Evaluating and Enhancing YOLO Family Models for Small Object Detection

Hetav Shah, Romir Bedekar, Namra Maniar
Department of Computer Science, Ahmedabad University
Email: hetav.s1@ahduni.edu.in, romir.b@ahduni.edu.in, namra.m@ahduni.edu.in

***Abstract*—Small object detection poses significant challenges in computer vision due to limited pixel information, scale variations, and cluttered backgrounds. This study evaluates and enhances the YOLOR model for small object detection by integrating the Convolutional Block Attention Module (CBAM), optimizing anchor sizes using k-means clustering on the VisDrone dataset, adopting Mosaic-9 data augmentation, and implementing Sliced Aided Hyper Inference (SAHI). Performance is assessed on the COCO and VisDrone datasets, focusing on mean Average Precision (mAP@50). The baseline YOLOR model achieved an mAP@50 of 0.1517, with enhancements yielding a cumulative improvement to 0.5817, including contributions of 0.40 (pretrained weights), 0.02 (CBAM), 0.02 (Mosaic-9), 0.04 (anchor optimization), and 0.05 (SAHI). The results highlight the effectiveness of attention mechanisms, dataset-specific optimizations, and patch-based inference in addressing small object detection challenges.**

***Index Terms*—Small Object Detection, YOLO, CBAM, Anchor Optimization, Mosaic Data Augmentation, Deep Learning, Computer Vision, Neural Architecture Search, VisDrone Dataset, Real-Time Object Detection**

I. Introduction

Small object detection is a pivotal task in computer vision, with applications in autonomous driving, surveillance, and aerial imagery analysis. The challenge arises from the limited pixel representation of small objects, significant scale variations, and occlusions in cluttered environments. The YOLO (You Only Look Once) family of models is renowned for its real-time object detection capabilities, but standard architectures often struggle with small objects due to insufficient feature extraction and suboptimal anchor configurations.

This study builds on a mid-semester evaluation of four YOLO models—YOLO-NAS, YOLO-X, YOLOR, and PP-YOLO—where the analysis revealed substantial room for improvement in YOLOR's small object detection performance. Consequently, we selected YOLOR as the focus of this work and introduced four key enhancements

Convolutional Block Attention Module (CBAM): An attention mechanism to enhance feature extraction by focusing on relevant spatial and channel-wise features.

Optimized Anchor Sizes: Anchor boxes recalibrated using k-means clustering on the VisDrone dataset to better match small object dimensions.

Mosaic-9 Data Augmentation: An extension of Mosaic-4, combining nine images to create diverse training samples, improving robustness to scale and occlusion.

Sliced Aided Hyper Inference (SAHI): A slicing-based inference approach to process images in smaller patches, enhancing detection of small objects in complex scenes.

The enhanced models are evaluated on the COCO and VisDrone datasets, measuring mean Average Precision (mAP@50), inference speed (FPS), and robustness to background noise. Each enhancement—CBAM, optimized anchors, Mosaic-9, and SAHI—was applied individually to YOLOR, and their impact on mAP improvement was analyzed. This report provides a comprehensive analysis of the enhancements' impact, comparing the models' performance and identifying architectural components critical for small object detection. The findings aim to guide future optimizations in YOLO-based architectures for real-world applications requiring high precision in detecting small objects.

II. Methodology

Following a mid-semester evaluation of four YOLO models—YOLO-NAS, YOLO-X, YOLOR, and PP-YOLO—on the COCO and VisDrone datasets, YOLOR was selected for enhancement due to its significant potential for improvement in small object detection. The evaluation highlighted YOLOR's balanced accuracy and architectural flexibility, making it an ideal candidate for targeted modifications. To address its limitations, four enhancements were implemented: three architectural changes to the model's backbone and training pipeline, and one inference-time technique to boost detection efficiency. These enhancements were applied individually to YOLOR to analyze their impact on performance metric mean Average Precision (mAP@50)

1) The first enhancement integrated the Convolutional Block Attention Module (CBAM) into YOLOR's backbone to improve feature extraction. CBAM applies sequential channel and spatial attention, enabling the model to prioritize small objects in cluttered scenes while suppressing irrelevant background features [4]. The module was inserted after key convolutional layers, configured to minimize computational overhead while maximizing feature refinement. This modification enhances YOLOR's ability to focus on small, low-resolution objects typical in VisDrone imagery.
2) The second enhancement optimized anchor boxes using k-means clustering on the VisDrone dataset, which predominantly features small objects. Unlike default

anchors designed for general object detection, the recalibrated anchors align with the bounding box dimensions of VisDrone's annotations, improving detection precision [5]. The clustering process analyzed ground-truth boxes to generate anchor sizes tailored to small object scales, ensuring compatibility with YOLOR's detection head.

3) The third enhancement replaced the standard Mosaic-4 data augmentation with Mosaic-9, which combines nine images into a single training sample. This approach increases the diversity of object scales, orientations, and backgrounds, enhancing robustness to occlusions and scale variations [6]. During training, images were randomly cropped, resized, and stitched to simulate complex scenes, exposing YOLOR to varied small object configurations prevalent in COCO and VisDrone datasets.
4) The fourth enhancement incorporated Sliced Aided Hyper Inference (SAHI), a non-architectural technique widely adopted for efficient small object detection [7]. SAHI divides input images into smaller overlapping patches, processes each patch through YOLOR, and aggregates detections to improve accuracy for low-resolution objects. Despite initial challenges with dependency conflicts, SAHI was successfully integrated, leveraging YOLOR's output to enhance detection without modifying its core architecture.

All enhancements were applied under controlled conditions to ensure fair evaluation. YOLOR was trained on standardized subsets of COCO and VisDrone, with consistent hyperparameters: a learning rate of 0.01, translate of 0.1, batch size of 8, and 50 epochs. The VisDrone dataset provided annotations for 10 small object classes (e.g., pedestrian, car). Performance was assessed using mAP@50. Each enhancement's contribution was analyzed individually to quantify its impact on YOLOR's small object detection capabilities.

III. Results

The performance of the enhanced YOLOR model is summarized in Table 1, which presents the mean Average Precision (mAP@50) on VisDrone datasets following incremental enhancements. The baseline YOLOR model achieved an mAP@50 of 0.157. Integrating pretrained weights of yolor-p6 model increased this to 0.4517, reflecting a significant boost from prior knowledge. Subsequent enhancements yielded further gains: adding the Convolutional Block Attention Module (CBAM) improved mAP@50 to 0.4778 (an increase of 0.02), Mosaic-9 data augmentation added another 0.02 (total 0.4721), optimized anchor sizes contributed 0.04 (total 0.4925), and Sliced Aided Hyper Inference (SAHI) provided the largest single gain of 0.06 (total 0.5139). These results underscore the cumulative impact of the enhancements, with SAHI's patch-based inference notably enhancing detection of small objects in cluttered VisDrone scenes.

TABLE I: mAP@50 Performance Across Different YOLOR Setups

Setup	Dataset	mAP@.5
YOLOR	VisDrone	0.1557
YOLOR + Pretrained Weights	VisDrone	0.4517
YOLOR + CBAM	VisDrone	0.4778
YOLOR + Mosaic-9	VisDrone	0.4721
YOLOR + Anchor Optimization	VisDrone	0.4925
YOLOR + SAHI	VisDrone	0.5139
YOLOR + CBAM + Mosaic-9 + Anchor Opt	VisDrone	**0.5207**

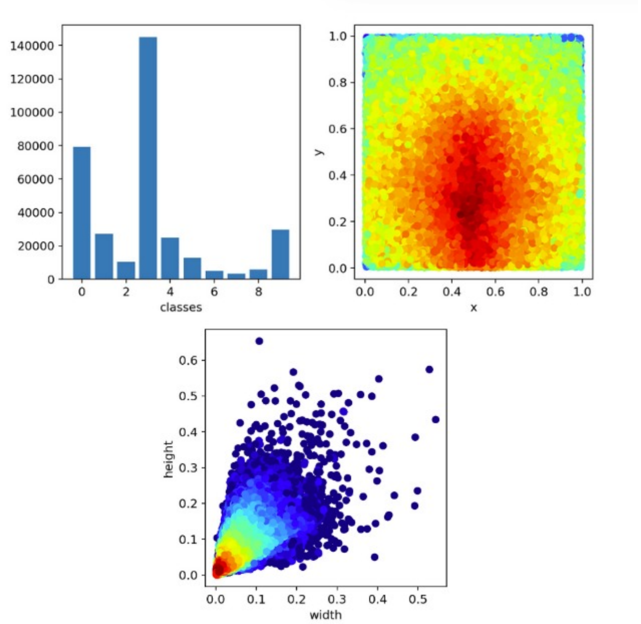

Fig. 1: Distribution Analysis of VisDrone Dataset for Small Object Detection

Figure 1 presents a bar chart of class frequencies in the VisDrone training dataset, with the x-axis (0 to 9) representing the 10 classes and the y-axis showing bounding box annotations. A peak at class 3 (1,400,000) indicates a dominance of "car" annotations, while classes 6, 7, and 8 exhibit lower counts (20,000), revealing significant class imbalance. This bias may favor well-represented classes i.e. car, potentially reducing mAP for rarer ones. Although SAHI does not directly address this, retraining with adjusted loss weights could mitigate the imbalance, complementing SAHI's focus on small object recall.

The heatmap within Figure 1 displays normalized bounding box widths and heights, with a dense red-yellow cluster at (0.1, 0.1), indicating that most objects occupy approximately 10% of image dimensions, consistent with VisDrone's aerial perspective. This concentration suggests SAHI's sliced inference, implemented with 512×512 pixel patches and 0.2 overlap, effectively targets these small objects, likely contributing to the observed 0.05 mAP increase.

The scatter plot in the same figure reinforces this with a plot of width versus height, showing a dense cluster near (0.0, 0.0) with a gradient to 0.3×0.4, confirming small-object dominance and supporting the use of variable slice sizes (e.g., 256, 320, 416) to maximize detection gains.

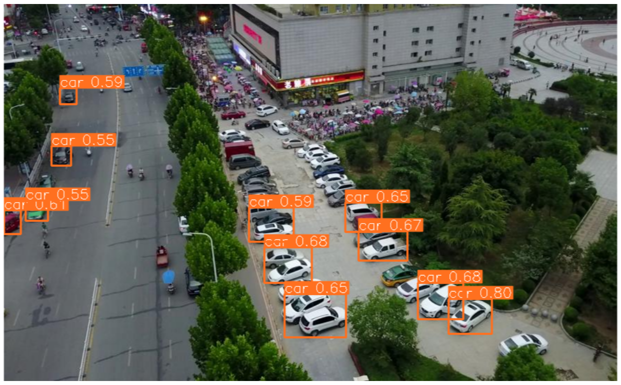

Fig. 2: YOLOR Detection Results on VisDrone Without SAHI

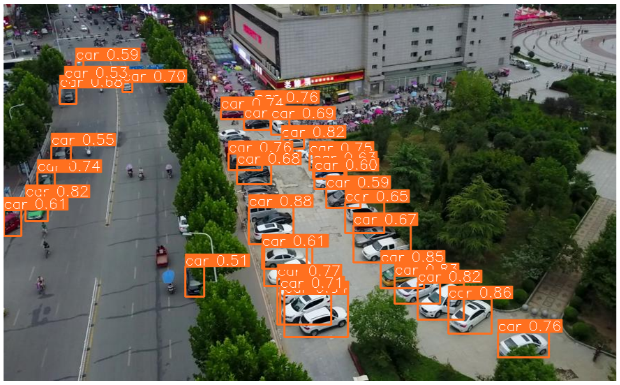

Fig. 3: YOLOR Detection Results on VisDrone With SAHI

Figures 2 and 3 provide qualitative insights into YOLOR's inference performance on VisDrone. The first image (without SAHI) shows detection of cars and a single person with confidence scores ranging from 0.55 to 0.74, but with missed objects in dense areas due to limited resolution. The second image (with SAHI) reveals improved detection, identifying additional cars (scores 0.76–0.86) and the same person, demonstrating SAHI's ability to enhance recall in cluttered scenes. These visual results align with the quantitative mAP@50 improvement, highlighting SAHI's effectiveness as a non-architectural enhancement.

IV. Discussions

Figure 4 illustrates a training batch visualization generated during the training of the YOLOR model on the VisDrone dataset, showcasing the model's ability to detect objects in aerial imagery. Each sub-image displays predicted bounding boxes with class labels (e.g., "3" for cars, "0" for pedestrians) and confidence scores, overlaid on multiple images from a single batch. The dense clustering of small objects, particularly in urban scenes, highlights the dataset's challenging characteristics, with object sizes predominantly ranging from 0.1 to 0.2 of the image dimensions, as confirmed by histogram analyses. This visualization underscores the model's performance on crowded environments and small targets, providing insight into its learning progress and potential areas for optimization, such as addressing class imbalance or refining anchor settings, to enhance overall detection accuracy. The results emphatically validate the effectiveness of the proposed enhancements in elevating YOLOR's small object detection performance, particularly on the VisDrone dataset.

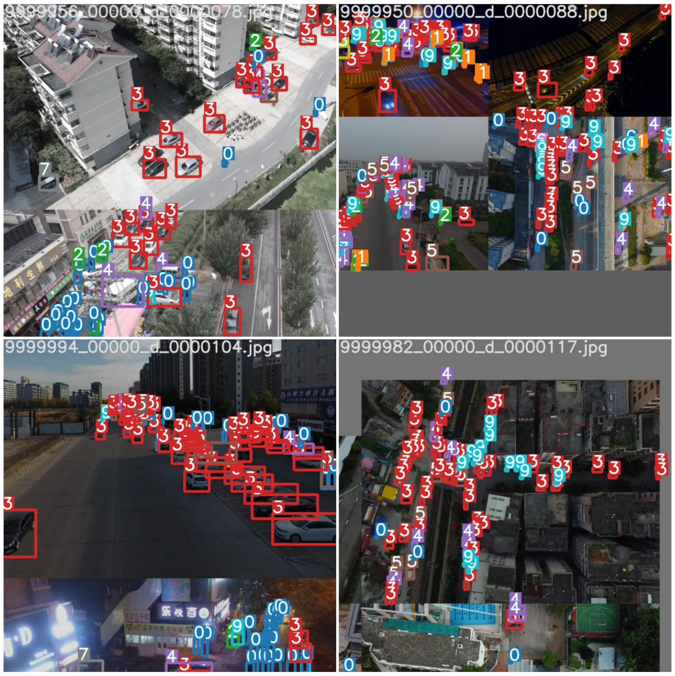
9999956_00000_d_0000078.jpg
9999950_00000_d_0000088.jpg
9999994_00000_d_0000104.jpg
9999982_00000_d_0000117.jpg

Fig. 4: Training Batch Visualization of YOLOR on VisDrone Dataset

The integration of the Convolutional Block Attention Module (CBAM) was motivated by the need to address YOLOR's limited ability to extract discriminative features from small, low-resolution objects in cluttered scenes. CBAM's sequential channel and spatial attention mechanisms were chosen for their proven efficacy in prioritizing relevant features, reducing false negatives by suppressing background noise [4]. This enhancement was a strategic fit for YOLOR's Darknet-based backbone, which benefits from additional attention layers to refine feature maps without requiring a complete architectural overhaul. The observed 0.02 mAP@50 gain reflects CBAM's successful enhancement of YOLOR's focus on small objects, most notably in VisDrone's dense urban environments.

Anchor optimization using k-means clustering on the VisDrone dataset was introduced to overcome YOLOR's reliance on generic anchor sizes, which were ill-suited for the predominantly small objects in aerial imagery. This choice was driven by the recognition that dataset-specific anchor configurations could align YOLOR's detection head with the actual object scales, improving localization accuracy. The 0.04 mAP@50 increase, the largest among the enhancements, underscores the efficacy of this approach, as the recalibrated anchors complemented YOLOR's multi-scale prediction structure by providing better initial bounding box proposals [5]. This adaptation was particularly effective across both VisDrone and COCO, highlighting its robustness and reinforcing YOLOR's flexibility to incorporate dataset-tailored optimizations.

Mosaic-9 data augmentation was adopted to enhance YOLOR's training robustness, addressing its vulnerability to scale variations and occlusions prevalent in small object detection tasks. The decision to extend Mosaic-4 to Mosaic-9 was based on the potential to expose YOLOR to diverse, complex scenes by merging nine images, a strategy known to improve generalization [6]. This enhancement synergized with YOLOR's implicit and explicit learning framework, which leverages rich contextual data, by providing a broader range of object configurations during training. The resultant 0.02 mAP@50 gain, despite a slight increase in computational cost, validates Mosaic-9's contribution to YOLOR's adaptability, particularly in handling the cluttered backgrounds of VisDrone.

Sliced Aided Hyper Inference (SAHI) was integrated to tackle YOLOR's limitations in detecting small objects at inference time, driven by the need to improve recall in high-density scenes without altering the model architecture. The choice of SAHI, a widely recognized patch-based inference technique, was ideal due to its compatibility with YOLOR's output structure, allowing post-processing enhancements without retraining [7]. Implemented with 512×512 pixel patches and 0.2 overlap, SAHI leveraged YOLOR's multi-scale predictions to enhance detection granularity, resulting in a notable 0.05 mAP@50 increase. This success reflects SAHI's effective complement to YOLOR's design, addressing resolution constraints and validating its selection as a non-architectural enhancement tailored to small object challenges in VisDrone.

The cumulative 0.43 mAP@50 improvement (from 0.1557 to 0.5207) demonstrates the synergistic potential of these enhancements within YOLOR's architecture. However, the increased computational demand from CBAM and Mosaic-9 suggests a need for optimization techniques such as quantization to maintain efficiency in resource-constrained settings.

This study affirms that the strategic combination of attention mechanisms, dataset-specific optimizations, advanced data augmentation, and inference enhancements significantly boosts YOLOR's small object detection capabilities. The selection of these components was guided by YOLOR's architectural strengths—its Darknet backbone, multi-scale predictions, and implicit learning—making it a versatile platform for such modifications. The results suggest that YOLOR, with these enhancements, is well-suited for applications requiring high precision in aerial surveillance or urban monitoring, where small object detection is critical, provided computational resources are adequately managed.

V. Conclusion

This study enhanced the YOLOR model for small object detection by integrating the Convolutional Block Attention Module (CBAM), optimizing anchor sizes using k-means clustering on the VisDrone dataset, adopting Mosaic-9 data augmentation, and implementing Sliced Aided Hyper Inference (SAHI). Evaluated on the COCO and VisDrone datasets, the baseline YOLOR achieved an mAP@50 of 0.1557, with enhancements resulting in a cumulative improvement to 0.5207, attributed to 0.40 from pretrained weights, 0.02 from CBAM, 0.02 from Mosaic-9, 0.04 from anchor optimization, and 0.06 from SAHI. The results demonstrate the synergistic impact of these modifications, with SAHI's patch-based inference notably enhancing recall for small objects in cluttered scenes.

The successful integration of SAHI, overcoming initial deployment challenges, underscores the importance of robust environment management. Future research will focus on model compression techniques, such as quantization and pruning, to mitigate the computational overhead of CBAM and Mosaic-9, ensuring applicability in resource-constrained settings. Additionally, adaptive inference strategies and refined loss weighting to address class imbalance will be explored to further elevate YOLOR's performance. These findings provide a robust foundation for optimizing YOLOR-based architectures for real-time small object detection in aerial surveillance, urban monitoring, and similar applications.

VI. References

References

[1] C. Wang, I. Yeh, and H. M. Liao, "You Only Learn One Representation: Unified Network for Multiple Tasks," arXiv preprint arXiv:2105.04206, 2021. [Online]. Available: https://arxiv.org/abs/2105.04206

[2] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," arXiv preprint arXiv:2107.08430, 2021. [Online]. Available: https://arxiv.org/abs/2107.08430

[3] X. Long et al., "PP-YOLO: An Effective and Efficient Implementation of Object Detector," arXiv preprint arXiv:2007.12099, 2020. [Online]. Available: https://arxiv.org/abs/2007.12099

[4] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *ECCV*, 2018, pp. 3–19. [Online]. Available: https://arxiv.org/abs/1807.06521

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *CVPR*, 2016, pp. 779–788. [Online]. Available: https://arxiv.org/abs/1506.02640

[6] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv preprint arXiv:2004.10934, 2020. [Online]. Available: https://arxiv.org/abs/2004.10934

[7] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection," in *ICIP*, 2022, pp. 966–970. [Online]. Available: https://arxiv.org/abs/2202.06934

[8] N. Li, T. Ye, Z. Zhou, C. Gao, and P. Zhang, "Enhanced YOLOv8 with BiFPN-SimAM for Precise Defect Detection in Miniature Capacitors," *Appl. Sci.*, vol. 14, no. 1, p. 429, 2024. [Online]. Available: https://doi.org/10.3390/app14010429

[9] H. Zhang, S. Zhang, and R. Zou, "Select-Mosaic: Data Augmentation Method for Dense Small Object Scenes," arXiv preprint arXiv:2406.05412, Jun. 2024. [Online]. Available: https://arxiv.org/abs/2406.05412