International Journal of Multidiscipline

TLEP – International Journal of Multidiscipline
(Technology, Language, Education, and Psychology)
ISSN: 2488-9342 (Print) | 2488-9334 (Online)

Open Access | Peer-Reviewed | Monthly Publication | Impact factor: 8.497 / 2025

Transformer-Based Object Detection in Natural Images. State-of-the-Art Architectures and Recent Algorithms

Asqarov Elbek Erkinjon o'g'li
Department of Digital Technologies and Mathematics,
Kokand University, Kokand, Uzbekistan
E-mail: e.e.askarov@kokanduni.uz

Abstract
Object detection in natural images is a fundamental computer-vision task that underpins applications ranging from autonomous driving to industrial inspection. Since the introduction of the DEtection TRansformer (DETR), the field has shifted from anchor-based convolutional pipelines toward end-to-end, attention-driven set-prediction frameworks that remove hand-crafted components such as anchor generation and non-maximum suppression. This paper presents a structured review of transformer-based object detectors, tracing their evolution from the original DETR through deformable attention, query-design refinements (DAB-DETR, DN-DETR), contrastive denoising (DINO), collaborative hybrid assignment (Co-DETR), and the most recent real-time variants (RT-DETR, RT-DETRv2, RF-DETR). We organise these methods into a taxonomy according to their core innovations in attention mechanisms, query formulation, and label-assignment strategies, and we compare their reported accuracy and inference speed on the MS-COCO benchmark. The analysis shows that contemporary detection transformers now match or surpass convolutional and YOLO-family detectors in both accuracy and real-time efficiency, with leading models exceeding 60 mean Average Precision on standard benchmarks. We further discuss persistent challenges—small-object localisation, training convergence, and computational cost—and outline promising research directions, including open-vocabulary detection and lightweight deployment.
Keywords: object detection; vision transformer; DETR; self-attention; set prediction; real-time detection; MS-COCO; deep learning.

Introduction

Object detection—the task of simultaneously localising and classifying every object instance in an image—is one of the central problems of computer vision and a key enabler of autonomous systems, robotics, surveillance, medical imaging, and visual search [1], [17]. For nearly a decade, the dominant solutions were convolutional neural network (CNN) detectors built around region proposals, predefined anchor boxes, and a non-maximum suppression (NMS) post-processing stage. Although highly successful, these pipelines rely on hand-crafted components and heuristics that complicate training and limit end-to-end optimisation [18].

The transformer architecture, originally proposed for sequence modelling in natural language processing [1], introduced the self-attention mechanism, which models long-range dependencies between all elements of an input. Its adaptation to images through the Vision Transformer (ViT) [2] demonstrated that purely attention-based models can rival CNNs on image classification. This success motivated researchers to reformulate detection itself as an attention-driven problem.

The DEtection TRansformer (DETR) [3] marked a turning point by casting object detection as a direct set-prediction problem. By combining a CNN backbone with a transformer encoder–decoder and a bipartite-matching loss, DETR predicts a fixed-size set of objects end to end, eliminating anchors and NMS entirely. This conceptual elegance, however, came at the cost of slow convergence and weak performance on small objects, which

prompted an intense wave of follow-up research [20].

This paper reviews the resulting family of transformer-based detectors with an emphasis on the most recent and highest-performing algorithms. The contributions are threefold: (i) we provide a concise taxonomy of detection transformers based on their core innovations; (ii) we compare representative models on the MS-COCO benchmark in terms of accuracy and inference speed using figures reported in the primary literature; and (iii) we identify open challenges and future directions. The remainder of the paper is organised as follows. Section 2 reviews the background of vision transformers and the set-prediction paradigm. Section 3 presents the DETR family and its evolution. Section 4 describes the comparison methodology, and Section 5 discusses the comparative results. Section 6 analyses open challenges, Section 7 surveys applications, Section 8 outlines future directions, and Section 9 concludes.

The self-attention mechanism computes a weighted sum of value vectors, where the weights are derived from the similarity between query and key vectors. Formally, attention is expressed as a scaled dot-product of queries Q, keys K, and values V. Unlike convolution, which aggregates information from a fixed local neighbourhood, self-attention allows every position to attend to every other position, enabling global context modelling in a single layer [1]. Multi-head attention extends this idea by learning several attention patterns in parallel.

The Vision Transformer [2] applies this mechanism to images by splitting an image into fixed-size patches, linearly embedding them, adding positional encodings, and processing the resulting token sequence with a standard transformer encoder. ViT showed that, given sufficient data, attention-only models can outperform CNNs on classification. Hierarchical variants such as the Swin Transformer [12] introduced shifted local windows to reduce the quadratic cost of global attention and to produce multi-scale feature maps suitable for dense prediction tasks, making them popular backbones for detection.

Classical detectors produce a large number of candidate boxes that are later filtered by NMS. DETR instead predicts a fixed set of N object queries, each of which is decoded into either an object or a 'no-object' label. Training uses the Hungarian algorithm to find a one-to-one matching between predictions and ground-truth objects, after which a combined classification and box-regression loss is applied [3]. Two innovations underpin the paradigm shift: the use of self-attention for global context, and an end-to-end learning framework that removes hand-crafted components such as anchors and NMS [18]. This unified formulation simplifies the detection pipeline but introduces optimisation difficulties that subsequent models address.

DETR [3] established the encoder–decoder set-prediction framework. With a ResNet-50 backbone it attains roughly 42 mean Average Precision (AP) on MS-COCO, but only after about 500 training epochs, and it underperforms on small objects because its global attention is uniform and slow to focus [20]. These two weaknesses—slow convergence and small-object accuracy—define the research agenda of the models that follow.

Deformable DETR [4] replaces dense global attention with deformable attention, which attends only to a small set of learnable sampling points around each reference location. This sparse sampling drastically reduces computation, integrates multi-scale features, and accelerates convergence from hundreds of epochs to roughly fifty, while substantially improving small-object detection. Deformable attention has since become a standard

building block of nearly all later detection transformers.

A second line of work re-examined what an object query should represent. DAB-DETR [5] formulates queries as dynamic anchor boxes—four-dimensional coordinates that are refined layer by layer—giving queries an explicit spatial meaning and improving interpretability and convergence. DN-DETR [6] identified the instability of bipartite matching as a major cause of slow training and introduced query denoising: noisy versions of ground-truth boxes are fed to the decoder, which learns to reconstruct them, thereby stabilising the matching process and further accelerating training.

DINO [7] consolidated these advances into a single high-performing framework. It combines a contrastive denoising (CDN) module that adds both positive and negative noised queries, a mixed query-selection strategy for anchor initialisation, and a 'look-forward-twice' box-refinement scheme. DINO reaches about 49 AP with a ResNet-50 backbone in only 12 training epochs [20]. When scaled to a Swin-Large backbone and pre-trained on the Objects365 dataset, DINO achieves 63.2 and 63.3 AP on COCO val2017 and test-dev, respectively, becoming the first end-to-end transformer detector to top the COCO leaderboard [16] (DINO). Stable-DINO [16] later improved the matching stability further, reporting up to 57.7–58.6 AP with a Swin-Large backbone under shorter schedules.

Co-DETR [8] addressed a limitation of one-to-one matching: the small number of positive samples per image weakens encoder supervision. It introduces parallel auxiliary heads that apply conventional one-to-many label assignment (as in Faster R-CNN and ATSS) during training, while keeping the one-to-one DETR head for inference. This collaborative hybrid assignment enriches gradient flow, accelerates convergence, and yields leading accuracy on COCO and LVIS, exceeding 64 AP with large backbones, all without altering the inference pipeline [8], [14].

Until recently, the high latency of attention prevented detection transformers from competing with the real-time YOLO family. RT-DETR [9] changed this by designing an efficient hybrid encoder that decouples intra-scale interaction from cross-scale fusion, together with IoU-aware query selection that initialises object queries from high-confidence regions. As the first real-time end-to-end detector, RT-DETR-L reaches 53.0 AP at 114 frames per second (FPS) on an NVIDIA T4 GPU, and RT-DETR-X reaches 54.8 AP at 74 FPS; the ResNet-50 variant attains 53.1 AP at 108 FPS, surpassing a comparable DINO-Deformable-DETR by 2.2 AP while running about twenty times faster [10], [12]. RT-DETRv2 [10] adds a set of training 'bag-of-freebies' that raise accuracy across all scales without reducing speed, for example improving the small model by 1.4 AP.

Lightweight and domain-robust real-time variants continued this trend. LW-DETR [14] streamlines the architecture for high accuracy at real-time speed, while RF-DETR [11], released in 2025, builds on the deformable-DETR foundation and adopts a DINOv2 vision backbone for strong transfer learning. RF-DETR reportedly became the first real-time detector to exceed 60 mAP on the RF100-VL domain-adaptation benchmark, highlighting the maturity of transformer detectors for diverse real-world deployment.

Several orthogonal directions enrich the design space. Salience DETR [15] introduces hierarchical salience filtering to select informative encoder tokens and reduce redundancy. YOLOS [13] demonstrates that a plain ViT, augmented with learnable detection tokens, can perform detection without any convolutional components, underscoring the flexibility of pure transformers. OWL-ViT [14] extends

detection to the open-vocabulary setting by aligning image features with text queries, enabling zero-shot recognition of categories described only in natural language. Sparse and conditional attention variants further trade accuracy for efficiency.

Methodology

To compare the reviewed models on an equal footing, we adopt the MS-COCO 2017 benchmark [17], the de-facto standard for natural-image object detection, and report mean Average Precision (AP) averaged over Intersection-over-Union thresholds from 0.50 to 0.95. For real-time models we additionally report inference throughput in frames per second (FPS) on an NVIDIA T4 GPU, and for convergence behaviour we note the number of training epochs, since this differs by an order of magnitude across models. All figures are taken from the respective primary publications rather than re-implemented, so absolute values should be read as indicative of architectural trends rather than as the outcome of a single controlled experiment. Where models use different backbones or large-scale pre-training (for example Objects365), this is stated explicitly because it strongly affects accuracy.

Table 1 summarises representative detection transformers. Two trends are evident. First, successive query and assignment innovations have compressed training from roughly 500 epochs for the original DETR to about 12 epochs for DINO, while simultaneously raising accuracy. Second, the hybrid-encoder real-time models have closed the latency gap with YOLO detectors, so that transformers now lead in both accuracy and speed at comparable scales [10].

Model	Year	Backbone	COCO AP	Speed / Epochs	Key innovation
DETR [3]	2020	ResNet-50	~42	500 ep	Set prediction; no anchors/NMS
Deformable DETR [4]	2021	ResNet-50	~46	50 ep	Deformable multi-scale attention
DAB-DETR [5]	2022	ResNet-50	~45	50 ep	Dynamic anchor-box queries
DN-DETR [6]	2022	ResNet-50	~48.6	50 ep	Query denoising for stable matching
DINO [7]	2023	ResNet-50	49.0	12 ep	Contrastive denoising; mixed queries
DINO [7]	2023	Swin-L + O365	63.2	—	Scaled SOTA on COCO leaderboard
Stable-DINO [16]	2023	Swin-L	57.7–58.6	—	Stable matching
Co-DETR [8]	2023	Large + O365	>64	—	Collaborative hybrid assignment
RT-DETR-L [9]	2024	HGNetv2	53.0	114 FPS	Real-time hybrid encoder
RT-DETR-X [9]	2024	HGNetv2	54.8	74 FPS	IoU-aware query selection
RT-DETRv2-L [10]	2024	ResNet-50	53.4	108 FPS	Bag-of-freebies
RF-DETR [11]	2025	DINOv2	>60*	real-time	Domain-robust; DINOv2 backbone

*Table 1. Representative transformer-based detectors on MS-COCO. Values are as reported in the cited primary sources; backbones and pre-training differ. *RF-DETR exceeds 60 mAP on the RF100-VL domain benchmark.*

The accuracy progression from DETR to DINO and Co-DETR confirms that the principal gains stem not from larger backbones alone but from better query formulation and label assignment, which improve the use of training signal. Meanwhile, RT-DETR and its successors show that careful encoder design can eliminate the historical speed penalty of attention, achieving over 100 FPS without sacrificing accuracy [10]. The 2025 release of RF-DETR, built on a self-supervised DINOv2 backbone, further indicates that strong pre-trained representations are becoming a decisive factor for robustness across domains [11].

Despite rapid progress, several challenges remain. Small-object detection continues to lag behind large-object accuracy; although deformable and multi-scale attention have narrowed the gap, dense scenes with tiny, overlapping instances remain difficult [19]. Training convergence, while greatly improved, still depends on careful query initialisation and denoising schedules. The quadratic complexity of attention makes high-resolution inputs and very deep encoders computationally expensive, motivating sparse, windowed, and hybrid designs [12], [15]. Finally, detection transformers are data-hungry: their strongest results rely on large-scale pre-training datasets such as Objects365, which are costly to assemble and may not transfer cleanly to specialised domains.

Transformer-based detectors are increasingly deployed across natural-image domains. In autonomous driving and robotics, the end-to-end formulation and global context modelling improve detection in cluttered scenes, while real-time variants meet strict latency budgets [9]. In remote sensing and aerial imagery, Swin- and DETR-based detectors handle scale variation and oriented objects [7]. In industrial inspection, medical imaging, and agriculture, domain-robust models such as RF-DETR transfer effectively to small, specialised datasets [11]. Open-vocabulary detectors such as OWL-ViT enable flexible recognition of novel categories described in text, broadening the practical scope of detection beyond fixed label sets [14].

Future research is likely to advance along four axes. The first is efficiency: lightweight and hardware-aware architectures that preserve accuracy under tight compute and energy budgets, extending the real-time line begun by RT-DETR and LW-DETR. The second is robustness and generalisation through self-supervised and foundation-model backbones such as DINOv2, reducing dependence on large labelled datasets. The third is unification, in which detection is integrated with segmentation, tracking, and vision–language understanding inside a single transformer, enabling open-vocabulary and promptable detection. The fourth is reliable small-object and dense-scene detection, where hierarchical salience filtering and improved multi-scale fusion are promising. Progress along these axes will determine how broadly detection transformers replace convolutional pipelines in deployed systems.

Conclusion

Transformer-based object detection has matured from an elegant but slow proof of concept into the state of the art for natural-image detection. Beginning with DETR's set-prediction formulation, successive innovations—deformable attention, anchor-box and denoising queries, contrastive denoising in DINO, and collaborative hybrid assignment in Co-DETR—have delivered large gains in both accuracy and training efficiency. The latest real-time models, RT-

DETR, RT-DETRv2, and RF-DETR, demonstrate that attention-based detectors can now match or exceed the YOLO family in speed while leading in accuracy, with top systems surpassing 60 AP on standard benchmarks. Remaining challenges in small-object accuracy, convergence, computational cost, and data efficiency define an active research frontier, but the trajectory is clear: detection transformers have become a unifying and increasingly practical paradigm for object detection in natural images.

References

A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017, pp. 5998–6008.

A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in Proc. ECCV, 2020, pp. 213–229.

X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in Proc. ICLR, 2021.

S. Liu et al., "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in Proc. ICLR, 2022.

F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query denoising," in Proc. CVPR, 2022, pp. 13619–13627.

H. Zhang et al., "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in Proc. ICLR, 2023.

Z. Zong, G. Song, and Y. Liu, "DETRs with collaborative hybrid assignments training," in Proc. IEEE/CVF ICCV, 2023, pp. 6748–6758.

W. Lv et al., "DETRs beat YOLOs on real-time object detection," in Proc. IEEE/CVF CVPR, 2024.

W. Lv, Y. Zhao, Q. Chang, et al., "RT-DETRv2: Improved baseline with bag-of-freebies for real-time detection transformer," arXiv:2407.17140, 2024.

Roboflow, "RF-DETR: A real-time, transformer-based object detection model," Technical Report, 2025.

Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF ICCV, 2021, pp. 10012–10022.

Y. Fang et al., "You only look at one sequence: Rethinking transformer in vision through object detection," in Proc. NeurIPS, 2021.

M. Minderer et al., "Simple open-vocabulary object detection with vision transformers," in Proc. ECCV, 2022, pp. 728–755.

X. Hou, M. Liu, S. Zhang, P. Wei, and B. Chen, "Salience DETR: Enhancing detection transformer with hierarchical salience filtering refinement," in Proc. IEEE/CVF CVPR, 2024.

S. Liu, T. Ren, J. Chen, et al., "Detection transformer with stable matching," in Proc. IEEE/CVF ICCV, 2023.

T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in Proc. ECCV, 2014, pp. 740–755.

T. Cao et al., "Object detection based on CNN and vision-transformer: A survey," IET Computer Vision, 2025.

A survey, "Transformers in small object detection: A benchmark and survey of state-of-the-art," ACM Computing Surveys, 2025.

T. Khan et al., "Object detection with transformers: A review," Sensors (MDPI), 2025.