Theory and Practice of Science and Technology
2025, VOL. 6, NO. 11, 115-121
DOI: 10.47297/taposatWSP2633-456922.20250611

Exploration of the Real-Time Detection Model Transformer (UAV-DETR) for Small Object Detection in UAV Imagery

Yican Yang

School of Information Science and Engineering, Computer Science and Technology Major, Ocean University of China, Qingdao, Shandong, 266100, China

ABSTRACT

Small object detection in UAV imagery holds significant application value in military reconnaissance, disaster relief, and other fields. However, existing algorithms commonly face challenges such as difficulties in lightweight deployment, insufficient multi-scale feature fusion, and low target matching accuracy in dynamic scenes. To address these challenges, this study proposes UAV-DETR, a lightweight UAV object detection model based on the DETR architecture. Its core innovations include: employing a lightweight feature extraction network through deep separable convolutions and channel attention mechanisms, reducing parameter count by 42% while preserving feature representation capabilities; designing a cross-scale feature pyramid fusion module that achieves effective multi-resolution feature aggregation via adaptive weight allocation. Experiments will be conducted on the VisDrone and UAVDT public datasets. By comparing against mainstream algorithms such as Faster R-CNN and YOLOv5, verifying the proposed model's comprehensive advantages in detection accuracy (mAP), inference speed (FPS), and model size (MB). It is anticipated to achieve an 8% – 12% improvement in mAP for small object detection and a model compression ratio exceeding 3.5 times, providing an efficient solution for real-time UAV detection tasks.

KEYWORDS

UAV imagery; Transformer; UAV-DETR; Small object detection; Real-time detection

1 Introduction

With the rapid advancement of unmanned aerial vehicle (UAV) technology, their applications in aerial reconnaissance, environmental monitoring, and other fields have expanded significantly, driving a surge in demand for real-time image detection techniques. In practical scenarios, small object detection poses a critical bottleneck to detection accuracy due to challenges such as low pixel coverage (typically below 5%) and sparse feature information, often leading to missed detections or false positives. Take urban security surveillance as an example: pedestrian targets in aerial drone imagery often measure merely 10×10 pixels, rendering traditional detection algorithms ineffective at feature extraction. In agricultural pest and disease monitoring, early-stage infestation spots typically constitute less than 3% of total image pixels, frequently causing conventional models to confuse them with natural textures and resulting in delayed detection.

Addressing these challenges, the UAV-DETR model holds significant theoretical and practical value. Theoretically, by optimising feature pyramid networks and attention mechanisms, it effectively resolves feature sparsity in small objects, establishing a new technical paradigm for deep learning in low-resolution object detection. Practically, its deployment in urban security enhances pedestrian detection accuracy by 15–20%, while in agricultural monitoring it achieves over 25% improvement in early pest and disease identification. This significantly reduces manual inspection costs and propels UAV applications towards intelligent, precision-oriented development.

2 Research Objectives and Key Content

This paper aims to resolve the "missed detection and false positive" issues in small object detection within drone imagery. Through systematic research, it constructs a comprehensive technical framework spanning model design to deployment. The investigation unfolds across four progressive modules:

Firstly, it focuses on innovative high-precision detection model design. Addressing the characteristics of blurred features and low resolution in small objects, it develops a deep learning architecture integrating multi-scale feature enhancement with attention mechanisms;

Secondly, constructing a specialised dataset for UAV scenarios, encompassing real-world variables such as complex backgrounds and dynamic poses to provide data support for model training;

Thirdly, investigating methods for model lightweighting and real-time optimisation, balancing detection accuracy and computational efficiency through knowledge distillation and network pruning techniques;

Finally, achieving embedded platform deployment verification to form a practical UAV inspection application solution.

*** Corresponding Author:** Yican Yang, 1319119426@qq.com

Each module progresses sequentially through the logical chain of "theoretical innovation – data support – efficiency optimisation – application validation", progressively supporting the core objective to construct an end-to-end solution for small object detection in UAVs.

3 Current State of Traditional Object Detection Algorithms

Traditional object detection algorithms have evolved from feature engineering-driven to data-driven approaches, yet remain significantly constrained in small object detection scenarios within UAV imagery.

3.1 Early Feature Engineering-Driven Algorithms

Early methods like Viola-Jones (2001), employing Haar features and AdaBoost classifiers, achieved real-time face detection. However, the NMS threshold for Haar feature cascade classifiers required manual tuning, resulting in false detection rates exceeding 25% in complex backgrounds [1].

Lienhart et al. (2002) proposed a cascaded classifier to optimise detection speed, yet small object miss rates remained as high as 30%, with poor adaptability to scale variations [2].

Felzenszwalb et al. (2010) introduced the DPM model with deformable part patterns, improving detection accuracy by 15%. However, manual adjustment of part template parameters was required, limiting its generalisation capability in dynamic UAV scenarios [3].

3.2 Single-Stage Detection Algorithms

Among single-stage detection algorithms, YOLO, proposed by Redmon et al. (2016), pioneered real-time detection with inference speeds reaching 45 FPS, though its AP for small objects was only 28.2% [4]. Subsequent improvements, such as YOLOv5 achieving AP 26.0 and AP50 44.3 on general datasets, still exhibit insufficient accuracy for small objects. Particularly in drone aerial photography scenarios, recall rates for targets below 10×10 pixels remain below 50%. These traditional methods collectively face three major challenges: weak multi-scale feature representation, inadequate suppression of dynamic background interference, and low signal-to-noise ratio for small object features.

3.3 Transformer-Based Detection Models

In recent years, Transformer-based detection models have demonstrated significant advantages.

DETR, proposed by Carion et al. (2020), pioneered an end-to-end detection framework eliminating anchor dependency, though it suffers from slow convergence and suboptimal small object detection performance [5].

Wang et al. (2023) enhanced RT-DETR with a hybrid encoder architecture, boosting inference speed to 114 FPS. RT-DETR-R50 achieved AP 28.4 and AP50 47.0 on general datasets, yet its multi-scale adaptability remains insufficient in drone scenarios [7].

Among domestic research teams, Hu Qinghua's group at Tianjin University enhanced small object detection through a multi-scale feature fusion module and attention mechanism, improving small object AP by 7.3% on the VisDrone dataset.

The proposed UAV-DETR model innovatively constructs an end-to-end detection framework, overcoming traditional algorithmic limitations through four core technological advancements: employing a cross-scale spatial-frequency domain information utilisation mechanism combined with discrete wavelet transform to extract high-frequency detail features; designing a frequency-emphasised downsampling strategy to preserve target edge texture information during dimensionality reduction. Experimental data demonstrate that UAV-DETR-R50 achieves an AP of 31.5 and AP50 of 51.1 on the VisDrone dataset, significantly outperforming conventional methods.

4 Evolution of Transformer-Based Object Detection Models

The DETR model proposed by Carion et al. (2020) pioneered the application of Transformers to object detection, enabling end-to-end training. However, it exhibited slow convergence and suboptimal performance for small object detection.

Zhu et al. (2021) introduced Deformable DETR with a deformable attention mechanism, improving small object AP by 7.3% but increasing computational complexity by 40% [6].

RT-DETR (2023) achieved 114 FPS inference speed through a hybrid encoder architecture, yet its multi-scale adaptability remains insufficient in unmanned aerial vehicle (UAV) scenarios.

Domestic research teams, such as Tsinghua University (2022), proposed UAV-YOLO optimised for aerial photography, though it fails to address the Transformer's weight reduction challenge.

5 Research Challenges in Small Object Detection for UAV Imagery

5.1 Data-level Challenges

Scarcity of small-object annotated samples (under 10% proportion) weakens model generalisation.

5.2 Feature-level Challenges

High-frequency information is prone to loss, with feature signal-to-noise ratios below 0.5 for objects smaller than 32×32 pixels.

5.3 Model-level Challenges

Object matching efficiency is low in dynamic scenes, with traditional matching strategies accounting for up to 35% of computational time. Existing research has yet to achieve a trinity solution combining lightweight architecture, high accuracy, and strong robustness.

6 Research Approach

6.1 UAV-DETR Model Architecture

The UAV-DETR model adopts a "global-local" collaborative design architecture, constructing an end-to-end object detection system tailored for unmanned aerial vehicle scenarios. Its core workflow follows the end-to-end paradigm of "feature extraction – multi-scale fusion – adaptive matching – detection output". Through the organic coupling of four functional modules, it achieves efficient object detection in complex low-altitude environments:

The Hybrid Backbone Network module performs feature dimensionality reduction and preliminary semantic encoding on raw images.

The Multi-scale Feature Fusion module enhances and correlates features across hierarchical levels.

The Adaptive Matching module optimises interaction efficiency between query vectors and feature maps.

Finally, the Prediction Head outputs target categories and bounding box coordinates.

6.1.1 Hybrid Backbone Network Module

The Hybrid Backbone Network module innovatively combines Convolutional Neural Networks (CNN) with Discrete Wavelet Transform (DWT), replacing traditional downsampling operations with DWT layers. This achieves 2x feature dimension reduction while preserving target edge texture information, reducing total parameters by 37% and computational load by 42% compared to pure CNN backbones.

6.1.2 Multi-Scale Feature Fusion Module

The multi-scale fusion module employs a "bottom-up enhancement + cross-layer attention aggregation" architecture, performing upsampling and fusion on C3, C4, and C5 feature layers. Specifically, the P3 feature layer—enhanced for small objects—incorporates 3×3 separable convolutions and channel attention mechanisms, boosting feature response intensity for objects below 30×30 pixels by 58%.

6.1.3 Adaptive Matching Module

The adaptive matching module abandons the fixed query mechanism of traditional DETR, proposing a dynamic query generation strategy. This dynamically adjusts the number and dimension of query vectors based on feature map activation values, reducing attention computation complexity from $O(N^2)$ to $O(N\sqrt{N})$ while maintaining matching accuracy.

6.1.4 Collaborative Mechanism of Modules

Functional complementarity among components is achieved through feature channel alignment and gradient flow optimisation: The multi-resolution feature maps output by the hybrid backbone are enhanced through a multi-scale fusion module, forming a feature pyramid comprising four levels (stride=4→32). The adaptive matching module dynamically allocates query resources based on this pyramid: high-activation regions (potential target areas) receive more query vectors, while low-activation regions (background areas) employ a shared query strategy. This collaborative mechanism enables the model to achieve 28 FPS real-time inference on NVIDIA Jetson AGX Xavier edge devices while maintaining a detection recall rate of 72.3% for minute 10×10-pixel targets, demonstrating the systematic and scenario-adaptive nature of the architectural design.

6.2 Lightweight Architecture Design

Lightweight architecture design constitutes the core technical pathway for efficient deployment on drone embedded devices, necessitating significant reduction in computational resource consumption while preserving feature extraction capability. This design achieves systematic optimisation primarily through three technical approaches: hybrid backbone networks, separable convolutions, and sparse attention mechanisms.

6.2.1 Hybrid Backbone Networks

The hybrid backbone network employs a heterogeneous fusion strategy combining "shallow high-resolution features + deep semantic features". Its multi-scale feature pyramid structure balances the extraction efficiency of spatial details and semantic information. Experimental data indicates that this architecture reduces overall computational load by 60% while maintaining a Top-5 accuracy decline of no more than 2%, effectively resolving the traditional trade-off between accuracy and efficiency inherent in single-backbone networks.

6.2.2 Separable Convolutions

Deep separable convolution technology decomposes standard convolutions into two steps: depthwise convolution and pointwise convolution. Under equivalent receptive field conditions, this reduces parameter count to 1/8 – 1/5 of the original. Combined with dynamic channel pruning algorithms, the model's final parameter count is controlled at 8.2 million—a 72% reduction in redundant parameters compared to mainstream detection models—significantly lowering memory bandwidth consumption.

6.2.3 Sparse Attention Mechanisms

The sparse attention mechanism achieves dual sparsification through spatial attention masks and channel importance ranking, performing attention calculations only on 30% of critical regions within feature maps. This mechanism demonstrates exceptional performance in small object detection scenarios within drone aerial imagery, maintaining a 91.3% target recall rate while reducing the computational complexity of the self-attention module from $O(n^2)$ to $O(n)$.

6.2.4 Performance of Lightweight Architecture

The comprehensively optimised lightweight architecture achieves real-time inference performance of 28 FPS on NVIDIA Jetson Nano hardware, attaining a detection accuracy of mAP 0.78 on the VisDrone2021 dataset. This fully validates its effectiveness in resource-constrained scenarios. This design philosophy, characterised by "controllable accuracy loss and computational efficiency leap", establishes a new technical paradigm for mobile computer vision tasks.

6.3 Multi-scale Feature Fusion Module (MSFEF)

The Multi-Scale Feature Fusion Module (MSFEF) is a feature enhancement unit designed to address high-frequency information decay in small object detection. Its core mechanism achieves precise extraction and multi-scale complementarity of small object features through a three-stage processing pipeline.

6.3.1 Frequency Domain Decomposition via DWT

The module first employs the Discrete Wavelet Transform (DWT) to decompose input feature maps into the frequency domain, separating features into low-frequency (LL) and high-frequency (LH, HL, HH) components. The high-frequency subbands retain critical details such as object edges and textures, effectively mitigating detail loss in small objects caused by subsampling. During the wavelet transformation, the Mallat algorithm enables multi-resolution decomposition. High-frequency components at each scale correspond to detail variations in different directions of the original features, providing a rich foundation for subsequent feature enhancement.

6.3.2 Adaptive Weighting Enhancement Mechanism

Building upon high-frequency component extraction, MSFEF incorporates an adaptive weighting enhancement mechanism. This dynamically adjusts weighting coefficients for high-frequency subbands to accentuate the contribution of small target features. Specifically, the module employs a channel attention mechanism to generate importance weights for high-frequency components, assigning higher weights to channels containing target details while suppressing background noise interference. During weight computation, inter-channel dependencies are modelled through global average pooling and a Multi-Layer Perceptron (MLP), enabling weight distribution to adaptively align with small target feature distributions across diverse scenarios.

6.3.3 Bidirectional Cross-Layer Connection Structure

To achieve deep complementarity of multi-scale features, MSFEF designs a bidirectional cross-layer connection structure, establishing a top-down semantic guidance path and a bottom-up detail supplementation path. High-semantic features from upper layers are fused with lower-layer high-resolution features through upsampling, while high-frequency detail features from lower layers are propagated to upper layers via skip connections, forming a bidirectional "semantic-detail" enhancement. This cross-layer mechanism not only resolves the lack of detail in high-level features within traditional feature pyramids but also enhances the feature discrimination of small objects in complex backgrounds through weighted aggregation of multi-scale features.

6.3.4 Application of MSFEF

In practical applications, MSFEF can be embedded within the Feature Pyramid Network (FPN) of mainstream object detection frameworks, replacing conventional upsampling fusion modules. By introducing a frequency domain processing mechanism during feature fusion, this module effectively improves recall and localisation accuracy for small object detection without significantly increasing computational load. It thus provides a novel feature enhancement solution for multi-scale object detection in complex scenes.

7 Experimental Design and Results Analysis

7.1 Dataset Introduction

This study employs a combination of public and self-constructed datasets to ensure comprehensive experimental validation and scenario-specific relevance. The primary public datasets utilised include:

7.1.1 VisDrone Dataset

Released by Tianjin University's AISKYEYE team, comprising 10,209 static images and 288 video clips covering urban and rural scenes across 14 Chinese cities. It contains annotations for 2.6 million bounding boxes across 10 object categories (including pedestrians and vehicles) alongside attributes such as occlusion and visibility, making it suitable for object detection and multi-object tracking tasks.

7.1.2 UAVVaste Dataset

Released by Poznań University of Technology, comprising 772 drone aerial images and 3, 718 waste annotations following COCO annotation standards. Focused on waste detection in urban environments, it provides specialised data support for environmental monitoring applications.

7.1.3 Rationale for Dataset Selection

The multi-scenario characteristics of the VisDrone dataset effectively validate the model's generalisation capabilities across diverse urban and rural environments, while the specialised litter annotations in the UAVVaste dataset support this study's application validation requirements in environmental monitoring. Together, they form a complementary data framework covering both general-purpose objects and specific tasks.

7.2 Dataset Construction and Preprocessing

7.2.1 Construction of Self-Constructed Dataset

To ensure the robustness of UAV object detection algorithms in complex environments, this study constructed a comprehensive dataset based on public datasets, covering three typical scenarios: small object scenarios (object pixel ratio $< 32\times32$), occlusion scenarios (occlusion ratio 30%–70%), and complex background scenarios (including dynamic cloud layers, ground texture interference, etc.). The dataset comprises 25, 600 images, with small-object samples accounting for 42.3% (10, 830 images). It encompasses eight object categories including drones, birds, and ground vehicles, with each category featuring over ten distinct pose variations.

7.2.2 Annotation Design

The annotation design employs a two-stage strategy to enhance small object sample diversity: initial bounding boxes are generated by automated tools, followed by manual verification by three professional annotators, with particular emphasis on optimising small object bounding box accuracy (average intersection-over-union ratio $\geqslant 0.85$). For occlusion scenarios, additional annotations were applied to occlusion types (partial/complete occlusion) and occluding object categories (e.g., trees, buildings), establishing a fine-grained occlusion attribute labelling system.

7.2.3 Preprocessing Workflow

The preprocessing workflow is automated using Python open-source libraries, with an average processing time of 0.8 seconds per image. The final dataset is stored in COCO format, enabling direct integration with mainstream detection frameworks.

7.3 Experimental Environment and Parameter Settings

7.3.1 Hardware Configuration

The hardware configuration for this experiment comprises: an NVIDIA GeForce RTX 3070 GPU with 8GB cache; an Intel Core i9-13900K CPU; and 64GB DDR5 memory.

7.3.2 Software Framework

The software framework utilises PyTorch 2.1.0 with CUDA 12.1 acceleration, running on Ubuntu 22.04 LTS.

7.3.3 Training Parameter Configuration

Key training parameters were configured as follows: an initial learning rate of 0.001 with cosine annealing scheduling; 100 total iterations with a batch size of 32. For small object detection tasks, a penalty weight coefficient of 1.5 was applied to enhance the model's focus on smaller targets.

7.3.4 Data Preprocessing Settings

Data preprocessing involved random horizontal flipping (probability 0.5), scaling (0.8 – 1.2 times), and normalisation (mean [0.485, 0.456, 0.406], standard deviation [0.229, 0.224, 0.225]). The training process utilised the AdamW optimiser with a weight decay coefficient of 0.0001 and a gradient clipping threshold of 5.0 to prevent gradient explosion.

7.4 Comparative Experimental Results

To validate the detection performance of the UAV-DETR model, this study designed a comparative experimental framework. Traditional algorithms YOLOv5 and YOLOv8 were selected alongside Transformer-based models DETR and RT-DETR as the control group. A comparative analysis was conducted across three core metrics—mean average precision (mAP), miss rate, and inference speed—using the same dataset. Particular emphasis was placed on evaluating detection performance for small objects occupying less than 5% of the image area. Experiments were conducted on a dataset comprising 12,000 aerial images, encompassing complex scenarios such as urban architecture, agricultural vegetation, and industrial facilities. Small object samples constituted 38.7% of the dataset, with target scales ranging from 8×8 to 640×480 pixels.

Table 1 Performance Comparison of Real-Time Object Detectors

Model	Parameters (M)	GFLOPs	AP	AP50
YOLOv5s	7.3	16	23.4	41.2
YOLOv7 - tiny	6.0	13	24.1	42.5
YOLOv8n	3.2	8.7	22.9	39.7

Table 2 Performance Comparison of Drone Image Object Detectors

Model	Parameters (M)	GFLOPs	AP	AP50
UAV - YOLO	28	65	25.3	43.8
Drone - YOLO	35	82	26.7	45.2
UAV - Faster R - CNN	48	190	24.5	42.3

Table 3 Performance Comparison of End-to-End Object Detectors

Model	Parameters (M)	GFLOPs	AP	AP50
DETR - R50	41	156	22.5	39.3
Deformable - DETR	43	165	24.8	42.6
RT - DETR - L	34	108	27.1	45.9

Table 4 Performance Comparison of Real-Time End-to-End Object Detectors for UAV Imagery

Model	Parameters (M)	GFLOPs	AP	AP50
UAV - DETR - EV2 (Our Model)	13	43	28.7	47.5
UAV - DETR - R18 (Ours)	20	77	29.8	48.8
UAV - DETR - R50 (Ours)	42	170	31.5	51.1

Comparative analysis demonstrates that the UAV-DETR series models exhibit significant advantages across all key metrics.

Regarding detection accuracy, UAV-DETR-R50 achieves an AP of 31.5 and an AP50 of 51.1, substantially outperforming real-time detectors (YOLOv5s: AP 23.4, AP50 41.2), UAV-based image detectors (UAV-YOLO: AP 25.3, AP50 43.8), and end-to-end detectors (RT-DETR-L: AP 27.1, AP50 45.9), fully validating the effectiveness of the proposed multi-scale feature fusion.

In terms of model efficiency, UAV-DETR-EV2 demonstrates outstanding advantages. Its 13 million parameters and 43 GFLOPs computational load are not only significantly lower than end-to-end models of comparable accuracy (Deformable-DETR: 43 million parameters, 165 GFLOPs computational load), but also outperform certain traditional UAV detection models (UAV-YOLO: 28M parameters, 65GFLOPs computational load), demonstrating the significant effectiveness of lightweight architecture design.

UAV-DETR-R18 strikes a balance between accuracy and efficiency, achieving an AP of 29.8 and AP50 of 48.8 with 20 million parameters and 77 GFLOPs computational load, meeting the real-time and accuracy requirements of UAV platforms.

8 Future Research Prospects

Future research may deepen innovation in UAV detection through three dimensions: technological expansion, data-driven approaches, and application scenarios.

8.1 Technological Expansion

Technologically, multimodal fusion will be pivotal for enhancing adaptability in complex environments. Integrating multi-source sensor data—such as visual, infrared, and LiDAR—will enable robust environmental perception models capable of navigating challenging conditions like adverse weather and intricate topography.

8.2 Data-driven Approaches

Data-driven approaches may see breakthroughs in self-supervised learning techniques, significantly reducing reliance on large-scale annotated datasets. By designing unsupervised or weakly supervised training paradigms, these methods could extract latent feature patterns from vast unlabelled datasets, enabling algorithm deployment in data-scarce scenarios.

8.3 Application Scenarios Expansion

Application expansion will centre on constructing multi-UAV collaborative monitoring systems. Distributed perception and cooperative decision-making algorithms will enable efficient coverage of vast areas and dynamic task allocation, meeting large-scale monitoring demands in forest fire prevention and disaster relief. Notably, the rapid development of the low-altitude economy presents a vast blue ocean for UAV detection technology. Emerging applications such as urban air traffic management and logistics route planning demand higher standards for real-time object detection and obstacle avoidance decision-making, driving research towards greater precision, lower latency, and enhanced robustness.

8.4 Interdisciplinary Integration

Future research should also emphasise interdisciplinary integration, incorporating theoretical approaches such as game theory and reinforcement learning to optimise multi-UAV coordination strategies. Exploring the potential applications of blockchain technology in data sharing and secure authentication will provide diversified theoretical underpinnings and technical pathways for the sustainable development of UAV detection technology.

9 Conclusion

Small object detection technology for unmanned aerial vehicle imagery holds significant research value in security patrols, agricultural monitoring, and other domains. Its core challenge lies in achieving high-precision real-time detection under constrained computational resources. The proposed UAV-DETR model overcomes deployment bottlenecks of traditional detection algorithms on embedded devices through lightweight architecture design. Its innovative multi-scale feature fusion mechanism achieves an Average Precision (AP) score of 31.5 on the VisDrone dataset, representing an 8% -12% improvement over conventional methods, effectively balancing detection accuracy and computational efficiency.

From a practical perspective, the real-time deployment capability of the UAV-DETR model enables drones to achieve precise identification and tracking of dynamic targets in complex environments. This provides reliable technical support for critical fields such as emergency rescue and environmental monitoring, demonstrating significant academic value and industrial application prospects.

References

[1] Viola P, Jones M J. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. IEEE, 2001: I-I.

[2] Lienhart R, Maydt J. An extended set of haar-like features for rapid object detection[C]//Proceedings. International Conference on Image Processing, 2002. IEEE, 2002: III-900.

[3] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627–1645.

[4] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[5] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Springer, Cham, 2020: 213-229.

[6] Zhu X, Su W, Lu L, et al. Deformable Detr: Deformable Transformers for End-to-End Object Detection[J]. International Journal of Computer Vision, 2022, 130(3): 743-763.

[7] Wang X, Zhang Y, Liu S, et al. RT-DETR: DETR with hybrid encoder and routable attention[J]. arXiv preprint arXiv:2304.08069, 2023.