Object Detection - Model layer

This section covers the **model layer** for object detection, focusing on architectures, key innovations, and practical implementations.

Canonical DETR family (Transformer-based, set prediction)

- **DETR (ResNet backbone)** — `facebook/detr-resnet-50`, `facebook/detr-resnet-101` *Bipartite matching (Hungarian), NMS-free, strong with multi-scale aug; slower to converge vs one-stage.*
- **DETR (ViT backbone, compact)** — `facebook/detr-resnet-50` (swap backbone in fine-tune) or **YOLOS** (see below).
- **Deformable DETR** — `SenseTime/deformable-detr` *Multi-scale deformable attention → much faster convergence and small-object gains.*
- **DN-/DAB-/DINO-DET** — (community ports) e.g., `IDEA-Research/dino-5scale` *Query denoising, anchor refinement, stronger training signals → SOTA-ish mAP on COCO.*

ViT-style single-stage

- **YOLOS** — `hustvl/yolos-tiny`, `hustvl/yolos-small`, `hustvl/yolos-base` *ViT adapted for detection; light and easy to fine-tune via 🤗 Transformers.*

Open-Vocabulary / Language-Grounded Detectors

- **OWL-ViT (zero-shot OD)** — `google/owlvit-base-patch32`, `google/owlvit-large-patch14` *Text queries → detect novel categories without box supervision.*
- **Grounding DINO** — `IDEA-Research/grounding-dino-base`, `groundingdino/swint-ogc` *Phrase grounding + detection; strong zero-shot and promptable OD.*
- **GLIP / OWLv2 (if needed)** — community checkpoints exist on HF Hub for open-vocab detection.

High-throughput one-stage (non-Transformer backbones, widely used)

- **YOLO family** — (Ultralytics exports on Hub; inference via `ultralytics` or ONNX) e.g., `ultralytics/yolov8n`, `ultralytics/yolov8l` *Real-time, strong engineering; train outside Transformers API or via custom loaders.*
- **RT-DETR** — `PaddlePaddle/RT-DETR-R50` (ports available) *Real-time DETR variant balancing accuracy/latency.*

Domain / Task-specific

- **Oriented/Rotated** — (DOTA/xView ports on Hub; e.g., Rotated-YOLO, Oriented-RCNN) *Adds angle to boxes; aerial/remote sensing.*

- **Instance Segmentation (box + mask)** — `facebook/mask2former-swin-large-coco-instance` (if boxes + masks needed).
- **Video OD (per-frame baseline)** — use above image detectors frame-wise; trackers (e.g., ByteTrack) add IDs externally.

Architectural Innovations (cheat-sheet)

- **Two-stage CNNs (Faster/Mask R-CNN):** region proposal → per-ROI heads; accurate, heavier, mature ecosystem.
- **One-stage CNNs (YOLO/RetinaNet):** dense predictions, focal loss; real-time, excellent engineering & tools.
- **Anchor-free (FCOS/CenterNet/YOLOX head):** predict centers/boxes directly; simpler label assignment.
- **Transformers for detection (DETR):** set prediction with **Hungarian matching**, global attention, **NMS-free**.
- **Deformable attention:** sparse, multi-scale sampling → faster training, better small-object recall.
- **Query tricks (DN-, DAB-, DINO-DETR):** denoising, anchor refinement, better query initialization → big mAP bumps.
- **Open-vocab / grounded:** align vision with text (CLIP-like) → **zero-shot** detection from prompts (OWL-ViT, Grounding DINO).
- **Real-time optimizations:** lightweight necks/heads, quantization, dynamic shapes, knowledge distillation.
- **Rotated boxes / oriented heads:** regress angle for aerial/OCR/logistics.
- **Video extensions:** temporal features or simple per-frame detect + tracker for strong baselines.
