Abstract:

Introduction:

1. Cite references in the 1st paragraph and give a clear discussion of the applications and background.
2. Small object detection is not a technical problem to be addressed in this paper. It is what will be achieved after we address the technical problem(s). Draw evidence from the literature and state the gaps.
3. Avoid citing multiple references in one place, e.g., [3,4,5].
4. Given the 8-page limit, we can omit the statements of contributions. Instead, state the novelty in the paragraph that describes the key ideas. In this version, such a paragraph is missing.

Related work:

5. Since it is unclear what challenges (or research problems) are to be addressed in this paper, it is difficult to say if the related work is on target or not. After revising the introduction, make sure to reorganize the discussion around the problems stated in the introduction.

Method:

6. Replace Figure 1 with a diagram of the architecture with functioning blocks instead of network layers.
7. **Complete the method before writing the results.**

Results:

Conclusion:

cropping based small object detection

anivchakravarty

October 2025

1 Introduction

Provide the application of small objects

Current state-of-the-art methods that are being compared with mainly YOLOv11/LRDS-YOLO, maintain neutral tone and do not use negative words to describe limitations

Contribution and new: Efficiency or better accuracy

Detecting small objects in remote sensing images is a significant challenge for object detectors, due to their limited feature sets and varying spatial resolutions. Such problems also can't be ignored in applications such as urban planning, navigation, and monitoring. Hence, object detectors develop specialized techniques to capture key details during feature extraction to improve detection effectively.

Current small object detectors rely on a feature pyramid network structure to efficiently refine and extract features. However, such methods can inadvertently miss small objects during convolutional downsampling, sacrificing higher precision for faster inference. The YOLO architecture is considered a staple of reference, with active research to further improve and optimize its computational efficiency while maintaining accurate detection of small objects. YOLOv11 [1] integrates higher resolution c3k2 blocks and better scale mixing in its aggregation neck, while LRDS-YOLO [2] adds new attention-based downsampling modules to improve semantic representations of deeper layers in shallow layers.

Cropping-based approaches aim to maintain the overall spatial resolution of regions while preserving the features of small objects. In addition, they inherently mitigate class imbalance by reducing background and unnecessary patches. But such methods add computational overhead to their pipeline from their cropping operations as a preprocessing step. Recent detection methods have sought to integrate cropping based on learnable parameters during their training phase to optimize computational overhead and accuracy [3, 4, 5].

We propose a hierarchical, cropping-based small-object detection method that fuses attention. Our key contributions are:

- New learnable cropping module that effectively integrated to the YOLO pipeline.
- A learnable attention-based aggregation logic in the neck.

2 Related work

what are the methods that use cropping for object detection? Traditional concepts such as sliding windows and cropping are leveraged in deep learning models for explicit region scanning rather than hierarchical feature extraction.

Cropping-based methods for small-object detection in remote sensing images have shown promise in recent years by leveraging a two-stage region-based approach with two separate networks. The first network identifies key regions in the images to crop, then passes the cropped features to the second network, which serves as the object detector. CRPN [6] used cropping to reduce large-scale images to smaller effective regions using a region proposal network that generates cropped images based on a scale threshold before passing them to a Faster R-CNN detector. Lin et. al [7] proposed two YOLOv5 networks. The initial network uses a sliding window of a fixed size, assuming the object sizes, to obtain the general location before cropping and passing the patches to a second YOLOv5 network that detects the small objects. DCLANet [8] proposed a crop-attention approach with a non-uniform density-based cropping during training before passing the patches to a YOLOv5 detector. A bottleneck attention module is integrated into the spatial pyramid pooling of the detection network to remove low-level features. Dai et al. [9] proposed a sampling strategy where a discriminative cropping network is connected to the backbone of the detector that uses probability heatmaps. The detector network uses heatmaps to adaptively crop regions, eliminating the need for cropping during pre-processing. Shen et al [10] leveraged cropping to mask easy-to-detect region on front car cameras after the second residual layer and Convolutional Block Attention Modules(CBAM) after the pooling layer of YOLOv3.

Such methods trade higher computational resources for higher accuracy during training, leading to methods that wrap cropping and stitching as pre- and post-processing during inference, such as SAHI [11]. However, such methods also risk cropping larger objects. YOLO based Confidence Adjustment (YOSCA) [12] proposed Gaussian Mixture Model readjustment in the post-processing to improve the detection of small objects and also ensure the detection of large objects.

Zang et al. [13] proposed enhancements to RT-DETR-R18 for small object detection with a dynamic cropping strategy to pre-process the images with dynamic scaling and boundary verification. Deng et al. [14] try to address distortions that appear from current cropping-based methods that rotate images and feed the cropped images into detectors by proposing an Adaptive Image Cropping

Method (AICM) by using an initial 800 x 800 window and tracking ground tracks with overlap ratio as citeria.

3 Method

method architecture and modules goes here

New CropModule layers are integrated into the backbone of the base YOLOv11s to perform spatial downsampling illustrated in the overall architectural workflow in Fig. 1. The layers use fixed window sizes and strides to crop and merge patches, maintaining spatial resolution while reducing feature map dimensions. Each patch also includes a positional embedding to better align with its location in the original image. The 1x1 convolution layers extract features, while the C3k2 layers refine them to yield a richer representation. The fast spatial pyramid pooling (SPPF) and C2 partial self-attention (C2PSA) help maintain a balance between local and global semantics during feature extraction, thereby mitigating the loss of spatial continuity during cropping by increasing the receptive fields in a few regions of the feature maps. The neck replaces the FPN/PAN upsampling and concatenation strategy with three AttentionFuse modules that adaptively fuse multi-scale feature maps with learnable attention to handle features with non-uniform receptive fields. The outputs of AttentionFuse modules are passed YOLO's detection head and non maximum suppression post process

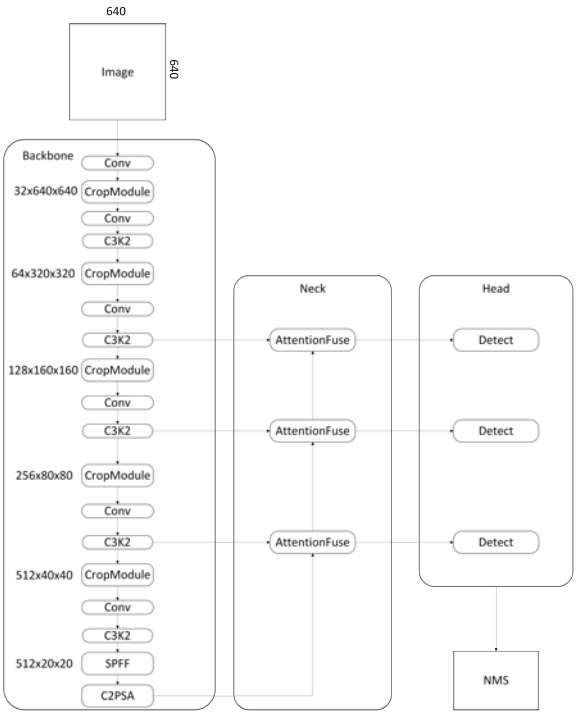

Figure 1: overall architectural workflow of method

3.1 CropModule

3.2 AttentionFusion

4 Experiment

Configuration, dataset, evaluation methodology

We evaluate the average precision of the cropped method relative to the baseline FPN-based YOLOv11s and cropping-based methods on the filtered VisDrone-VID dataset using 3-fold cross-validation. Followed by the impact of single or hierarchical cropping on detecting small objects and the performance impact of each module.

4.1 Metrics

Maintain metrics for small object detection

We use average precision at IoU thresholds 50 and 50 to 95 across each class.

4.2 Dataset

Refine dataset for small object

Our initial dataset is derived from the individual frames of the visdrone-VID [15] dataset that are cropped to a size of 640x640 and are filtered to objects less than 500 square pixels. This leads to a total of 45,967 consistent patches used in a 3-fold cross-validation for 30,645 training and 15,322 validation. Fig. 2 represents the number of instances per class across the filtered dataset, which predominantly consists of pedestrian, car, motor, and people.

plot of instances per class(bar plot)

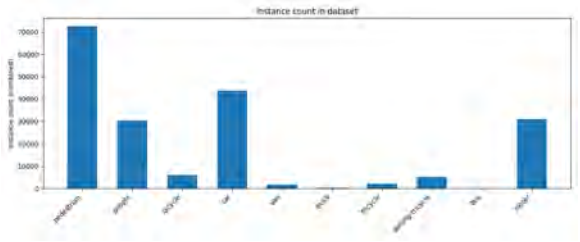

Figure 2: Number of instances of small objects per class in dataset

Add k fold data splits Add class-wise object size distribution and maintain only small objects

5 Results and Discussion

discuss with the results on key insights observed, any future work in optimistic tone

Table 1 reports mean and standard deviation scores for the dataset from YOLOv11 and the proposed method.

Table 1: Comparative results for 3 fold cross validation

Method	mAP 50		mAP 50:95	
	mean	std dev	mean	std dev
YOLOv11s	0.89	0.01	0.54	0.01

6 Conclusion

References

[1] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024.

[2] Yuqi Han, Chengcheng Wang, Hui Luo, Huihua Wang, Zaiqing Chen, Yuelong Xia, and Lijun Yun. LRDS-YOLO enhances small object detection in UAV aerial images with a lightweight and efficient design. *Sci. Rep.*, 15(1):22627, July 2025.

[3] Changlin Li, Taojiannan Yang, Sijie Zhu, Chen Chen, and Shanyue Guan. Density map guided object detection in aerial images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 737–746, 2020.

[4] Akhil Meethal, Eric Granger, and Marco Pedersoli. Cascaded zoom-in detector for high resolution aerial images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2046–2055, 2023.

[5] Zhe Guo, Guoling Bi, Hengyi Lv, Yang Feng, Yisa Zhang, and Ming Sun. No-extra components density map cropping guided object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

[6] Qifeng Lin, Jianhui Zhao, Qianqian Tong, Guian Zhang, Zhiyong Yuan, and Gang Fu. Cropping region proposal network based framework for efficient object detection on large scale remote sensing images. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1534–1539, 2019.

[7] Long Lin, Cuncun Shi, Neng Wan, Weijiang Lu, and Kunlun Gao. Research and implementation of small object detection algorithm for power embedded devices. In *2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, volume 2, pages 178–183, 2021.

[8] Xiangqing Zhang, Yan Feng, Shun Zhang, Nan Wang, and Shaohui Mei. Finding nonrigid tiny person with densely cropped and local attention object detector networks in low-altitude aerial images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4371–4385, 2022.

[9] Honghao Dai, Shanshan Gao, Hong Huang, Deqian Mao, Chenhao Zhang, and Yuanfeng Zhou. An adaptive sample assignment network for tiny object detection. *IEEE Transactions on Multimedia*, 26:2918–2931, 2024.

[10] Lingzhi Shen, Hongfeng Tao, Yuanzhi Ni, Yue Wang, and Vladimir Stojanovic. Improved yolov3 model with feature map cropping for multi-scale road object detection. *Measurement Science and Technology*, 34(4):045406, jan 2023.

[11] Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 966–970, 2022.

[12] Loi Nguyen and Khang Nguyen. Yosca: Confidence adjustment for better object detection in aerial images. *Vietnam Journal of Computer Science*, 12(04):469–488, 2025.

[13] Panpan Zang, Jinxin He, Yongbin Yang, Yu Li, and Hanya Zhang. Enhanced rt-detr with dynamic cropping and legendre polynomial decomposition rockfall detection on the moon and mars. *Remote Sensing*, 17(13), 2025.

[14] Zhiming Deng, Tianyu Zhang, Cheng Wei, and Xibin Cao. Fast object detection and localization in ultrawide swath rotating scan remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:4767–4779, 2025.

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