Small Object Detection on Drone Dataset

Jash Parikh [AU2140108], Kathan Dave [AU2140113], Paridhi Jain [AU2120226], Saahil Doshi [AU2140106]

***Abstract*—Detecting small objects in large scenes, such as those captured by drones, is a significant hurdle in computer vision, affecting applications in surveillance, environmental monitoring, and beyond. This research explores the enhancement of small object detection on the VisDrone dataset, which encompasses a comprehensive collection of aerial images. By studying the application of advanced deep learning models—YOLOv8, YOLOv5, and EfficientDet—this work highlights the effectiveness of model architecture and preprocessing techniques, such as image slicing, in improving detection rates. Notably, the integration of the SAHI slicing method with YOLOv5 demonstrates a remarkable improvement in performance, underlining the potential of image slicing in small object detection tasks. Challenges related to EfficientDet's annotation format conversion and preliminary results across models are discussed. Existing approaches often suffer from high computational costs, hindering real-time deployment. In response, we analyzed HICYOLOv5, an enhanced YOLOv5 model which introduces a dedicated prediction head for small objects, incorporates involution blocks to enhance feature map information, and integrates the CBAM attention mechanism to reduce computation while emphasizing crucial information. Experimental results on the VisDrone-DET2019 dataset demonstrate improvements of 6.42% in mAP@0.95 and 9.38% in mAP@0.5 compared to baseline models.**

***Index Terms*—Object detection, Aerial imagery, Small objects, Deep Learning, Visdrone 2019 detection, VisDrone dataset, YOLO (You Only Look Once), ConvMixer architecture.**

I. Introduction

In computer vision, especially for UAVs and drone detection of small objects aerial imaging remains a major problem. This challenge becomes more complicated with the natural difficulties in the contrast between small objects and the background, the large range in the sizes of the objects, and the complexities added on because of the occlusions and varying lighting conditions.[1] VisDrone dataset, a complete set of aerial images, is a useful tool for supporting the evaluation of object detection models and improving small object detection models.

The YOLO (You Only Look Once) series of models, such as YOLOv5 and YOLOv8, have been gaining huge popularity for their efficiency and accuracy in object detection tasks. However, these models are quite appealing because of their simplicity, ease of training, and high-performance capabilities, making them a preferred choice for researchers and practitioners.[2] Nevertheless, the distinction of small objects is still a critical issue and calls for developing new approaches through which detection rates will improve. In literature, researchers have improved the existing YOLOv5 by utilizing the ConvMixer architecture at the head of a YOLOv5-like architecture with an additional prediction head. The addition of ConvMixer to the prediction head enables us to find the spatial and channel-wise relationships of the feature, which is extracted by the body and delivered by the neck of the architecture. In the ConvMixers, these spatial and channel-wise relationships are extracted by depthwise and pointwise convolutions. Pointwise convolution improves the prediction heads' ability to detect small objects better since it deals with individual data-point-level information.[3] Other works add several attention blocks to the backbone. However, the computation cost is high among the previous methods, and there is still improvement in the performance.[4]. Compared to these methods, we aim to add a lightweight CBAM block at the end of the backbone, resulting in less computation cost and focusing more on essential information when extracting features.

II. Methodology

A. Dataset Overview

The VisDrone-DET2019 dataset builds upon the VisDrone-DET2018 dataset, utilizing the same core data. It comprises 8,599 images captured by drones at varying heights and locations. These images are richly annotated with over 540,000 bounding boxes for ten predefined object categories. The dataset primarily focuses on everyday objects encountered daily, including pedestrians, people, cars, vans, buses, trucks, motorcycles, bicycles, awning tricycles, and tricycles. [5]
The VisDrone dataset serves as the foundation for this study. It offers approximately 6,000 training images, 500 validation images, and 1,500 test images. This dataset presents a unique challenge for object detection algorithms due to the diverse aerial image resolutions and densely placed objects within the frames. As a valuable resource for researchers and practitioners in drone-based computer vision, the VisDrone dataset provides a comprehensive data collection for training and evaluating models in various tasks.

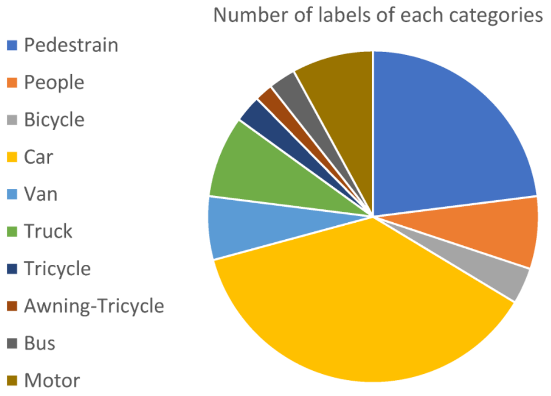

Fig. 1. Number of instances of labels for each category.

B. Model Selection and Preprocessing

The selection of YOLOv8, YOLOv5 and HIC-YOLOv5 as the focus of this study stems from their proven capabilities in object detection tasks. They have demonstrated effectiveness in object detection tasks.

The YOLOv5 stands out in this regard since its capacity to detect small objects from the drone's aerial pictures is well recognizable. This is supported by research such as ”Small Target-YOLOv5: Improvement of the Algorithm for Small-scale Object Detection by using YOLOv5 in Drone Aerial Imagery”. The research shows the YOLOv5's ability to enhance the detection accuracy of small objects, which is a key element needed in applications like surveillance, agriculture, and wildlife monitoring.[4]

Objects are usually very small in drone-based scenes. Extracting more contextual semantic information for the discriminative representation of small objects is necessary for better performance. The VisDrone data has imbalanced categories of objects. As presented in Tables 1 and 2, every detection method performs less in awning tricycles and bicycles than in cars and pedestrians. To deal with this issue, it was essential to adjust the weights of different object classes in the loss function or perform data augmentation for the category with small data.[5]

In YOLOv5, it includes Mosaic, Copy paste, Random affine, MixUp, HSV augmentation and Cutout. In addition, centre cropping has been added to the data augmentation techniques mentioned above as many small people and cars are in the centre of a picture Visdrone2019.

C. HIC-YOLOv5's Approach

HIC-YOLOv5 builds upon YOLOv5, a single-stage object detection model known for its balance between accuracy and real-time processing capabilities. YOLOv5 achieves this efficiency by employing a three-stage architecture:

- Backbone Network: This network extracts features from the input image. YOLOv5 utilizes a powerful backbone network, such as CSPDarkNet53, to capture crucial visual information from the UAV imagery.
- Neck Network: The neck network is responsible for feature fusion. It merges feature maps extracted at different resolutions by the backbone, providing a comprehensive representation of the image for object detection.
- Prediction Heads: YOLOv5 employs multiple prediction heads tailored for objects of varying sizes. These heads analyze the fused feature maps and predict the detected objects' bounding boxes, confidence scores, and class probabilities.

Additional Prediction Head for Enhanced Small Object Detection Since small objects often occupy a smaller portion of the image in UAV data, their features might be less prominent within the standard YOLOv5 feature maps. HIC-YOLOv5 introduces an Additional Prediction Head specifically designed for small object detection to address this. This head operates on higher-resolution feature maps compared to the standard prediction heads in YOLOv5. A higher resolution allows the model to capture finer details crucial for accurately detecting small objects. The SODH leverages the same ConvMixer architecture used in the primary prediction heads, ensuring consistency and efficiency.

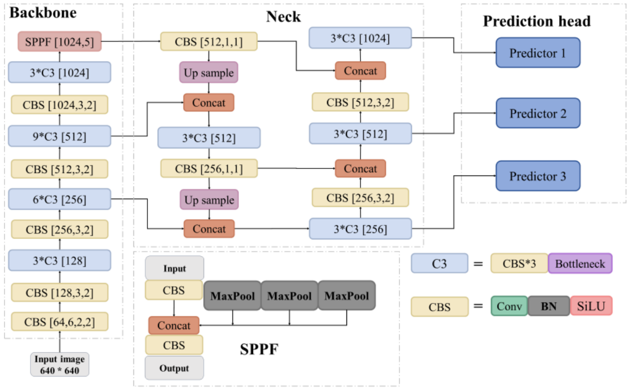

Fig. 2. Structure of YOLOv5-6.0(Retrived from Shiyi Tang et al.,2023)

The backbone of YOLOv5 firstly extracts features from the input image and generates different sizes of feature maps. These feature maps are then fused with the feature maps in the neck. Finally, three different feature maps generated from the neck are sent to the prediction head.

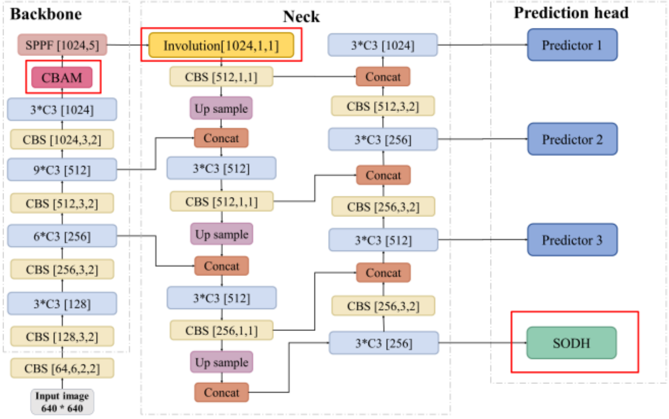

Fig. 3. Structure of HIC-YOLOv5(Retrived from Shiyi Tang et al.,2023)

Based on the work of Shiyi Tang et al., we implement three modifications:
1) Adding an additional prediction head to detect layers with high-resolution feature maps for small and tiny objects specifically
2) An Involution block at the beginning of the neck to improve the performance of PANet
3) Incorporate the Convolutional Block Attention Module (CBAM) into the backbone network.

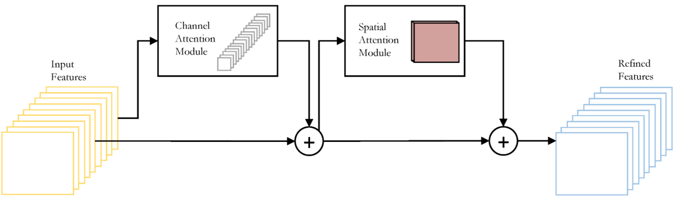

Fig. 4. Structure of CBAM(Retrived from Shiyi Tang et al.,2023)

CBAM [6] is a simple, lightweight attention module that can be fused into most of the general CNN architectures. This module sequentially deduces the channel-wise and spatial attention maps, which are used to obtain the refined features by performing a product between the input features and the obtained attention maps. CBAM helps capture significant target objects in the UAV-captured images which tend to contain regions covering large areas in real life; wherein performing object detection is tricky.

III. Results

From Tables 1 and 2, we can see that compared with the YOLOv8 model, the mAP@[0.5:0.95] has been improved by 5.11%, and mAP@0.5 has been improved by 18%. The small object detection head greatly helps retain small objects' features. Involution effectively amplifies the channel information, while the CBAM block selectively emphasizes crucial features during their extraction within the backbone. The detection effect between YOLOv5s and HIC-YOLOv5 is shown in Fig 5 and 6. It visually indicates that more small objects can be detected when using the improved method.

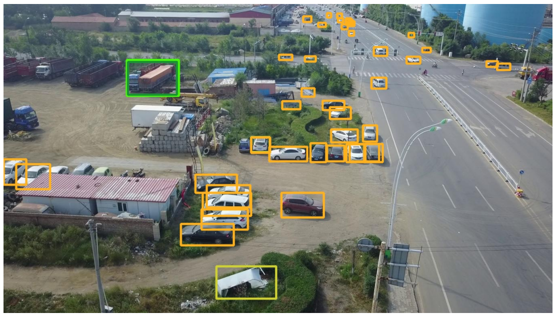

Fig. 5. YoloV5

Object Class	YOLOv8	YOLOv5	HIC-yolov5
All	25.4%	29.3%	30.51%
Pedestrian	28.4%	42.7%	46.14%
People	22.7%	36%	27.65%
Bicycle	5.61%	18.7%	2.86%
Car	66.5%	71%	81.97%
Van	31.9%	28.5%	36.84%
Truck	22.8%	25.4%	22.84%
Tricycle	16.4%	0%	12.65%
Awning-tricycle	8.68%	0%	7.31%
Bus	22.2%	31.8%	24.74%
Motor	28.8%	39.1%	42.06%

TABLE I
COMPARISON OF OBJECT DETECTION MODELS ON VAL DATASET - MAP@0.5

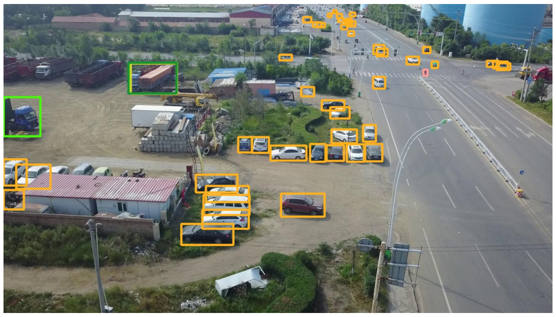

Fig. 6. HIC-YoloV5

Object Class	YOLOv8	YOLOv5	HIC-yolov5
All	14.6%	16.8%	16.18%
Pedestrian	12.4%	19.9%	18.77%
People	7.97%	14.1%	8.80%
Bicycle	2.16%	8.45%	1.05%
Car	44.9%	48.2%	54.76%
Van	21.5%	22%	23.98%
Truck	15.2%	18%	12.9%
Tricycle	9.2%	0%	6.5%
Awning-tricycle	5.77%	0%	4.6%
Bus	16.2%	20.7%	15.24%
Motor	11.2%	16.7%	15.15%

TABLE II
COMPARISON OF OBJECT DETECTION MODELS ON VAL DATASET - MAP@0.5:0.95

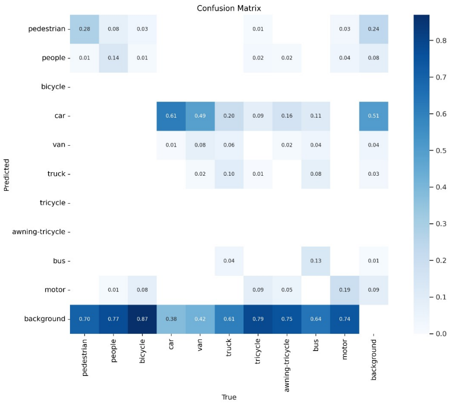

Fig. 7. Confusion matrix for our method with the number of epochs during training and validation.

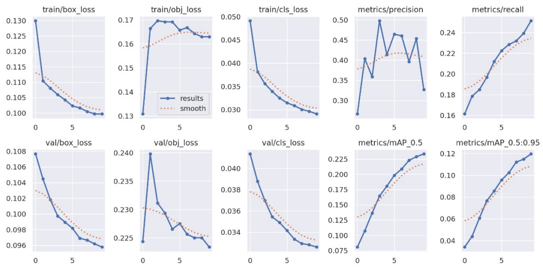

Fig. 8. Visualization of various metrics (box loss, objective loss, class loss, precision, recall, mAP@0.5, mAP@0.5:0.95) with the number of epochs during training and validation.

A. Evaluation Criterion

The common criteria used to evaluate the performance of an object detection algorithm include IoU, Precision, Recall, and mAP. The detailed definitions are listed below.

1) IoU: The Intersection over Union (IoU) is calculated by taking the overlap area between the predicted region (A) and the actual ground truth (B) and dividing it by the combined area of the two. The value of IoU ranges from 0 to 1. The larger the value, the more precise the model. Particularly, a lower numerator value indicates that the prediction failed to accurately predict the ground truth region. On the other hand, a higher denominator value indicates a larger predicted region, resulting in a lower IoU value.

$$IoU = \frac{|A \cap B|}{|A \cup B|}$$

Where: $|A \cap B|$ denotes the area of overlap between the predicted region A and the actual ground truth B. $|A \cup B|$ denotes the combined area of the two regions.

2) Precision: Precision represents the proportion of samples predicted correctly in the set of samples predicted positively.

$$Precision = \frac{TP}{TP + FP}$$

Where: TP is the number of true positive samples (correctly predicted samples). FP is the number of false positive samples (incorrectly predicted samples).

3) Recall: Recall represents the proportion of samples that are positive and predicted to be correct.

$$Recall = \frac{TP}{TP + FN}$$

where: FN is the number of false negative samples (positives missed by the model).

4) mAP: The Average Precision (AP) is a measure of the Precision scores at different thresholds along the Precision-Recall (PR) curve and is calculated as a weighted mean. Mean Average Precision (mAP) is the mean value of the AP for all classes. Specifically, mAP@0.5 represents the mAP when IoU is 0.5, and mAP@[.5:.95] is the mean mAP when IoU ranges from 0.5 to 0.95.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i$$

where: N is the total number of classes. AP_i is the Average Precision for class i.

IV. Discussions

HIC-YOLOv5 can be further enhanced for pedestrian detection by focusing on small object detection, improving accuracy, and maintaining efficiency, making it a valuable tool for real-time applications. Focusing on refining small object detection, improving accuracy, and maintaining efficiency could significantly elevate its utility, especially for real-time applications. This suggests a pathway for future research and development to optimize the model's performance across diverse scenarios. Future research endeavours involve exploring the integration of diverse model architectures tailored to address the challenges posed by lower-resolution images.

V. Conclusion

HIC-YoloV5 underscores the promise of combining advanced deep learning models with preprocessing methods to improve small object detection in the VisDrone dataset. Future research can be done to explore the integration of further models with different architectures and seek solutions to the specific challenges posed by lower-resolution images. By refining these methods, we aim to contribute valuable insights and techniques to small object detection in aerial and similar complex imaging scenarios.

The study highlights the potential of leveraging advanced deep learning models with preprocessing methods to advance small object detection in aerial imagery, particularly on datasets like VisDrone. By integrating approaches such as the ConvMixer architecture and lightweight CBAM blocks, HIC-YOLOv5 demonstrates notable improvements in detection accuracy.

VI. References

[1] Tang, S., Zhang, S., Fang, Y. (2023, September 28). HIC-YOLOV5: Improved YOLOV5 for small object detection. arXiv.org. https://arxiv.org/abs/2309.16393

[2] Keleş, M. C., Salmanoğlu, B., Güzel, M. S., Gürsoy, B., Bostancı, G. E. (2023). Evaluation of YOLO Models with Sliced Inference for Small Object Detection. arXiv preprint arXiv:2203.04799. Retrieved from https://arxiv.org/pdf/2203.04799

[3] Ultralytics. (2023, November 22). SAHI tiled inference. Ultralytics YOLOv8 Docs. https://docs.ultralytics.com/guides/sahi-tiled-inference/

[4] A. Wang, T. Peng, H. Cao, Y. Xu, X. Wei, and B. Cui, "Tia-yolov5. An improved yolov5 network for real-time detection of crop and weed in the field," Frontiers in Plant Science, vol. 13, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2022.1091655

[5] Zhu, Pengfei and Wen, Longyin and Du, Dawei and Bian, Xiao and Fan, Heng and Hu, Qinghua and Ling, Haibin(2021). Detection and Tracking Meet Drones Challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence. doi: 10.1109/TPAMI.2021.3119563.

[6] Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.