Journal of Computer Science and Cybernetics, V.42, N.2 (2026), 153–165
DOI no. 10.15625/1813-9663/21899

SMALL OPEN VOCABULARY OBJECT DETECTION FROM DRONE IMAGES USING OWL-VIT COMBINED WITH SAHI

NGUYET NGUYEN[1], CONG TRAN[1], MICHAEL NEFF[2], CUONG PHAM[1*]

[1] *Posts and Telecommunications Institute of Technology, Km 10, Nguyen Trai Street, Mo Lao Ward, Ha Noi, Viet Nam*
[2] *University of California, Davis, One Shields Avenue, Davis, CA 95616, United States*

Abstract. The demand for precise and efficient object detection in aerial imagery has surged, driven by applications in agriculture, surveillance, disaster management, and environmental monitoring. However, detecting small objects in drone-captured images remains challenging due to factors like low resolution, occlusion, and varying scales. This research explores a novel approach to small, open vocabulary object detection by combining the OWL-ViT (OpenWorld Vision Transformer) model with the SAHI (Slicing Aided Hyper Inference) technique. OWL-ViT, known for its ability to handle open vocabulary object detection, is leveraged for its robust feature extraction and generalization capabilities across diverse object categories. SAHI is integrated to address the small object detection challenge by slicing high resolution drone images into smaller patches, enabling more focused and detailed inference. In a comprehensive evaluation, our combined method achieves significant improvements in mAP@50 for small-scale object detection, with an average increase of +6.8% on the VisDrone dataset.

Keywords. Closed-set object detection, open-vocabulary object detection, drone imagery, vision transformer, small object detection.

1. INTRODUCTION

Object detection is a fundamental task in computer vision that involves identifying and localizing objects within an image or video frame. It plays a critical role in various applications, including autonomous driving, surveillance, robotics, and aerial imagery analysis. The ability to accurately detect and classify objects in different environments and conditions is crucial for the success of these applications. As the field advances, there has been an increasing focus on improving detection accuracy, especially for challenging scenarios such as small object detection, which is often encountered in drone-captured images.

Traditional object detection approaches, often referred to as closed-set detection, rely on predefined classes that the model is trained on. These models, like YOLO (You Only Look Once) [1] and Faster R-CNN [2], are highly effective when the objects of interest are well-defined and belong to a fixed set of categories. However, closed-set models face significant

*Corresponding author.
E-mail addresses: nguyetnt.ncs@ptit.edu.vn (N. Nguyen), congtt@ptit.edu.vn (C. Tran), pmneff@ucdavis.edu (M. Neff), cuongpv@ptit.edu.vn (C. Pham).

limitations when applied to real-world scenarios where the variety of objects is vast and constantly evolving. The inability to detect objects outside the predefined classes restricts the adaptability of these models, making them less suitable for dynamic environments, such as those captured by drones, where objects can be diverse and unpredictable.

To address the limitations of closed-set approaches, open-set object detection has emerged as a promising alternative. Open-set models are designed to recognize and detect objects that were not present during the training phase, making them more flexible and adaptable to new environments. The Open-World Vision Transformer (OWL-ViT) [3] is one such model, leveraging the power of transformers to generalize across a broad range of object categories. However, while OWL-ViT excels at handling a wide vocabulary of objects, it still faces challenges in detecting small objects, particularly in high-resolution images typical of drone photography. The small object detection problem remains a bottleneck, as the fine-grained details required to identify these objects are often lost in the model's broader feature extraction process.

In this paper, we explore ways to address the challenge individually and combine them to propose a combined approach between OWL-ViT [3] and SAHI [4]. On the one hand, we conduct analyses based on the VisDrone data to edit labels and remove low-resolution images which may affect the model's performance. On the other hand, we investigate the ability of combining SAHI [4] with OWL-ViT [3]. The contributions of our paper are as follows:

- For the first time, we solved the Small Open-Vocabulary Object Detection problem. To handle such a challenging task, we integrate OWL-ViT [3], a popular Open Vocabulary Vision Transformer Model, smoothly into the SAHI [4] algorithm to robustly detect small objects within an image given a query text prompt.
- We analyze data, edit labels, and remove images of insufficient quality to ensure the model's performance.
- We perform empirical experiments and assess OWL-ViT [3] with/without SAHI [4] and integrate other models with the SAHI [4] method to compare results.

2. RELATED WORKS

2.1. Close-set object detection

Object detection has been a critical area of research in computer vision, evolving rapidly over the past decade. Popular model architectures can be broadly categorized into two types: 1-stage detectors, which prioritize speed by predicting bounding boxes and class labels in a single step, and 2-stage detectors, which first generate region proposals and then classify them for higher accuracy, albeit at the cost of speed. Both of these approaches represent the trade-off between real-time performance and detection precision in closed-set object detection tasks.

In 1-stage detectors, such as YOLO [1], SSD [5], RetinaNet [6], skip the proposal generation step and directly predict bounding boxes and class scores from the input image to preserve their efficiency and speed inference. The YOLO family architecture, particularly YOLOv4 [7], YOLOv5 [8], and YOLOv8 [9], have seen significant performance improvements

due to advanced architectural modifications. YOLOv4 introduced CSPDarknet53, a backbone that optimizes the gradient flow and reduces computational complexity, while PANet (Path Aggregation Network) enhances feature fusion for better object localization at various scales. YOLOv5 focused on model size reduction and deployment optimization, making it lightweight and easy to train. YOLOv8, in particular, combines these innovations with further enhancements like anchor-free detection and auto-learning of anchor boxes, which streamline the training process and improve accuracy, making it highly efficient for real-time applications without sacrificing precision.

These advancements enable YOLOv8 to achieve state-of-the-art performance with lower latency, ideal for time-sensitive tasks such as autonomous driving or drone vision. While 1-stage detectors like YOLO focus on speed and efficiency, 2-stage models such as Faster R-CNN [2] prioritize accuracy by introducing an additional step of region proposal generation, which allows for more precise object detection localization and classification, making them suitable for tasks where detection precision is critical. These 2-stage based models first generate region proposals and then classify those proposals. Faster R-CNN stands out as a highly influential architecture for closed-set object detection due to its remarkable balance of precision and robustness. It utilizes a Region Proposal Network (RPN) that streamlines the region proposal generation, allowing the model to efficiently identify potential object locations in an image. This end-to-end training approach not only improves the model's speed compared to its predecessors but also maintains high accuracy in detecting objects across various scales and conditions. Moreover, the extension to Mask R-CNN [10] introduces an additional branch for instance segmentation, further enhancing its capabilities by enabling the model to delineate object boundaries, which is particularly beneficial in complex scenes where objects may overlap. Overall, the combination of accuracy, efficiency, and versatility in handling different detection tasks makes Faster R-CNN and its extensions invaluable in the realm of object detection, especially for applications requiring detailed analysis and high fidelity.

2.2. Open-vocabulary object detection

Unlike closed-set detectors, which have been limited by the fixed set of categories they can detect, open-vocabulary object detection allows models to recognize and localize objects that belong to novel, unseen categories. This approach is crucial for applications in autonomous systems, medical diagnostics, and other fields where new or rare objects frequently emerge, and retraining models is either impractical or inefficient.

Open-vocabulary detectors tend to leverage techniques like zero-shot learning, such as in Zero-Shot Object Detection (ZSD) [11], where in its zero-shot paradigm, it was trained to recognize new object categories by learning from a set of predefined attributes or textual descriptions. This approach allows ZSD to leverage pre-trained language models to bridge the gap between seen and unseen categories. Besides, CLIP [12], DETR [13], and OWL-ViT [3] take advantage of the power of multiple large-scale models. CLIP learns to align visual features with corresponding textual descriptions, enabling it to identify objects described in natural language, while DETR combines its transformer-based architecture with pre-trained language models, which allows it to understand textual input. OWL-ViT, a cutting edge vision transformer, combines image-text processing to handle open-vocabulary tasks effectively.

In this research, we are using OWL-ViT [3] to solve an open vocabulary problem. OWL-ViT is a computer vision model designed to recognize and understand objects that have never appeared in its training dataset, similar to how humans learn new vocabulary through context. Instead of being limited by a fixed set of labels, OWL-ViT can identify objects based on flexible textual descriptions by combining a Vision Transformer (ViT) model with text embeddings. Specifically, the model learns open vocabulary using zero-shot learning, allowing it to recognize new objects by comparing image features with textual descriptions it has never encountered before. For example, if the system has never been trained on images of a "Tesla electric car" but is given a description such as "a car without an exhaust pipe, featuring a Tesla logo," OWL-ViT can infer and correctly identify the object in the image. Additionally, the model uses contrastive learning to optimize its ability to distinguish between objects based on contextual differences, enabling it to adapt flexibly to new concepts without requiring an expanded training dataset. As a result, OWL-ViT not only enhances image recognition capabilities but also unlocks powerful applications in text-based image search, security surveillance, and object recognition in constantly evolving real-world environments.

2.3. Small object detection

Detecting small objects in an image is a challenging task in computer vision, as small objects often visually contain less information, making them harder to distinguish from the background. One of the pioneering approaches to enhance models' detection abilities is to design a pyramid-based model or Multiple Scale Feature Extraction. FPN [14], one of the first models using the pyramid architecture, builds feature maps at multiple scales by integrating low-level, fine-grained features from earlier layers with high-level, semantic-rich features from deeper layers. This design architecture is followed by RetinaNet [6] where authors use an additional Focal Loss to address the issue of class imbalance by down-weighting the loss for easy-to-classify examples, improving detection of hard-to-detect ones. The YOLO family architecture integrates this multiple scale technique with PANet, which enhances the feature pyramid by providing richer semantic information through a multi-scale feature extraction process.

In addition, Kisantal et al. [15] also indicate the potential of using data augmentation to help model robustness in detecting small objects. They mention that the lack of samples for small objects in the dataset, also known as data imbalance between classes and among shapes within each class, causes the model to be difficult to learn. To deal with the problem, they propose to over-sample images with small objects and augment each of those images by copy-pasting small objects to trade off the quality of the detector on large objects with that on small objects. Furthermore, other augmentations such as random crop or scale are also potential options in data generalization and diversity, making models more robust to environmental changes or variations in appearance. However, a risk of computational overhead from distorting small objects should be considered, as generating and processing augmented images can make the training process more resource intensive or distorting the appearance of small object due to their original small size.

Sliding window techniques, a simple and easy-to-implement algorithm, use a fixed-size window to scan through the entire image at various scales and locations. This exhaustive search helps in detecting small objects that might be overlooked by other methods. However, modern algorithms have evolved beyond the traditional sliding window by incorporating more

efficient strategies. For example, the SAHI [4] algorithm builds upon the sliding window approach by slicing images into smaller sections and applying object detection to each slice, effectively improving the detection of small and densely packed objects. This approach leverages the benefits of sliding windows, such as thorough coverage, while reducing the computational overhead by focusing on relevant areas and ensuring that even tiny objects are detected with higher precision. SAHI's integration of sliding windows allows for better scalability and efficiency in tasks like drone image analysis, where small objects in large images are common.

3. PRELIMINARIES

3.1. OWL-ViT in open vocabulary object detection

OWL-ViT (Open-world learning vision transformer) [3] stands out as a highly effective model for small object detection due to several key strengths. Leveraging the transformer architecture, OWL-ViT excels at capturing global contextual information across images, making it particularly well-suited for detecting small objects that are often sparsely distributed or embedded in complex backgrounds. One of its most notable features is its open-vocabulary capability, which allows it to detect objects based on visual-textual embeddings. This innovative approach operates by projecting both image features and textual descriptions into a shared embedding space. The model uses a dual-encoder architecture that processes visual and textual inputs separately before merging them into a unified representation. By doing so, OWL-ViT can compute the similarity between image regions and textual labels, enabling it to generalize to unseen or novel small objects without requiring extensive retraining.

This capability is particularly advantageous for small object detection tasks, as it allows the model to leverage rich semantic information from text to inform its predictions, ensuring that even objects not present in the training dataset can be accurately recognized. Additionally, the model's ability to perform zero-shot detection enables recognition without the need for retraining on specific classes, which is highly beneficial in scenarios where small objects may not be well-represented in the training data. In addition, OWL-ViT with its multi-scale feature extraction within the transformer layers helps preserve the fine details crucial for small object detection. This ensures that even objects occupying a tiny portion of the image are accurately identified. By combining advanced semantic understanding with fine-grained spatial recognition, OWL-ViT offers a cutting-edge approach to small object detection, making it ideal for applications like drone imagery or surveillance where such challenges are prevalent.

3.2. Slicing aided hyper inference

Slicing aided hyper inference (SAHI) [4] is a specialized technique designed to improve object detection, particularly for small objects or in scenarios where objects are densely packed within an image. SAHI works by dividing or slicing large images into smaller, more manageable sections, or tiles, and then applying object detection algorithms to each of these smaller slices individually. This approach is beneficial when dealing with high-resolution images, where small objects might get overlooked by traditional detection methods due to scale differences or limited resolution.

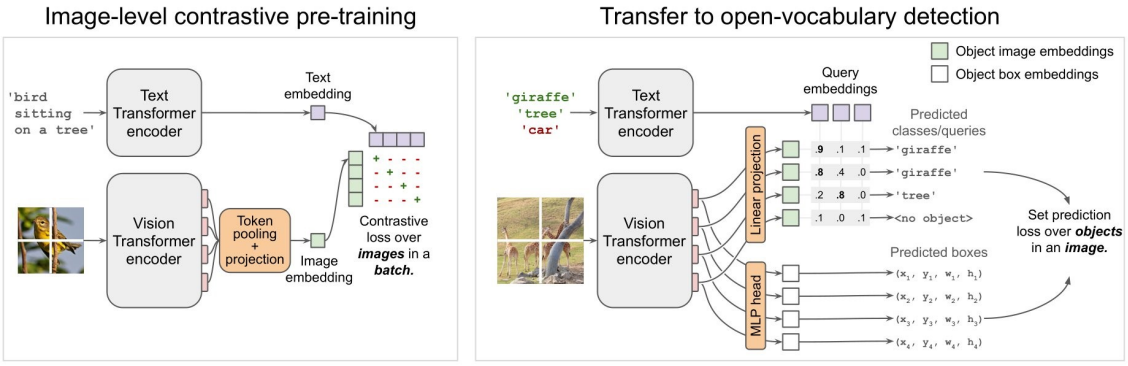

Figure 1: OWL-ViT [3], combining the power of vision transformers and natural language processing.

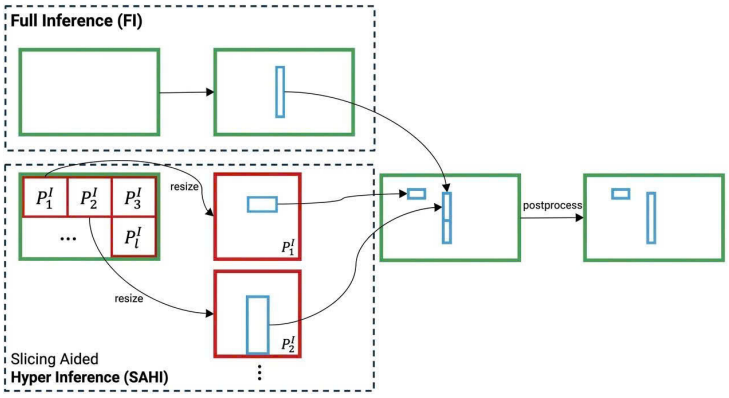

Figure 2: Slicing Aided Hyper Inference for small object detection

By processing these slices independently as shown in Fig. 2, SAHI ensures that even the smallest objects are detected, as the model can focus on smaller portions of the image with greater detail. The technique addresses the challenge of detecting small or distant objects in large images, which is often difficult with standard object detection algorithms due to their limited field of focus or resolution constraints.

SAHI also enables hyper-efficient inference by only processing slices where objects are more likely to be present, reducing computational overhead compared to processing the entire image at once. The sliding window technique plays a crucial role in this process, as it systematically scans the sliced images, ensuring no object is missed. The SAHI method is particularly useful in fields like aerial imagery, medical imaging, satellite data, and processing images from drones, where detecting small objects in large scenes is critical.

4. METHODOLOGY

4.1. General approach

Although the OWL-ViT [3] model is trained on a large and diverse dataset, its performance in detecting small-sized objects, particularly those in drone imagery, remains suboptimal. This limitation may stem from a lack of small-object data or inadequate training on such objects, as exemplified by datasets like VisDrone2019. To overcome this challenge, the

project utilizes the SAHI [4] technique to augment the training dataset with additional small-object examples. By leveraging OWL-ViT's extensive pre-training on large-scale data, the model is fine-tuned on this enriched dataset. Furthermore, adjustments to the loss function are made during fine-tuning to enhance detection accuracy.

A detailed breakdown of each step in the object detection method, combining OWL-ViT and SAHI, is illustrated in Fig. 3.

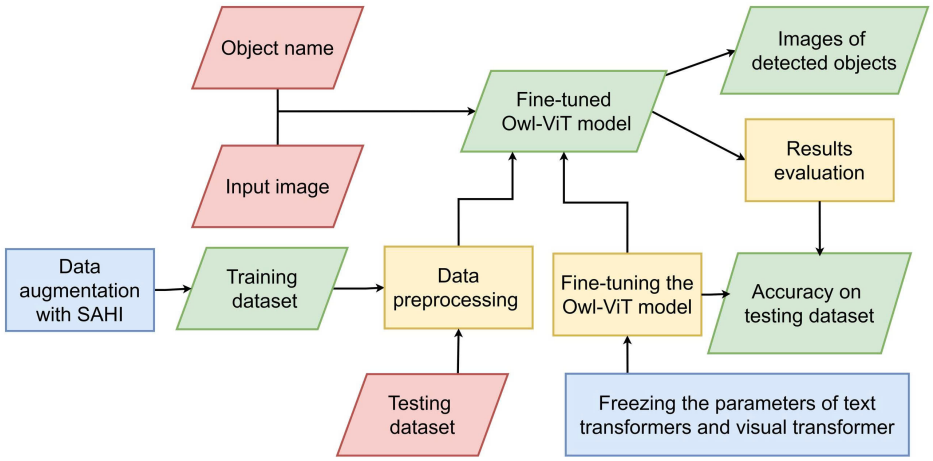

Figure 3: Object detection method combining OWL-ViT and SAHI

4.2. Increase data generalization using SAHI

In the model fine-tuning process, the first step is data preparation. To enhance the dataset, the project employs the SAHI technique [4]. SAHI divides the training images from the VisDrone2019 dataset into smaller sections. By combining these smaller sections with the original dataset, a new training set with more diverse image sizes is created. This approach increases the exposure to small objects of different image scales, improving the model's ability to detect them.

As illustrated in Fig. 4, the original images are divided into smaller segments while preserving the integrity of the object labels. In these cropped images, the small objects appear larger relative to the overall image size, allowing the model to learn their features more effectively. Additionally, the increased number of images enables the model to repeatedly learn object characteristics, enhancing detection accuracy.

Given that the VisDrone2019 dataset contains high-resolution, large-scale images, cropping them does not compromise image quality. After applying the SAHI technique, the smaller images retain their clarity, ensuring the quality of the training data and improving the model's performance.

4.3. Fine-tuning OWL-ViT

Following data preparation, the next step is to fine-tune the OWL-ViT model [3]. Given OWL-ViT's extensive network size, training the entire model from scratch demands significant computational resources and vast amounts of data. To optimize resources, the project focuses on fine-tuning using the SAHI-enhanced dataset.

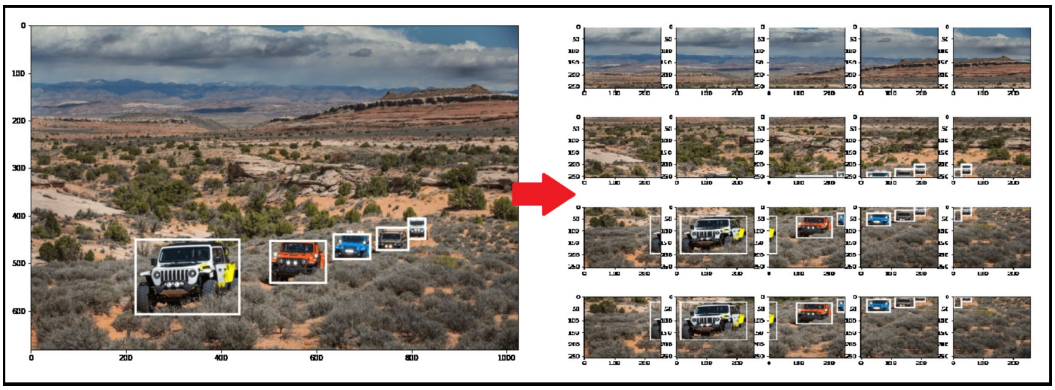

Figure 4: Illustration of data augmentation using SAHI: original images are sliced into smaller segments, increasing the relative size of small objects

Table 1: VisDrone2019-Detection dataset details

Class	Training	Validation	Test
Car	144,866	14,064	28,074
Pedestrian	79,337	8,844	21,006
People	27,059	5,125	6,376
Motor	29,647	4,886	5,845
Van	24,956	1,975	5,771
Bicycle	12,875	1,287	1,302
Tricycle	4,812	1,045	530
Truck	12,875	750	2,659
Awning-tricycle	3,246	532	599
Bus	5,926	251	2,940

Leveraging the model's pre-training on large-scale datasets, we freeze the Text Encoder and Vision Encoder components, allowing us to fine-tune the remaining sections of the network. This includes two crucial components: the bounding box prediction head and the class name prediction module. During fine-tuning, the OWL-ViT Adaptation loss function is applied to guide the training process. The AdamW [16] optimizer, an improved variant of Adam [17], is used for parameter optimization, ensuring more efficient and effective learning.

5. EXPERIMENTS

5.1. Dataset

We choose VisDrone2019-Detection, which is moderate in size and is one of the most common benchmarks for small object detection. In addition, to improve the generalization of small object detection, the SAHI [4] technique is applied using two window sizes: 840×840 and 960×960, with overlap ratios of 0 and 0.25, respectively. This process generates four smaller datasets: 840×840 with overlap ratios of 0 and 0.25, and 960×960 with overlap ratios of 0 and 0.25.

Table 2: VisDrone2019-dataset + SAHI: object count per class after slicing

Class	Orig.	840-0	840-0.25	960-0	960-0.25
Car	144,866	280,028	299,263	328,313	359,478
Pedestrian	79,337	145,894	154,630	167,470	179,868
Motor	29,647	59,054	62,814	65,710	75,213
People	27,059	52,687	57,093	57,047	66,631
Van	24,956	45,081	48,579	52,629	57,374
Truck	12,875	25,088	26,231	28,988	31,362
Bicycle	10,480	20,641	23,538	23,077	26,387
Bus	5,926	10,087	10,335	11,750	12,538
Tricycle	4,812	9,612	10,331	9,290	11,597
Awning-tricycle	3,246	6,965	7,470	6,997	9,259

As observed in Table 1, the original dataset exhibits a significant imbalance, with the Car and Pedestrian classes containing far more instances than others like Awning-tricycle and Tricycle, which are underrepresented. This imbalance could lead to the model favoring the majority classes and under-performing on the minority ones, particularly during object detection tasks. To address this, techniques such as class weighting, data augmentation, and the SAHI [4] method could be employed to improve the detection of small and under-represented classes, ensuring a more balanced performance across all categories.

After applying the SAHI technique to create smaller images containing small objects, we focus on augmenting the underrepresented classes. For instance, the Car class already has sufficient representation in the training set, so no additional images are generated for it. In contrast, the Van and Truck classes require more images to balance them with the Car class, as vehicles like Trucks are often misclassified as Cars, particularly when viewed from the front.

Similarly, the Motorbike, Bicycle, Tricycle, and Awning-tricycle classes are prone to misclassifications due to visual similarities. Therefore, their image counts are increased to achieve balance with the other vehicle classes. Table 2 reveals that the classes Car, Pedestrian, Motor, and People remain unchanged due to their already abundant representation in the training set. In contrast, the classes Bus and Truck experience the most significant boost, as they are notably rare and frequently overshadowed by the car class, which is often misidentified. Following the enhancement process, certain classes now boast over ten times their original object counts, effectively addressing the imbalance and improving the model's overall detection capabilities for these less frequent categories.

5.2. Comparison to the Baseline model

The OWL-ViT model [3], which integrates the SAHI technique, is evaluated in conjunction with several other models that also utilize SAHI [4] on the test set of the VisDrone2019 dataset. This comprehensive evaluation allows for a comparative analysis of their performance, highlighting the strengths and potential improvements offered by the OWL-ViT architecture in the context of small object detection. By examining the results across these models, we can gain valuable insights into the effectiveness of the SAHI technique in enhancing detection capabilities in challenging scenarios.

Table 3: Evaluation results of object detection using OWL-ViT + SAHI on the test set

Metric	Result
mAP@0.5 (whole objects)	28.5
mAP@0.5s (small objects)	17.5
mAP@0.5m (medium objects)	43.2
mAP@0.5l (large objects)	48.7

Table 4: Comparison between OWL-ViT + SAHI and original OWL-ViT

Model	mAP@0.5	mAP@0.5s	mAP@0.5m	mAP@0.5l
OWL-ViT	21.3	10.7	38.5	45.2
OWL-ViT + SAHI	28.5	17.5	43.2	48.7

The evaluated baselines include:

- **FCOS** [18] **+ SAHI**: Fully Convolutional One-Stage Object Detection, a fully convolutional architecture to perform object detection in a single pass.
- **VFNet** [19] **+ SAHI**: An IoU-aware Dense Object Detector, utilizing an IoU aware loss function to improve object localization.
- **TOOD** [20] **+ SAHI**: Task-aligned One-stage Object Detection, aligning object detection tasks with the inherent characteristics of the data.

The evaluation of the OWL-ViT model with SAHI demonstrates a notable enhancement in object detection performance compared to the original OWL-ViT model, especially without fine-tuning (Table 4). The overall mAP@0.5 for the entire test set shows an impressive increase of 7.2%, indicating a significant improvement in the model's ability to detect objects accurately. The performance on small objects is also commendable, with an increase of 6.8% in mAP@0.5s, which reflects the effectiveness of the SAHI technique in addressing the challenges associated with detecting smaller and less prominent objects.

These results underscore the potential of combining OWL-ViT with SAHI to enhance detection capabilities, especially in scenarios where fine-tuning is not applied. The subsequent images provide a visual comparison of detection results between the original OWL-ViT and the improved OWL-ViT + SAHI model, highlighting the advancements achieved through this integration.

Additionally, we conducted an experiment running YOLOv8 on the same VisDrone dataset and obtained lower performance compared to OWL-ViT when combined with SAHI, as shown in Table 5.

Based on Table 5, the TOOD [20] + SAHI model consistently outperforms the other models across all evaluation metrics (mAP@0.5, mAP@0.5s, mAP@0.5m, and mAP@0.5l). This suggests that TOOD, when combined with SAHI, is particularly effective for object detection tasks. Although OWL-ViT + SAHI does not achieve the top performance overall, it still demonstrates strong results, particularly in detecting small and large objects. FCOS +

Table 5: Performance comparison between the proposed method and other detectors on the VisDrone-2019 dataset

Model	mAP@0.5	mAP@0.5 (Small)	mAP@0.5 (Medium)	mAP@0.5 (Large)
FCOS + SAHI	25.8	14.2	39.6	45.1
VFNet + SAHI	28.8	16.8	44.0	47.5
TOOD + SAHI	29.4	18.1	44.1	50.0
OWL-ViT + SAHI	28.5	17.5	43.2	48.7
YOLOv8 + SAHI	24.2	10.4	37.6	41.5

SAHI and VFNet + SAHI show competitive performance in certain metrics but are generally outperformed by OWL-ViT + SAHI and TOOD + SAHI. YOLOv8 + SAHI performs well in detecting large objects but struggles significantly with small object detection.

To gain a deeper understanding of the models' performance, further analysis, such as visualizing the results, analyzing failure cases, and experimenting with different hyperparameters, is recommended. By the way, TOOD (Task-aligned One-stage Object Detection) [20] is not specifically designed for open-set detection, as it focuses on recognizing objects from a predefined set during training. This limitation makes it unsuitable for detecting unseen objects. In contrast, OWL-ViT is well-suited for open-set detection due to its open-vocabulary learning and ability to link images with natural language descriptions. It excels at recognizing new objects and generalizing, making it more flexible than TOOD for detecting unknown objects.

6. CONCLUSIONS

In this paper, we successfully addressed the challenges of object detection, specifically focusing on open vocabulary and small-sized object detection in drone images through the innovative integration of the OWL-ViT model architecture [3] and the SAHI technique [4]. Our experiments demonstrated significant improvements in mAP accuracy, particularly in detecting small objects, validating the effectiveness of our approach. Moving forward, we aim to enhance the speed and accuracy of object detection models for video applications and explore the applicability of our methods in various real-world contexts, thereby contributing to the advancement of object detection technologies and their practical implementation.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016. [Online]. Available: https://doi.org/10.1109/CVPR.2016.91

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

[3] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf,

and N. Houlsby, "Simple open-vocabulary object detection with vision transformers," *European Conference on Computer Vision (ECCV)*, 2022. [Online]. Available: https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136700714.pdf

[4] F. C. Akyon, O. Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 966–970, 2022. [Online]. Available: https://doi.org/10.1109/ICIP46576.2022.9897990

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," *European Conference on Computer Vision (ECCV)*, pp. 21–37, 2016. [Online]. Available: https://doi.org/10.1007/978-3-319-46448-0_2

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2858826

[7] 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

[8] M. Hussain, "Yolov5, yolov8 and yolov10: The go-to detectors for real-time vision," *arXiv preprint arXiv:2407.02988*, 2024. [Online]. Available: https://arxiv.org/abs/2407.02988

[9] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time flying object detection with yolov8," *arXiv preprint arXiv:2305.09972*, 2023. [Online]. Available: https://arxiv.org/abs/2305.09972

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2844175

[11] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," *European Conference on Computer Vision (ECCV)*, pp. 384–400, 2018. [Online]. Available: https://www.ecva.net/papers/eccv_2018/papers_ECCV/papers/Ankan_Bansal_Zero-Shot_Object_Detection_ECCV_2018_paper.pdf

[12] A. R. et al., "Learning transferable visual models from natural language supervision," *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[13] N. C. et al., "End-to-end object detection with transformers," *European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-58452-8_13

[14] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, 2017. [Online]. Available: https://doi.org/10.1109/CVPR.2017.106

[15] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," *arXiv preprint arXiv:1902.07296*, 2019. [Online]. Available: https://arxiv.org/abs/1902.07296

[16] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

[18] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 9627–9636, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/9010746

[19] H. Zhang, Y. Wang, F. Dayoub, and N. Sünderhauf, "Varifocalnet: An iou-aware dense object detector," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9609–9618, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9578034

[20] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3490–3499, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9710724

[21] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8759–8768, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8579011

Received on November 05, 2024
Accepted on December 31, 2025