Journal of Electronics and Informatics (ISSN: 2582-3825)
www.irojournals.com/iroei/

RepViT-Based UAV Small Object Detection on VisDrone and UAVDT

Jiva Bharathi M.[1], Sathya M.[2]

Department of Computer Science and Engineering, Pondicherry University, Puducherry, India.

E-mail: [1]Jivabharathi17@gmail.com, [2]msathya.csc@pondiuni.ac.in

Abstract

The Unmanned Aerial Vehicles (UAVs), popularly referred to as drones, have gained increasing significance with respect to their use in applications like traffic monitoring, surveillance, disaster management, environmental observation, and precision agriculture. Despite substantial progress in aerial image sensing technologies, accurate object detection in UAV imagery has remained elusive owing to the issues related to small object sizes, object scale variance, motion blur, occlusions, and complexity of the background. The proposed study describes a RepViT architecture to enhance the detection of small objects. The proposed framework makes use of RepViT architecture, Feature Pyramid Network (FPN), and attention mechanism for enhancing discriminative features and suppressing background noise. The effectiveness of the proposed model was evaluated using benchmark UAV object detection datasets such as VisDrone and UAVDT using widely adopted object detection metrics. The proposed framework has sustained the inference rate of 48 frames per second (FPS).

Keywords: UAV Object Detection, RepViT, Feature Pyramid Network (FPN), VisDrone, UAVDT (Unmanned Aerial Vehicle Detection and Tracking).

1. Introduction

UAVs emerged as a revolutionary technology over the last few years by leveraging high-resolution aerial images across large and difficult to reach geographical locations. The versatility, mobility, and cost-effectiveness of UAVs have led to their increasing use in a number of practical scenarios ranging from surveillance, environmental monitoring, disaster

DOI: https://doi.org/10.36548/jei.2026.3.004
Extended version of article received from **6th International Conference on Artificial Intelligence, 6G Communications and Network Technologies 2026**
Received: 07.05.2026, received in revised form: 14.06.2026, accepted: 02.07.2026, published: 11.07.2026

management, traffic monitoring, border control, agricultural management to smart cities infrastructure. Overall, UAVs enable the generation of vital visual information that aids in decision making and automation. Object detection is a core component of vision-based UAV technology, whereby identification and localization of the objects of interest in images or frames is done. Effective object detection is a prerequisite for various functions including vehicle tracking, pedestrian monitoring, damage assessment, and anomaly detection. However, unlike conventional ground level images, UAV aerial images possess some unique challenges.

One of the most significant challenges in the UAV enabled object detection is the existence of small objects. Images captured from a high altitude will only have very small objects occupying just a few pixels. Such objects are thus difficult for detection algorithms to differentiate from its background. Another challenge that comes with UAV images is the existence of dense object distributions in many cases. Variation in scale is yet another challenge since there will be different sizes of objects in the image depending on the altitude of the UAV, the camera angle, and distance from the object.

Earlier approaches for object detection employed feature extraction approaches like HOG, SIFT, Haar, etc. With these approaches and classifiers like SVM and AdaBoost, they performed exceptionally well in controlled environments but had difficulties when it comes to adaptability and robustness in complex aerial images. Existing approaches fail to capture the high-level semantic cues and were thus not effective in UAV detection. Deep Learning approaches have revolutionized the field of object detection, where models can automatically learn feature representations from large data sets. With the advent of CNN, new models for object detection like Faster R-CNN, SSD, and YOLO were developed. Despite its advantages, CNNs fail to detect small objects with limited receptive fields and multiresolution representation.

Recently vision transformers (ViTs) and attention-based architectures have become popular by modeling global relationships in images. In contrast to CNNs, which mainly rely on capturing local features, transformers utilize self-attention to model long-range dependencies. It makes them better at detecting small and fine-grained objects. Hybrid approaches, which incorporate CNNs and transformers in one network, including light architectures, such as RepViT, attempt to find a balance between efficiency and better detection performance. They are applicable in different UAV-based scenarios.

Apart from developments in models, availability of benchmark datasets, such as VisDrone and UAVDT has played a crucial role in the development of object detection for UAVs. They provide a variety of aerial scenarios, diverse in object density, size, environment, which allows to test detection models. However, despite recent developments, UAV-based small object detection remains an active research area. There are trade-offs between model accuracy, computation costs, and real-time performance.

2. Related Works

2.1 Traditional and CNN-Based Object Detection

Initial approaches to object detection depended on features like HOG, SIFT, and Haar descriptors with machine learning classifiers. The approach demonstrated limited robustness with regard to scale and viewpoint changes in UAV images. The progress in deep learning increased detection precision; in particular, two-stage detectors like Faster R-CNN [8] and Mask R-CNN [7] have achieved impressive localization precision using region proposal mechanism. Although being efficient, their heavy computational load prevented them from real-time operation on UAVs.

Single-stage detectors have resolved this issue by speeding up inference process. YOLO [1] represented object detection as a single regression problem providing real-time object detection capability. Its successors such as YOLO9000 [2], YOLOv3 [3], YOLOv4 [4], YOLOv5 [6], and YOLOv7 [5] have improved accuracy, features representation, and speed of inference. They are actively used in UAVs because of their combination of efficiency and precision, but detection of small objects is still difficult due to lack of context.

2.2 Multi-Scale Feature Learning

Various feature analysis methods have been explored to enhance detection performance for small objects. The FPN [9] uses spatial information from low levels and semantics from high levels, giving the detectors the ability to detect objects at various scales. The FPN architectures have shown impressive results in UAV images, where small objects consist of only a few pixels. The ResNet architecture [13] has contributed significantly to feature extraction abilities by allowing deep neural networks while avoiding gradient degradation issues.

2.3 Transformer-Based Object Detection

The development of transformers has led to a new generation of architectures in computer vision. Vaswani et al.'s [10] transformer model showcased the power of self-attention mechanism in capturing long-range dependencies. Vision Transformer (ViT) [11] leveraged transformer architectures for image recognition tasks and achieved comparable results to those achieved using CNN architectures. Swin Transformer [12] achieved higher computational efficiency through hierarchical feature extraction and shifted window attention mechanism. Transformer models are capable of offering better global context understanding and thus are more suited for UAV images in which there is strong correlation between the appearance of objects and context. However, the high computational cost associated with transformers limits their use in UAV applications.

2.4 Hybrid CNN-Transformer Approaches

Recent researches have focused to hybridization between CNNs and transformers. Hybrid approaches capitalize on the efficiency of local features extracted using CNN and global relationships captured using transformer attentions. MobileNetV2 [15] has proposed lightweight convolutional layers for embedded devices, creating a good starting point for efficient feature extraction. The approach of RepViT [14], which revisited the design of mobile CNNs under the viewpoint of Vision Transformers and proposed a lightweight and yet effective architecture, seems a promising approach for real-time UAV small object detection task. Considering the existing literature, there is still a requirement for lightweight object detection architectures that can detect small objects with high accuracy in real time. In light of this research gap, a RepViT-based UAV object detection framework is proposed in this study.

3. Proposed Methodology

3.1 Framework Overview

This study presents an efficient UAV-based framework for small object detection that utilizes the RepViT network architecture. The developed model intends to enhance the performance of object detection. The proposed framework comprises five main elements, including pre-processing of inputs, RepViT network, feature fusion from different scales, attention module, and detection head. Fig. 1 shows the overall workflow of the proposed framework.

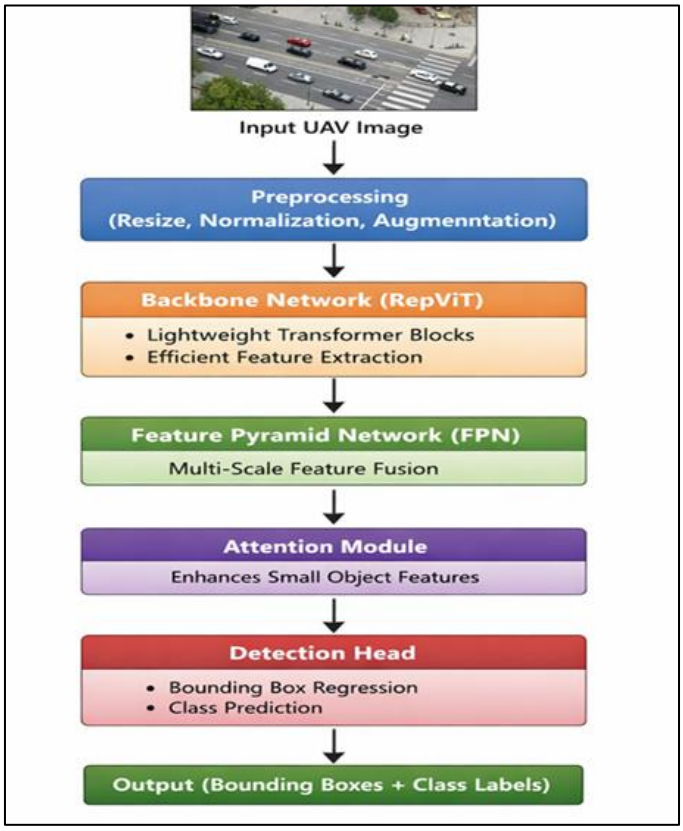

Figure 1. Proposed RepViT-based UAV small-object detection framework

The architecture of the proposed framework based on RepViT for UAV small object detection is depicted in Figure 1. The framework has been designed to enhance the detection of small objects present densely in the aerial images. The proposed framework comprises of five stages namely input processing, RepViT based feature extraction, multi-scale feature fusion, attention based feature refinement, and object detection.

Initially, the input UAV image undergoes preprocessing, including resizing, normalization, and Mosaic augmentation to improve model robustness under varying environmental conditions.

The preprocessing operation is represented as.

$$I_{\text{aug}} = T(I) \tag{1}$$

where I denotes the input image and $T(\cdot)$ represents the preprocessing and augmentation function. The augmented image is then passed through the RepViT, which serves as the primary feature extraction network. RepViT combines efficient convolutional operations

with transformer-inspired representations to detect both local spatial features and global contextual information. The extracted feature representation is given by

$$F_b = f_{\text{RepViT}}(I_{\text{aug}}) \tag{2}$$

where f_{RepViT} denotes the feature extraction function of the RepViT and F_b represents the learned feature maps.

Subsequently, the obtained hierarchical features are fed into a FPN to produce multi-scale feature representations. In this way, both fine spatial details and high-level semantics are preserved to achieve better small object detection at multiple scales.

The overall feature representation is expressed as

$$F_{\text{fused}} = \sum_{i=1}^{N} F_i \tag{3}$$

where F_i shows the obtained feature map from the i^{th} stage and N shows the total number of selected feature levels.

To ensure dimensional compatibility among feature maps of different spatial resolutions, the multi-scale fusion operation can be more precisely represented as

$$F_{\text{fused}} = \sum_{i=1}^{N} U_i\big(C_i(F_i)\big) \tag{4}$$

where $C_i(\cdot)$ denotes the channel alignment to the i^{th} feature map and $U_i(\cdot)$ denotes the corresponding upsampling or spatial alignment operation.

An attention module is then employed to refine the fused features by emphasizing informative object regions and suppressing irrelevant background information. The attention map is generated as

$$A = f_{\text{att}}(F_{\text{fused}}) \tag{5}$$

where $f_{\text{att}}(\cdot)$ denotes the attention-generation function and A represents the learned attention map. The attention-enhanced feature map is subsequently computed as

$$F_{\text{att}} = A \odot F_{\text{fused}} \tag{6}$$

where $\odot$ represents element-wise multiplication. This operation strengthens discriminative responses associated with small objects.

The refined output is then forwarded further for simultaneous bounding-box regression and object classification. The final detection output is represented as

$$D = H(F_{\text{att}}) \tag{7}$$

where $H(\cdot)$ denotes the detection head and D represents the final set of detection results. For each detected object, the detector predicts an object category, confidence score, and bounding-box coordinates. Bounding box outcome is represented as

$$B_p = (x_p, y_p, w_p, h_p) \tag{8}$$

where x_p and y_p indicate the center coordinates of the predicted bounding box, while w_p and h_p denote its width and height, respectively.

The corresponding ground-truth bounding box is defined as

$$B_g = (x_g, y_g, w_g, h_g) \tag{9}$$

where x_g and y_g represent the center coordinates of the ground-truth bounding box, while w_g and h_g denote its width and height, respectively.

The localization performance is evaluated using the Intersection over Union (IoU) metric

$$\text{IoU} = \frac{\text{Area}(B_p \cap B_g)}{\text{Area}(B_p \cup B_g)} \tag{10}$$

where B_p represents the predicted bounding box and B_g denotes the corresponding ground-truth bounding box. The numerator measures the overlapping area, whereas the denominator represents their total union area.

3.2 Dataset Description

The proposed RepViT-based UAV object detection model was tested on two widely accepted benchmarking datasets, such as VisDrone and UAVDT. These datasets contain a variety of aerial images with objects that have different scales, density, views, and

environmental settings and hence serve the purpose of testing the small object detection performance of UAVs.

VisDrone is a large-scale UAV benchmarking dataset consisting of 10,209 still images and 263 video clips taken from urban and suburban areas using different types of drones. This dataset consists of 10 object classes. With the dense object distribution, scale variability, occlusion, and clutter background, VisDrone is widely used for the evaluation of small object detection algorithms in the real-world scenario [16].

UAVDT is a dataset that has been created specifically for UAV traffic monitoring and surveillance purposes. This dataset consists of 100 video sequences consisting of about 80,000 frames taken under various weather conditions, viewpoints, and altitudes of flight. This dataset features three main types of vehicles including car, bus, and truck. UAVDT comes with such difficulties as motion blur, changes in illumination, changes in scale, and partial occlusions, which makes it an ideal dataset to test the performance of object detection systems [17].

3.3 Data Pre-processing and Augmentation

Data pre-processing and augmentation play an essential role in making models more robust and generalized. Considering the specific features of UAV imagery, such as variable sizes of objects and variable environmental factors, it was decided to apply a comprehensive pre-processing technique. In addition, data augmentation techniques were used to improve the model generalization and prevent overfitting: A data augmentation technique that was used in this research includes random horizontal and vertical flipping, random scaling and cropping, rotation augmentation, and color jittering.

The application of random flipping allows the model to become invariant to the orientation of the object, random scaling and cropping help the network to learn scale variations that occur in UAV imagery. Rotation augmentation increases the model's robustness against changes in viewpoint and orientation of the object, and color jittering helps to introduce changes in brightness, contrast, and saturation to make the model more generalized in terms of different illumination conditions. Along with these common data augmentation techniques Mosaic Augmentation technique was applied.

3.4 Training Strategy

The architecture was built based on a supervised learning technique using annotated UAV images, whereby the data was split into 80% training set and 20% validation set. The training data optimized the model's parameters, while the validation data evaluated the generalization ability and determined whether there is any overfitting. This partitioning technique was used consistently in the experiments to ensure consistency in training procedures. Several optimization techniques were applied to improve the efficiency and stability of the training process. The RepViT architecture was initialized with the pretrained weights of the ImageNet-1K dataset that gave robust visual features. The FPN, attention module, and detection head were initialized individually and then optimized in some particular training procedures. The fine-tuning process included joint optimization of the pretrained weights and other detection framework weights to make the architecture adapt to small and dense objects detected in UAV images. Gradient clipping was used to stabilize the process of parameter update during training. The continuous evaluation during training was done by monitoring such metrics as training/ validation loss, precision, recall, and mAP.

4. Implementation

The experiments were performed in an environment, where the use of GPUs was facilitated in order to enhance the learning and inference process for the high-resolution UAV imagery. Hardware acceleration was used in order to cut down the computational time as well as to ease the process of learning the deep object detection model. Table 1 lists the hardware and software configuration used for the implementation of the proposed framework.

Table 1. Hardware and software configuration

Component	Specification
Framework	PyTorch
GPU	NVIDIA RTX 3060 (12 GB VRAM)
CPU	Intel Core i7 Processor
RAM	16 GB
Operating System	Windows 10

A uniform configuration was used to train all the models to ensure convergence and fairness in the evaluations. Training parameters are listed in Table 2.

Table 2. Training configuration parameters

Parameter	Value
Input Image Size	640 × 640
Batch Size	16
Optimizer	AdamW
Initial Learning Rate	0.001
Number of Epochs	100
Learning Rate Scheduling	Cosine Decay Schedule
Algorithms Used	Faster R-CNN, SSD, YOLOv5, YOLOv8 and RepViT
Loss Function	Multi-Component Loss (Classification, Localization and Objectness)

AdamW optimization algorithm was selected for this reason because the decoupled weight decay method is used in the algorithm to regularize parameters. The architecture utilizes a multi-objective loss function that optimizes classification, localization, and objectness tasks simultaneously.

The total training loss is formulated as:

$$\mathcal{L}_{total} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{loc}\mathcal{L}_{loc} + \lambda_{obj}\mathcal{L}_{obj}, \quad (11)$$

where $\mathcal{L}_{cls}$denotes the classification loss used to penalize incorrect object-category predictions, $\mathcal{L}_{loc}$represents the localization loss, and $\mathcal{L}_{obj}$denotes the objectness loss used to estimate the confidence associated with the presence of an object. The weighting coefficients λ_{cls}, λ_{loc}, and λ_{obj}control the relative contribution of each loss component during model optimization.

RepViT was embedded as the feature extraction component, which uses effective convolutional layers and context modeling inspired by transformers. This allows for the incorporation of both local and global information, and therefore enhances contextual feature representation, especially for small objects with less visual information. After multi-scale

feature fusion, an attention mechanism was added before the detector head to highlight informative features related to objects.

Given a fused feature map $F \in \mathbb{R}^{C\times H\times W}$, the module generates an attention map to emphasize informative object-related features while suppressing irrelevant background responses. The attention map is computed as

$$A = \sigma\big(f_{\text{att}}(F)\big), \tag{12}$$

where $f_{\text{att}}(\cdot)$denotes a learnable attention transformation and $\sigma(\cdot)$represents the sigmoid activation function. The attention-refined feature representation is then obtained as

$$F_{\text{att}} = F \odot A, \tag{13}$$

where $\odot$ denotes element-wise multiplication. The refined feature map $F_{\text{att}}$is subsequently forwarded for object classification, objectness estimation, and bounding-box regression. By applying attention after multi-scale feature fusion, the framework emphasizes discriminative responses associated with small and densely distributed objects while reducing interference from complex backgrounds in UAV imagery.

Practical implementation issues, such as class imbalance, were also considered during the training process. Due to some objects being more abundant in UAV data sets than others, a strategy of balancing was used to prevent the bias towards the majority classes.

5. Results and Discussion

Evaluation of the UAV object detection system based on RepViT on complex aerial images having dense object distribution and different sizes of objects is shown. In Figure 2, the initial interface of the UAV Object Detection Dashboard before uploading the image and performing inference is presented. After uploading the image, preprocessing, feature extraction using RepViT model, multi-scale feature fusion, feature refinement through attention mechanism, and object detection are performed.

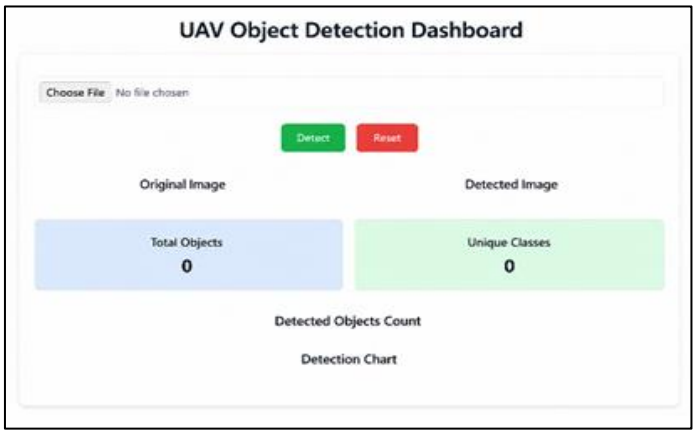

Figure 2. UAV object detection dashboard

Figure 3. Detection results with class labels

Figure 3 shows the capability of the model to precisely localize and classify various object types such as cars, pedestrians, vans, motorcycles, tricycles, and humans.

Figure 4. Summary of detected objects and class counts

Figure 4 presents a summary of the detected object classes and their corresponding counts. In total, 74 objects classified into six categories were found within the test image. Among these categories, cars comprised the majority of objects found, considering that the test image had high density traffic scenes.

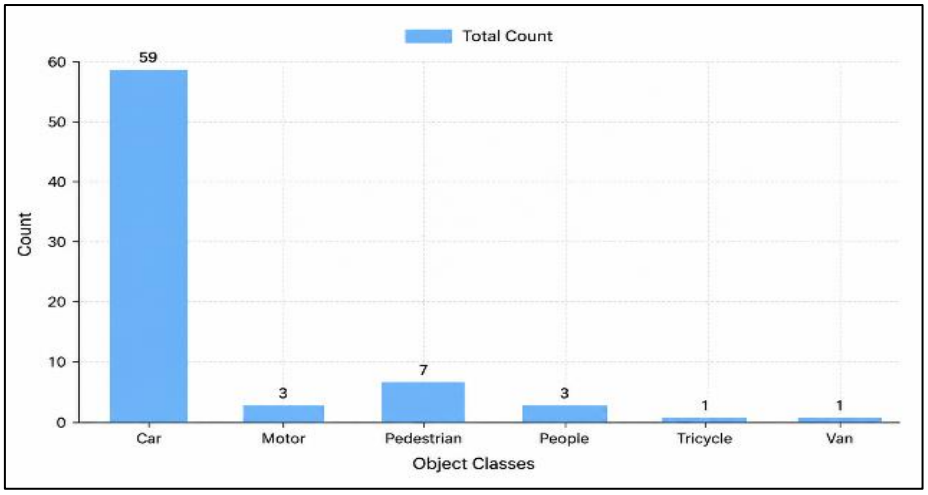

Figure 5. Distribution of detected object classes

Figure 5 depicts the distribution of the detected classes of objects by means of a bar chart visualization. This visualization allows for easy comparison of object frequencies within various classes and quick understanding of scene composition.

Table 3. Performance comparison with existing models

Model	Precision (%)	Recall (%)	F1-Score (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	FPS
Faster R-CNN	88.2	85.6	86.9	87.4	68.5	7
SSD	81.4	78.9	80.1	82.3	61.8	22
YOLOv5	91.6	89.8	90.7	92.1	74.6	45
YOLOv8	93.4	91.5	92.4	94.2	77.8	50
Proposed RepViT	94.8	92.7	93.7	95.3	79.6	48

Performance Comparison of Proposed Framework with Existing Object Detection Models using Quantitative Approach is provided in Table 3. Also, the proposed RepViT framework has yielded a maximum mAP@0.5 of 95.3% and mAP@0.5:0.95 value of 79.6%, indicates the proposed model is effective in the detection of small and dense objects in UAV

images. Even though YOLOv8 had an inference speed of 50 frames per second, the proposed RepViT based model has an inference speed of 48 frames per second.

6. Conclusion

In this study, a framework for small object detection in UAV images using lightweight RepViT model along with FPN-based multi-scale feature fusion and attention mechanism. The proposed model increases the capability of feature extraction, context awareness, and localization of small and dense objects. Evaluation of the proposed framework on VisDrone and UAVDT datasets shows that the framework obtains a good trade-off between the accuracy and efficiency for object detection in complex aerial scenes. These results reveal that an approach which utilizes the combination of lightweight feature extraction, multi-scale representation, and attention-based refinement works well for solving the challenges associated with the detection of objects from aerial images. Further efforts will be dedicated to enhancing the performance on detecting very small and occluded objects.

References

[1] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You Only Look Once: Unified, Real-Time Object Detection." In Proceedings of the IEEE conference on computer vision and pattern recognition 2016, 779-788.

[2] Redmon, Joseph, and Ali Farhadi. "YOLO9000: Better, Faster, Stronger." In Proceedings of the IEEE conference on computer vision and pattern recognition 2017, 7263-7271.

[3] Redmon, Joseph. "Yolov3: An Incremental Improvement." arXiv preprint arXiv:1804.02767 (2018).

[4] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal Speed And Accuracy of Object Detection." arXiv preprint arXiv:2004.10934 (2020).

[5] Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable Bag-Of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2023, 7464-7475.

[6] Jocher, Glenn, Liu Changyu, Adam Hogan, Lijun Yu, Prashant Rai, and Trevor Sullivan. "ultralytics/yolov5: Initial Release." Zenodo 2020.

[7] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask R-Cnn." In Proceedings of the IEEE international conference on computer vision 2017, 2961-2969.

[8] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." IEEE transactions on pattern analysis and machine intelligence 2016, vol. 39, no. 6, 1137-1149.

[9] Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature Pyramid Networks for Object Detection." In Proceedings of the IEEE conference on computer vision and pattern recognition 2017, 2117-2125.

[10] Ashish, Vaswani. "Attention is All You Need." Advances in neural information processing systems 2017, vol. 30, I.

[11] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani "An Image is Worth 16x16 Words: Transformers For Image Recognition at Scale." arXiv preprint arXiv:2010.11929 (2020).

[12] Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows." In Proceedings of the IEEE/CVF international conference on computer vision 2021, 10012-10022.

[13] Kaiming, He, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. "Deep Residual Learning for Image Recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition 2016, vol. 34, 770-778.

[14] Wang, Ao, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. "Repvit: Revisiting Mobile Cnn From Vit Perspective." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2024, 15909-15920.

[15] Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted Residuals and Linear Bottlenecks." In Proceedings of the IEEE conference on computer vision and pattern recognition 2018, 4510-4520.

[16] VisDrone Dataset - https://www.kaggle.com/datasets/banuprasadb/visdrone-dataset

[17] UAVDT Dataset - https://www.kaggle.com/datasets/shakaibkaggle/uavdt-dataset