Evaluating the Performance of YOLO Family on 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*—Object detection plays a crucial role in various applications, with small object detection being particularly challenging due to resolution constraints and background noise. This study evaluates the performance of four YOLO family models: YOLO-NAS, YOLO-X, YOLOR, and PP-YOLO, based on their accuracy, speed, and robustness in detecting small objects. Experimental results indicate that YOLO-NAS achieves the highest accuracy, followed by YOLO-X. YOLOR and PP-YOLO exhibit relatively lower performance on small objects. The analysis highlights key architectural components contributing to detection efficiency, such as feature pyramids, quantization-aware modules, and anchor-free mechanisms. The study concludes that integrating certain features from higher-performing models into YOLOR and PP-YOLO could enhance their effectiveness in small object detection tasks.**

***Index Terms*—Small Object Detection, YOLO, Deep Learning, Computer Vision, Object Detection Models, Neural Architecture Search (NAS), Anchor-Free Detection, Feature Extraction, Model Performance Evaluation, Real-Time Object Detection.**

I. Introduction

Small object detection remains a significant challenge in computer vision due to limited pixel representation, scale variations, and occlusions in images. The YOLO (You Only Look Once) family has demonstrated substantial advancements in real-time object detection, making it a preferred choice for various applications such as autonomous driving, surveillance, and remote sensing. However, detecting small objects effectively requires specialized architectural enhancements.

This paper evaluates the performance of five YOLO models—YOLO-NAS, YOLO-X, YOLOR, PP-YOLO, and YOLO-World—with a focus on their effectiveness in detecting small objects. Each model incorporates distinct architectural improvements, such as feature pyramids, anchor-free mechanisms, quantization-aware modules, and unified implicit-explicit learning strategies. YOLO-World further enhances detection capabilities by leveraging large-scale vision-language pretraining and adaptive tokenization strategies. By analyzing their performance on the COCO dataset, this study highlights their strengths and limitations, providing insights into architectural components that influence small object detection accuracy.

II. Methodology

The evaluation process began with an initial performance analysis of five YOLO models: YOLO-NAS, YOLOR, YOLO-X, PP-YOLO, and YOLO-World, using the COCO dataset. The final selection of these models was based on their mean Average Precision (mAP) scores for small object detection. Among these, YOLO-World demonstrated the highest accuracy, followed by YOLO-NAS and YOLO-X, while YOLOR and PP-YOLO showed lower performance in detecting small objects.

To understand the varying performance of these models, a detailed analysis was conducted. YOLO-NAS had the Neural Architecture Search (NAS), enabling the model to automatically optimize its layers, resulting in enhanced feature extraction for small object detection [4] YOLO-X, on the other hand, utilized an anchor-free design and a decoupled detection head, which contributed to better generalization and adaptability in complex scenes. Additionally, SimOTA label assignment improved training efficiency by replacing the traditional IoU-based assignment [2]. YOLOR, a unified network, combined explicit and implicit learning, where the CSP-based scaled YOLOv4 backbone represented explicit knowledge, while manifold space reduction and kernel alignment facilitated implicit learning. However, despite its strong feature learning capabilities, YOLOR struggled with small-scale object detection due to limitations in implicit knowledge representation [1]. PP-YOLO incorporated feature pyramid enhancements, path aggregation networks, adaptive spatial fusion, and DropBlock regularization, improving robustness in detecting small objects in cluttered backgrounds[3].

YOLO-World introduced additional advancements by leveraging large-scale vision-language pretraining and adaptive tokenization strategies, significantly enhancing its small object detection capabilities. By incorporating multi-scale feature aggregation and cross-domain knowledge transfer, YOLO-World achieved superior detection performance, particularly in occluded and low-resolution scenarios. Its ability to generalize across diverse small object categories was a key differentiator, making it the most effective model in this study.

Each model was trained under identical conditions and tested on a standardized small object subset of the COCO dataset to ensure a fair comparison. The evaluation measured key performance metrics, including mean Average Precision (mAP), inference speed (FPS), and robustness to occlusion and background noise. These metrics allowed for a comprehensive assessment of each model's capability in small object detection, highlighting the strengths of YOLO-World, YOLO-NAS, and YOLO-X while identifying the weaknesses in YOLOR and PP-YOLO, which could be improved by incorporating

advanced architectural refinements.

III. Results

Table Ipresents the key performance metrics of the four selected YOLO models: YOLO-NAS, YOLO-X, YOLOR, and PP-YOLO, evaluated on the COCO dataset for small object detection. Among them, YOLO-NAS achieved the highest mean Average Precision (mAP), demonstrating superior detection accuracy. YOLO-X provided a trade-off between accuracy and inference speed, making it well-suited for real-time applications. YOLOR and PP-YOLO exhibited competitive performance but lagged behind in detecting extremely small objects. These results provide a foundation for further discussion on the architectural strengths and limitations of each model.

TABLE I
Performance Comparison of YOLO Models on Different Datasets

Model	Dataset	mAP@50
YOLO-X	COCO	40.5
PP-YOLO	COCO	44.4
YOLOR	COCO	35.08
YOLO-NAS	COCO	47.03
YOLO-World	COCO	51.0
YOLO-X	VisDrone	30.6
PP-YOLO	VisDrone	27.4
YOLOR	VisDrone	20.5
YOLO-NAS	VisDrone	30.6
YOLO-World	VisDrone	33.4

IV. Discussions

The evaluation results highlight YOLO-NAS as the best-performing model in terms of accuracy, making it highly suitable for applications requiring precise small object detection. Its NAS-based architecture optimizes feature extraction and enhances recognition capabilities for minute details in images. The multiple upsampling and downsampling operations in its detection neck contribute significantly to its ability to detect small objects with high precision[4].

YOLO-X, with its anchor-free design and strong data augmentation techniques, demonstrated a balanced trade-off between detection performance and speed. Its SimOTA label assignment improved object matching efficiency, while its decoupled head architecture allowed for independent classification and localization, enhancing performance. While its accuracy is slightly lower than YOLO-NAS, its faster inference speed makes it a preferred choice for applications requiring real-time processing[2].

YOLOR, leveraging both explicit and implicit knowledge representation, showed potential for small object detection but did not perform as well as YOLO-NAS and YOLO-X. Its unified network architecture helped in integrating deep feature learning, but it struggled with optimizing detection in cluttered environments. However, its ability to learn manifold space reduction and kernel alignment suggests that further refinements could improve its small object detection capabilities[1].

PP-YOLO, despite not outperforming the other models, exhibited robust small object detection in cluttered and complex backgrounds. Its performance was enhanced by feature aggregation mechanisms like Spatial Pyramid Pooling (SPP) and Matrix NMS, which improved detection stability. However, its mAP remained lower, indicating that additional improvements in its backbone and feature fusion strategies could enhance its effectiveness in detecting small objects[3].

Overall, this study confirms that selecting an appropriate YOLO model for small object detection depends on the specific requirements of an application—whether prioritizing accuracy, speed, or computational efficiency. While YOLO-NAS remains the best in accuracy and robustness, YOLO-X balances speed and detection quality, whereas YOLOR and PP-YOLO require further enhancements to improve their small object detection performance. Future work could focus on integrating advanced feature fusion techniques, quantization improvements, and label assignment refinements to further optimize small object detection across all YOLO architectures.

V. Conclusion

This study evaluated the performance of YOLO-NAS, YOLO-X, YOLOR, PP-YOLO, and YOLO-World for small object detection, selecting them from an initial pool of models based on their mAP scores on the COCO dataset. The findings indicate that YOLO-World emerged as the most accurate model, leveraging large-scale vision-language pretraining and adaptive tokenization strategies to enhance small object detection. YOLO-NAS followed closely, benefiting from Neural Architecture Search (NAS) and optimized feature extraction mechanisms. YOLO-X provided an optimal balance between speed and precision, making it well-suited for real-time applications. YOLOR, despite its unified network architecture combining explicit and implicit deep learning, lagged in performance compared to YOLO-World, YOLO-NAS, and YOLO-X, indicating a need for further refinements. PP-YOLO, while incorporating advanced techniques such as Spatial Pyramid Pooling and Matrix NMS, demonstrated competitive but lower performance, suggesting that additional architectural improvements could enhance its effectiveness.

Future research will focus on modifying the architectures of these models to further improve their small object detection capabilities. Possible enhancements include refining feature extraction layers, incorporating transformer-based attention mechanisms, and optimizing model compression techniques to maintain high accuracy while reducing computational overhead. Additionally, investigating hybrid approaches that combine YOLO-based architectures with advanced multi-scale feature fusion strategies could further enhance detection accuracy, particularly for small and occluded objects in real-world scenarios. YOLO-World's success highlights the potential of integrating vision-language modeling into object detection, paving the way for future advancements that further leverage cross-modal learning techniques.

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] "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," arXiv preprint arXiv:2304.00501, n.d. [Online]. Available: https://arxiv.org/html/2304.00501v6

[5] M. C. Keles, B. Salmanoglu, M. S. Guzel, B. Gursoy, and G. E. Bostanci, "Evaluation of YOLO Models with Sliced Inference for Small Object Detection," arXiv preprint arXiv:2203.04799, 2022. [Online]. Available: https://arxiv.org/abs/2203.04799

[6] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-World: Real-Time Open-Vocabulary Object Detection," arXiv.org, Jan. 30, 2024. https://arxiv.org/abs/2401.17270