Transformer-based End-to-End Object Detection in Aerial Images

Nguyen D. Vo, Nguyen Le, Giang Ngo, Du Doan, Do Le, Khang Nguyen*
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

***Abstract*—Transformer models have achieved significant milestones in the field of Artificial Intelligence in recent years, primarily focusing on text processing and natural language processing. However, the application of these models in the domain of image processing, particularly on aerial images data, is actively research. This study concentrates on the experimental evaluation of Transformer-based models such as DETR, DAB-DETR, and DINO on the challenging Visdrone dataset, which is also essential for aerial image data processing. The experimental results indicate that Transformer-based models exhibit substantial potential, especially in object detection on aerial image data. Nevertheless, their application is not without challenges, including low resolution, dense object occurrences, and environmental noise. This work provides an initial glimpse into both the capabilities and limitations of Transformer-based approaches within this domain, with the aim of stimulating further development and optimization for practical applications, including traffic monitoring, environmental protection, and various other domains.**

Keywords*—*Object detection; aerial images; end-to-end; transformer-based; DETR; DAB-DETR; DINO

I. Introduction

One of the foundational tasks in the computer vision field is untangling the Object Detection problem. The purpose of this task is to predict the location and classify various objects in an image, thereby fostering an enhanced understanding of visual content. This serves as a critical cornerstone for numerous computer vision applications and various practical technology domains, including healthcare, security, transportation, education, etc. In recent years, the emergence of unmanned aerial vehicles (UAVs, drones, and flycams) has resulted in a surge of aerial data, presenting abundance of advantageous opportunities that conventional sources cannot provide, such as diverse perspectives and panoramic views (Fig. 1). Successfully tackling this task holds significant potential for enhancing and broadening intelligent applications like security monitoring or smart transportation. Hence, object recognition in aerial image data is a subject of paramount importance and a vigorously researched area. However, this task presents a myriad of challenges, including but not limited to small object dimensions, high object densities, and low image resolutions [1], [2].

Research on object detection in recent years can be categorized into three major divisions: two-stage methods, known for their high accuracy, with Faster R-CNN [3] serving as a representative example; one-stage methods, with YOLO [4] algorithm as a prominent representative, known for its fast inference time; and end-to-end methods. End-to-end methods have gained popularity within the research community in recent years due to their simplicity, efficiency, ease of integration, utilization of global information, and time and cost savings during setup and training phases (see Fig. 2). For these reasons, conducting research on end-to-end methods for object detection is essential to enhance performance and facilitate their integration into real-world applications in the field of computer vision [5].

An abundance of end-to-end methods have been proposed to address object detection, including several Transformer-based end-to-end models, such as DETR [6], DAB-DETR [7], and DINO [8]. These methods are evaluated on general objects across standard datasets, such as Pascal VOC [9] and MS-COCO [10]. Each method has its own strengths and weaknesses. In contrast, there is still a limitation in evaluation of these methods in the aerial data domain. Exploring and analyzing the advantages and disadvantages of end-to-end methods promises to provide valuable information for future research.

Therefore, this study focuses on surveying and analyzing three representative end-to-end models, namely DETR [6], DAB-DETR [7], and DINO [8]. Experiments are conducted on standard aerial image datasets VisDrone2019 [11]. Challenges in the aerial image data domain will be highlighted and discussed, along with potential approaches to address the difficulties encountered by these models.

The remaining part of the paper is organized as follows: In Section II, we present related research. Three Transformer-based end-to-end object detection methods, including DETR, DAB-DETR, and DINO, will be described in Section III. The detailed experimental results of the Transformer-based end-to-end method on the VisDrone dataset are reported and discussed in Section IV, along with provided evaluations. Finally, Section V will conclude this paper and suggest directions for future research.

II. Related Works

Object detection represents a foundational task within the field of computer vision, requiring the precise classification and localization of objects of interest within both images and video content.This task holds an essential position in a variety of practical applications, ranging from traditional utilizations like image annotation to modern applications such as autonomous vehicles, robots, surveillance systems, and augmented reality [12]. Over the past decade, object detection methods based on deep learning have garnered significant

*Corresponding author

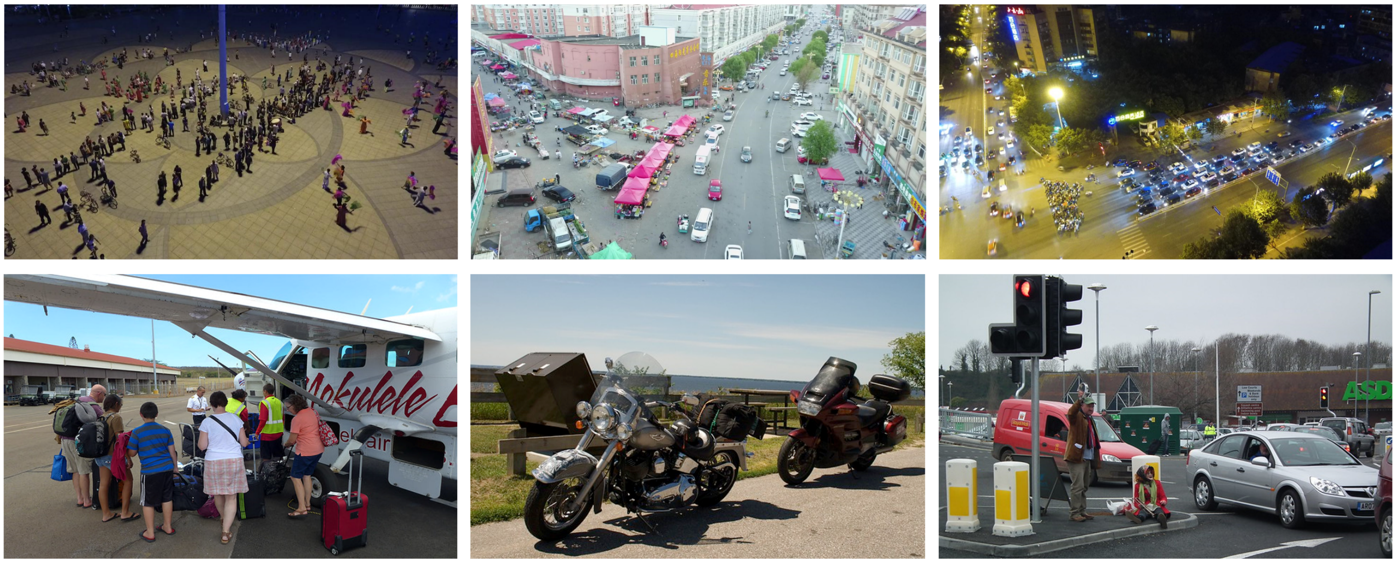


Fig. 1. Images acquired from ground-level vantage points (below) and those obtained from aerial perspectives (above) provide a comparison in terms of perspective and scope. Aerial images (above) offer a wide field of view, encompassing diverse angles and densely distributed small objects. Ground-captured images (below) showcase finer details and focus on more distinct objects.

attention due to the rapid advancements in deep learning techniques [13]. However, there are still a substantial number of challenges, including balancing accuracy and efficiency, handling multi-scale objects, and developing lightweight models.

Traditional object detection methods have primarily relied on convolutional neural networks (CNNs), including Faster R-CNN [3], SSD [14], and YOLO [4]. Some YOLO-based methods, including PH-YOLOv5 [15], AVS-YOLO [16], YOLOv7-Drone [17], and so on, have been specifically developed for object detection in aerial images. Leveraging the considerable success of Transformer in natural language processing (NLP), researchers have been striving to apply Transformer architectures to computer vision tasks. As a result, an extensive number of vision models based on Transformer have emerged in recent years, achieving comparable or even superior performance compared to CNN-based variants.

Transformer architecture [18], which were initially proposed as a self-attention mechanism for machine translation tasks, have increasingly gained attention in object detection, especially in the last three years. High-performance models such as DETR [6], DAB-DETR [7], DINO [8], and many others have been proposed. Currently, Transformer-based models have become a novel approach to object detection, making systematic analysis and evaluation of these models essential for future research.

In recent years, unmanned aerial vehicles (UAVs) have been steadily developing, becoming more affordable, capable of longer flights, and highly maneuverable. Researchers leverage these advantages to employ drones in supporting various daily activities, including rapid delivery services, security surveillance, traffic monitoring, border patrols, and even military use. This has led to the generation of a vast number of images and videos, posing new challenges for object detection. While an abundance of object detection methods have been proposed and have achieved high effectiveness on common datasets like Pascal VOC [9] and MS-COCO [10], they often yield inferior results when tested on non-standard datasets, particularly in the aerial domain. This underscores the need for object detection algorithms and models capable of handling diverse object sizes, densities and viewing angles, as well as adapting to noisy images and low-resolution data resulting from remote sensing. Evaluating Transformer-based methods in the aerial domain is important, as it can provide valuable insights into the challenges of this unique data domain.

III. Methodology

A. *DETR [6]*

In two-stage object detection models, bounding boxes are estimated based on proposals using Region of Interest (RoI), while one-stage detector rely on anchors. Research has shown that the model's performance is significantly influenced by how initial predictions are generated. In mid-2020, Nicolas Carion and Francisco Massa along with other colleagues introduced a completely new approach to the object detection problem. The DETR model (Fig. 3) considers object detection as a set-based matching problem, performs detection and classification in an end-to-end pipeline harnessing the Transformer architecturea distinct paradigm when juxtaposed with one-stage modelsfor comprehensive image processing. DETR also does not generate RoIs or other intermediate steps (e.g., anchor boxes) as in two-stage models.

Two key factors that contribute to DETR's direct object detection capability are a loss function called bipartite matching loss, which ensures a unique match between predictions and ground truth; a network architecture capable of predicting sets of objects and modeling the relationships between them (Fig. 3). DETR is renowned for its revolutionary architecture that reduces the complexity of object detection while achieving strong performance in various scenarios.

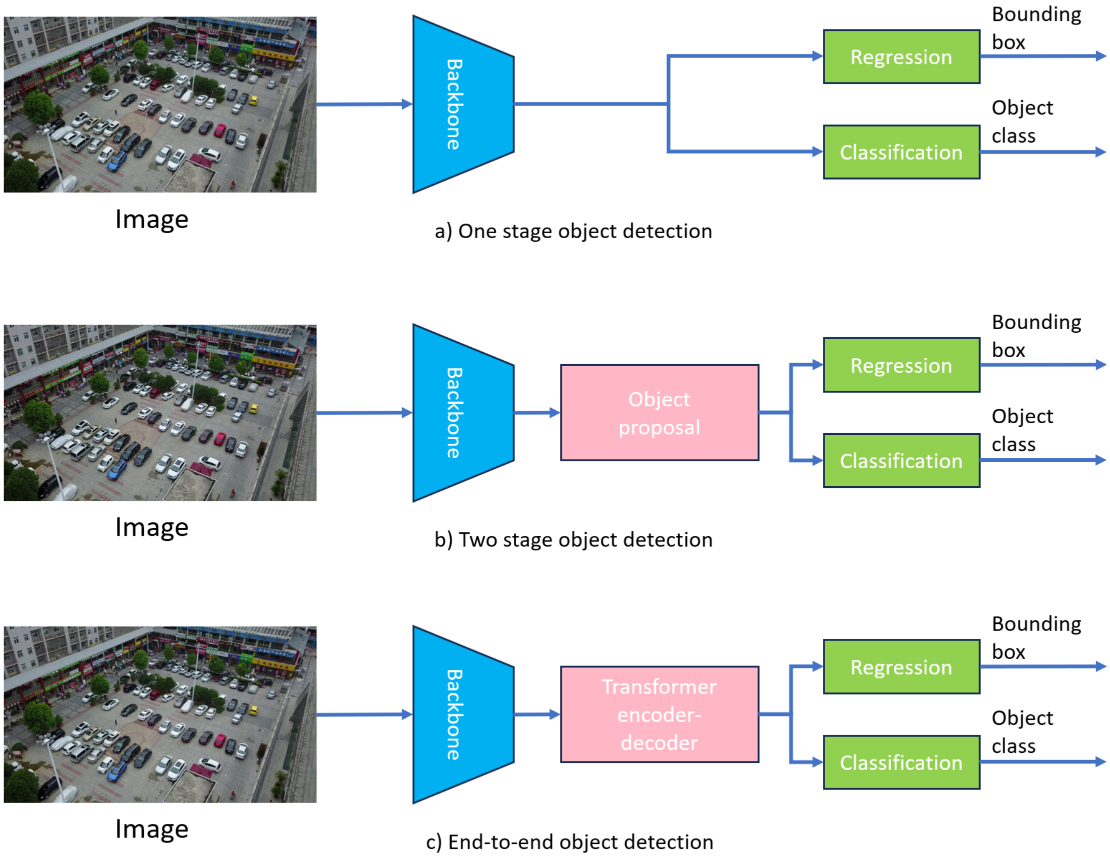


Fig. 2. The framework of the three common object detection methods. Figure a) represents a one-stage object detection model. Figure b) represents a two-stage object detection model. Figure c) represents an end-to-end object detection model.[6]

In contrast, DETR has several limitations, including slow training convergence time compared to other object detection models like Faster R-CNN and subpar performance when detecting small-sized objects. The underlying reason of these problems can be attributed to the absence of components in the Transformer architecture while processing a set of object features. Initially, the attention modules assign random weights to all pixels in the object feature set. A considerable number of epochs is necessary during the training process to allow the attention weights to be learned, focusing on important and sparse pixels [19].

B. DAB-DETR [7]

The DAB-DETR introduces a new query formulation, which is applied within the DETR (Detection Transformer) model, aiming to enhance the understanding of the role of queries in DETR. This new query formulation directly utilizes the coordinates of bounding boxes as queries during the decoding process of the Transformer and dynamically updates them across model layers. This approach has led to significant improvements in the similarity between queries and bounding box features, simultaneously solving the slow convergence issue during DETR training. By using bounding box coordinates as queries, the authors have been able to integrate explicit location information into the querying process and adjust attention maps' positions based on the width and height information of each bounding box.

This representation allows the deployment of queries in DETR as a soft Region of Interest (ROI) aggregation and layer-wise classification stacking process. Specifically, the DAB-DETR method utilizes 4D anchor box coordinates (x, y, w, h) as queries in DETR, as shown in Fig. 4, and updates them across layers. With the information about the size of each anchor box (w, h), Gaussian positional constraints can be adapted to better fit objects of different scales. Additionally, shaping queries as anchor boxes allows for using the center position (x, y) of anchor boxes for feature extraction, increasing the similarity between queries and features and eliminating the slow convergence issue during training. This provides a simpler implementation and a deeper understanding of the role of queries in DETR.

C. DINO [8]

Research directions stemming from DETR are increasingly receiving attention and continuously evolving. The weaknesses of DETR have been addressed and improved continuously. However, most of the improvements have been focused on individual modules and have not resulted in a significant breakthrough. DINO synthesizes prior advancements and introduces superior methods, which have led to a significant leap forward for end-to-end approaches. DINO is a model similar to DETR, with an end-to-end architecture comprising a backbone, a multi-layer Transformer encoder, a multi-layer Transformer decoder, and multiple prediction heads. The overall pipeline is

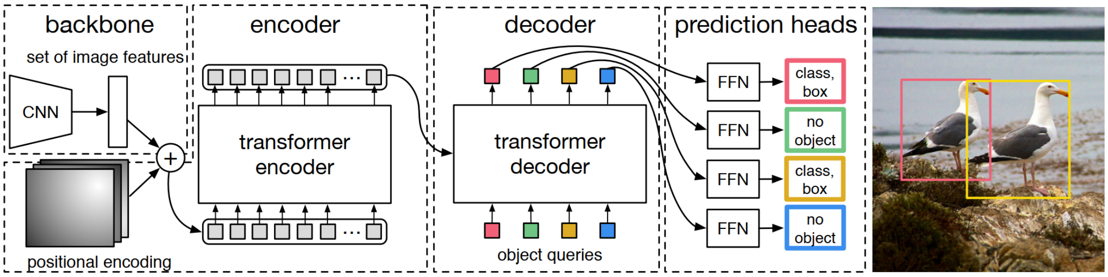


Fig. 3. The architecture of DETR consists of three main components: a CNN network serving as the backbone to extract image features, a Transformer encoder-decoder architecture, and a feed-forward network (FFN) to generate the final predictions [6].

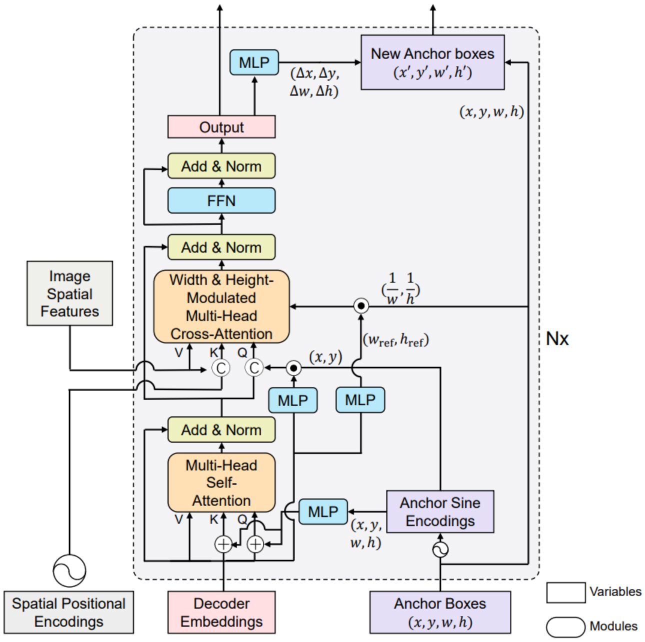

MLP
$(\Delta x, \Delta y, \Delta w, \Delta h)$
New Anchor boxes (x', y', w', h')
(x, y, w, h)
Output
Add & Norm
FFN
Add & Norm
Width & Height-Modulated Multi-Head Cross-Attention
$(\frac{1}{w}, \frac{1}{h})$
$(w_{\text{ref}}, h_{\text{ref}})$
(x, y)
V
K
Q
C
C
Image Spatial Features
Nx
MLP
MLP
Add & Norm
Multi-Head Self-Attention
V
K
Q
MLP
Anchor Sine Encodings
(x, y, w, h)
Spatial Positional Encodings
Decoder Embeddings
Anchor Boxes (x, y, w, h)
Variables
Modules

Fig. 4. DAB-DETR directly uses dynamically updated anchor boxes to provide both a reference query point (x, y) and a reference anchor size (w, h) to improve the cross-attention computation [7].

depicted in Fig. 5.

In this approach, multi-level features are extracted from input images using backbones like ResNet or Swin Transformer. These features, along with positional embeddings, are then processed through a Transformer encoder. A unique query selection strategy is introduced to initialize anchors as position queries for the decoder, while content queries are left for learning. The model utilizes these initialized anchors and learnable content-based queries in conjunction with a deformable attention mechanism to merge features from encoder outputs and update queries at each layer and stage. The ultimate output is generated from adjusted anchor boxes, and classification results are predicted using fine-tuned content features.

Just like DN-DETR, this model incorporates a denoising branch (DN) to carry out noise reduction during training. Beyond the conventional DN technique, a novel noise-contrastive reduction training method is introduced, taking into account challenging negative samples. To maximize the utilization of information from the modified boxes in later stages, which aids in optimizing neighboring stage parameters, a unique "look forward twice" technique is introduced to facilitate the gradient propagation between adjacent layers.

IV. Results and Discussion

A. Dataset

In this work, the VisDrone-DET (object detection in images) dataset [11] is utilized. This dataset comprises images collected through drones in various real-world scenarios, using different types of drones, across multiple locations (14 cities in China spanning thousands of kilometers), and under various weather and lighting conditions. VisDrone-DET contains a total of 8,629 images, with 6,471 for training, 1,610 for test-dev, and 548 for validation (Table I). The dataset also includes over 350,000 bounding boxes for labeled objects across 12 classes: *ignored regions, pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, motor*, and *others*. Excluding the 2 classes, *ignored regions* and *others*, the study delves into the remaining 10 object classes. Some images of the dataset are shown in the Fig. 6

B. Experimental Configuration

Experiments were carried out on Detrex Toolbox [20], Ubuntu 20.04.1 LTS operating system (Linux 5.8.0-53-generic x86-64), Python version 3.8.17, CUDA 11.3, PyTorch 2.0.1, and 2 NVIDIA GeForce RTX 2080 Ti GPUs. Pretrained models are employed for both the training and evaluation processes of three methods: DETR, DAB-DETR, and DINO, all utilizing the R50 backbone. The Average Precision (AP) metric introduced in MS-COCO [10] is used in the object detection process.

C. Discussion

After training the DETR model with a ResNet-50 backbone, the best mAP result obtained was 7.64%. This data reveals that DETR struggles with objects of small or very small sizes. Table II has been presented, showcasing the AP results for each class of interest. Upon analysis, the classes *bus* (19.29%) and *car* (21.70%) achieved the highest scores. Overall, individual class scores remain limited, displaying significant variation. In comparison to other models in this study, DETR's performance on the VisDrone dataset remains notably low.

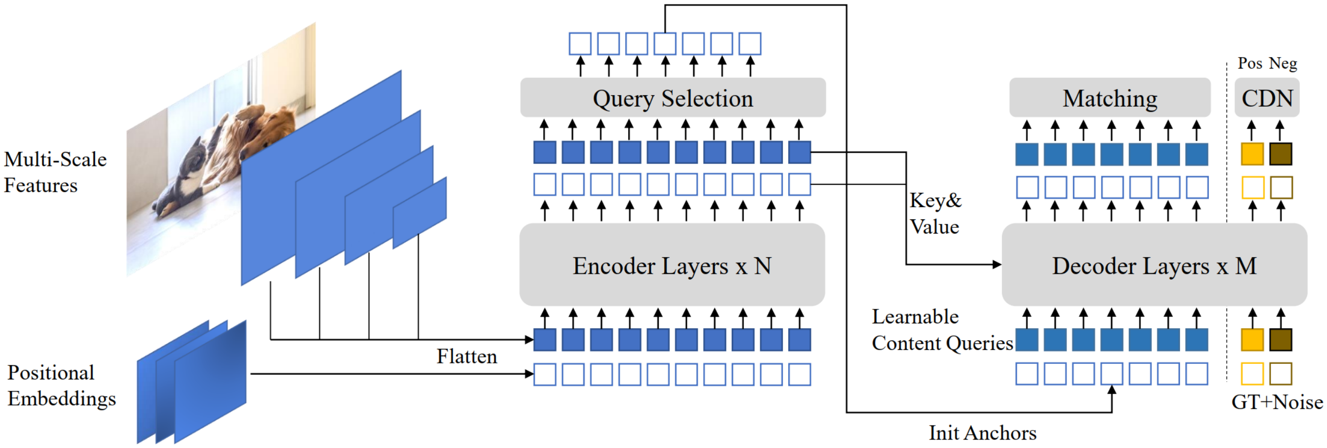


Fig. 5. Illustration of the DINO framework. The main improvements mainly focus on the Transformer encoder and Transformer decoder. The top-K encoder features in the last layer are selected to initialize query positions for the Transformer decoder, while content queries are retained as learnable parameters. The decoder also includes a DeNoising Contrastive component with both positive and negative samples [8].

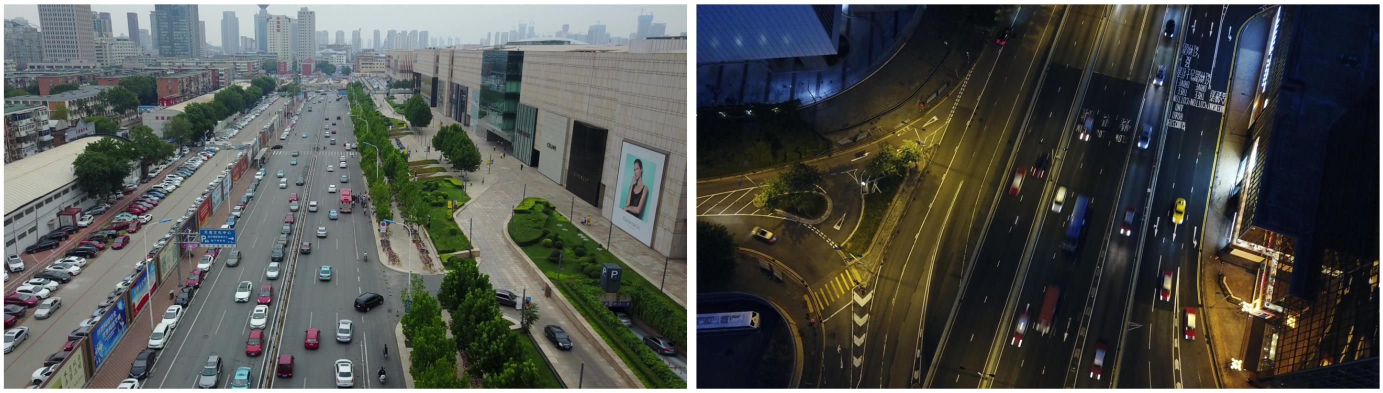


Fig. 6. Some sample images from the VisDrone dataset [11].

TABLE I. STATISTICAL INFORMATION ABOUT THE VISDRONE2019 DATASET

Class/ Subset	Ignore	Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	Awning-tricycle	Bus	Motor	Others
Train	8,813	79,337	27,059	10,480	144,867	24,956	12,875	4,812	3,246	5,926	29,647	1,532
Validation	1,378	8,844	5,125	1,287	14,064	1,975	750	1,045	532	251	4,886	32
Test-dev	2,180	21,006	6,376	1,302	28,074	5,771	2,659	530	599	2,940	5,845	265
Total	12,371	109,187	38,560	13,069	187,005	32,702	16,284	6,387	4,377	9,117	40,378	1,829

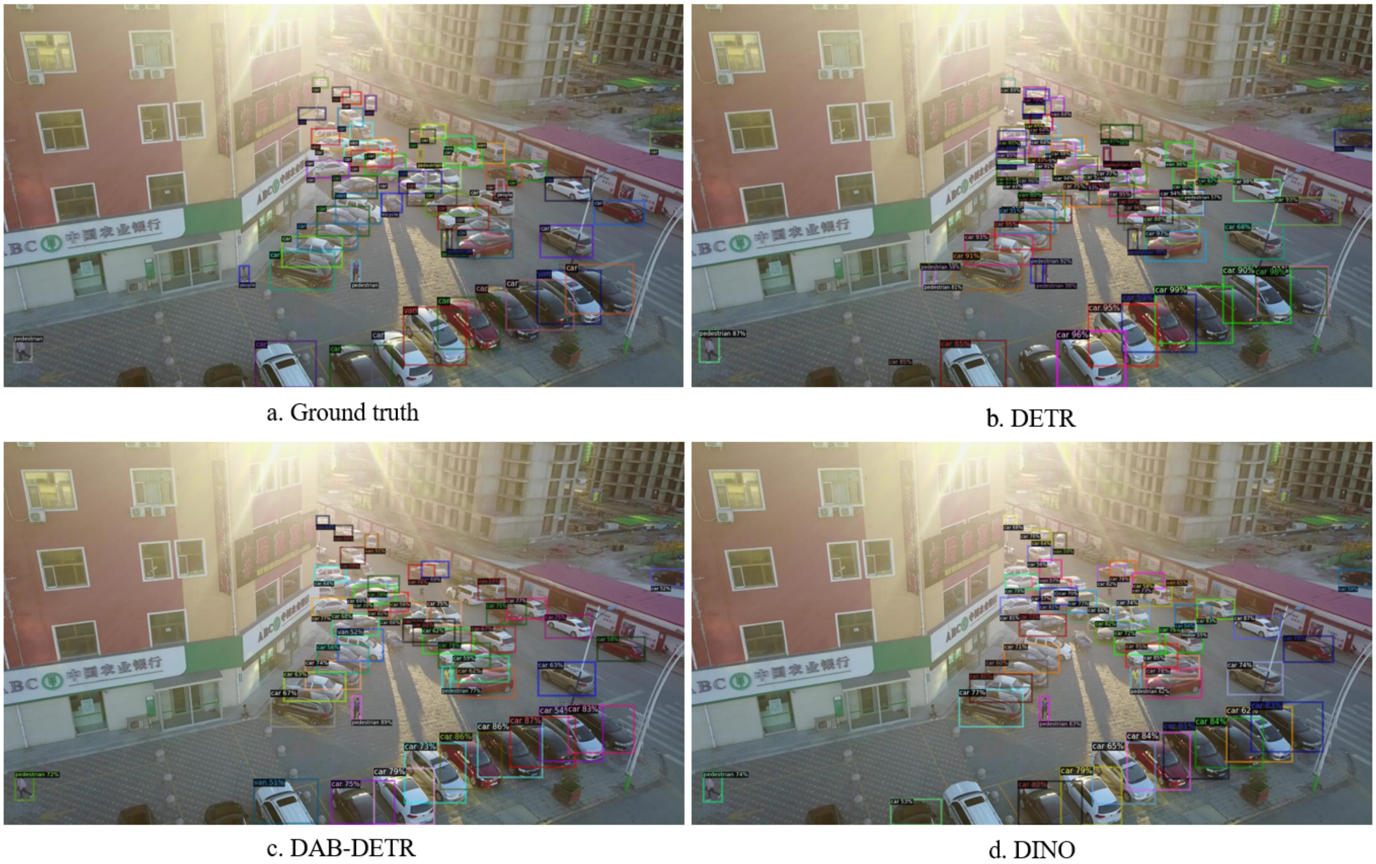
Upon examining the AP scores for each class in DAB-DETR, it is evident that the classes *car* and *bus* attained the highest scores at 36.40 and 34.86 AP points, respectively. Interestingly, the *car* class has the highest number of labels in the training dataset, with 144,867 labels, whereas the *bus* class has significantly fewer labels, specifically 5,926 labels. However, the AP score for the *bus* class is nearly on par with that of the *car* class.

Two other classes, *van* and *truck*, also achieved relatively good AP scores, with 22.31 and 20.81 AP points, respectively. The remaining classes, including *motor*, *bicycle, tricycle, awning-tricycle, people* and *pedestrian*, all had AP scores less than half of those for the *car, bus, van,* and *truck* classes. *Bicycle* and *people* had the lowest AP scores, with only 5.36 and 4.35, respectively.

According to the results presented in Table II, a prominent observation arises, demonstrating DINO's distinction as the top-performing object detector, boasting an mAP score of 24.83%. This achievement can be partly attributed to DINO's notable performance in discerning object categories characterized by resemblances, for instance, *pedestrian* and *people*, with respective scores of 15.60 and 9.38, as well as *tricycle* and *awning-tricycle*, exhibiting scores of 17.25 and 16.76, which appear relatively subdued compared to other object classes. However, it is worth noting that DINO still exhibits relatively weaker performance when compared to other popular methods,

TABLE II. The Evaluation Results of the DETR, DAB-DETR, and DINO Methods on the VisDrone Dataset using the AP Metric

Class / Method	**Pedestrian**	**People**	**Bicycle**	**Car**	**Van**	**Truck**	**Tricycle**	**Awning-tricycle**	**Bus**	**Motor**	**mAP (%)**
DETR	2.41	1.23	1.01	21.70	10.13	8.92	2.71	1.32	19.29	2.57	7.64
DAB-DETR	7.88	4.35	5.36	36.40	22.31	20.81	10.01	7.04	34.86	9.26	16.56
DINO	15.60	9.38	9.98	47.71	31.14	30.56	17.25	16.76	45.05	17.57	24.83


a. Ground truth

b. DETR

c. DAB-DETR

d. DINO

Fig. 7. Visualizing the results of the three object detection methods in challenging scenarios with occlusion and truncation.

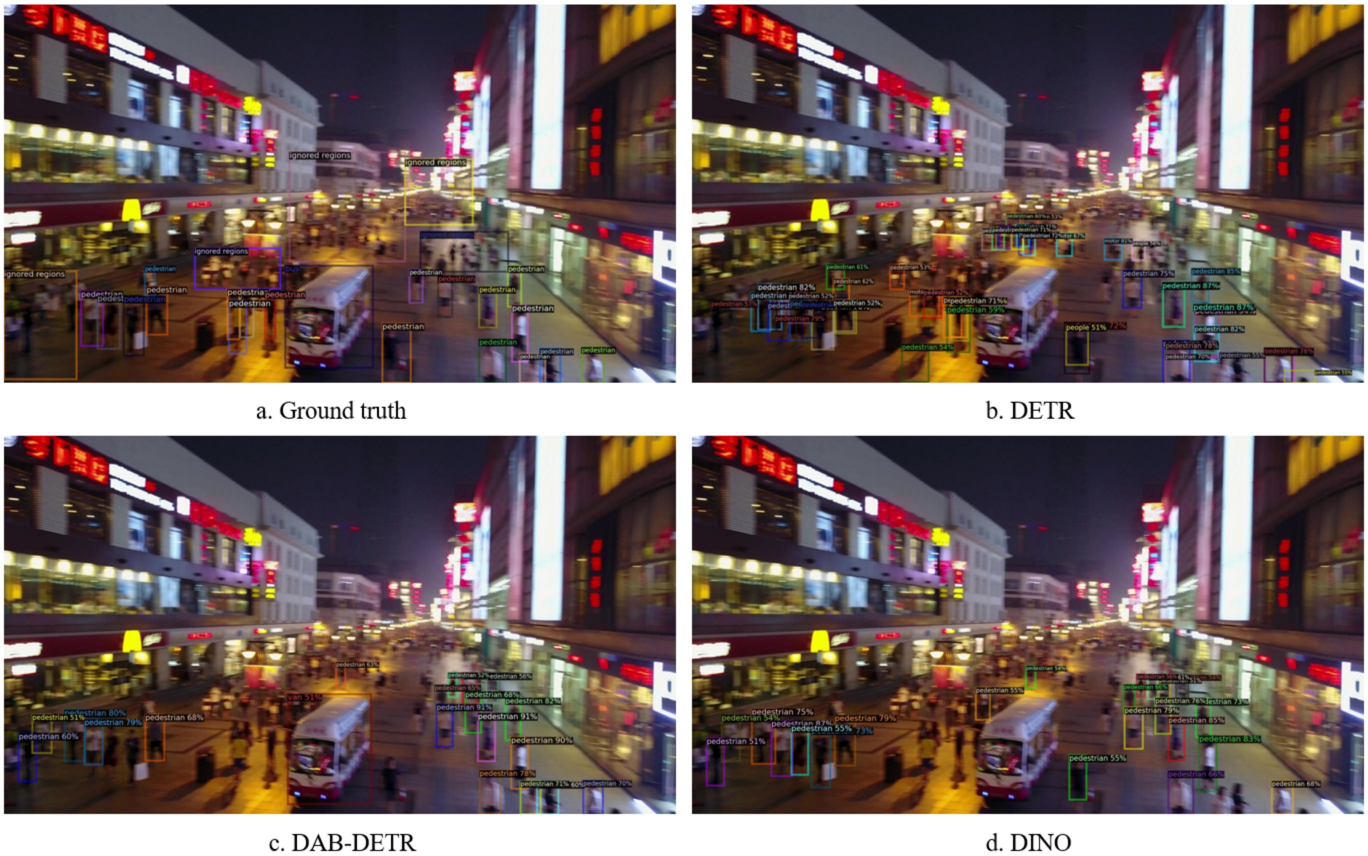
such as YOLOv4 and YOLOv5 [1].

When using three models namely DETR, DAB-DETR, and DINO for inference on challenging images from the VisDrone dataset, each model yielded different results. Specifically, in Fig. 7, which shows a scenario with multiple small-sized, densely packed, and heavily occluded cars, the DETR model exhibited issues with multiple occlusion bounding boxes, imprecise positioning and sizing. On the other hand, the DAB-DETR and DINO models showed fewer instances of occlusion bounding boxes compared to DETR. While some objects were only detected when using a specific model, however, all three models failed to detect a partially obscured car in the distant corner.

Fig. 8 is captured in a more challenging scenario characterized by low lighting conditions, blurred and out-of-focus elements, and a higher density of both smaller and larger objects. The DETR model, while still experiencing bounding box overlap, managed to detect more objects in this context. Both DAB-DETR and DINO yielded relatively similar results, particularly in detecting small pedestrian objects. DAB-DETR outperformed the other two models by detecting a bus object in the center of the image, which remained undetected by the other two methods. Due to the significant number of missed object detections, all three models have not yet achieved satisfactory results when confronted with blurred images, small objects, and high object density.

In Fig. 9, where the object density is not too high but the scene is considerably darker, causing objects to appear more blurred, the DETR model managed to detect most objects in the image, although a few objects were mislabeled, and there was an instance of bounding box overlap. DAB-DETR and DINO, on the other hand, detected fewer objects but provided more accurate results.

From the three examples above, it is evident that the DETR model excels in detecting more objects when images are dark and blurred. However, it faces challenges with a significant number of overlapping bounding boxes, and its accuracy in terms of bounding box size and position is not very high.



a. Ground truth

b. DETR

c. DAB-DETR

d. DINO

Fig. 8. Visualizing the results of the three methods for dark, blurry, and densely populated image scenarios.

This often results in mislabeling of objects. Conversely, the DAB-DETR and DINO models exhibit higher accuracy and stability. However, they are less effective when operating on dark, blurred, or fuzzy images.

V. Conclusion

To offer a fresh perspective on the task of object detection in aerial image domains, we conducted experiments using three novel end-to-end object detection methods based on the Transformer architecture. These methods include DETR, DAB-DETR, and DINO, and they were evaluated on the well-known VisdroneDET2019 dataset. When using mAP as the evaluation metric, we observed that these end-to-end Transformer-based models achieved promising performance. While DETR was a pioneering method in tackling end-to-end object detection, it achieved a modest mAP score of 7.64. In contrast, DAB-DETR achieved a higher mAP score of 16.56 by employing the Dynamic Anchor Boxes technique. Specifically, the model achieving the highest mAP score among the three experimental methods is DINO, with an AP of 24.83. This is attributed to the application of several advanced techniques compared to DETR and DAB-DETR, such as Contrastive denoising training, Mixed query selection, and “Look forward twice". This is a stable and promising result for the object detection task using the end-to-end Transformer-based approach. This paper represents a crucial milestone for us to undertake more effective improvements in future research.

Acknowledgment

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2023-26-01. We also would like to show our gratitude to the UIT-Together Research Group for sharing their pearls of wisdom with us during this research.

References

[1] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, “Detection and tracking meet drones challenge,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.

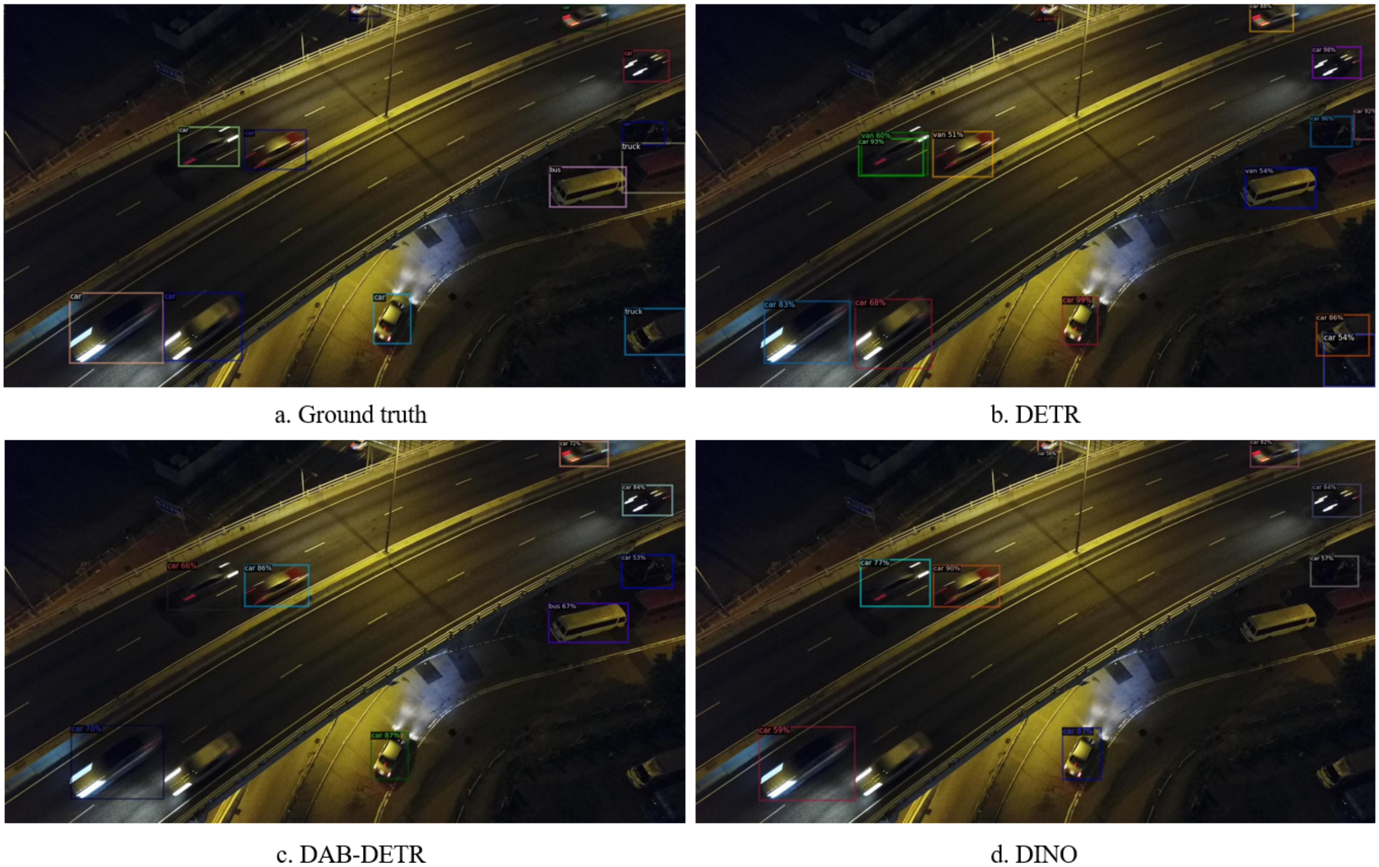
[2] P. Nguyen, T. Truong, N. D. Vo, and K. Nguyen, “Rethinking classification of oriented object detection in aerial images,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, 2022.

[3] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[4] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, “A review of yolo algorithm developments,” *Procedia Computer Science*, vol. 199, pp. 1066–1073, 2022.

[5] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, “A survey on vision transformer,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European conference on computer vision*. Springer, 2020, pp. 213–229.



a. Ground truth

b. DETR

c. DAB-DETR

d. DINO

Fig. 9. Visualizing the results of the three methods in low-light, blurred, and fuzzy conditions on an overpass.

[7] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, “Dab-detr: Dynamic anchor boxes are better queries for detr,” *arXiv preprint arXiv:2201.12329*, 2022.

[8] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, “Dino: Detr with improved denoising anchor boxes for end-to-end object detection,” *arXiv preprint arXiv:2203.03605*, 2022.

[9] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” *International journal of computer vision*, vol. 111, pp. 98–136, 2015.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[11] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang *et al.*, “Visdrone-det2019: The vision meets drone object detection in image challenge results,” in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[12] K.-D. Nguyen, K. Nguyen, D.-D. Le, D. A. Duong, and T. V. Nguyen, “Yada: you always dream again for better object detection,” *Multimedia Tools and Applications*, vol. 78, no. 19, pp. 28 189–28 208, 2019.

[13] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: A survey,” *Proceedings of the IEEE*, 2023.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[15] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, “Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2778–2788.

[16] Y. Ma, L. Chai, L. Jin, Y. Yu, and J. Yan, “Avs-yolo: Object detection in aerial visual scene,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 36, no. 01, p. 2250004, 2022.

[17] X. Fu, G. Wei, X. Yuan, Y. Liang, and Y. Bo, “Efficient yolov7-drone: An enhanced object detection approach for drone aerial imagery,” *Drones*, vol. 7, no. 10, p. 616, 2023.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[19] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” *arXiv preprint arXiv:2010.04159*, 2020.

[20] T. Ren, S. Liu, F. Li, H. Zhang, A. Zeng, J. Yang, X. Liao, D. Jia, H. Li, H. Cao *et al.*, “detrex: Benchmarking detection transformers,” *arXiv preprint arXiv:2306.07265*, 2023.