



Computer-Vision Tasks, Metrics and Models

Perception and generation | Classification, detection (IoU, NMS, mAP), semantic / instance / panoptic segmentation, keypoints, tracking, OCR; YOLO, DETR, SAM / SAM 2 as of 2026

#ComputerVision #Detection #Segmentation #YOLO #SAM

Seeing as a set of tasks

Computer-vision task

×1

Overview

A question you ask about an image: what is it, where is it, which pixels are it, how does it move.

×10

Close up

Each question fixes the output format, the label you need, the loss and the metric.

- label → accuracy
- box → mAP
- mask → mIoU / PQ
- track → HOTA, IDF1

×100

Under the microscope

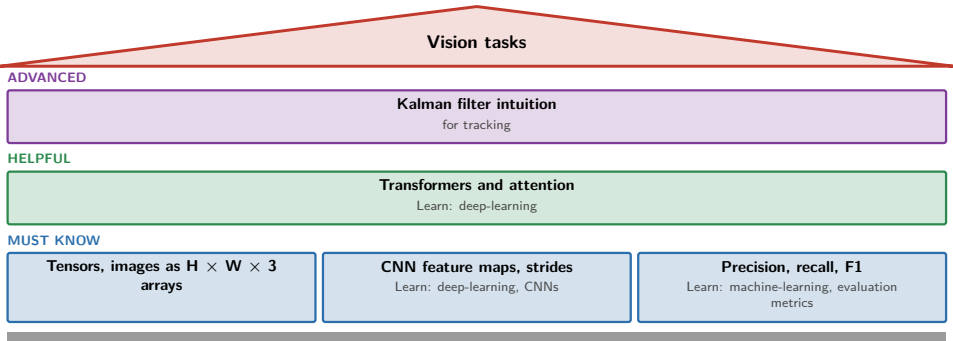
Modern systems share one pretrained backbone and swap heads; foundation models (SAM, DINOv2, CLIP) let you skip most labelling.

- open-vocabulary detectors take class names as text
- a VLM answers free-form questions but localises less precisely

ANALOGY

A doctor reading an X-ray: is there a fracture (classification), where (detection), exactly which bone area (segmentation), and is it healing over visits (tracking).

Computer vision at three magnifications



Task	Output	Metric	Typical models	Datasets
Classification	one label (or top-k)	top-1 / top-5 accuracy	ResNet, ConvNeXt, ViT	ImageNet-1k (1000 classes)
Object detection	boxes + class + score	mAP (COCO AP@[.5:.95])	YOLO, RT-DETR, Faster R-CNN	COCO (80 classes), LVIS
Semantic seg-mentation	class per pixel	mIoU	U-Net, DeepLab, Seg-Former	ADE20K (150), Cityscapes

Instance seg-mentation	mask per object	mask AP	Mask R-CNN, YOLO-seg	COCO
Panoptic seg-mentation	class + id per pixel	PQ (panoptic quality)	Mask2Former	COCO panoptic
Keypoints / pose	named points per object	OKS-based AP	HRNet, ViTPose, RTM-Pose	COCO (17 body points)
Multi-object tracking	id per box over time	HOTA, IDF1, MOTA	ByteTrack, BoT-SORT	MOT17, MOT20
OCR	text + its location	CER, WER	PaddleOCR, Tesseract, VLM OCR	ICDAR sets

The task ladder: each row adds information and labelling cost

Semantic

class per pixel, cars merged

vs

Instance

one mask per car

Tell: Can you count the objects from the output?

Instance

things only

vs

Panoptic

things + stuff, every pixel

Tell: Is the sky labelled too?

One-stage (YOLO)

boxes in one pass

vs

Two-stage (Faster R-CNN)

proposals, then refine

Tell: Is there a region-proposal step?

Closed-set detector

fixed class list

vs

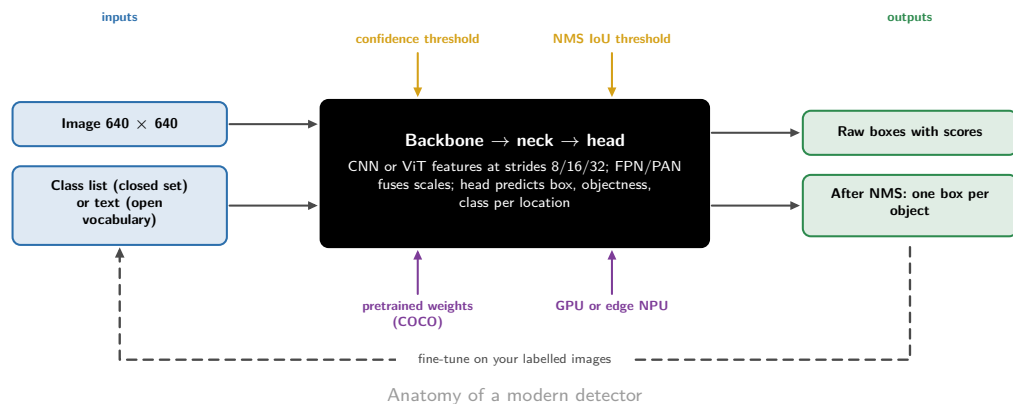
Open-vocabulary

Grounding DINO, OWLv2, YOLO-World

Tell: Can you type a new class name at run time?

Pairs beginners confuse, and the one question that separates them

© Detection: IoU, NMS, mAP



Intersection over Union

FORMULA $\text{IoU} = I / (A + B - I)$

WHERE

- IoU = overlap score of two boxes (0-1)
- I = area of the intersection rectangle (px²)
- A = area of box A = (x2 - x1)(y2 - y1) (px²)
- B = area of box B (px²)

WORKED

I = 400
A = 1600
B = 1600
IoU = 400 / (1 600 + 1 600 - 400)
IoU = 0.1429

RESULT IoU = 0.1429

READ IT Boxes (10,10,50,50) and (30,30,70,70): each 40 × 40 = 1600, overlap 20 × 20 = 400, IoU ≈ 0.14, a miss at the usual 0.5 threshold.

MISTAKES

- Intersection width = max(0, min(x2a, x2b) - max(x1a, x1b)): clamp at 0 for disjoint boxes
- Mixing (cx, cy, w, h) and corner formats
- Two boxes shifted by half their width have IoU 1/3, not 1/2

PROBLEM Run NMS (IoU threshold 0.5) on four 'person' boxes (x1, y1, x2, y2)

GIVEN

- B1 (100,100,200,200) score 0.95
- B2 (110,110,210,210) score 0.90
- B3 (300,100,400,200) score 0.80
- B4 (105,95,205,195) score 0.60
- each box is 100 × 100 = 10 000 px²

1 Sort by score, keep the best

keep B1 (0.95)

Why: the most confident box survives

2 IoU(B1, B2)

I = 90 × 90 = 8100; U = 20000 - 8100 = 11900; IoU = 0.68

Why: 0.68 > 0.5 → delete B2 (duplicate)

3 IoU(B1, B3)

x ranges 100-200 and 300-400 do not overlap → IoU = 0

Why: different object → keep

4 IoU(B1, B4)

I = 95 × 95 = 9025; U = 10975; IoU = 0.82

Why: duplicate → delete B4

5 Repeat with the next survivor

B3 has nothing left to suppress

Why: loop ends when the list is empty

ANSWER Kept: B1 and B3, two people

CHECK

✓ Lowering the threshold to 0.3 would not change this result; raising it to 0.75 would keep B2 (0.68) as a false duplicate

Rank	Match	Precision	Recall
1	TP	1/1 = 1.00	1/3 = 0.33
2	TP	2/2 = 1.00	2/3 = 0.67
3	FP	2/3 = 0.67	0.67
4	TP	3/4 = 0.75	3/3 = 1.00
5	FP	3/5 = 0.60	1.00

AP by hand for one class: 5 detections ranked by score, 3 ground-truth objects, match = IoU ≥ 0.5

Interpolated AP takes, at each recall, the best precision at that recall or higher: 1.00 on 0-0.67 and 0.75 on 0.67-1.0, so AP = 0.33 × 1 + 0.33 × 1 + 0.33 × 0.75 ≈ **0.92**. mAP averages AP over classes; **COCO AP** also averages over 10 IoU thresholds 0.50, 0.55 ... 0.95 (AP50 = loose, AP75 = strict localisation; APs/APm/API by object size).

EXAM TIP

Report the **IoU threshold** with every mAP: 'mAP 0.62' on VOC (AP50) and 'AP 0.45' on COCO (AP@[.5:.95]) are not comparable. A detector with high AP50 but low AP75 finds objects but draws sloppy boxes.

Model families (as of 2026)

YOLO family	DETR family	Open-vocabulary
one-stage CNN, dense predictions + NMS YOLOv5, v8, YOLO11 (Ultralytics, AGPL-3.0 or paid licence); v9, v10, v12 by academic groups fastest on edge devices; huge tooling ecosystem	transformer decoder with object queries, set loss, no NMS DETR (2020), Deformable DETR, RT-DETR (2023, real time), D-FINE, RF-DETR (2025) simpler pipeline, strong on crowded scenes; real-time variants now rival YOLO	image-text alignment: boxes for any typed phrase OWL-ViT / OWLv2, Grounding DINO, YOLO-World no relabelling for new classes; slower and less precise than a fine-tuned closed-set model

© SAM (Apr. 2023)

ViT image encoder run once per image, light prompt encoder and mask decoder answer clicks, boxes or rough masks in milliseconds; trained on SA-1B (11 M images, 1.1 B masks). Returns masks, not class names.

🔄 SAM 2 (Jul. 2024)

Extends SAM to video: a memory bank of past frames and prompts propagates the mask through the clip (streaming), with corrections on any frame. Also faster than SAM on images.

🔍 SAM 3 (Nov. 2025)

Adds concept prompts: a short noun phrase or an example image returns masks for every matching instance, in images and video, closing the gap with open-vocabulary detectors.

👤 Labelling with SAM

Typical 2026 workflow: an open-vocabulary detector proposes boxes, SAM turns each box into a mask, a human fixes the few errors, then you fine-tune a small fast model (YOLO-seg, RT-DETR) on the result for production. The big models label; the small model serves.

🔍 Keypoints, tracking, OCR

- **Keypoints** COCO body pose = 17 points (nose, eyes, ears, shoulders, elbows, wrists, hips, knees, ankles). **Top-down**: detect each person, then regress points in the crop (accurate, cost grows with people). **Bottom-up**: find all points, then group them (OpenPose; constant cost). Metric: AP over OKS thresholds.
- **Tracking-by-detection** Detect per frame → predict each track's next box with a **Kalman filter** → match by IoU (and appearance embeddings) with the **Hungarian** algorithm. SORT = IoU only; DeepSORT adds a re-identification embedding; **ByteTrack** also matches low-score boxes, recovering occluded objects; BoT-SORT adds camera-motion compensation.
- **Tracking metrics** **MOTA** counts misses, false positives and identity switches; **IDF1** measures how long ids stay correct; **HOTA** balances detection and association (the current default).

- **OCR pipeline** Text detection (DB, CRAFT) → line recognition (CRNN with CTC loss, or a transformer such as TrOCR) → layout and reading order. Tesseract (LSTM engine since v4) is fine for clean scans; PaddleOCR for scene text; 2025-2026 VLM-based document parsers read tables and forms end to end but can hallucinate text.

⚠️ VLM OCR can invent text

A classic OCR engine outputs garbled characters when unsure; a vision-language model outputs **fluent but wrong** words, which is harder to spot. For invoices, IDs and medical records, keep character-level confidence, cross-check totals, or use a classic engine as a second opinion.

🚩 Myths and recap

YOU MAY HEAR

"Higher confidence threshold = better detector"

BUSTED

✅ ACTUALLY

It trades recall for precision

mAP integrates over all thresholds; choose the operating point from the cost of a miss vs a false alarm.

YOU MAY HEAR

"DETR-style models always beat YOLO"

BUSTED

✅ ACTUALLY

It depends on latency budget, hardware and data

Benchmark both on your images at your input size and batch size.

YOU MAY HEAR

"SAM tells you what the object is"

BUSTED

✅ ACTUALLY

SAM returns masks, not classes

Pair it with a detector, a classifier or SAM 3 concept prompts for labels.

Task fixes output and metric

label, box, mask, track

$\text{IoU} \geq 0.5 = \text{match}$

COCO averages 0.50-0.95

NMS removes duplicates

DETR does not need it

Foundation models label, small models serve

SAM + open-vocab → YOLO / RT-DETR

Q1 IoU of two 40×40 boxes overlapping on a 20×20 square?

A $400 / (1600 + 1600 - 400) \approx 0.14$: a union is not a sum.

Q2 What does NMS remove and what threshold does it use?

A Lower-scored boxes whose IoU with a kept box exceeds the NMS threshold (often 0.45-0.7).

Q3 Why does DETR not need NMS?

A It is trained with one-to-one Hungarian matching, so each object gets exactly one query.

Q4 COCO AP vs AP50?

A AP averages IoU thresholds 0.50 to 0.95 in 0.05 steps; AP50 uses only 0.50, so it is easier.

Q5 Semantic vs instance vs panoptic in one line each?

A Class per pixel; mask per object; every pixel gets class + instance id (stuff and things).

Q6 What does SAM 2 add over SAM?

A Video: a memory of past frames propagates and refines the mask across the clip.

Q7 What makes ByteTrack better under occlusion?

A It also associates low-confidence detections instead of discarding them.

Q8 How many keypoints in COCO body pose, and what metric?

A 17; AP computed over OKS thresholds.

Q9 What licence trap does the Ultralytics YOLO code carry?

A AGPL-3.0: closed-source commercial use needs an enterprise licence.