LECTURE NOTES

Foundations of Computer Vision: Image Classification, Object and Mass Detection, and Image Segmentation

Definitions, Taxonomies, Methods, Evaluation Metrics and Applications, with Worked Examples and Figures

Dr. Md. Atiqur Rahman
Associate Professor
Department of Computer Science and Engineering
Islamic University, Kushtia, Bangladesh

Contents

1. Introduction to Visual Recognition 4

1.1 Digital Image Representation 4

1.2 Why Recognition is Difficult 4

2. Image Classification 6

2.1 Types of Image Classification 6

2.2 Classical (Hand-Crafted) Pipeline 7

2.3 Deep Learning Pipeline: Convolutional Neural Networks 7

2.4 Landmark Classification Architectures 8

2.5 Evaluation Metrics for Classification 9

2.6 Applications of Image Classification 10

3. Object Localization, Object Detection and Mass Detection 11

3.1 Bounding Box Representations 11

3.2 Intersection over Union (IoU) 11

3.3 Anchor Boxes and Non-Maximum Suppression 12

3.4 Taxonomy of Object Detectors 13

3.5 Loss Functions for Detection 15

3.6 Evaluation: Precision-Recall, AP and mAP 15

3.7 Other Types of Detection 16

3.8 Mass and Lesion Detection in Medical Imaging 16

3.9 Applications of Object Detection 17

4. Image Segmentation 19

4.1 Types of Image Segmentation 19

4.2 Classical Segmentation Techniques 20

4.3 Deep Learning for Segmentation 21

4.4 Loss Functions for Segmentation 22

4.5 Evaluation Metrics for Segmentation 22

4.6 Applications of Image Segmentation 23

5. Comparison, Related Tasks and Applications 24

5.1 Summary Comparison of the Three Tasks 24

5.2 Related Vision Tasks 24

5.3 Applications by Domain 25

5.4 Benchmark Datasets 25
5.5 Practical Workflow for a Vision Project 26
6. Annotation and Dataset Design **27**
6.1 Annotation Formats 27
6.2 Annotation Quality 27
6.3 Dataset Splitting and Data Leakage 27
7. Choosing a Task and a Model Family **28**
7.1 Practical Trade-offs 28
8. Mini Case Studies **29**
9. Common Pitfalls **29**
10. Glossary **31**
11. Summary **31**
12. Review Questions **31**
Further Reading 32

1. Introduction to Visual Recognition

Computer vision is the field of study that enables machines to extract meaningful information from digital images and videos. Among its many problems, three are considered the core recognition tasks because almost every practical vision system is built on top of one or more of them: **image classification**, **object detection** and **image segmentation**. These tasks differ mainly in the *granularity* of the answer they produce: a single label for the whole image, a set of labelled boxes, or a label for every pixel.

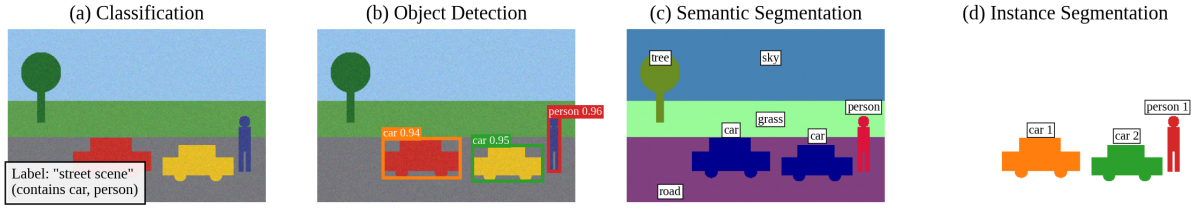


Figure 1. *The same input image processed by four recognition tasks. (a) Classification assigns one label to the whole image. (b) Detection outputs a class label, a confidence score and a bounding box for each object. (c) Semantic segmentation colours every pixel by its class. (d) Instance segmentation separates individual objects of the same class.*

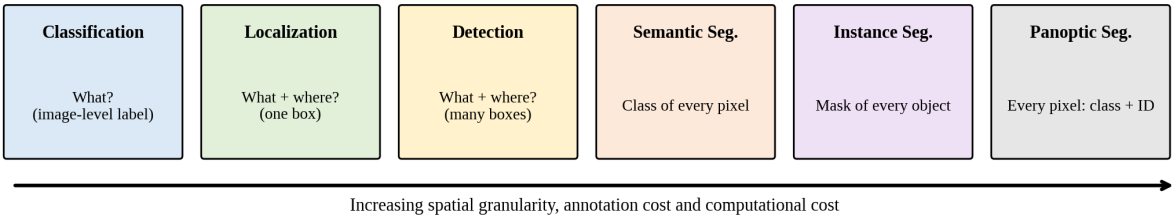


Figure 2. *Hierarchy of recognition tasks ordered by output granularity. Moving to the right requires richer annotations and more computation, but yields more detailed scene understanding.*

1.1 Digital Image Representation

A digital image is a discrete function $I(x, y)$ defined on a grid of pixels. A grayscale image of height H and width W is a matrix $I \in \mathbb{R}^{H \times W}$, and a colour image is a tensor $I \in \mathbb{R}^{H \times W \times C}$ with $C = 3$ channels (red, green, blue). Pixel values are usually integers in [0, 255] for 8-bit images and are normalised to [0, 1] or standardised (zero mean, unit variance) before being fed to a learning algorithm.

> **Example**
> A 224 × 224 RGB image, the standard input size for ImageNet models, contains 224 × 224 × 3 = 150,528 numbers. A recognition model must map this high-dimensional input to a compact, semantically meaningful output such as the word *cat*.

1.2 Why Recognition is Difficult

The same object may look very different across images. A robust recognition system must therefore be *invariant* or *robust* to the following sources of variation:

- **Viewpoint variation:** an object seen from the front and from the side produces different pixel patterns.
- **Scale variation:** objects appear large when near the camera and tiny when far away.
- **Illumination:** shadows, over-exposure and night scenes change pixel intensities drastically.
- **Occlusion:** only part of an object may be visible.
- **Deformation:** non-rigid objects (humans, animals, cells) change shape.
- **Background clutter:** the object may blend into a textured background.
- **Intra-class variation:** two chairs may look nothing alike, yet both are *chair*.

2. Image Classification

Definition (Image Classification)
Image classification is the task of assigning one or more category labels from a predefined set $Y = \{1, 2, ..., K\}$ to an entire image. Formally, a classifier learns a function $f_\theta : \mathbb{R}^{H\times W\times C} \rightarrow Y$ (or a probability distribution over Y) from a labelled training set $\{(I_n, y_n)\}$.

Classification answers the question *"What is in this image?"* but not *"Where is it?"*. It is the oldest and simplest of the three tasks and serves as the foundation for the others: the backbone networks used in detection and segmentation (for example ResNet or ViT) are almost always first trained for classification.

2.1 Types of Image Classification

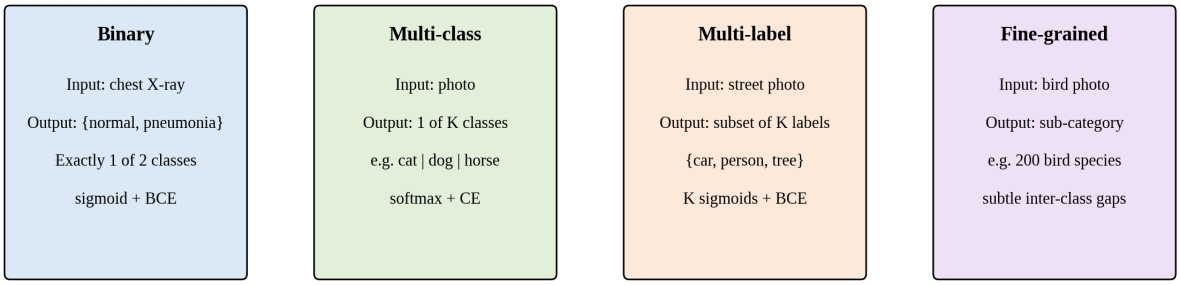


Figure 3. *Four common types of image classification and their typical output layer and loss function.*

Table 1. *Taxonomy of image classification problems.*

Type	Definition	Example
Binary	Exactly one of two classes is assigned.	Chest X-ray: normal vs. pneumonia; e-mail image: spam vs. not spam.
Multi-class	Exactly one of $K > 2$ mutually exclusive classes is assigned.	CIFAR-10: airplane, car, bird, cat, ..., truck. Handwritten digit 0 to 9 (MNIST).
Multi-label	Any subset of the K labels may be assigned; labels are not exclusive.	A street photo tagged {car, person, tree}; satellite tile tagged {cloud, river, farmland}.
Hierarchical	Classes are organised in a tree; prediction follows parent to child.	Animal → Bird → Raptor → Eagle; ICD disease codes.
Ordinal	Classes have a natural order, so some errors are worse than others.	Disease severity: mild, moderate, severe; diabetic retinopathy grade 0 to 4; age group.
Fine-grained	Discriminates visually similar sub-categories of one super-class.	200 bird species (CUB-200), car make and model, plant cultivars, sperm head morphology classes.
Few-shot / Zero-shot	Recognises classes with very few (or no) labelled examples, using prior knowledge or text descriptions.	CLIP labelling an image as "a photo of a zebra" without zebra training images.
Open-set	Must also reject inputs that belong to none of the known classes.	Face recognition rejecting an unknown visitor; detecting an unseen disease.

Note on ordinal classification. In ordinal problems, predicting *severe* for a *mild* case is a larger error than predicting *moderate*. Ordinary cross-entropy ignores this ordering, so ordinal tasks are often trained with ordinal regression losses (for example cumulative-link or CORAL losses) and evaluated with the quadratic weighted kappa or mean absolute error between grades.

Annotation required. Classification needs the cheapest supervision: one class ID per image (binary, multi-class, ordinal), a set of IDs per image (multi-label), or a path in a label tree (hierarchical).

2.2 Classical (Hand-Crafted) Pipeline

Before deep learning (roughly before 2012), classification was a two-step process: first a human-designed **feature extractor** converted the image into a fixed-length vector, then a standard **classifier** such as a Support Vector Machine (SVM), k-Nearest Neighbour (k-NN) or Random Forest predicted the label.

Table 2. *Common hand-crafted features.*

Feature	Idea	Typical use
Colour histogram	Distribution of pixel colours, ignoring spatial layout.	Image retrieval, scene type.
HOG (Histogram of Oriented Gradients)	Histograms of edge directions in small cells.	Pedestrian detection and classification.
SIFT / SURF	Scale and rotation invariant keypoint descriptors.	Object recognition, image matching, Bag-of-Visual-Words.
LBP (Local Binary Patterns)	Encodes each pixel by comparing it with its neighbours.	Texture and face classification.
Gabor / Haar filters	Responses to oriented frequency or rectangular patterns.	Texture analysis; Viola-Jones face detection.

2.3 Deep Learning Pipeline: Convolutional Neural Networks

A Convolutional Neural Network (CNN) learns the features and the classifier jointly, end-to-end, from data. It is composed of the following building blocks:

- **Convolution layer:** slides a set of learnable $k \times k$ filters over the input to produce feature maps. Weight sharing makes it efficient and translation equivariant.
- **Activation (ReLU):** introduces non-linearity, $\mathrm{ReLU}(z) = \max(0, z)$.
- **Pooling layer:** reduces spatial resolution (for example 2×2 max-pooling halves width and height), giving some translation invariance.
- **Normalisation:** Batch Normalisation or Layer Normalisation stabilises training.
- **Fully connected (FC) layer and softmax:** map the final feature vector to class probabilities.

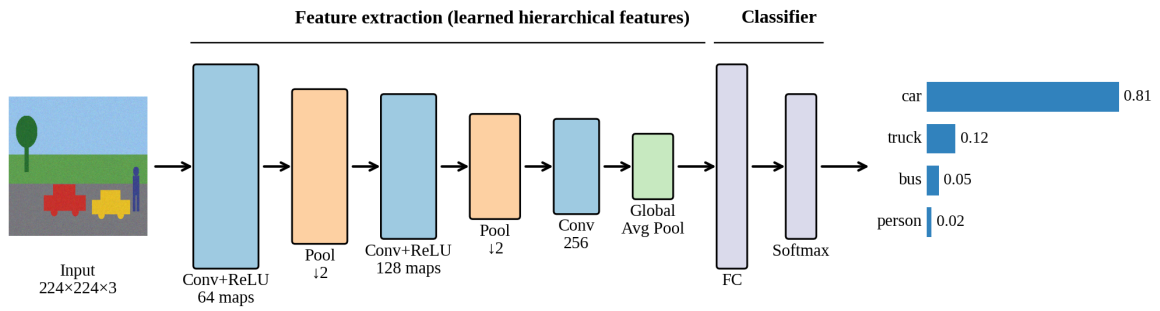


Figure 4. *A typical CNN for image classification. Early layers detect edges and colours, middle layers detect textures and parts, and deep layers detect whole objects. The softmax layer outputs a probability for every class.*

For K classes, the network produces logits $z \in \mathbb{R}^K$, which the softmax converts into probabilities, and training minimises the cross-entropy loss:

$$p_k = \exp(z_k) / \Sigma_j \exp(z_j), \qquad L_{CE} = -\Sigma_k y_k \log p_k$$

For multi-label problems each class gets an independent sigmoid $\sigma(z_k)$ and the binary cross-entropy is summed over classes.

Example
Suppose a 3-class network outputs logits $z = (2.0, 1.0, 0.1)$ for (cat, dog, horse). Softmax gives $p \approx (0.66, 0.24, 0.10)$. If the true class is cat, the loss is $-\log(0.66) \approx 0.42$. If the true class were horse, the loss would be $-\log(0.10) \approx 2.30$, a much larger penalty.

2.4 Landmark Classification Architectures

Table 3. *Milestone architectures for image classification.*

Model (Year)	Key idea	Significance
LeNet-5 (1998)	Conv + pool + FC for digits.	First practical CNN (cheque reading).
AlexNet (2012)	Deep CNN, ReLU, dropout, GPU training.	Won ImageNet 2012 by a large margin; started the deep learning era.
VGG-16/19 (2014)	Stacks of small 3 × 3 convolutions.	Showed depth matters; simple, uniform design.
GoogLeNet / Inception (2014)	Parallel multi-scale filters in one module.	High accuracy with few parameters.
ResNet (2015)	Residual (skip) connections $y = F(x) + x$.	Enabled 100+ layer networks; most common backbone.
DenseNet (2017)	Each layer connected to all later layers.	Feature reuse, parameter efficiency.
MobileNet / EfficientNet (2017, 2019)	Depthwise separable conv; compound scaling.	Accurate models for mobile and edge devices.

Model (Year)	Key idea	Significance
Vision Transformer, ViT (2020)	Image split into 16 × 16 patches treated as tokens; self-attention.	Transformers match or beat CNNs with large data.
Swin Transformer, ConvNeXt (2021, 2022)	Hierarchical windowed attention; modernised CNN.	Strong general-purpose backbones for all tasks.

2.5 Evaluation Metrics for Classification

Let TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives for a given class. The standard metrics are:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{Precision} = TP / (TP + FP), \qquad \text{Recall} = TP / (TP + FN)$$

$$F1 = 2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall})$$

Top-*k* accuracy counts a prediction as correct if the true label appears among the *k* highest-scoring classes (Top-5 is standard on ImageNet). For imbalanced data, accuracy is misleading, so precision, recall, F1, balanced accuracy and the area under the ROC curve (AUC) are preferred. The **confusion matrix** summarises all errors at once.

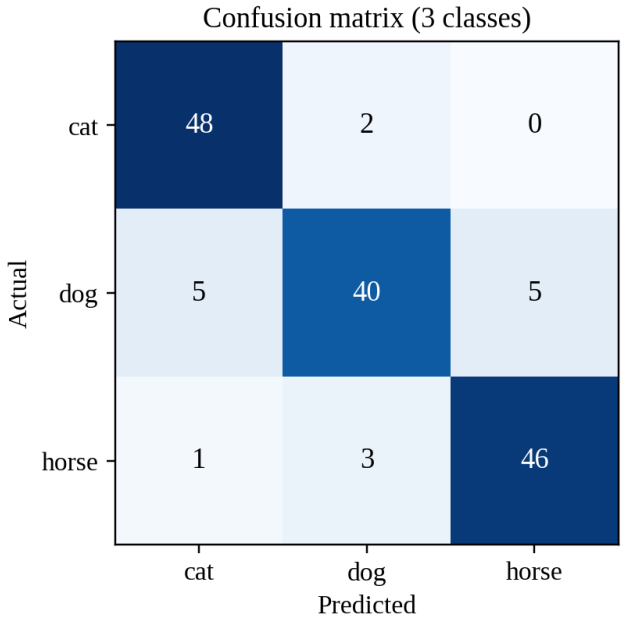


Figure 5. *Confusion matrix for a 3-class problem. Diagonal cells are correct predictions. Here 5 dogs were misclassified as cats and 5 as horses, so dog recall is 40/50 = 0.80.*

Example
From Figure 5: accuracy = (48 + 40 + 46) / 150 = 0.893. Precision for *cat* = 48 / (48 + 5 + 1) = 0.889, recall for *cat* = 48 / 50 = 0.96, so F1(cat) = 2(0.889)(0.96) / (0.889 + 0.96) ≈ 0.923.

Binary metrics for screening and diagnosis

In medical and safety-critical binary problems, the following metrics are reported in addition to accuracy and F1:

$$\text{Sensitivity (Recall)} = TP / (TP + FN), \quad \text{Specificity} = TN / (TN + FP)$$

$$\text{PPV (Precision)} = TP / (TP + FP), \quad \text{NPV} = TN / (TN + FN)$$

- **Sensitivity** is the fraction of diseased cases that are caught; **specificity** is the fraction of healthy cases correctly cleared.
- **NPV** (negative predictive value) tells how trustworthy a negative result is; it matters when a negative result means a patient is sent home.
- **ROC curve** plots sensitivity against (1 − specificity) as the decision threshold varies; **ROC-AUC** is the probability that a random positive is scored higher than a random negative.
- **PR curve and PR-AUC** plot precision against recall. When positives are rare, PR-AUC is more informative than ROC-AUC, because ROC-AUC can stay high even when most positive predictions are false alarms.
- **Threshold choice:** 0.5 is not automatically correct. A screening tool usually lowers the threshold to gain sensitivity; a confirmatory test raises it to gain specificity.

Example

A pneumonia screening model is tested on 200 X-rays: TP = 45, FN = 5, FP = 15, TN = 135. Accuracy = 180/200 = 0.90, sensitivity = 45/50 = 0.90, specificity = 135/150 = 0.90, PPV = 45/60 = 0.75 and NPV = 135/140 = 0.964. Although accuracy looks high, one in four positive alarms is false, which is only visible through PPV.

2.6 Applications of Image Classification

- **Medical diagnosis:** diabetic retinopathy grading from fundus images, skin lesion (melanoma) classification, pneumonia and tuberculosis detection from chest X-rays, histopathology tumour grading.
- **Agriculture:** crop and plant disease identification from leaf images (for example rice blast, tomato leaf curl), fruit quality grading.
- **Industry:** defective vs. non-defective product inspection on assembly lines.
- **Remote sensing:** land-use and land-cover classification of satellite tiles.
- **Biometrics and security:** face verification, fingerprint and iris recognition.
- **Content moderation and search:** tagging photos, filtering unsafe content, visual search in e-commerce.
- **Biology:** cell type and sperm morphology classification from microscopy images.

3. Object Localization, Object Detection and Mass Detection

Definition (Object Localization)
Object localization is the task of predicting the class of the *single* dominant object in an image together with a bounding box that encloses it. It is classification plus regression of four box coordinates.

Definition (Object Detection)
Object detection is the task of finding *all* instances of objects from a set of known classes in an image. For each instance the detector outputs a triplet (class label *c*, confidence score *s*, bounding box *b*). The number of objects is not known in advance and can be zero.

Detection answers both *"What?"* and *"Where?"*, and also *"How many?"*. It is harder than classification because the model must handle a variable number of outputs, objects of very different sizes, and overlapping objects.

3.1 Bounding Box Representations

Table 4. *Bounding box formats.*

Format	Parameters	Used by
Corner (xyxy)	$(x_{min}, y_{min}, x_{max}, y_{max})$	PASCAL VOC, torchvision
Corner + size (xywh)	$(x_{min}, y_{min}$, width, height)	MS COCO annotations
Centre + size (cxcywh)	(c_x, c_y, w, h), often normalised to [0, 1]	YOLO label files, DETR
Oriented box	(c_x, c_y, w, h, θ)	Aerial imagery (DOTA), text detection

Example
A car occupying pixels from (90, 150) to (200, 205) in a 360 × 240 image has xyxy = (90, 150, 200, 205), xywh = (90, 150, 110, 55), and normalised YOLO format cxcywh = (145/360, 177.5/240, 110/360, 55/240) = (0.403, 0.740, 0.306, 0.229).

3.2 Intersection over Union (IoU)

IoU measures how well a predicted box B_p matches a ground-truth box B_g. It is used in training (to assign positive and negative samples), in post-processing (NMS) and in evaluation (deciding whether a detection is correct).

$$\text{IoU}(B_p, B_g) = \text{Area}(B_p \cap B_g) / \text{Area}(B_p \cup B_g)$$

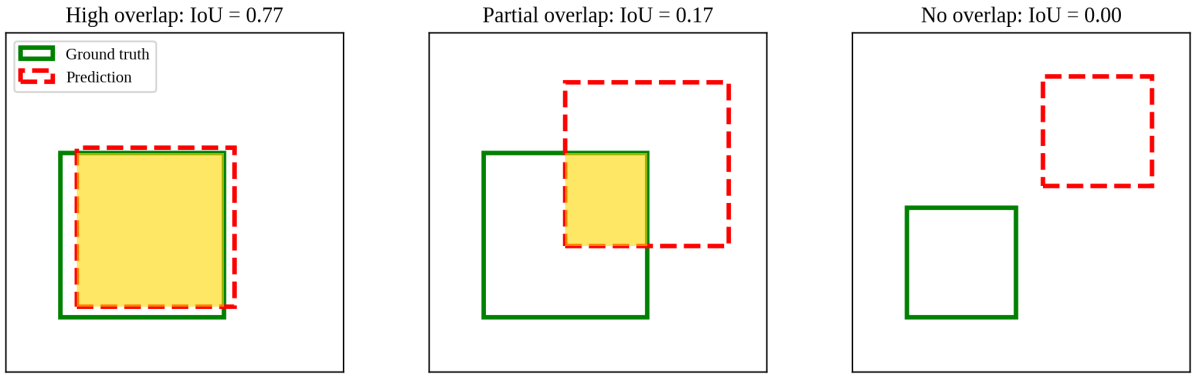


Figure 6. *IoU for three cases. The yellow region is the intersection. A detection is usually counted as a true positive when IoU ≥ 0.5.*

Example
Ground truth (1, 1, 4, 4) and prediction (2, 2, 5, 5). Intersection = (2, 2, 4, 4), area 4. Each box has area 9, so union = 9 + 9 − 4 = 14 and IoU = 4/14 ≈ 0.29, which would be a false positive at the 0.5 threshold.

3.3 Anchor Boxes and Non-Maximum Suppression

Anchor boxes are a fixed set of reference boxes of different scales and aspect ratios (for example 1:1, 1:2, 2:1) placed at every location of a feature map. Instead of predicting coordinates from scratch, the network predicts small offsets from each anchor plus a class score. Anchors make it easy to detect objects of many shapes.

Because neighbouring anchors fire on the same object, a detector produces many duplicate boxes. **Non-Maximum Suppression (NMS)** removes them:

- Sort all boxes of a class by confidence score.
- Select the highest-scoring box and keep it.
- Remove every remaining box whose IoU with the kept box exceeds a threshold (typically 0.45 to 0.5).
- Repeat with the remaining boxes until none are left.

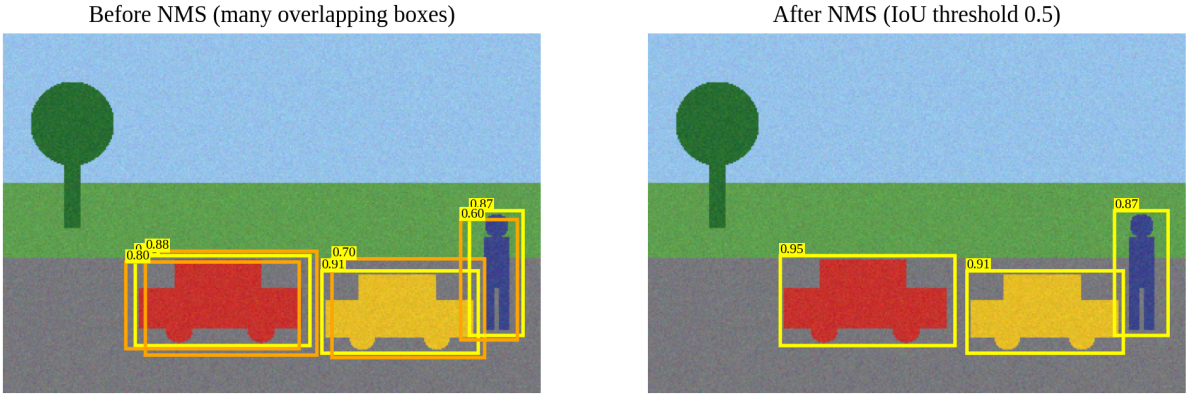


Figure 7. *Non-Maximum Suppression. Left: raw detector output with several overlapping candidates per object. Right: after NMS only the highest-scoring box for each object remains.*

3.4 Taxonomy of Object Detectors

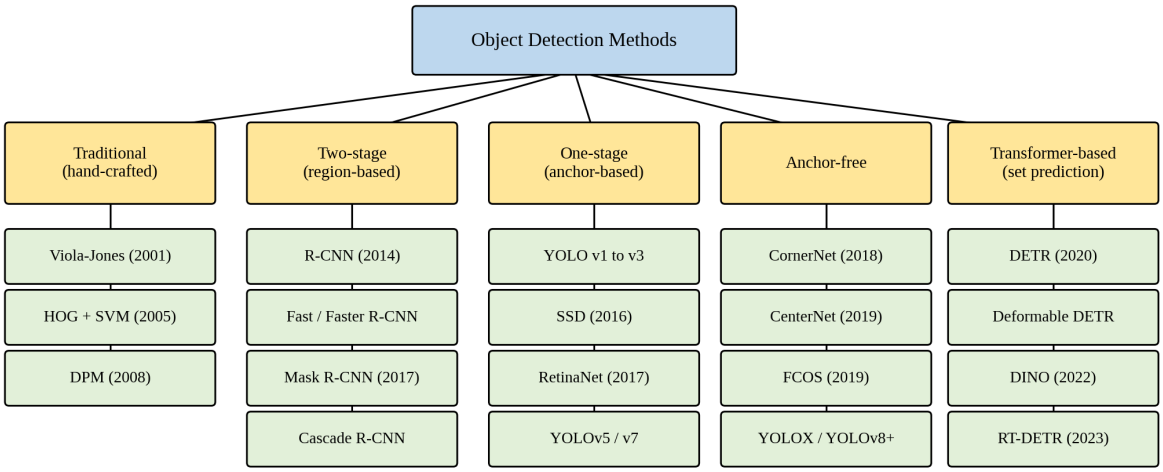


Figure 8. *Taxonomy of object detection methods with representative models.*

(a) Traditional detectors

Use a sliding window over an image pyramid and classify each window with hand-crafted features. **Viola-Jones** (2001) used Haar features with a boosted cascade and achieved real-time face detection. **HOG + linear SVM** (Dalal and Triggs, 2005) became the standard pedestrian detector. **Deformable Part Models** (DPM, 2008) represented objects as flexible arrangements of parts.

(b) Two-stage detectors

Stage 1 generates a sparse set of candidate *region proposals* likely to contain objects; stage 2 classifies each proposal and refines its box. **R-CNN** ran a CNN on ~2000 Selective Search proposals (slow). **Fast R-CNN** shared one CNN pass and introduced RoI pooling. **Faster R-CNN** replaced Selective Search with a learnable **Region Proposal Network (RPN)**. **Mask R-CNN** adds a mask branch for instance segmentation. Two-stage methods are generally more accurate, especially for small objects, but slower.

(c) One-stage detectors

Predict class scores and boxes directly from dense feature-map locations in a single forward pass. **YOLO** (You Only Look Once) divides the image into an $S \times S$ grid; the cell containing an object's centre is responsible for predicting it. **SSD** predicts from multiple feature maps for different scales. **RetinaNet** introduced the **focal loss** to handle the extreme imbalance between background and object locations. One-stage detectors are fast and dominate real-time applications.

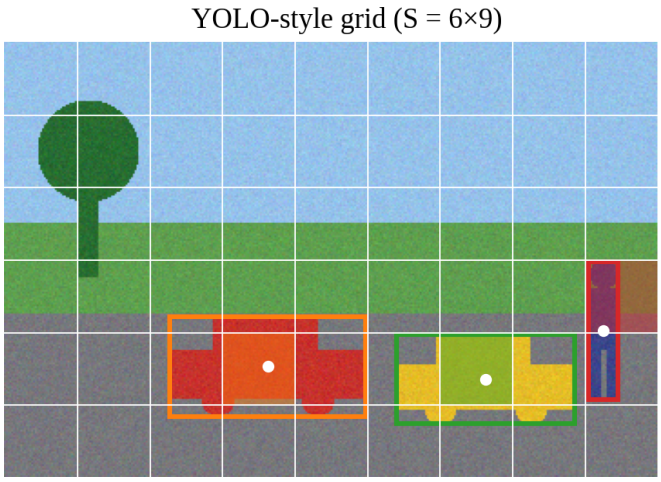


Figure 9. *YOLO-style grid. The cell (shaded) that contains an object's centre (white dot) is responsible for predicting that object's box and class.*

(d) Anchor-free detectors

Remove hand-designed anchors. **CornerNet** predicts box corners as keypoints, **CenterNet** predicts object centres plus width and height, and **FCOS** predicts distances from each pixel to the four box sides. Modern YOLO versions (YOLOX, YOLOv8 and later) are anchor-free.

(e) Transformer-based detectors

DETR (DEtection TRansformer) treats detection as direct *set prediction*: a fixed number of learned object queries attend to image features, and a Hungarian matching loss assigns predictions to ground truth one-to-one. This removes both anchors and NMS. **Deformable DETR**, **DINO** and **RT-DETR** improve convergence speed and achieve real-time performance.

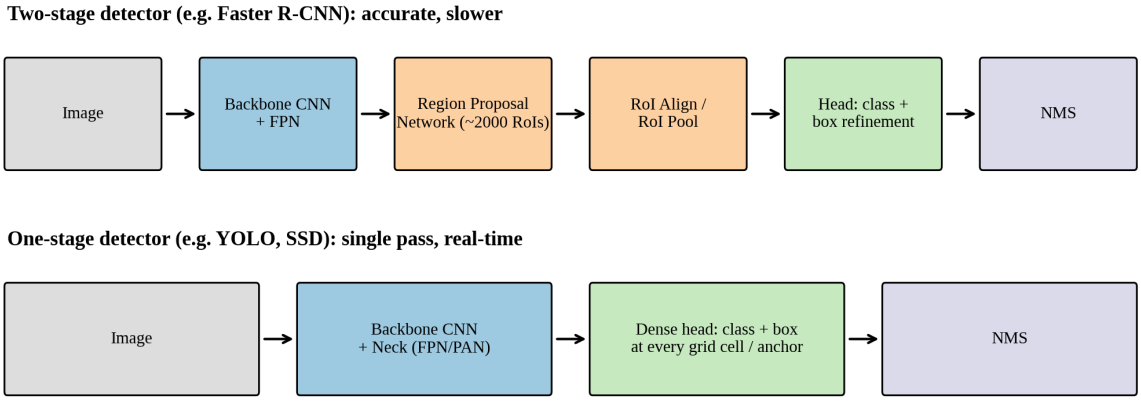


Figure 10. *Comparison of the processing pipeline of two-stage and one-stage detectors.*

Table 5. *Two-stage versus one-stage detectors.*

Aspect	**Two-stage**	**One-stage**
Pipeline	Proposals, then classify and refine	Dense prediction in one pass
Speed	Slower (about 5 to 20 FPS)	Fast (30 to 150+ FPS)

Aspect	Two-stage	One-stage
Accuracy	Traditionally higher, good for small objects	Now comparable with modern designs
Examples	Faster R-CNN, Mask R-CNN, Cascade R-CNN	YOLO family, SSD, RetinaNet, FCOS
Best suited for	Offline analysis, medical images, high precision	Autonomous driving, drones, video, edge devices

3.5 Loss Functions for Detection

A detector minimises a weighted sum of a classification loss and a localisation (box regression) loss:

$$L = L_{\text{cls}} + \lambda L_{\text{box}}$$

- **Classification:** cross-entropy, or focal loss $\text{FL}(p_t) = -\alpha(1 - p_t)^{\gamma} \log p_t$, which down-weights easy background examples.
- **Box regression:** Smooth L1 on box offsets, or IoU-based losses such as IoU, GIoU, DIoU and CIoU loss (for example $L_{\text{IoU}} = 1 - \text{IoU}$), which directly optimise overlap.

3.6 Evaluation: Precision-Recall, AP and mAP

Detections are sorted by confidence. Each is marked TP if it matches an unmatched ground-truth box of the same class with IoU ≥ threshold, otherwise FP. Sweeping the confidence threshold gives a precision-recall curve. **Average Precision (AP)** is the area under the (interpolated) precision-recall curve for one class, and **mean Average Precision (mAP)** is the mean of AP over all classes.

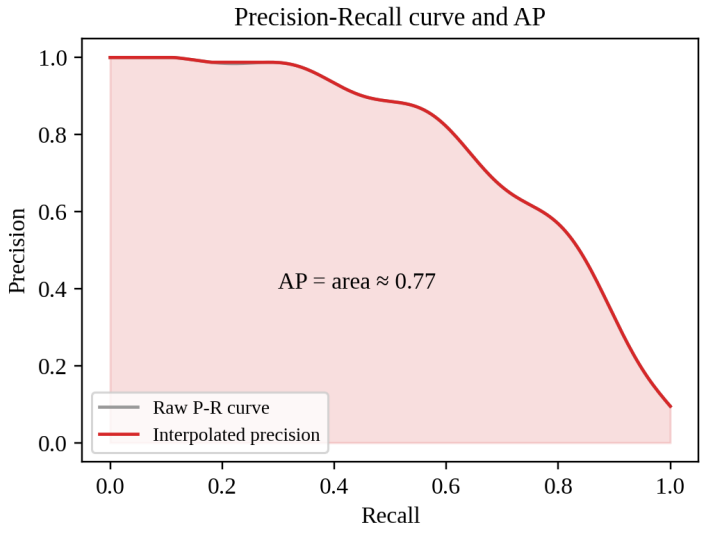


Figure 11. *Precision-recall curve. AP is the area under the interpolated (monotonically non-increasing) curve.*

- **mAP@0.5** (PASCAL VOC): IoU threshold 0.5.
- **mAP@[0.5:0.95]** (MS COCO primary metric): mAP averaged over IoU thresholds 0.50, 0.55, ..., 0.95, rewarding precise localisation.

- **AP_S, AP_M, AP_L**: AP for small, medium and large objects.
- **FPS / latency**: inference speed, essential for real-time systems.

Example
A detector finds 8 boxes for 5 ground-truth persons. Sorted by score the outcomes are TP, TP, FP, TP, TP, FP, FP, TP. After 5 detections precision = 4/5 = 0.80 and recall = 4/5 = 0.80; after all 8, precision = 5/8 = 0.625 and recall = 5/5 = 1.0. Plotting all such pairs and integrating gives the AP for class *person*.

3.7 Other Types of Detection

Table 6. *Specialised forms of object detection.*

Type	Description	Example
Face / pedestrian detection	Specialised single-class detectors.	Camera autofocus, crowd counting.
Oriented (rotated) detection	Boxes with an angle θ.	Ships and planes in aerial images.
3D object detection	3D boxes from LiDAR or stereo cameras.	Cars and cyclists for self-driving.
Video object detection	Uses temporal context across frames.	Traffic surveillance, sports analytics.
Small / tiny object detection	Objects of a few pixels.	Drone imagery, cells and sperm in microscopy.
Open-vocabulary detection	Detects classes described by free text.	Grounding DINO, OWL-ViT: "find the red helmet".
Salient object / anomaly detection	Finds the most noticeable or abnormal region.	Defect localisation in manufacturing.

3.8 Mass and Lesion Detection in Medical Imaging

Definition (Mass Detection)
In medical imaging, a *mass* is a space-occupying abnormality, such as a tumour or cyst, that appears as a localised region of different density or intensity. *Mass (or lesion) detection* is the application of object detection to locate every suspicious mass in a mammogram, CT, MRI or ultrasound image, outputting a bounding box (or centre point) and a suspicion score for each finding.

The same clinical image can be posed as three different tasks, and each answers a different clinical question (Figure 12). Classification tells *whether* a suspicious finding exists, detection tells *where* it is, and segmentation gives its *exact boundary*, from which size, area and volume are measured for staging and follow-up.

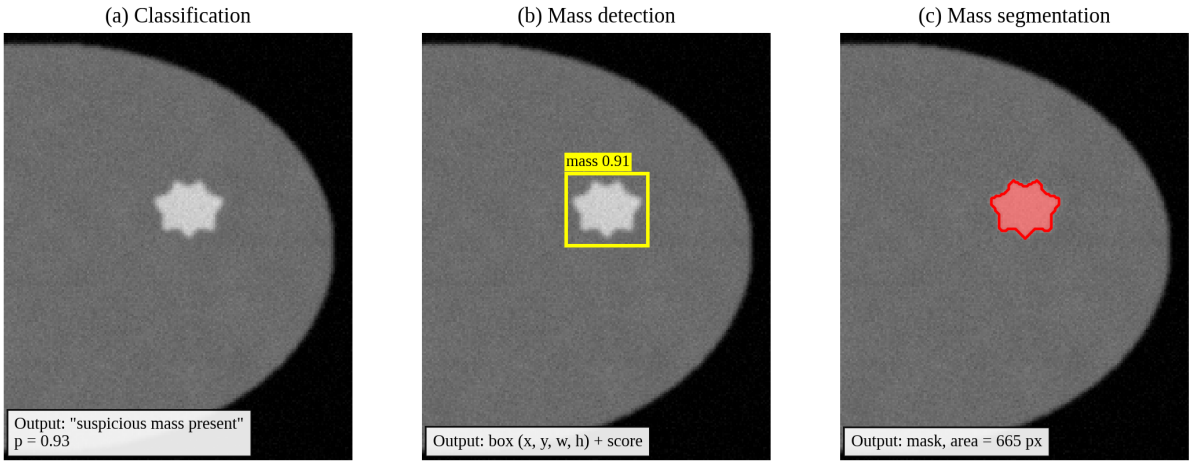


Figure 12. *The same synthetic mammogram-like image as (a) image-level classification, (b) mass detection with a bounding box and score, and (c) mass segmentation with a pixel mask from which area is computed.*

Mass detection differs from detection in natural images in several ways:

- **Small targets in large images:** a 1 cm mass may occupy less than 0.1% of a full-field mammogram (often over 3000 × 4000 pixels), so images are processed at high resolution or in tiles (patches).
- **Low contrast and ambiguous boundaries:** masses may be hidden by dense tissue; spiculated (star-shaped) margins are clinically important and must not be lost by downsampling.
- **Severe imbalance:** most images and almost all pixels are normal, so focal loss and hard-negative mining are common.
- **Multiple views and time points:** mammography uses two views per breast (CC and MLO), and radiologists compare with prior scans; models may fuse views or register prior images.
- **Evaluation by FROC:** instead of mAP, medical detection is often evaluated with the **Free-Response ROC (FROC)** curve, which plots lesion-level sensitivity against the average number of false positives per image (for example, sensitivity at 0.5, 1 and 2 FP per image). A detection is often counted as correct when its centre falls inside the true lesion rather than by IoU ≥ 0.5.

Example
A lung CT screening system reports a sensitivity of 0.92 at 1 false positive per scan. This means it finds 92% of true nodules while showing, on average, one false alarm per patient for the radiologist to dismiss. This is far more clinically meaningful than an mAP value.

3.9 Applications of Object Detection

- **Autonomous driving and ADAS:** detecting vehicles, pedestrians, cyclists, traffic signs and lights.
- **Surveillance and security:** intrusion detection, abandoned baggage, weapon detection, people counting.

- **Traffic management:** vehicle counting, number-plate localisation, helmet-violation detection for motorbikes.
- **Retail:** shelf monitoring, cashier-less checkout, product counting.
- **Medical imaging:** lung nodule, polyp, fracture and lesion localisation.
- **Agriculture:** fruit counting for yield estimation, weed and pest detection, livestock monitoring.
- **Robotics and drones:** grasping, obstacle avoidance, search and rescue.
- **Document analysis:** detecting tables, figures and text lines in scanned pages.

4. Image Segmentation

Definition (Image Segmentation)
Image segmentation is the task of partitioning an image into regions by assigning a label to *every pixel*. Formally, it learns a mapping $f_\theta : \mathbb{R}^{H\times W\times C} \to Y^{H\times W}$, producing a label map of the same spatial size as the input.

Segmentation gives the most precise description of a scene: exact object shape, boundaries and area, which bounding boxes cannot provide. This is essential when shape matters, such as measuring a tumour's volume or finding the drivable road surface.

4.1 Types of Image Segmentation

To understand the types, distinguish two kinds of categories. **Things** are countable objects with well-defined shapes (person, car, cell). **Stuff** refers to amorphous background regions (sky, road, grass, water).

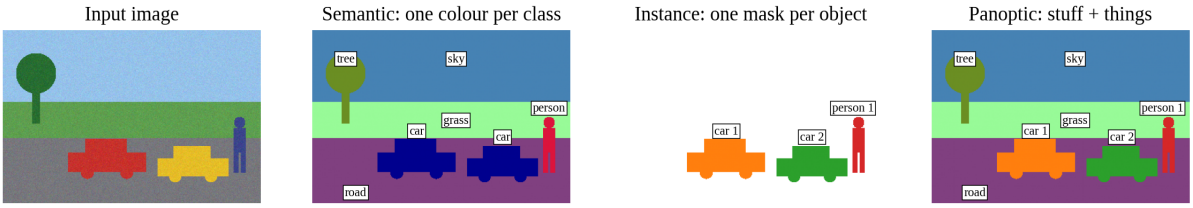


Figure 13. *Semantic, instance and panoptic segmentation of the same scene. Semantic segmentation gives both cars the same colour; instance segmentation separates them but ignores background; panoptic segmentation labels every pixel and separates every object.*

Definition (Semantic Segmentation)
Assigns a class label to every pixel. All pixels of the same class receive the same label, so individual objects of the same class are *not* distinguished.

Definition (Instance Segmentation)
Detects each object instance (things only) and outputs a separate pixel-level mask for it, so two cars receive two different masks. It can be seen as object detection with masks instead of boxes.

Definition (Panoptic Segmentation)
Unifies the two: every pixel receives a class label, and every pixel belonging to a thing also receives an instance ID. Stuff classes are labelled semantically; thing classes are labelled per instance.

Table 7. *Comparison of the three main segmentation types.*

Type	Labels every pixel?	Separates instances?	Typical model
Semantic	Yes	No	FCN, U-Net, DeepLab, SegFormer

Type	Labels every pixel?	Separates instances?	Typical model
Instance	No (things only)	Yes	Mask R-CNN, YOLACT, SOLO
Panoptic	Yes	Yes (things)	Panoptic FPN, Mask2Former

Other important variants include:

- **Binary (foreground-background) segmentation:** two classes only, for example lesion vs. healthy skin.
- **Medical and 3D volumetric segmentation:** voxel-wise labelling of CT or MRI volumes (organs, tumours) using 3D U-Net or nnU-Net.
- **Interactive / promptable segmentation:** the user provides clicks, boxes or scribbles; the Segment Anything Model (SAM) segments any object from such prompts.
- **Referring segmentation:** segment the object described by a sentence, such as "the man in the blue shirt".
- **Video object segmentation:** track and segment objects consistently across frames.
- **Part segmentation:** segment parts of an object (head, arms, legs of a person).

4.2 Classical Segmentation Techniques

Table 8. *Classical (non-learning) segmentation methods.*

Method	Principle	Strength / Weakness
Thresholding (global, Otsu, adaptive)	Pixels above threshold T are foreground. Otsu chooses T to maximise between-class variance.	Very fast; fails with uneven lighting or overlapping intensities.
Edge-based	Detect boundaries with Sobel, Canny or Laplacian operators, then link edges.	Sharp boundaries; edges are often broken.
Region growing / splitting and merging	Start from seed pixels and add neighbours with similar intensity.	Connected regions; sensitive to seed choice.
Clustering (k-means, mean-shift)	Group pixels by colour and position in feature space.	Unsupervised; no semantic meaning.
Watershed	Treat gradient image as a landscape and flood from markers.	Separates touching objects (cells); prone to over-segmentation.
Graph-based (Graph Cut, GrabCut, superpixels)	Minimise an energy over a pixel graph.	Good boundaries; computationally heavier.

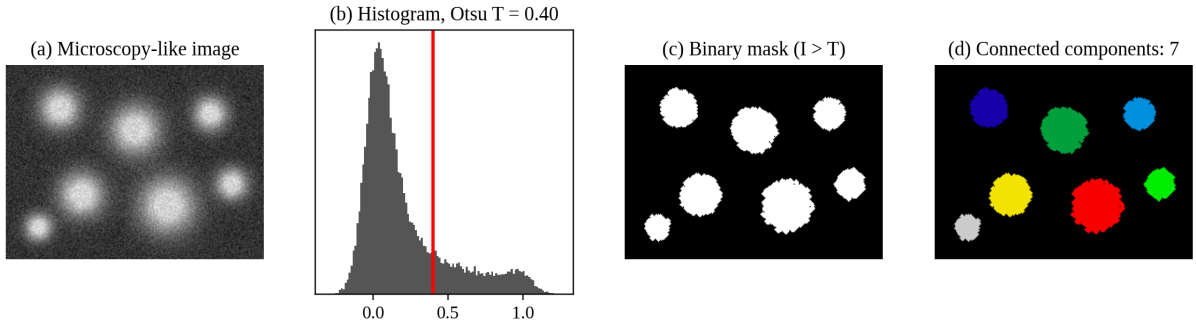


Figure 14. *Classical segmentation pipeline on a microscopy-like image: Otsu thresholding of the intensity histogram, followed by morphological opening and connected-component labelling, which counts the objects.*

Example

Otsu's method examines every candidate threshold T, splits pixels into two groups with weights w_0, w_1 and means μ_0, μ_1, and selects the T that maximises $\sigma_B^2(T) = w_0\, w_1\, (\mu_0 - \mu_1)^2$. In Figure 14 the optimum is $T \approx 0.40$, which separates the bright cells from the dark background.

4.3 Deep Learning for Segmentation

Deep segmentation networks follow an **encoder-decoder** design. The encoder (often a classification backbone) progressively reduces resolution while extracting semantic features; the decoder upsamples the features back to full resolution to produce a per-pixel prediction.

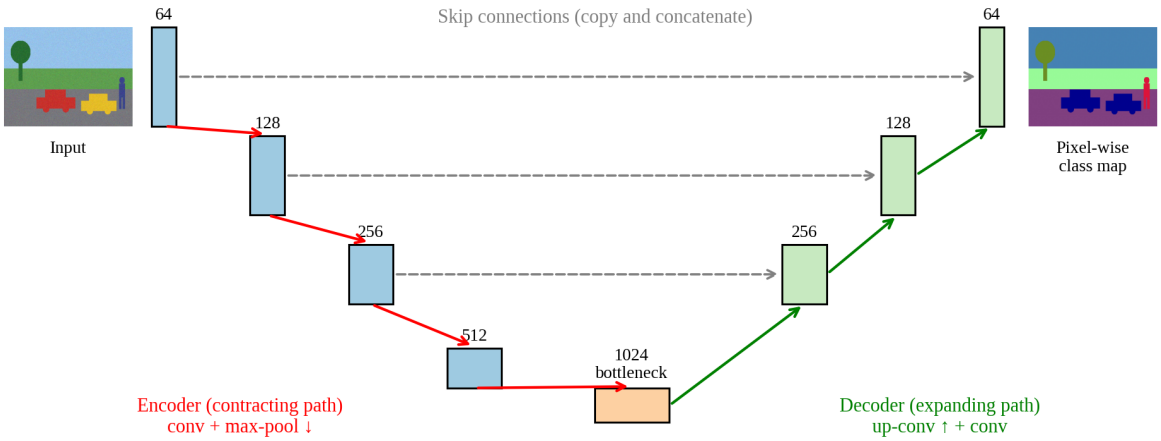


Figure 15. *Simplified U-Net architecture. The encoder captures context, the decoder recovers spatial detail, and skip connections pass fine-grained features from encoder to decoder at the same resolution.*

Table 9. *Milestone deep segmentation models.*

Model (Year)	Key idea
FCN (2015)	Replaced FC layers with convolutions so a classifier outputs a dense map; first end-to-end semantic segmentation.

Model (Year)	**Key idea**
U-Net (2015)	Symmetric encoder-decoder with skip connections; works with few training images; standard in biomedical imaging.
SegNet (2015)	Decoder reuses max-pooling indices from the encoder for upsampling.
DeepLab v1 to v3+ (2015 to 2018)	Atrous (dilated) convolution and Atrous Spatial Pyramid Pooling (ASPP) for multi-scale context.
PSPNet (2017)	Pyramid pooling module for global scene context.
Mask R-CNN (2017)	Faster R-CNN plus a small FCN mask head per RoI; standard for instance segmentation.
YOLACT, SOLO, YOLOv8-seg (2019 onward)	Real-time one-stage instance segmentation.
SegFormer (2021)	Hierarchical transformer encoder with a lightweight MLP decoder.
Mask2Former (2022)	One masked-attention transformer for semantic, instance and panoptic segmentation.
SAM / SAM 2 (2023, 2024)	Foundation model trained on over one billion masks; segments anything from prompts, including in video.

4.4 Loss Functions for Segmentation

- **Pixel-wise cross-entropy:** the classification loss averaged over all pixels.
- **Weighted cross-entropy:** larger weights for rare classes to address imbalance.
- **Dice loss:** $L_{\text{Dice}} = 1 - 2|P \cap G| / (|P| + |G|)$; directly optimises overlap and is robust when the foreground is small (common in medical images).
- **Focal and Tversky losses:** emphasise hard pixels or trade off false positives against false negatives.
- **Combined losses:** cross-entropy + Dice is a widely used default.

4.5 Evaluation Metrics for Segmentation

$$\text{Pixel Accuracy} = \text{correctly labelled pixels} / \text{total pixels}$$

$$\text{IoU (Jaccard)} = |P \cap G| / |P \cup G| = \text{TP} / (\text{TP} + \text{FP} + \text{FN}), \qquad \text{mIoU} = (1/K)\, \Sigma_k\, \text{IoU}_k$$

$$\text{Dice (F1)} = 2|P \cap G| / (|P| + |G|) = 2\text{TP} / (2\text{TP} + \text{FP} + \text{FN})$$

$$\text{PQ} = \Sigma_{(p,g)\in \text{TP}}\, \text{IoU}(p, g) / (\, |\text{TP}| + \tfrac{1}{2}|\text{FP}| + \tfrac{1}{2}|\text{FN}|\,)$$

Here P is the predicted mask and G the ground-truth mask. Mean IoU (mIoU) is the standard metric for semantic segmentation, mask AP for instance segmentation, Dice for medical segmentation, and Panoptic Quality (PQ) for panoptic segmentation. Pixel accuracy can be misleading when the background dominates.

Example
A ground-truth tumour mask has 100 pixels. The prediction has 90 pixels, of which 80 overlap the tumour. Then TP = 80, FP = 10, FN = 20. IoU = 80 / (80 + 10 + 20) = 0.727 and Dice = 160 / (160 + 10 + 20) = 0.842. Note that Dice is always at least as large as IoU, with Dice = 2·IoU / (1 + IoU).

Boundary-based metrics

Overlap metrics such as Dice and IoU are dominated by the interior of large objects and may barely change when a thin structure (vessel, membrane, crack) is broken or a boundary is shifted. **Boundary metrics** measure contour quality directly:

- **Hausdorff distance (HD):** the largest distance from any point on one boundary to the nearest point on the other boundary; it captures the worst-case error.
- **HD95:** the 95th percentile of those distances, which is robust to a few outlier pixels and is widely reported in medical segmentation.
- **Average Surface Distance (ASD) and Boundary F1 / NSD:** the mean boundary distance, or the fraction of boundary points lying within a tolerance τ of the true boundary.

4.6 Applications of Image Segmentation

- **Medical imaging:** organ, tumour and lesion segmentation in CT, MRI and ultrasound; cell and nucleus segmentation in microscopy; retinal vessel segmentation; radiotherapy planning.
- **Autonomous driving:** drivable area, lane and road-scene parsing (Cityscapes classes).
- **Remote sensing:** building footprints, flood mapping, deforestation and crop field boundaries.
- **Agriculture:** leaf disease area estimation, fruit and weed segmentation.
- **Photo and video editing:** background removal, portrait mode, virtual backgrounds in video calls.
- **Augmented reality and robotics:** scene understanding, object grasping, human body segmentation.
- **Industrial inspection:** pixel-level crack, scratch and corrosion detection.

5. Comparison, Related Tasks and Applications

5.1 Summary Comparison of the Three Tasks

Table 10. *Side-by-side comparison of the core recognition tasks.*

Aspect	Classification	Detection	Segmentation
Question answered	What is in the image?	What and where (box)?	What is at each pixel?
Output	Label(s) + probabilities	List of (class, score, box)	Label map $H \times W$ (masks)
Annotation	Image-level label (seconds)	Boxes (tens of seconds)	Pixel masks (minutes per image)
Typical models	ResNet, EfficientNet, ViT	Faster R-CNN, YOLO, DETR	U-Net, DeepLab, Mask R-CNN, SAM
Main loss	Cross-entropy	Cls. loss + box regression	Pixel CE + Dice
Main metric	Accuracy, F1, AUC	mAP@[0.5:0.95]	mIoU, Dice, PQ
Example	Is this X-ray normal?	Where are the cars?	Exact shape of the tumour

5.2 Related Vision Tasks

Table 11. *Vision tasks built on top of classification, detection and segmentation.*

Task	Definition
Keypoint / pose estimation	Localises predefined points, such as human joints or facial landmarks (OpenPose, HRNet).
Object tracking	Detects objects and associates them across video frames with consistent IDs (SORT, DeepSORT, ByteTrack).
Action recognition	Classifies the activity in a video clip (walking, falling, fighting).
Depth estimation	Predicts the distance of every pixel from the camera.
Image captioning and VQA	Generates a sentence describing an image, or answers questions about it.
Optical character recognition (OCR)	Detects and reads text in images.
Image retrieval	Finds images similar to a query image in a large database.
Anomaly detection	Learns what *normal* looks like and flags deviations, useful when defect classes are rare or unknown (PatchCore, autoencoders).
Optical flow	Estimates the motion vector of every pixel between two consecutive frames (RAFT).
Image registration	Spatially aligns two images, for example a patient's follow-up scan to the earlier scan, so changes can be measured.
Super-resolution	Reconstructs a high-resolution image from a low-resolution input (SRCNN, ESRGAN).
Denoising and restoration	Removes noise, blur or artefacts, for example low-light photos or low-dose CT.

Task	Definition
Generative vision	Synthesises or edits images (GANs, diffusion models); also used for data augmentation.

5.3 Applications by Domain

Table 12. *Example applications of each task across domains.*

Domain	Classification	Detection	Segmentation
Healthcare	Disease vs. normal X-ray; skin cancer type	Lung nodules, polyps, fractures	Tumour boundary, organ volume, cell nuclei
Autonomous driving	Weather / scene type	Cars, pedestrians, signs	Drivable road, lanes
Agriculture	Leaf disease type	Fruit counting, pest detection	Diseased leaf area
Remote sensing	Land-use of a tile	Ships, aircraft, vehicles	Buildings, water bodies, floods
Security	Face verification	Weapons, intruders, helmets	Crowd density regions
Manufacturing	Defective / OK part	Missing components	Crack and scratch shape
Retail	Product category	Items on shelves	Virtual try-on
Biology	Sperm / cell morphology class	Cell and sperm localisation for motility tracking	Cell membrane, nucleus, tissue regions

5.4 Benchmark Datasets

Table 13. *Widely used benchmark datasets.*

Dataset	Task(s)	Size and content
MNIST / Fashion-MNIST	Classification	70,000 28 × 28 grayscale images, 10 classes
CIFAR-10 / CIFAR-100	Classification	60,000 32 × 32 colour images, 10 or 100 classes
ImageNet (ILSVRC)	Classification, detection	About 1.28 million training images, 1000 classes
PASCAL VOC	Classification, detection, segmentation	About 11,000 images, 20 classes
MS COCO	Detection, instance and panoptic segmentation, captioning	About 330,000 images, 80 thing classes, 1.5 million instances
Cityscapes	Semantic / instance / panoptic segmentation	5,000 finely annotated urban driving images, 30 classes
ADE20K	Semantic segmentation	About 25,000 scene images, 150 classes
Open Images, LVIS, SA-1B	Large-scale detection and segmentation	Hundreds to over 1,000 classes; SA-1B has 1.1 billion masks

5.5 Practical Workflow for a Vision Project

- **Define the task** precisely: does the application need a label, a box or a mask? Choose the simplest task that solves the problem.
- **Collect and annotate data** using tools such as LabelImg, CVAT, Labelme or Roboflow; keep train, validation and test sets independent (for example split by patient or by video, not by frame, to avoid data leakage).
- **Pre-process and augment:** resizing, normalisation, flips, rotations, colour jitter, mosaic and mix-up.
- **Start from a pretrained model** (transfer learning) and fine-tune.
- **Evaluate** with the correct metric, then perform **error analysis**: which classes, object sizes, devices or sites fail?
- **Tune the decision threshold and check calibration** so that a score of 0.9 really means about 90% confidence.
- **Deploy and optimise** with quantisation, pruning, ONNX or TensorRT for real-time or edge hardware.
- **Monitor in production** for data drift, new failure modes and changing false positive / false negative rates, and feed errors back into the dataset.

6. Annotation and Dataset Design

The more detailed the required output, the more expensive the annotation. In practice, a well-designed, consistently labelled and leakage-free dataset often matters more than the choice of architecture.

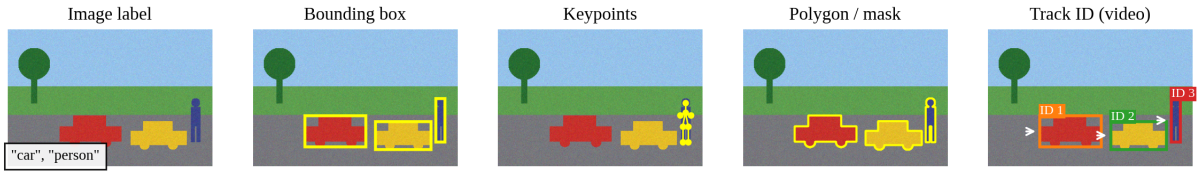


Figure 16. *Annotation types for the same scene, in increasing order of effort: image-level labels, bounding boxes, keypoints, polygon masks, and boxes with persistent track IDs across video frames.*

6.1 Annotation Formats

Each task needs its own kind of label. The common formats are:

- **Classification:** one class ID per image, or a set of labels per image for multi-label tasks (stored as folder-per-class or a CSV file).
- **Detection:** one bounding box and class ID per object, with an extra angle θ for rotated boxes (PASCAL VOC XML, COCO JSON or YOLO TXT).
- **Segmentation:** a binary or multi-class pixel mask, or a polygon outlining each region or instance (PNG masks, COCO polygons or RLE).
- **Pose / keypoints:** (x, y) coordinates of each landmark plus a visibility flag (COCO keypoints).
- **Tracking:** a box or mask for each object in every frame plus a track ID that stays the same across frames (MOT format).
- **OCR:** a box or polygon around each text region plus its transcription (ICDAR format).
- **3D / volumetric:** voxel-wise masks for CT or MRI volumes (NIfTI, DICOM-SEG).

6.2 Annotation Quality

- Write an **annotation guideline** with visual examples: how tight a box should be, how to handle occluded or truncated objects, what to do with ambiguous cases.
- Measure **inter-annotator agreement** (Cohen's kappa for labels, IoU or Dice between annotators for boxes and masks). If two experts agree only at Dice 0.80, a model Dice of 0.95 is not meaningful.
- For medical data, use expert annotators and, where possible, a consensus or pathology-confirmed reference standard.

6.3 Dataset Splitting and Data Leakage

Data is split into training, validation (for model and threshold selection) and test (for the final, untouched estimate) sets. The split must be done at the level of the **independent unit**, not the individual image.

Data leakage warning
If images of the same patient, the same video, the same product or overlapping satellite tiles appear in both training and test sets, the model can memorise the subject instead of learning the task, and the reported performance becomes unrealistically high. Use **group-aware splits** (by patient, video, subject, site or geographic region), and ideally validate on an **external** dataset from a different hospital, camera or location.

Example
A video dataset of 50 traffic clips yields 100,000 frames. A random frame-level split puts nearly identical consecutive frames in training and test sets and might report 98% mAP. Splitting by clip (40 clips for training, 10 for testing) gives an honest estimate that may be much lower.

7. Choosing a Task and a Model Family

Searching for *the best model* in general is not productive. The choice should follow the output the application needs, the size of the dataset, the annotation budget, and the latency and memory available on the deployment hardware.

Table 14. *Quick task and model selection guide.*

Question to be answered	Recommended task	Typical model family
Only the image category is needed	Classification	ResNet, EfficientNet, ViT
Where are the objects, and how many?	Object detection	YOLO, Faster R-CNN, RetinaNet, DETR
Exact boundary of a region or class	Semantic segmentation	U-Net, DeepLab, SegFormer
Separate masks for objects of the same class	Instance segmentation	Mask R-CNN, YOLO-seg
Whole scene plus individual objects	Panoptic segmentation	Panoptic FPN, Mask2Former
Human joints or anatomical landmarks	Pose / keypoint estimation	HRNet, OpenPose, YOLO-pose
Follow objects through a video	Tracking (detector + tracker)	ByteTrack, DeepSORT
Read text in the image	OCR (text detection + recognition)	DBNet + CRNN, TrOCR
Unknown or rare defect types	Anomaly detection	PatchCore, autoencoders
Compare two scans of the same patient	Image registration	VoxelMorph, classical B-spline

7.1 Practical Trade-offs

- **Accuracy vs. latency:** a real-time system on a drone or phone may need a smaller, faster model at some cost in accuracy.
- **Annotation cost:** masks are most informative but most expensive; if a box answers the question, do not pay for masks.
- **Small objects:** require higher input resolution, feature pyramids (FPN) or tiled (sliced) inference.
- **Class imbalance:** use class-balanced sampling, weighted or focal losses, and threshold tuning; report per-class metrics.
- **Domain shift:** performance often drops with a new scanner, camera, lighting or site; external validation and domain adaptation are essential.

8. Mini Case Studies

Real systems usually combine several tasks. The following cases show how one application decomposes into classification, detection, segmentation and related tasks.

Table 15. *Three applications decomposed into vision tasks.*

Task	Lung nodule (CT)	Autonomous driving	Manufacturing inspection
Classification	Does the scan contain a suspicious nodule?	Road and weather condition	Product pass / fail
Detection	Nodule location (box or centre)	Cars, pedestrians, cyclists, traffic signs	Location of the defect
Segmentation	Exact nodule boundary and volume	Drivable road, lanes, sidewalks	Exact area of a scratch or crack
Related tasks	Registration and tracking of the nodule across follow-up scans to measure growth	Depth estimation for distance; tracking for trajectory prediction	Anomaly detection for new, unseen defect patterns

9. Common Pitfalls

- **Data leakage:** the same patient, video or product in both training and test sets.
- **Misleading accuracy on imbalanced data:** 95% accuracy is trivial if 95% of samples are negative; check per-class recall.
- **Wrong metric for the purpose:** medical screening needs sensitivity, specificity, NPV and PR-AUC, not accuracy alone.
- **Inconsistent annotation:** differing box or mask conventions across annotators give the model conflicting supervision.
- **Untuned threshold:** a score threshold of 0.5 is rarely optimal.
- **Resolution mismatch:** resizing a large image to 224 × 224 can make a small lesion or object disappear entirely.
- **Distribution shift:** a model that works on a curated lab dataset may fail in the field.

- **Metric-only thinking:** errors with high real-world cost (a missed tumour, a missed pedestrian) must be analysed separately, not averaged away.

10. Glossary

Table 16. *Key terms.*

Term	Meaning
Ground truth	The reference label, box or mask provided by a human expert.
Backbone	The feature-extraction network (e.g. ResNet) shared by classification, detection and segmentation heads.
Inference	Running a trained model on new data to obtain predictions.
Epoch / batch	One full pass over the training set / a group of samples processed together.
Augmentation	Random transformations (flip, crop, colour jitter) applied to training images to increase variability.
Transfer learning	Initialising a model with weights pretrained on a large dataset, then fine-tuning it on the target task.
Overfitting	High performance on training data but poor performance on new data.
Anchor box	A predefined reference box from which a detector predicts offsets.
IoU / Dice	Overlap measures between prediction and ground truth.
NMS	Post-processing that removes duplicate overlapping detections.
mAP	Mean of per-class average precision; the standard detection metric.
FROC	Sensitivity versus false positives per image; the standard for medical lesion detection.
Things / stuff	Countable objects (car, cell) versus amorphous regions (sky, road).
Data leakage	Information from the test set influencing training, causing over-optimistic results.

11. Summary

Image classification, object detection and image segmentation form a hierarchy of increasingly detailed visual understanding. Classification maps an entire image to one or more labels and is the foundation of all modern backbones. Detection extends classification by localising every object with a bounding box, using two-stage, one-stage, anchor-free or transformer-based designs, and is evaluated with IoU-based mAP. Segmentation assigns a label to every pixel; semantic segmentation labels classes, instance segmentation separates objects, and panoptic segmentation does both. The choice among them should follow the application's need: use the coarsest task that provides the information the decision requires, since finer tasks cost more in annotation and computation.

12. Review Questions

1 Define image classification, object detection and image segmentation. How do their outputs differ for the same image?

2 Distinguish between multi-class and multi-label classification. Which output activation and loss function does each use?

3 List five sources of variation that make visual recognition difficult, with an example of each.

4 Explain the role of residual connections in ResNet.

5 Compute the IoU of boxes (0, 0, 6, 4) and (3, 2, 9, 6) given in xyxy format.

6 Describe the NMS algorithm step by step. What happens if the IoU threshold is set too low or too high?

7 Compare two-stage and one-stage object detectors in terms of pipeline, speed and accuracy.

8 What problem does the focal loss solve? Write its formula and explain the role of γ.

9 Explain the difference between mAP@0.5 and mAP@[0.5:0.95].

10 Differentiate among semantic, instance and panoptic segmentation using the terms *things* and *stuff*.

11 Draw the U-Net architecture and explain why skip connections are needed.

12 A predicted mask has 120 pixels, the ground truth 100 pixels, and they overlap in 90 pixels. Compute IoU and Dice.

13 Why is pixel accuracy a poor metric for lesion segmentation? Suggest better metrics.

14 What is mass detection? Why is FROC preferred over mAP for evaluating it?

15 A screening model gives TP = 30, FN = 10, FP = 20, TN = 340. Compute sensitivity, specificity, PPV and NPV, and comment on the result.

16 Explain data leakage with an example, and describe how a group-aware split prevents it.

17 For each of the following, choose classification, detection or segmentation and justify: (a) counting vehicles at a junction, (b) measuring a brain tumour's volume, (c) screening leaf images for disease presence.

Further Reading

- R. Szeliski, *Computer Vision: Algorithms and Applications*, 2nd ed., Springer, 2022.
- I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press, 2016 (Chapter 9: Convolutional Networks).
- K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," CVPR 2016.
- S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN," NeurIPS 2015.
- J. Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," CVPR 2016.
- N. Carion et al., "End-to-End Object Detection with Transformers (DETR)," ECCV 2020.
- O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," MICCAI 2015.
- K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," ICCV 2017.
- A. Kirillov et al., "Panoptic Segmentation," CVPR 2019; "Segment Anything," ICCV 2023.