small			mAP Medium			mAP Large			mAP Overall
	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	
SPAR	13.81	27.12	12.40	31.24	42.14	36.20	12.67	14.19	13.92	37.9
SPAR + Ours	**15.56**	**29.97**	**14.04**	**31.67**	**42.64**	**36.95**	9.24	10.36	10.16	**39.4**
YOLOv11	14.24	27.45	12.95	31.49	42.49	36.74	23.82	26.55	26.44	38.97
YOLOv11 + Ours	**16.06**	**30.19**	**14.90**	**31.61**	**42.49**	36.72	**25.22**	**27.90**	**27.68**	**40.62**

4.2.3 Category-wise analysis. The category-wise analysis, shown in Tables 2 and 3, provides further insight into where the improvements occur. For VisDrone, the largest relative gains are observed in the *People*, *Pedestrian*, and *motor* categories, where the average size of the objects is 21.40 pixels, 23.19 pixels and 26.16 respectively, which is much smaller than the average object size of the dataset (38.89 pixels) in the validation set. For AI-TODv2, the largest gains are observed in *Storage-tank*, *ship*, *wind-mill*, and *person* categories, where the average size of the objects is 16.52 pixels, 16.23 pixels, 12.80, and 13.38 pixels, respectively, in the test set. This shows that the introduction of an ultra-small object detection head directly benefits classes that predominantly appear at very small scales.

4.2.4 Visual Comparison. Qualitative comparisons, illustrated in Figure 5 and Figure 6 for VisDrone and AI-TODv2, respectively, further validate these findings. In oblique aerial views, where objects in the far field often occupy fewer than 20 pixels in height, the baseline YOLOv11 frequently fails to detect them, whereas our method succeeds — even identifying some instances absent in the ground truth annotations. In orthographic (top-down) views, improvements are

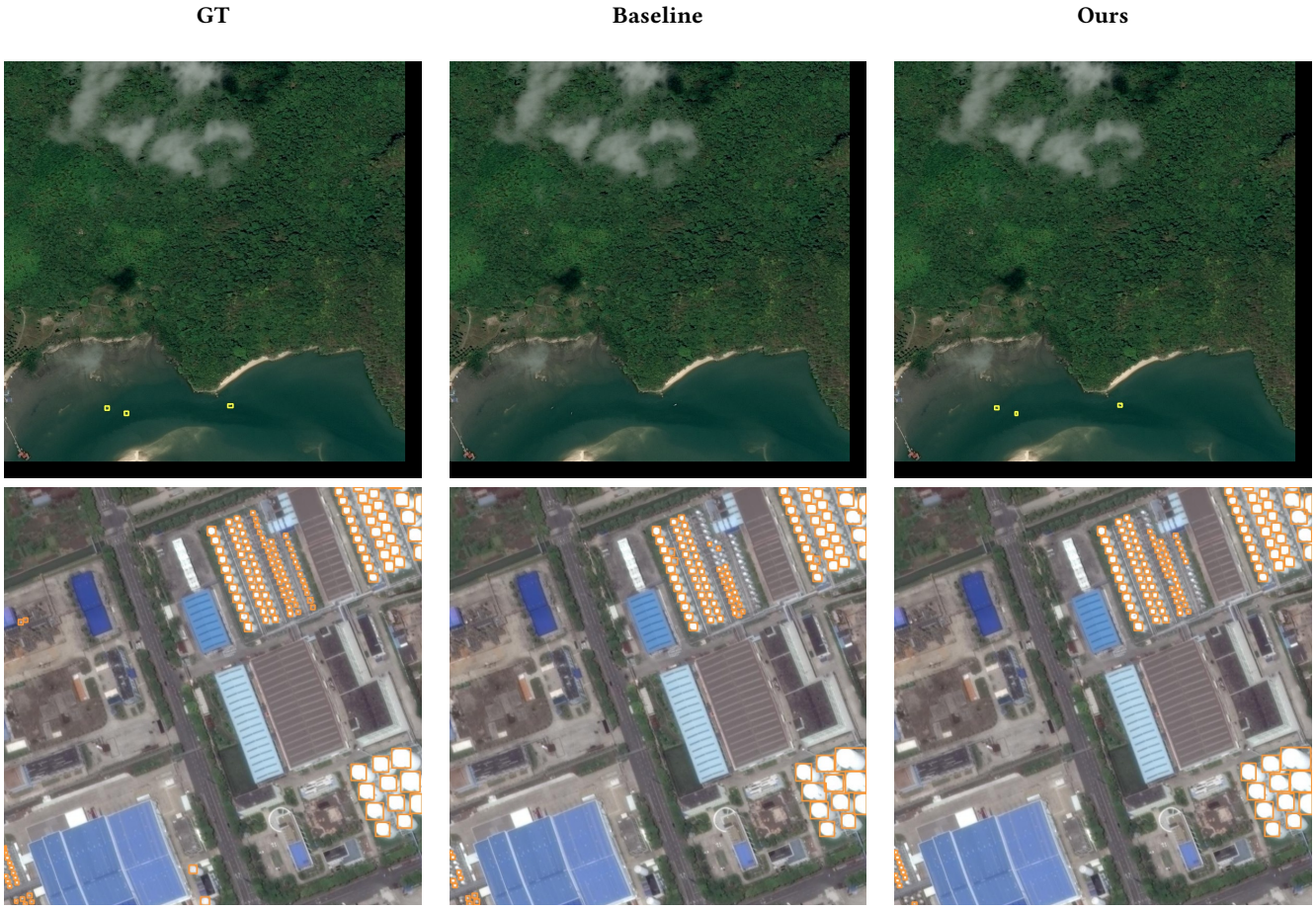


Figure 6: Comparison of Ground Truth (GT), Baseline (SPAR-YOLO), and Our method (SPAR-YOLO + Proposed) on the AI-TODv2 dataset. In the top row, the baseline does not give any prediction whereas our model detects all three instances of the ship(yellow). In the bottom row, multiple instances of the storage-tank(orange) are missed at the top right and bottom left of the image which our detected by the proposed model.

particularly visible for pedestrian and people detection, where our model is able to resolve extremely low-contrast silhouettes that are often difficult to identify even with the human eye.

4.3 Performance Across Object Sizes

Table 6 reports detection performance across three object size categories —*small*, *medium*, and *large* — following COCO-style area thresholds: $< 32 \times 32$ pixels, 32×32–96×96 pixels, and $> 96 \times 96$ pixels, respectively. For each category, evaluation is performed by retaining only the corresponding ground-truth instances and computing mAP, AP_{50}, and AP_{75}.

Across both SPAR and YOLOv11 backbones, the proposed modifications yield the most pronounced gains in the small-object category. For instance, mAP improvements of +1.75 (SPAR) and +1.82 (YOLOv11) are observed, accompanied by consistent AP_{50} and AP_{75} gains. These results confirm the effectiveness of the additional ultra-small-object detection head in capturing fine-scale spatial details, which are critical in aerial imagery where the majority of targets occupy only a few dozen pixels.

Performance on medium-sized objects remains largely unchanged, indicating that the proposed modifications do not compromise detection of more moderately scaled targets. A slight decrease is observed in large-object detection for SPAR, consistent with the removal of the large-object head, while YOLOv11 retains comparable performance for large objects. Overall, the results substantiate our design choice: enhancing the architecture's sensitivity to ultra-small objects yields substantial benefits in UAV-based detection scenarios without materially harming performance on other scales.

4.4 Ablation Analysis

Table 4 presents the ablation results for both YOLOv11 and SPAR-YOLO baselines, evaluating the effect of removing the large-object detection head, and adding the proposed ultra-small-object detection head along with removing the large head. Across both architectures, removing the large object head leads to a marginal drop in mAP, confirming that it contributes little in aerial imagery where large objects are rare. Adding the Ultra-small Head together with removing the large head consistently improves mAP, indicating its

effectiveness in capturing ultra-small targets. The table also reports average inference times (ms/image) evaluated on the NVIDIA V100 GPU, showing that the mAP gains due to the ultra-small head come with only a modest increase in processing time due to the high-resolution feature map used by the Ultra-small Head. This trade-off remains well within the bounds for real-time UAV applications.

4.5 Deployment on Edge Device

To assess practical viability in real-world scenarios, we deployed the YOLO-v11 model, enhanced with our proposed modifications, on the NVIDIA Jetson Orin Nano Developer Kit, which offers 8 GB of memory and delivers up to 100 TOPS of AI performance, making it well-suited for drone-borne inference. The Orin Nano's small form factor and low weight enable seamless integration into drone payloads without significantly affecting flight dynamics or endurance.

To maximize inference efficiency, we leveraged the device's native TensorRT acceleration by converting our trained PyTorch model into an optimized TensorRT engine using NVIDIA's SDK. Post optimization, the model achieved an average inference speed of 12.2 FPS on images of resolution 640×640, ensuring smooth, real-time operation for drone-based object detection tasks.

5 Conclusion

We addressed the unique challenges of object detection in UAV imagery, where the prevalence of small and ultra-small objects, coupled with varying altitudes and environmental conditions, hinder conventional detectors. Building upon a YOLOv11 baseline, we introduced a dedicated ultra-small object detection head to enhance fine-grained spatial feature representation and removed the large-object detection head to better align the architecture with the scale distribution of aerial datasets. These targeted modifications yielded a favorable balance between accuracy and efficiency, achieving 40.62% mAP on the VisDrone validation set. Deployment on an NVIDIA Jetson Orin Nano confirmed practical viability, sustaining 12.2 FPS for real-time operation in moderate-speed UAV scenarios.

Our results show that lightweight detectors can be effectively adapted to UAV constraints through scale-aware design. Future work will aim to boost inference speed for high-velocity flights, enhance robustness to varied aerial conditions, and improve generalization across different UAV platforms and sensors.

Acknowledgments

The authors gratefully acknowledge the financial support provided by Botlab Dynamics.

References

[1] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS journal of photogrammetry and remote sensing* 140 (2018), 20–32.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).

[3] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6154–6162.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*. Springer, 213–229.

[5] Dario Cazzato, Claudio Cimarelli, Jose Luis Sanchez-Lopez, Holger Voos, and Marco Leo. 2020. A survey of computer vision methods for 2d object detection from unmanned aerial vehicles. *Journal of Imaging* 6, 8 (2020), 78.

[6] Qiang Chen, Jian Wang, Chuchu Han, Shan Zhang, Zexian Li, Xiaokang Chen, Jiahui Chen, Xiaodi Wang, Shuming Han, Gang Zhang, et al. 2022. Group detr v2: Strong object detector with encoder-decoder pretraining. *arXiv preprint arXiv:2211.03594* (2022).

[7] Bowei Du, Yecheng Huang, Jiaxin Chen, and Di Huang. [n. d.]. Adaptive Sparse Convolutional Networks with Global Context Enhancement for Faster Object Detection on Drone Images Supplementary Material. ([n. d.]).

[8] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2918–2928.

[9] Jennifer N Hird, Alessandro Montaghi, Gregory J McDermid, Jahan Kariyeva, Brian J Moorman, Scott E Nielsen, and Anne CS McIntosh. 2017. Use of unmanned aerial vehicles for monitoring recovery of forest vegetation on petroleum well sites. *Remote Sensing* 9, 5 (2017), 413.

[10] Yecheng Huang, Jiaxin Chen, and Di Huang. 2022. UFPMP-Det: Toward accurate and efficient object detection on drone imagery. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 1026–1033.

[11] Yi-Xin Huang, Hou-I Liu, Hong-Han Shuai, and Wen-Huang Cheng. 2025. DQ-DETR: DETR with Dynamic Query for Tiny Object Detection. In *European Conference on Computer Vision*. Springer, 290–305.

[12] Glenn Jocher. 2020. *Ultralytics YOLOv5*. doi:10.5281/zenodo.3908559

[13] Glenn Jocher and Jing Qiu. 2024. *Ultralytics YOLO11*. https://github.com/ultralytics/ultralytics

[14] Benjamin Kellenberger, Michele Volpi, and Devis Tuia. 2017. Fast animal detection in UAV images using convolutional neural networks. In *2017 IEEE international geoscience and remote sensing symposium (IGARSS)*. IEEE, 866–869.

[15] Changlin Li, Taojiannan Yang, Sijie Zhu, Chen Chen, and Shanyue Guan. 2020. Density map guided object detection in aerial images. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 190–191.

[16] Nianxin Li, Mao Ye, Lihua Zhou, Song Tang, Yan Gan, Zizhuo Liang, and Xiatian Zhu. 2025. Self-Prompting Analogical Reasoning for UAV Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 18412–18420.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 740–755.

[19] Chenguang Liu, Guangshuai Gao, Ziyue Huang, Zhenghui Hu, Qingjie Liu, and Yunhong Wang. 2024. YOLC: You Only Look Clusters for Tiny Object Detection in Aerial Images. *IEEE Transactions on Intelligent Transportation Systems* 25, 10 (2024), 13863–13875.

[20] Hou-I Liu, Yu-Wen Tseng, Kai-Cheng Chang, Pin-Jyun Wang, Hong-Han Shuai, and Wen-Huang Cheng. 2024. A denoising fpn with transformer r-cnn for tiny object detection. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[21] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. 2022. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329* (2022).

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[23] You Ma, Lin Chai, and Lizuo Jin. 2023. Scale decoupled pyramid for object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–14.

[24] Akhil Meethal, Eric Granger, and Marco Pedersoli. 2023. Cascaded zoom-in detector for high resolution aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2046–2055.

[25] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. 2021. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3651–3660.

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[28] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. 2022. Drone-based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning. *IEEE Transactions on Circuits and Systems for Video Technology* (2022), 1–1. doi:10.1109/TCSVT.2022.3168279

[29] Rejin Varghese and Sambath M. 2024. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*. 1–6. doi:10.1109/ADICS58448.2024.10533619

[30] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. 2024. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems* 37 (2024), 107984–108011.

[31] Jinwang Wang, Wen Yang, Haowen Guo, Ruixiang Zhang, and Gui-Song Xia. 2021. Tiny Object Detection in Aerial Images. 3791–3798.

[32] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. 2018. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[33] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).

[34] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. arXiv:2203.03605 [cs.CV] https://arxiv.org/abs/2203.03605

[35] Qi Zhao, Binghao Liu, Shuchang Lyu, Chunlei Wang, and Hong Zhang. 2023. TPH-YOLOv5++: Boosting Object Detection on Drone-Captured Scenarios with Cross-Layer Asymmetric Transformer. *Remote Sensing* 15, 6 (2023). doi:10.3390/rs15061687

[36] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. 2021. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 7380–7399.

[37] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. 2021. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2778–2788.

[38] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. 2021. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2778–2788.

[39] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).

[40] Zhuofan Zong, Guanglu Song, and Yu Liu. 2023. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6748–6758.