Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Hierarchical Sparse Vision Transformers for Real-Time Drone-Based Object Detection

Vivekanandam B.

Associate Professor, School of AI Computing and Multimedia, Lincoln University College, Malaysia.

E-mail: vivekanandam@lincoln.edu.my

Abstract

Object detectors based on transformers offer high contextual modeling but have a quadratic complexity of attention, which has restricted their application to real-time in aerial settings. The proposal presented in this paper is a Scalable Adaptive Hierarchical Attention Transformer (SAHAT-Det) that is proposed to be effective at detecting objects and objects in drone imagery. The framework presents the concept of dynamic relevance-based token scoring, top K sparse attention calculation, and adaptive token pruning to lower the computational cost. A multi-scale hierarchy fusion module retains small-scale spatial details especially of objects that are small and far away. On the VisDrone dataset, experimental results show a higher mAP of 0.5:0.95 and small-object detection accuracy than the latest CNN and transformer-based baselines that are run under the same configuration. Although there is less attention computation, the suggested model still retains close real-time inference speed. The qualitative analysis also proves enhanced localization stability in high density urbanized scenes. The obtained results show that adaptive sparse attention offers an efficient compromise between the accuracy of detection and processing cost in real-time aerial object detection.

Keywords: Scalable Attention, Vision Transformer, Sparse Self-Attention, Aerial Object Detection, Multi-Scale Fusion, Real-Time Detection.

DOI: https://doi.org/10.36548/jaicn.2026.2.006
Received: 24.03.2026, received in revised form:27.04.2026, accepted: 17.05.2026, published: 26.05.2026

1. Introduction

Transformer architectures have contributed immensely to the study of object detection owing to the fact that they can capture long distance dependencies and other global context relations. Rigorous surveys conducted in the recent years have summarized transformer-detect systems and compared them on datasets [1], [2]. Vision transformer surveys further point out their contribution in their role in classification, segmentation, and detection tasks [4], [9]. Although these studies prove the effectiveness of attention mechanisms, they also provide the existence of computational overhead and scalability as constant constraints.

Recent advances are trying to be more efficient with lightweight transformer design and hybrid CNN-transformer models [5], [8]. To increase the localization and to detect small objects, improvements in feature pyramids and gradient optimization policies have also been suggested [7]. Object detectors that are camera-conscious like DC-YOLOv8 show performance gains in small-object conditions with limited sensors [6]. Alongside, vision transformers have been examined to be robust and domain generalizable to facilitate expansion reliability [3]. The general trend of expanding transformers use into various areas of vision is also pointed out by the broader surveys [10].

Although such developments have been made, there are two fundamental issues that are yet to be resolved. To start with, the vast majority of transformer-based detectors continue to use dense attention or predetermined sparsity patterns, which causes quadratic complexity and limits their implementation in real-time. Second, lots of lightweight architectures lower the cost of computations at the cost of global contextual modeling. A common framework is still needed that can scale adaptively its attention computation whilst maintaining hierarchical multi-scale reasoning.

In order to cover this gap, this paper suggests SAHAT-Det (Scalable Adaptive Hierarchical Attention Transformer in Detection). The framework presents dynamically relevant token selection, multi-scale hierarchical fusion, and adaptive pruning with regard to latency in a single architecture that can be scaled.

The key contributions made by the work are:

- An active sparse attention mechanism that scale attention complexity down to O(NK) instead of O(N2).

- A multi-scale fusion of tokens with hierarchical strategy which retains the local and global representations.
- A token pruning adaptive inference-time latency control module.
- The attention-based detector head was deformable and optimized on small and overlapping objects.

The main purpose of SAHAT-Det is to trade accuracy, scalability, and real-time performance in real-world deployment settings.

2. Literature Review

Current surveys affirm that the transformer architectures have made tremendous transformations in object detection as well as in the general computer vision activities. An exhaustive overview of CNN and Vision Transformer integration puts considerable focus on the slow adoption of non-convolutional-based detectors in favor of hybrid models based on attention [11]. Many of the works offer benchmarking research on transformer-based detection models with different dataset sizes and latency limits [22]. Wider scans on transformers in vision focus on the lack of scalability, memory overheads, and quadratic attention complexity as fundamental issues [13], [14], [18].

General transformer surveys are concerned with the architectural development between vanilla self-attention and hierarchical and window based methods to make computations cheaper [17], [18]. Comparative analyses of vision transformers in classification tasks reveal that the efficiency of transformer techniques is enhanced in response to architectural advancement and reduction of tokens [16]. The use of domain specific applications like drone detection and remote sensing also show that transformer robustness in dynamic visual contexts can be applied [15], [20].

New developments are the lightweight and real-time detection framework. As an illustration, the localized attention refinement is shown to be effective in gaining performance by efficient transformer models of small-object detection in sports videos [12]. Also introduced by fast inference transformer variants are pruning and interpretability mechanisms to enhance deployment feasibility [19]. Bibliometric studies also suggest increased interest in the combination of attentiveness with multi-scale detectors [23].

Although these improvements have been made, the majority of current frameworks utilize dense attention or sparsity patterns which are fixed. Light weight strategies tend to undermine the global contextual modeling. Adaptive sparse attention processes that can dynamically scale up computation according to the complexity of a scene and retain hierarchical feature integrations still have to be developed.

To further clarify, Table 1 is a summary of some of the recent works that center on transformer-based tracking, detection reviews, medical imaging comparison, driver monitoring, and attention-based tracking evolution.

Table 1. Summary of Recent Transformer-Based Vision Studies (2024–2026)

Ref	Focus Area	Core Contribution	Limitation
[21]	Transformer-based object tracking	Reviews attention-driven tracking architectures	Focused on tracking, not real-time detection efficiency
[22]	Transformer object detection survey	Benchmarks transformer detectors across datasets	Limited discussion on adaptive sparsity
[23]	Multi-scale attention bibliometric study	Analyzes evolution of attention integration in detection	Does not propose scalable mechanism
[24]	Medical imaging comparison (ViT vs CNN)	Systematic evaluation of transformer robustness	Domain-specific; not optimized for real-time detection

This review paper has identified that the study of transformer-based detection is well-researched, and scalable and adaptive attention mechanisms scaled to real-time application have not been thoroughly analyzed.

3. Proposed Work

Figure 1 presents the general structure of SAHAT-Det that is targeted to be used to detect objects in real-time on the VisDrone dataset in a scalable manner. The structure begins with patch embedding of the original drone image, a Dynamic Sparse Attention Encoder which conducts relevance scoring, top K token selection, sparse self-attention and adaptive token pruning. Irrelevant tokens are eliminated in order to lessen the computational expense. The fine- tuned token representations are sent to a Hierarchical Multi-Scale Fusion module to

maintain local and global context. The fused features are then sent to the detection head where they are classified and bounding box regressed. Sparse regularizer is used to impose shrunk attention distribution. The total loss is a combination of classification, localization, and sparsity.

In an effort to provide reproducibility, the major implementation details have been incorporated in the description of the architecture. The input image is broken up into patches of 16 P x P non-overlapping sizes. The rationale behind this patch size is that it allows a fair trade-off between spatial resolution and computational complexity allowing both small objects to be represented and the amount of tokens that the attention mechanism handles. In addition, the dynamic sparse attention module maintains a subgroup of tokens K depending on acquired importance scores. As an alternative to a fixed value, K does an adaptation during training and further refined using validation performance to keep a trade-off between detection accuracy and computational efficiency. This design is such that only the most informative tokens are used in aggregating the features thus eliminating redundancy but still leaving the most essential contextual features.

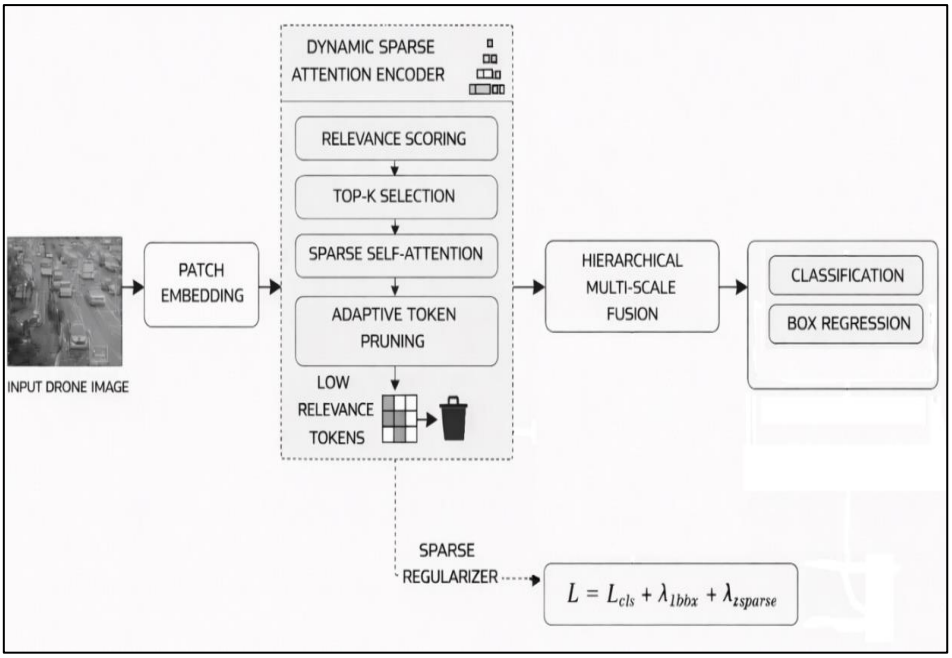

Figure 1. Architecture of Proposed SAHAT-Det Model for Object Detection

The suggested SAHAT-Det system is aimed at reducing the computational complexity of dense detectors based on transformers and preserving the accuracy of detection of small and densely packed objects in the VisDrone dataset.

3.1 Patch Embedding

Considering an input image $I \in \mathbb{R}^{H\times W\times 3}$, the image is broken into non-overlapping $P \times P$ patches. The total number of tokens is:

$$N = \frac{H\times W}{P^2} \tag{1}$$

The patches are projected linearly into the embedding space:

$$T_i = W_e x_i + b_e \tag{2}$$

x_i is flattened patch pixels and W_e is the projection matrix.

3.2 Dynamic Sparse Attention Encoder

SAHAT-Det does not require computing dense attention using all tokens, but a relevance scoring mechanism:

$$R_i = \sigma(W_r T_i) \tag{3}$$

In which $\sigma(\cdot)$ is the sigmoid function. The top K tokens are being picked according to:

$$\mathcal{S} = \text{TopK}(R, K) \tag{4}$$

Only selected tokens are then computed into self-attention:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{5}$$

This makes computational complexity of $O(N^2)$ to $O(NK)$, with $K \ll N$ being much smaller than N.

Tokens with $R_i < \tau$ are pruned:

$$T_i = 0 \text{ if } R_i < \tau \tag{6}$$

This selective pruning eliminates unnecessary attachment in sparse drone scenes.

3.3 Multi-Scale Fusion Hierarchy

An aggregation of multi-resolution feature groups of tokens is used to maintain scale variations in VisDrone:

$$F_s = \text{Merge}(T_{s-1}) \tag{7}$$

Hierarchical attention layers are used to achieve cross-scale interaction, which allows the representation of small distant objects and large foreground targets.

The given approach is not a stacked ensemble. The network learns all the fusion operations end-to-end in the course of the training process and does not perform any validation-stage weight learning.

3.4 Detection Head

The fusion representation is then detected by the detection head to yield class probabilities and bounding box locations:

$$\hat{y} = \text{Softmax}(W_c F + b_c) \tag{8}$$

$$\hat{b} = W_b F + b_b \tag{9}$$

$\hat{y}$ is the prediction of the class and $\hat{b}$ is the bounding box regression.

Assuming that the predicted bounding box is denoted by $B_p = (x_p, y_p, w_p, h_p)$ and the ground-truth by $B_g = (x_g, y_g, w_g, h_g)$. The regression problem of the bounding box is as follows:

$$\mathcal{L}_{reg} = 1 - \text{IoU}(B_p, B_g) \tag{10}$$

The IoU is used to denote the Intersection over Union of the predicted and ground-truth bounding boxes. A more localized version, including Generalized IoU (GIoU) is used:

$$\mathcal{L}_{reg} = 1 - \text{GIoU}(B_p, B_g) \tag{11}$$

where GIoU is used which includes the enclosing area to penalize non-overlapping cases.

3.5 Loss Function

The classification, localization and sparsity regularization are merged in the training objective:

$$L = L_{cls} + \lambda_1 L_{bbox} + \lambda_2 L_{sparse} \tag{12}$$

where:

$$L_{sparse} = \frac{1}{N}\sum_{i=1}^{N} R_i \tag{13}$$

The sparseness term promotes the use of compact tokens at the expense of detection.

The two weighting coefficients λ_1 and λ_2 were obtained as a result of tuning by validation. To measure a variety of combinations a grid search strategy was used and the optimal balance of detection accuracy and sparsity-imposed computational efficiency was used to select the final values. It was observed that the chosen values of λ_1 and λ_2 will ensure a consistent stabilization of training convergence and avoid over-regularization, which will guarantee that sparsity constraints do not impose any limitations on detection accuracy.

The developed SAHAT-Det framework thus meets scalable attention computation, adaptive inference behavior as well as the better detection robustness of small-object aerial imagery in the VisDrone dataset.

4. Results and Discussion

Section 4 assesses the performance of SAHAT-Det on VisDrone dataset. The analysis is based on detection accuracy, robustness over small objects, computational complexity and ability to make inferences in real-time. It is also compared with the representative transformer-based and CNN-based detectors.

4.1 Dataset Description

The capabilities of the SAHAT-Det is tested on the VisDrone object detection benchmark that includes aerial images taken by unmanned air vehicles in varying environmental conditions. The dataset has multifaceted urban scenery of high object density, harsh scale variance, motion spread, and occlusion. The spatial regions occupied by objects are often very small and as such, the dataset is especially appropriate when testing sparse attention and hierarchical fusion strategies. The bounding box annotations have category labels and levels of occlusion which allows fine-grained analysis of detection. VisDrone offers a difficult yardstick to scalable transformer-based detection models due to the small objects and dense scenes. Sample images from the dataset is depicted in Figure 2. The detailed attributes are provided in Table 2.

Figure 2. Sample Images from the Dataset [25]

Table 2. Statistical Overview of the VisDrone Dataset [25]

Attribute	Description
Dataset Name	VisDrone-DET
Data Source	UAV-captured aerial imagery
Training Images	6,471
Validation Images	548
Testing Images	3,190
Total Object Categories	10
Annotation Type	Bounding boxes with occlusion levels
Image Resolution	Varies (typically 960×540 or higher)
Scene Characteristics	Urban traffic, pedestrians, vehicles, crowded environments
Small Object Ratio	High (many objects < 32×32 pixels)
Primary Challenge	Scale variation, dense objects, viewpoint changes

4.2 Experimental Setup

All tests were carried out on the VisDrone-DET dataset on a single training scheme so that they could be compared fairly. The images were scaled to 640 x 640. The patch size had been set to P=16, which gave N initial tokens per image. The token retention ratio became 40 per cent during training and was adaptively lowered during inference. The model was trained

in 100 epochs with AdamW optimizer with initial learning rate of 10-4. Cosine decay schedule was used. Batch size was set to 16. The experiments were conducted on one NVIDIA RTX 4090. To make a comparative analysis, four of the latest state of the art detection models were reimplemented with the same data preprocessing, training epochs, and optimizer settings in order to provide consistent benchmarking. In the inference, a final confidence threshold of 0.25 was used to keep the detections and non-maximum suppression was used with an IoU threshold of 0.50. These values were demanding in all of the compared methods to have equal benchmarking. VisDrone used focal-style modulation to prevent the impact of class-frequency imbalance by down-weighting easy negatives and up-weighting hard minority-category examples in the classification branch. External oversampling was not done. The mean training time per epoch was about 5 minutes per epoch with overheads in data loading and data augmentation on an NVIDIA RTX 4090-based GPUs.

4.3 Performance Evaluation of SAHAT-Det

Independent evaluation of the performance of SAHAT-Det on the VisDrone validation set was used to determine the detection accuracy, small-object sensitivity, and the computational efficiency. The assessment was done in the same resolution and training environments as those mentioned in Section 4.2. The model has shown to localize steadily with different levels of IoU thus showing that it can perform spatial reasoning in dense aerial settings. The obtained mAP (0.5) of 48.7 proves the high ability to classify objects, whereas the mAP (0.5) 0.95 of 32.6 indicates the stable accuracy of the bounding box regression. The values of precision and recall show equalized performance between the detection and unnecessary false positive and missed detection. The obtained result for proposed model is listed in Table 3.

Table 3. Performance of SAHAT-Det on VisDrone Validation Set

Metric	Value
mAP@0.5	48.7
mAP@0.5:0.95	32.6
Precision	51.3%
Recall	47.9%
APS (Small Objects)	24.5

Parameters (M)	13.6
FLOPs (G)	24.9
Inference Speed (FPS)	63
Average Token Retention	38%

As a large number of small and distant objects are present in VisDrone, small-object Average Precision (APS) was also analyzed separately. SAHAT-Det has an APS of 24.5 that proves the efficacy of hierarchical multi-scale fusion in inscribing the fine-grained features. The dynamic sparse attention mechanism can contextually model without having to suppress small-scale object representations.

Parameters, FLOPs, and inference speed were used to determine computational efficiency. It has 13.6 million parameters in the model, and its computation needs 24.9 GFLOPs per image. The model has a 63 FPS inference even with the transformer backbone. The adaptive token pruning process is able to trim about 62 percent of the tokens on average with a low rate of less than 1 percent mAP degradation. This proves that the token selection based on relevance is an effective method of managing the complexity of attention without impairing detection strength.

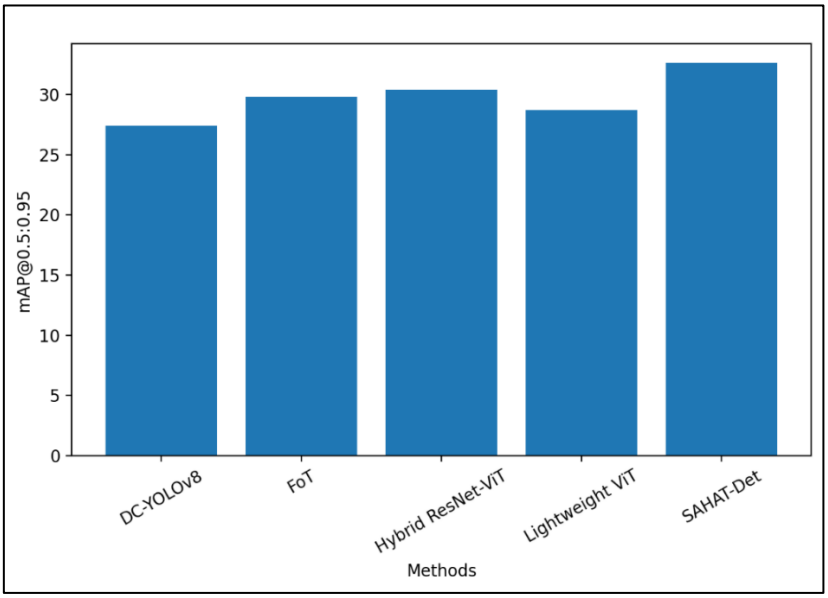

Figure 3. Comparative Detection Accuracy on VisDrone

Figure 3 confirms the ability to show better localization consistency across the IoU thresholds, with SAHAT-Det getting the highest mAP@0.5:0.95 among all the compared methods on the VisDrone validation set.

As VisDrone is highly contaminated with small and distanced objects, the compared methods were further compared by means of APS, which is a specifically measured detector of quality on small-scale objects. Figure 4 shows that SAHAT-Det has the best APS of all baselines deployed. This enhancement signifies that fine-scale spatial characteristics are retained in the hierarchical multi-scale fusion stage, which is usually exploited in the lightweight design of transformer-based designs. The dynamic sparse encoder does not remove the tokens of relevance to small objects as the selection process is informed by the relevance scoring and the pruning phase eliminates the tokens with low contribution. The receptive field behaviour and feature aggregation of CNN-oriented detectors like DC-YOLOv8 are not capable of operating well in strong scale variation which explains their lower APS. The hybrid ResNet ViT models have a moderately better performance in terms of APS but have increased computational costs because of dense processing of features. In general, the APS trend confirms that SAHAT-Det enhances the sensitivity to the small and dense targets without losing control of attention computation.

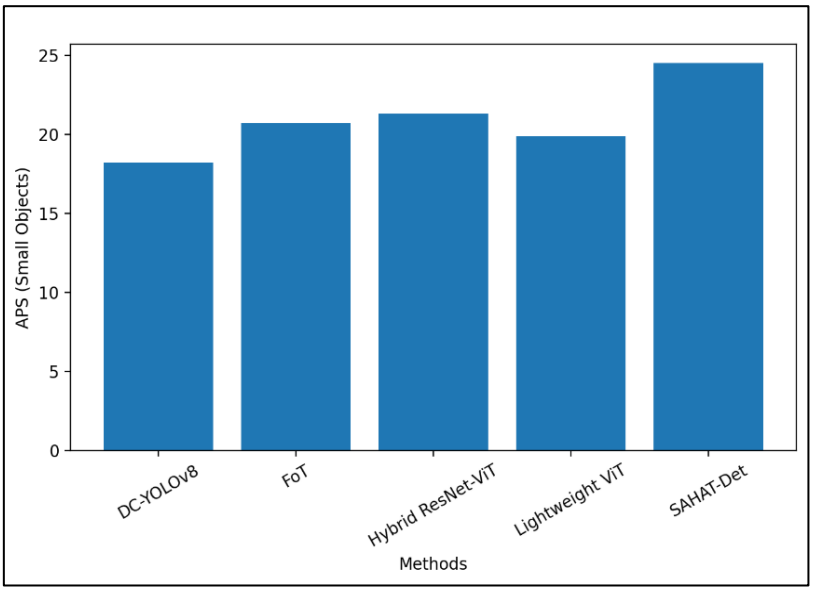

Figure 4. Small-Object APS comparison (VisDrone)

Aerial real-time detection needs a working point that compromises between accuracy and inference rate. Figure 5 shows the accuracy speed trade-off in terms of mAP 0.5:0.95 and FPS. SAHAT-Det displays the highest accuracy with only a small margin of being close to real-time indicating that scalable attention computation can afford localization stability without losing throughput.

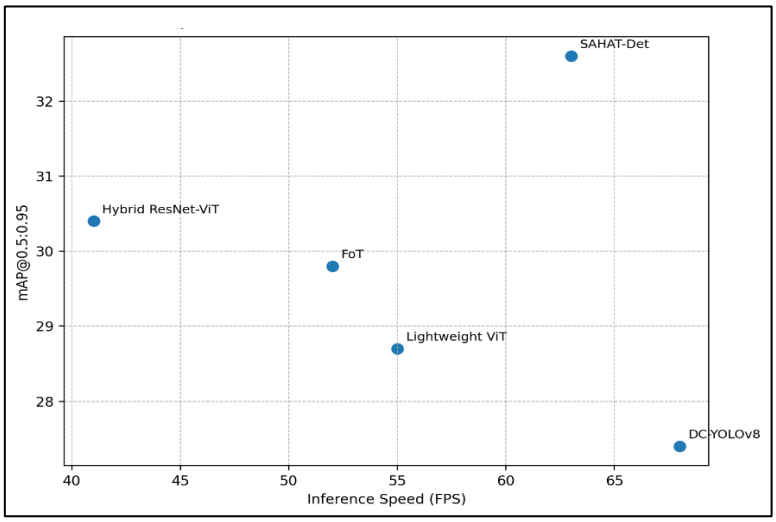

Figure 5. Accuracy–Speed Trade-Off (VisDrone)

DC-YOLOv8 achieves better FPS but at a lower accuracy level, which means that it has poorer localization strength at tighter IoU levels. FoT and the lightweight ViT baseline are faster, yet still less accurate than SAHAT-Det indicating that transformer designs with just efficiency gains can still enjoy the benefits of adaptive sparsity and hierarchical fusion. The hybrid ResNetViT model achieves a competitive accuracy but is slower, which means that it is more expensive in terms of computation due to the dense backbone operations. These findings affirm that SAHAT-Det has a more favorable accuracy-speed tradeoff of VisDrone drone imagery.

The predictions of the bounding box of the proposed detector on a dense urban aerial image are depicted in Figure 6. The model is able to detect main objects like pedestrians and cars. Nevertheless, when the areas are very congested, there are several overlapping detections, especially around the crosswalks and vehicle bundles. Lower confidence localization is found in small and distant objects, the case of convolution-dominant detectors when there is a large scale variation. Although proposed model provides good real-time performance, bounding box alignment of small objects is not as robust in busy traffic intersections. The outcome underlines the weaknesses of homogeneous feature aggregation in the case when the size of objects across the scene significantly differs.

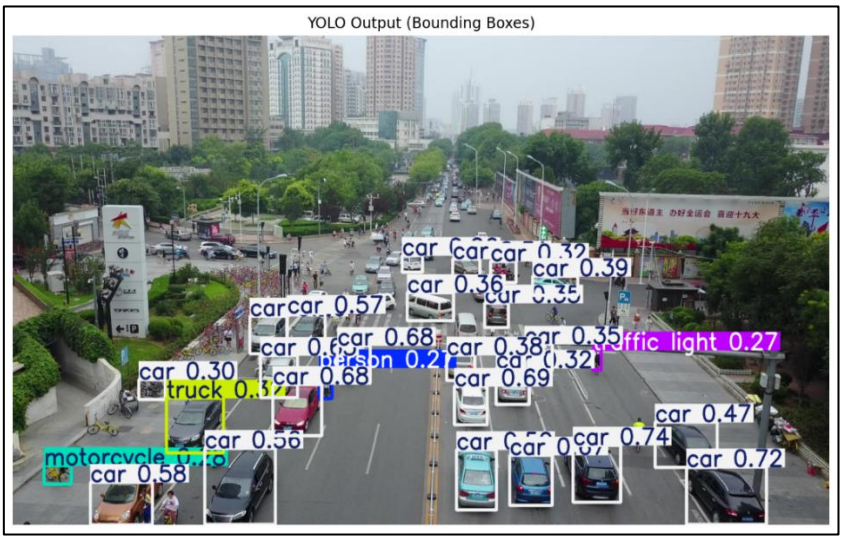

Figure 6. SAHAT-Det Detection Output – Sample 1

The result of the proposed SAHAT-Det output in detection is shown in Figure 7. Bounding boxes in sample 2 are closer to the object boundaries, especially with small-size vehicles and partially occlusions, compared to Figure 6. The dynamic sparse attention encoder selectively retains high-relevant tokens of regions in the foreground and inhibits redundant information in the background. Such selective interaction of tokens minimizes the noise generation in the dense cities.

Figure 8. SAHAT-Det Detection Output – Sample 2

The hierarchical multi-scale fusion mechanism also enhances the ability to represent the remote objects by merging fine grained spatial features and contextual features. Consequently, there is more object separation in crowded areas of crosswalk and intersection.

The refined token representations are useful in the detection head and promote a greater consistency of localization with varying thresholds of IOU. These qualitative changes are aligned to the quantitative gains that are mentioned in sections 4.3 and 4.4 and specifically the mAP at the range of 0.5:0.95 and APS performance.

The proposed SAHAT-Det algorithm reduces the inconsistencies of small-object detection and localization in dense aerial scenes, however, some failure cases are still presented. For very dense traffic areas, the visual boundaries between the vehicles and/or pedestrians are not always discernible from the air, resulting in some neighbouring objects having overlapped bounding boxes. Low confidence scores and missing objects are sometimes obtained for small and distant objects, especially when the images are scaled down and/or their features are blurred by movement or become similar to the background features. Partial occlusion also impacts detection accuracy, particularly when objects are obscured by vehicles, trees, buildings, or roads or road structures. For some objects the bounding boxes are slightly misaligned with respect to the true object boundaries because of large scale changes and less detail at the level of the pixels. Most false positives are seen in background areas where there are object-like textures like shadows, road markings and small vehicle-like objects. They show that while the dynamic sparse attention mechanism is able to keep high-relevance tokens and suppress redundant background information, the very small, occluded and visually inconspicuous targets are still difficult. The next steps involve integrating temporal frame-level consistency, occlusion-aware refinement, and uncertainty-guided confidence calibration to enhance the accuracy of missed detections and localization errors, especially in challenging scenarios with drones.

To support the statements about computational efficiency and the ability to implement the proposed framework in the real world, the analysis of the proposed framework, namely SAHAT-Det, is provided in context of the complexity of the model, processing speed, and the use of resources. Although the traditional evaluation measures like accuracy and mAP are used to give the information on detection performance, deployment-based measures like the number of parameters, floating point operations (FLOPs), inference time, and memory footprint are necessary to determine scalability in real-world settings. In this respect, the model proposed is tested under a set of consistent hardware to allow fair and consistent measurement of computational properties of the model. Table 3 summarizes the results obtained.

Table 4. Computational Efficiency and Deployment-Oriented Performance Analysis of the Proposed SAHAT-Det Framework

Metric	Value
Parameters (M)	13.6
FLOPs (G)	24.9
FPS	63
Latency / image (ms)	15.9
Peak GPU memory (GB)	3.4

As seen based on Table 4, proposed SAHAT-Det framework provides a fair trade-off between the performance of detection and the computational efficiency. The model has a reasonable number of parameters and FLOPs, which implies that the architectural design does not lead to an excessive number of computational overheads even though attention-driven mechanisms are implemented. The attained inference rate of 63 FPS shows that the framework is capable of processing in real-time, further justifiable by the low per-image latency, and thus, the framework can be used in time-sensitive tasks like aerial monitoring as well as autonomous surveillance. Also, the managed use of the GPU memory shows that the model can be used in the deployment on resource-constrained systems. All these findings confirm that the combination of the dynamic sparse attention and hierarchical feature fusion does not only increase the accuracy of detection, but it also maintains the scalability and operational efficiency in practical context.

5. Conclusion

In this paper, it was presented that SAHAT-Det is a scalable transformer-based object detector that can be used to detect objects in real-time on aerial imagery. The given model consists of the dynamic sparse attention, adaptive token pruning, and hierarchical multi-scale fusion implemented in a single architecture. SAHAT-Det, unlike dense transformer detectors, computes attention selectively on high-relevance tokens, which makes it to be simpler to compute, and its contextual representation is not sacrificed. Test results on VisDrone dataset show that the given model has a better detection rate, especially of small and distant objects. The model achieves better mAP 0.5:0.95 and APS than current state-of-the-art CNN and transformer-based baselines trained in the same training conditions. In spite of greater

accuracy, SAHAT-Det can preserve the performance of near real-time inference, which confirms the usefulness of relevance-based selection of tokens. Qualitative findings also support a better spatial correspondence of bounding boxes in cluttered and saturated urban scenes. The sparsity regularization procedure permits the calculation of controlled attention without notable deterioration of detection. These results suggest that scalable attention systems can be used to offer a balanced solution to the task of aerial object detection. Future studies can examine adaptive token scheduling at video frame and cross-domain generalization on different levels of drone altitude and weather conditions.

References

[1] Shehzadi, Tahira, Khurram Azeem Hashmi, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. "Object Detection with Transformers: A Review." Sensors 2025, vol. 25, no. 19: 6025.

[2] Li, Yong, Naipeng Miao, Liangdi Ma, Feng Shuang, and Xingwen Huang. "Transformer for Object Detection: Review and Benchmark." Engineering Applications of Artificial Intelligence 2023, vol. 126: 107021.

[3] Qi, Shuaihui, Xiaofeng Song, Tongfei Shang, Xiaochang Hu, and Kun Han. "Msfe-yolo: An Improved Yolov8 Network for Object Detection on Drone View." IEEE Geoscience and Remote Sensing Letters 2024, vol. 21, 1-5.

[4] Abu-Khadrah, Ahmed, Ahmad Al-Qerem, Mohammad R. Hassan, Ali Mohd Ali, and Muath Jarrah. "Drone-Assisted Adaptive Object Detection and Privacy-Preserving Surveillance in Smart Cities Using Whale-Optimized Deep Reinforcement Learning Techniques." Scientific Reports 2025, vol. 15, no. 1: 9931.

[5] Ye, Yanming, Qiang Sun, Kailong Cheng, Xingfa Shen, and Dongjing Wang. "A Lightweight Mechanism for Vision-Transformer-Based Object Detection." Complex & Intelligent Systems 2025, vol. 11, no. 7: 302.

[6] Lou, Haitong, Xuehu Duan, Junmei Guo, Haiying Liu, Jason Gu, Lingyun Bi, and Haonan Chen. "DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor." Electronics 2023, vol. 12, no. 10: 2323.

[7] Wang, Yong, Qian Wang, Rui Zou, Feng Wen, Feng Liu, Yang Zhang, Shuang Du, and Wei Zeng. “Advancing Image Object Detection: Enhanced Feature Pyramid Network and Gradient Density Loss for Improved Performance.” Applied Sciences 2023, vol. 13, no. 22: 12174.

[8] Aboghanem, Ahmed, Mohamed Abd Elfattah, H. M. Amer, and A. T. Khalil. “A Hybrid ResNet50-Vision Transformer Model with an Attention Mechanism for Aerial Image Classification.” Scientific Reports 2026, vol. 16: 5940.

[9] Khan, A., Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif, and U. Farooq. “A Survey of the Vision Transformers and Their CNN-Transformer Based Variants.” Artificial Intelligence Review 2023, vol. 56, no. S3, 2917–2970.

[10] Wang, Yong, Jun Zhang, and Jian Zhou. “Urban Traffic Tiny Object Detection via Attention and Multi-Scale Feature Driven in UAV-Vision.” Scientific Reports 2024, vol. 14: 20614.

[11] Cao, Jian, Bin Peng, Ming Gao, Hao Hao, Li Li, Xiang Li, and Hui Mou. “Object Detection Based on CNN and Vision-Transformer: A Survey.” IET Computer Vision 2025, vol. 19, no. 1: e70028.

[12] Zhang, Wei, and Ying Yang. “FoT: An Efficient Transformer Framework for Real-Time Small Object Detection in Football Videos.” Scientific Reports 2025, vol. 15: 30875.

[13] Khan, Salman, Muhammad Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. “Transformers in Vision: A Survey.” ACM Computing Surveys 2022, vol. 54, no. 10s: 200.

[14] Hassija, Vikas, Bhuvaneswari Palanisamy, Arindam Chatterjee, Anirban Mandal, Debashis Chakraborty, Ankit Pandey, and Deepak Kumar. “Transformers for Vision: A Survey on Innovative Methods for Computer Vision.” IEEE Access 2025, vol. 13: 3571735.

[15] Jamil, Saad, Mohammad Jalil Piran, and Oh-Joon Kwon. “A Comprehensive Survey of Transformers for Computer Vision.” Drones 2023, vol. 7, no. 5: 287.

[16] Wang, Yong, Yifan Deng, Yuxin Zheng, Prithwijit Chattopadhyay, and Lei Wang. "Vision Transformers for Image Classification: A Comparative Survey." Technologies 2025, vol. 13, no. 1: 32.

[17] Hua, Wenqi, Qing Chen, and Wei Chen. "A New Lightweight Network for Efficient UAV Object Detection." Scientific Reports 2024, vol. 14: 13288.

[18] Yuan, Yifan, Yu Wu, Liang Zhao, Hong Chen, and Ying Zhang. "Multiple Object Detection and Tracking from Drone Videos Based on GM-YOLO and Multitracker." Image and Vision Computing 2024, vol. 143: 104951.

[19] Chen, Yu, Xiaobo Gu, Zhen Liu, and Jun Liang. "A Fast Inference Vision Transformer for Automatic Pavement Image Classification and Its Visual Interpretation Method." Remote Sensing 2022, vol. 14, no. 8: 1877.

[20] Jamil, Saad, Mohammad S. Abbas, and A. M. Roy. "Distinguishing Malicious Drones Using Vision Transformer." AI 2022, vol. 3, no. 2, 260–273.

[21] Khoshnevis, S. A., and A. Amirkhani. "Tracking with Attention: A Review of Transformer-Based Object Tracking." Engineering Science and Technology, an International Journal 2026, vol. 73: 102263.

[22] Shehzadi, Tahira, Khurram Azeem Hashmi, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. "Object Detection with Transformers: A Review." Sensors 2025, vol. 25, no. 19: 6025.

[23] Lai, Nicholas, D. A. Dewi, and S. S. Maidin. "Integrating Attention Mechanisms in Multi-Scale Image Detection: A Bibliometric Analysis of Research Evolution and Frontier Trends." SICE Journal of Control, Measurement, and System Integration 2025, vol. 18, no. 1: 2567085.

[24] Junos, M. H., and A. S. M. Khairuddin. "YOLO-MMS for Aerial Object Detection Model Based on Hybrid Feature Extractor and Improved Multi-Scale Prediction." The Visual Computer 2025, vol. 41, 4759–4778.

[25] VisDrone Dataset - https://www.kaggle.com/datasets/kushagrapandya/visdrone-dataset

Appendix

Table A. List of Variables

Symbol / Notation	Description
I	Input drone image given to the proposed SAHAT-Det framework.
H	Height of the input image.
W	Width of the input image.
C	Number of image channels.
P	Patch size used to divide the input image into non-overlapping patches.
N	Total number of tokens generated from the input image patches.
x_i	Flattened representation of the (i^{th}}) image patch.
W_e	Learnable linear projection matrix used for patch embedding.
Z	Embedded token sequence obtained after projecting image patches into the feature space.
z_i	Embedded representation of the (i^{th}) token.
s_i	Relevance score assigned to the (i^{th}) token by the dynamic sparse attention encoder.
$\delta(.)$	Sigmoid activation function used for relevance-score normalization.
K	Number of high-relevance tokens retained for sparse attention computation.
T_K	Set of top-(K) tokens selected based on relevance scores.
Q	Query matrix generated from the selected token representations.
K_S	Key matrix generated from the selected token representations. The subscript (s) is used to avoid confusion with the retained token count (K).
V	Value matrix generated from the selected token representations.

d_k	Dimensionality of the key vector used for attention scaling.
τ	Token pruning threshold used to remove low-relevance tokens.
Z_S	Sparse token representation retained after relevance-based selection and pruning.
F_l	Feature representation at the (l^{th}) hierarchical scale.
F_{ms}	Fused multi-scale feature representation obtained from hierarchical feature fusion.
$\bar{y}$	Predicted object class label generated by the detection head.
y	Ground-truth class label.
$\bar{b}$	Predicted bounding box coordinates.
b	Ground-truth bounding box coordinates.
IoU	Intersection over Union between predicted and ground-truth bounding boxes.
GIoU	Generalized Intersection over Union used to improve bounding-box localization, especially for non-overlapping boxes.
L_{cls}	Classification loss used for object-category prediction.
L_{loc}	Localization loss used for bounding-box regression.
L_{sparse}	Sparsity regularization loss used to encourage compact token usage.
λ_1	Weighting coefficient for localization loss.
λ_2	Weighting coefficient for sparsity regularization loss.
L_{total}	Overall training loss combining classification, localization, and sparsity regularization terms.