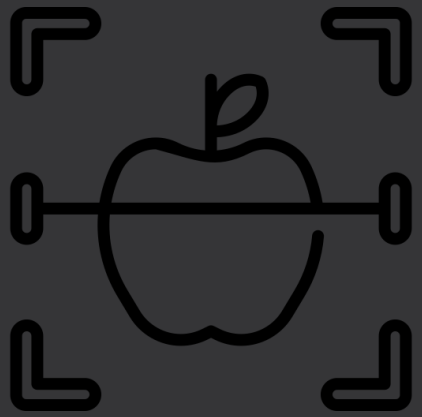


Computer Vision Applications

Classification, Detection, and Segmentation

Lecture 19 – HCCDA-AI

Imran Nawar

10 August 2025

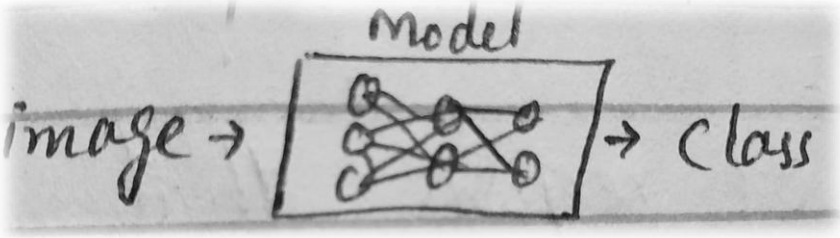
Overview

- **Image Classification**
  - Classification vs Localization
- **Object Detection**
- **Approaches to Object Detection**
  - Two-Stage Detectors
  - One-Stage Detectors
- YOLOv1
- YOLOv12
- State-of-the-Art Models (2025)
- Recent Advancements in Object Detection Advancements (2025)
- Real-world Benchmarks (2025)
- **Segmentation**
- **Types of Image Segmentation**
  - Semantic Segmentation
  - Instance Segmentation
  - Panoptic Segmentation
- YOLO-based Instance Segmentation
- Segment Anything (SAM) – Foundation Model
- Segmentation Advancements (2025)

Image Classification

Image Classification

- Assigns a label or category to an entire image.
- Predicts the class or classes present in the image without specifying its location.
- **Output:** Single label or probability distribution over predefined classes.
- **Example:** Classifying an image as “dog” or “cat.”

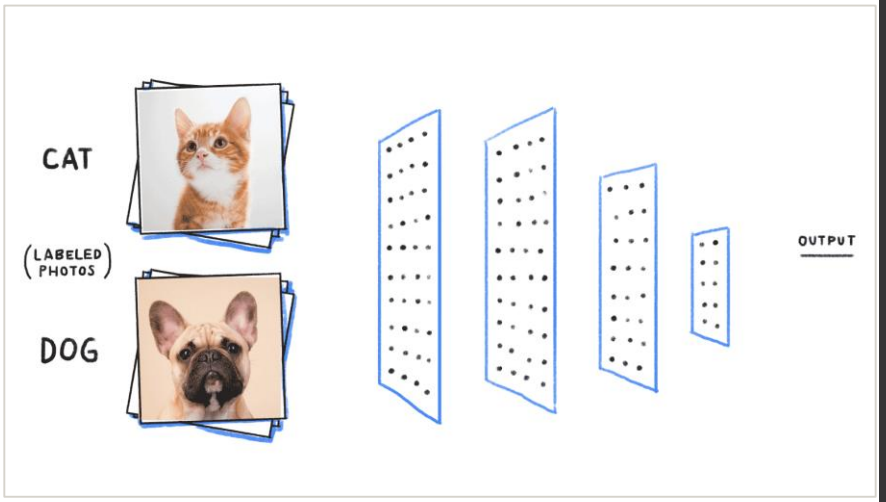


Applications:

- Medical image diagnosis
- Facial recognition
- Autonomous vehicles
- Satellite image analysis

State-of-the-art Models (2025):

- EfficientNetV2
- ConvNeXtV2
- Vision Transformers (ViTs) (e.g., Swin Transformers V2)



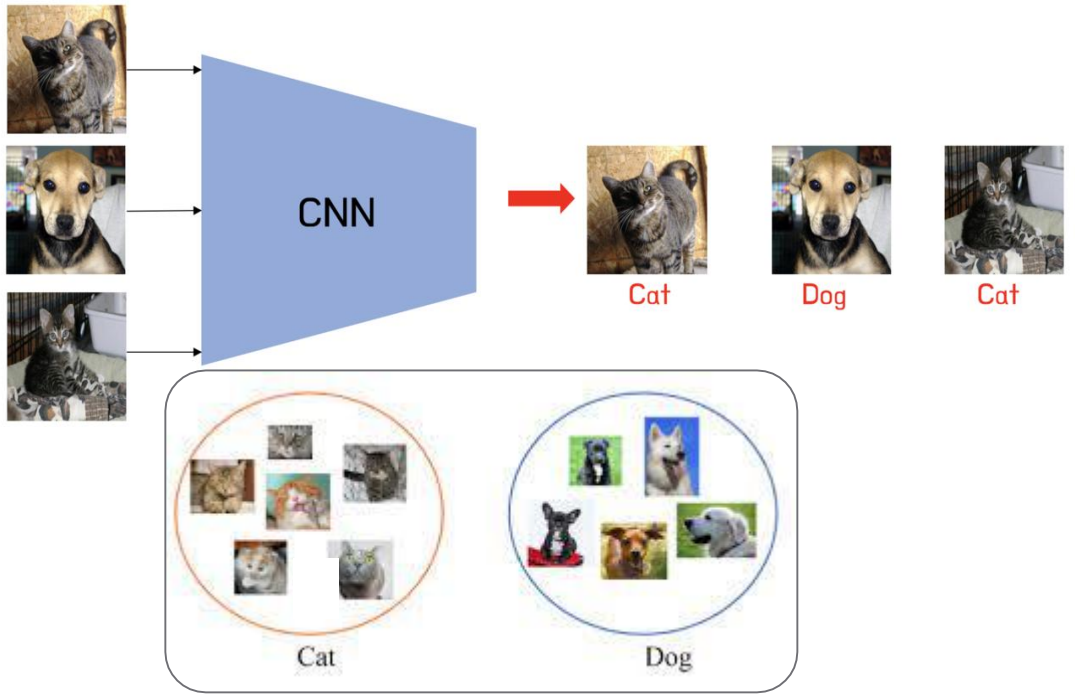
Classification vs Localization

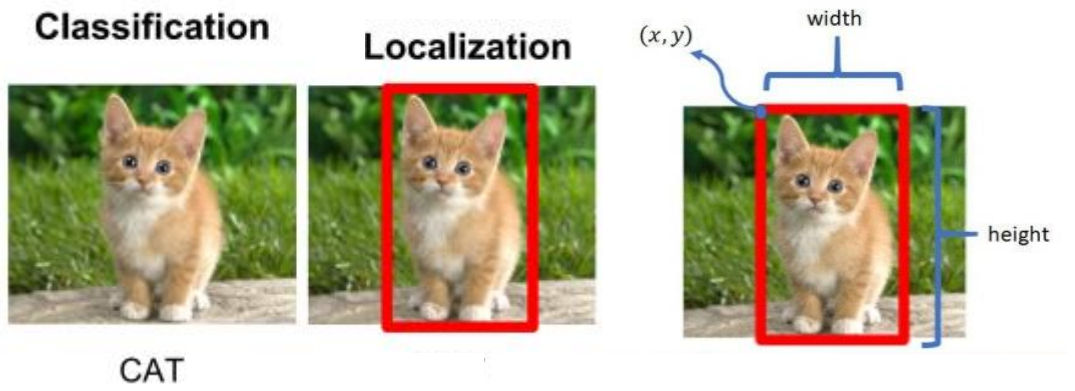
Classification

- Determines what is present in the image without specifying location.
- **Output:** Single label or probability distribution.
- **Example:** System classifies an image of a cat as "cat."
- **Applications:** Categorizing images, plant species classification, recognizing scenes.

Localization

- Predicts object locations within an image using bounding boxes *(x_min, y_min, width, height)*.
- **Output:** Bounding box + class label.
- **Example:** System predicts coordinates of a bounding box around a cat in an image.
- **Applications:** Detecting object positions, medical imaging (tumor localization).

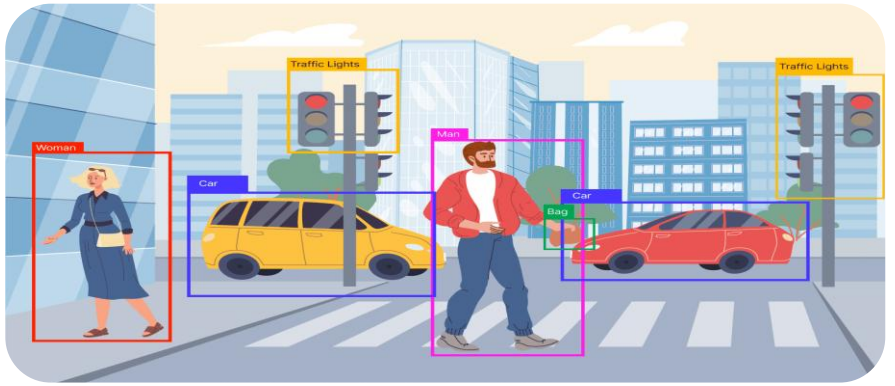




Object Detection

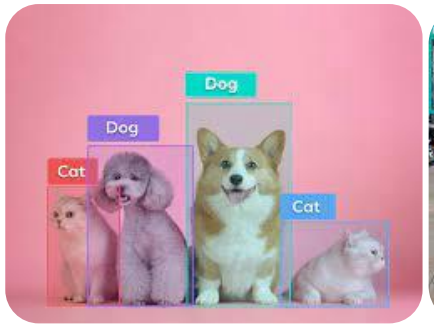
Object Detection

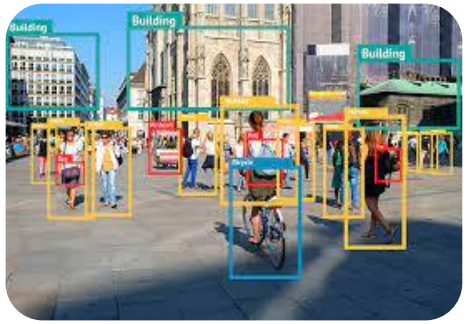
- Identifies multiple objects in an image and their locations.
- Predicts class labels and bounding box coordinates.
- **Example:** Detecting pedestrians, vehicles, and traffic signs in a single image.

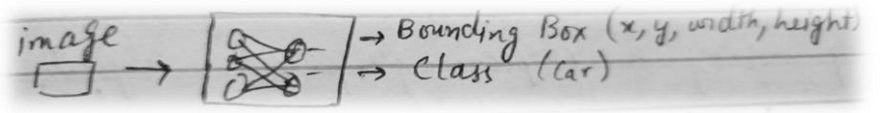


Applications:

- **Autonomous Driving**: Detects road signs, pedestrians, vehicles.
- **Medical Imaging:** Detects tumors or anomalies in X-rays.
- **Surveillance**: Identifies people or suspicious objects.
- **Retail:** Automated checkout systems and shelf inventory tracking.







Autonomous Driving

Surveillance

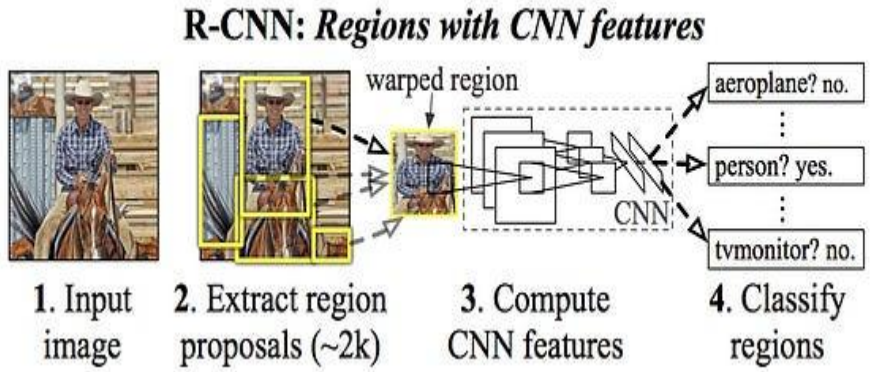
Healthcare

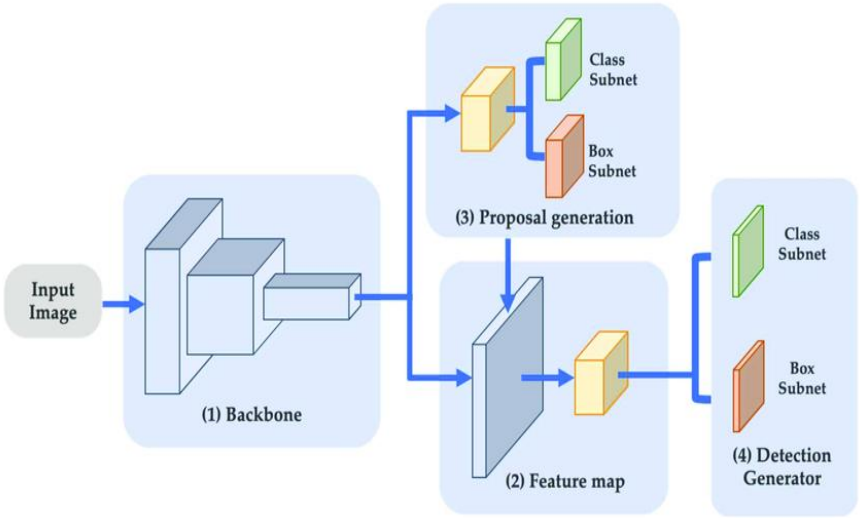
Sports Analytics

Approaches to Object Detection

1) Two-Stage Detectors

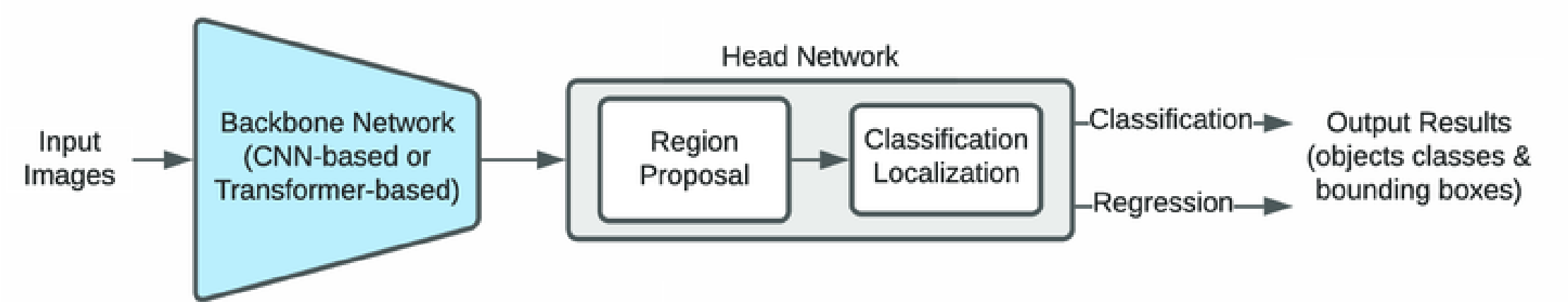
- **Process:**
  1. **Region Proposal Generation**: Identify possible object regions.
     - **Techniques**: Selective search, edge boxes, or Region Proposal Networks (RPNs).
  2. **Classification & Bounding Box Refinement**: Use CNN features for each proposed region to classify the object and refine its bounding box coordinates through a classification layer and a regression layer.
     - **Techniques**: CNN feature extraction → classification layer → regression layer.
- **Models:** Faster R-CNN, Cascade R-CNN, Mask R-CNN.
- **Advantages:**
  - High localization accuracy.
  - Handles occlusion and small objects well.
  - Can be extended to tasks like instance segmentation and keypoint detection.





Approaches to Object Detection

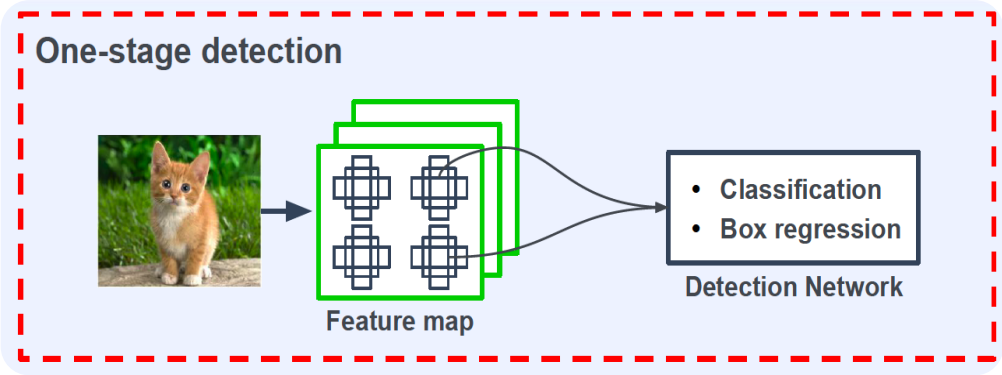
Two-Stage Detectors



Approaches to Object Detection

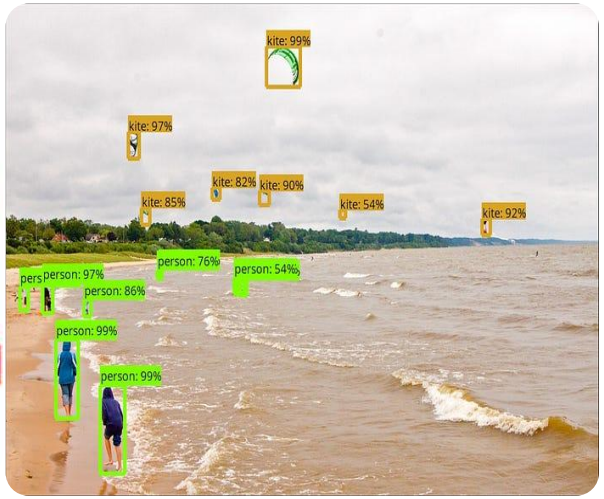
2) One-Stage Detectors

- **Process:**
  - Predicts bounding boxes and classes in a single pass.



- **Models:**
  - YOLOv12, RT-DETR, RetinaNet, EfficientDet.

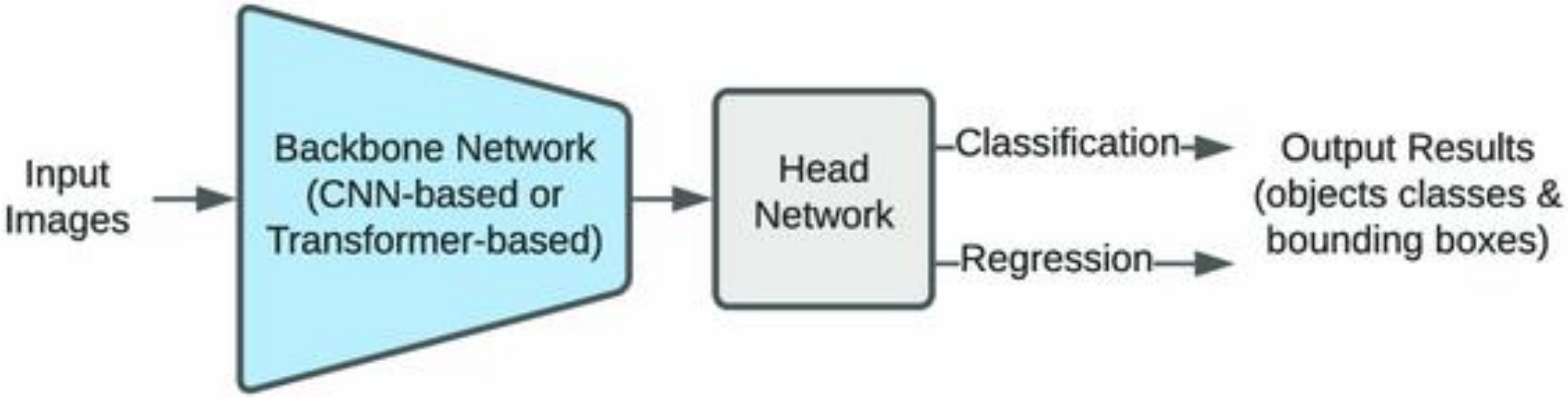
- **Advantages:**
  - Faster inference, suitable for real-time and embedded devices.
  - Simpler architecture for deployment.



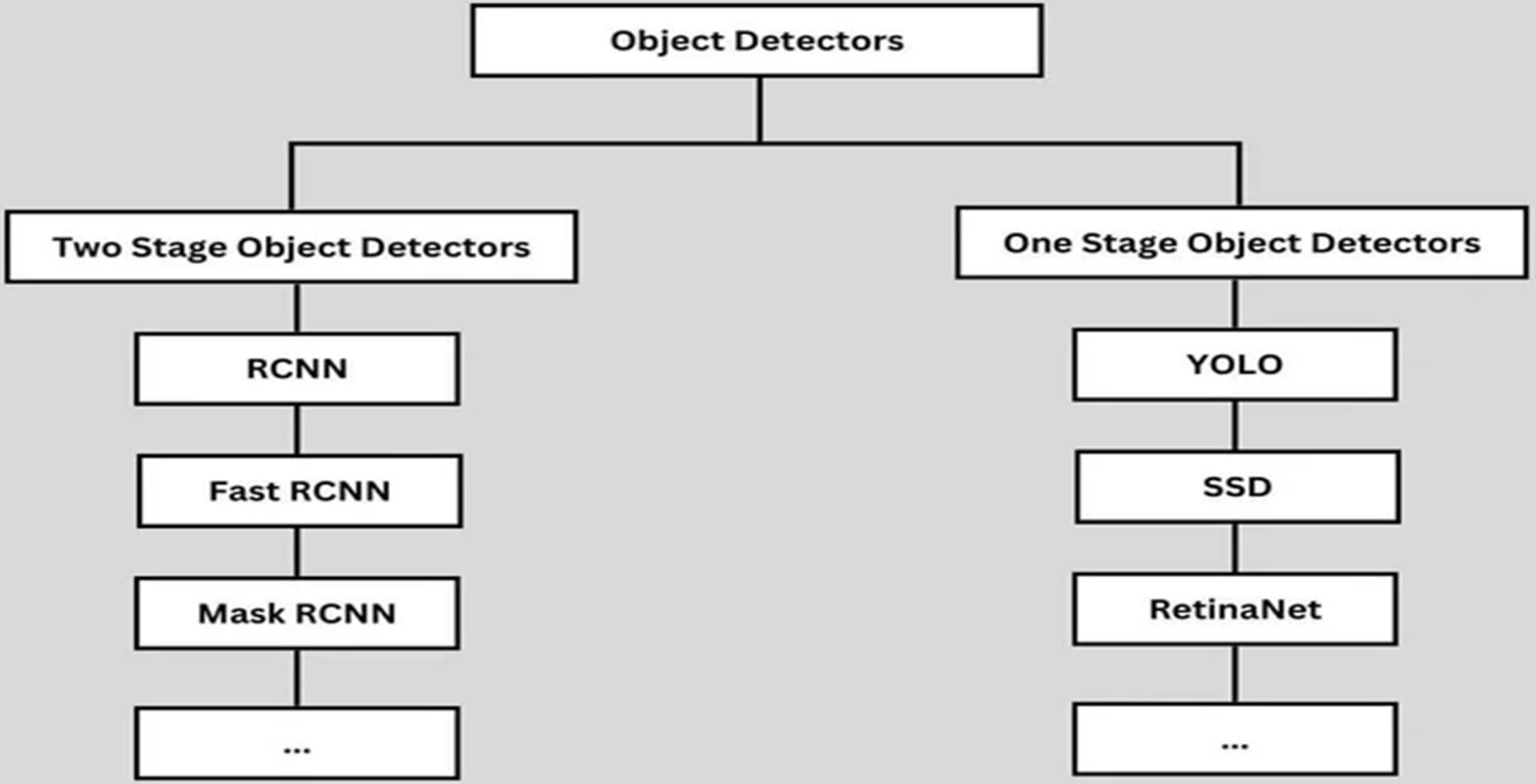
- **Challenges:**
  - Small object detection, precision trade-offs.

Approaches to Object Detection

One-Stage Detectors



One-stage and two-stage detectors



One-Stage Detector

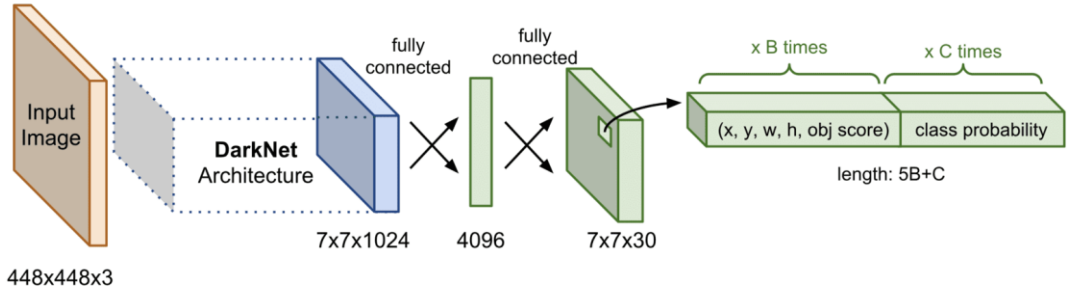
YOLO (You Only Look Once)

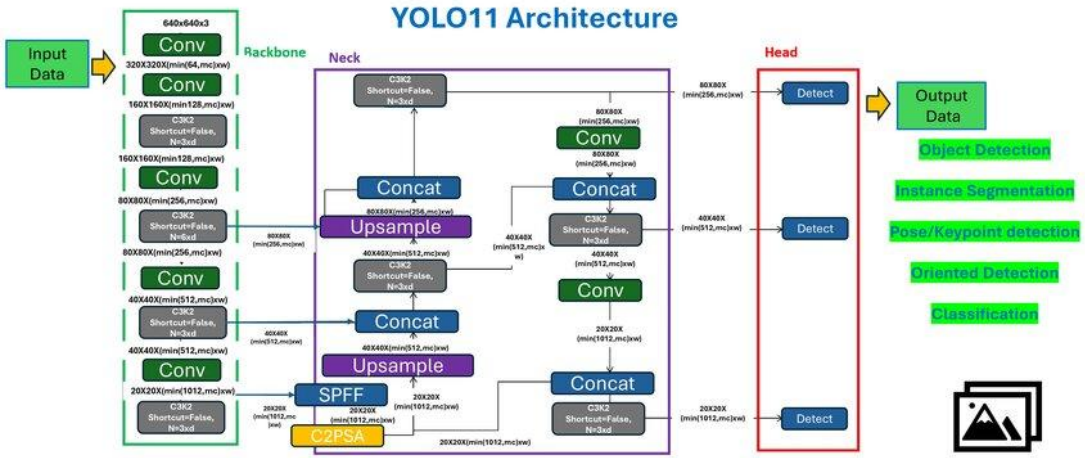
- YOLO is a pioneering one-stage detector that predicts bounding boxes and class probabilities directly from the image in a single pass, enabling real-time object detection..

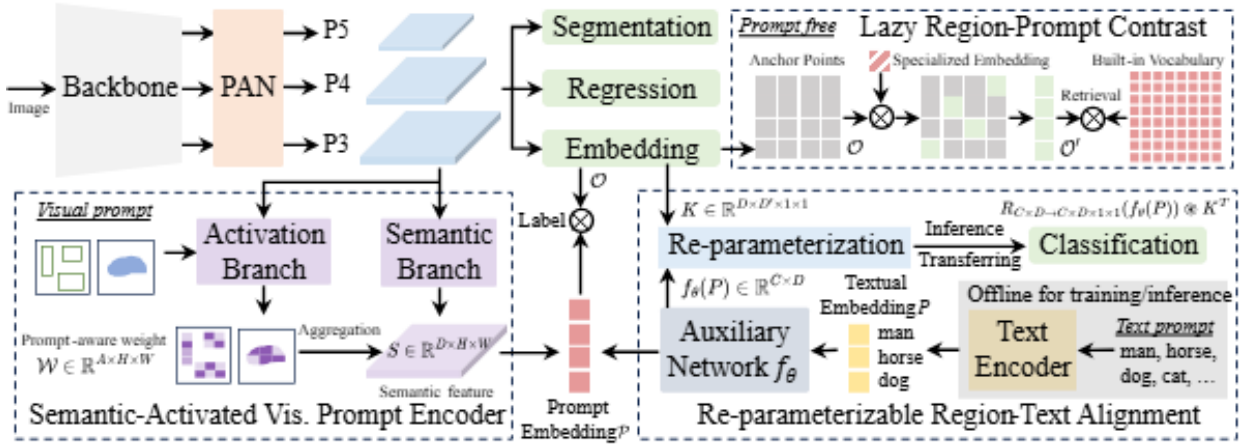
- **Key Components:**
  - **Grid Division**: Divides the input image into a grid of cells.
  - **Prediction**: Each grid cell predicts bounding box coordinates and class probabilities.
  - **Single Network**: Perform both localization and classification in one step (no separate region proposal stage).

- **Advantages:**
  - Real-time performance.
  - Works well on embedded and mobile devices.

- **State-of-the-art**: YOLOv12, YOLOE.







One-Stage Detector

YOLO v1

- YOLOv1 (2015) was the first widely recognized end-to-end one-stage object detector.

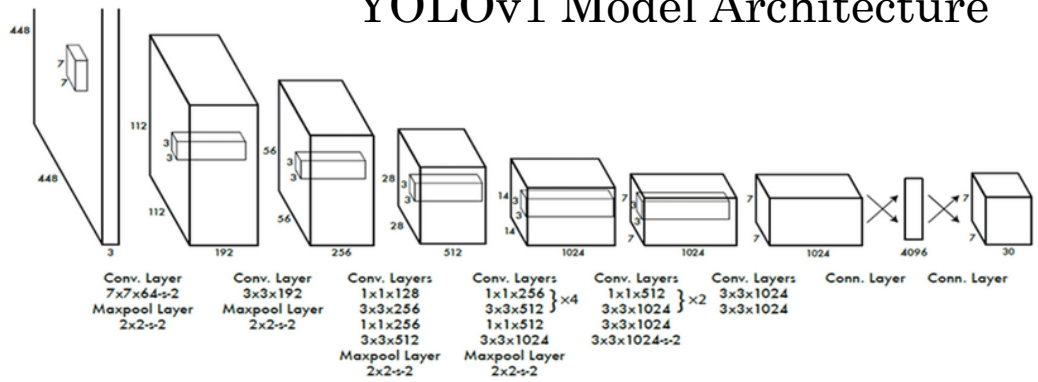
- **YOLOv1 Backbone:**
  - Used **GoogLeNet-inspired custom CNN** (not exactly GoogLeNet, but similar design with 24 convolutional layers + 2 fully connected layers).

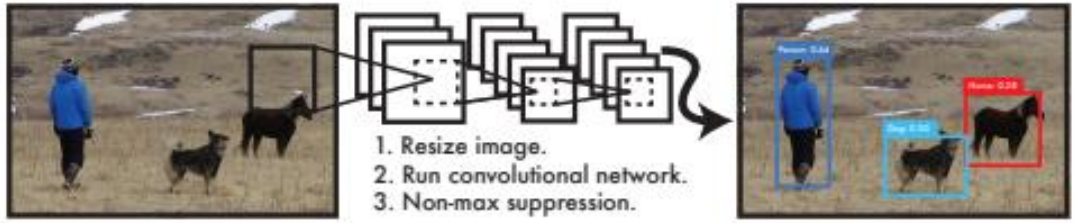
- **YOLOv1 Key Metrics:**
  - **Dataset:** Trained on PASCAL VOC 2007 + 2012 datasets.
  - **mAP (Mean Average Precision):** ~63.4% on VOC 2007 test set.
  - **Speed:** ~45 FPS (real-time) on Titan X GPU; "Fast YOLO" variant ran at ~155 FPS with slightly lower accuracy.
  - **Breakthrough**: 1000x faster than R-CNN, 165x faster than Fast R-CNN

You Only Look Once:
Unified, Real-Time Object Detection

Joseph Redmon*, Santosh Divvala*†, Ross Girshick¶, Ali Farhadi*†
University of Washington*, Allen Institute for AI†, Facebook AI Research¶
http://pjreddie.com/yolo/

YOLOv1 Model Architecture



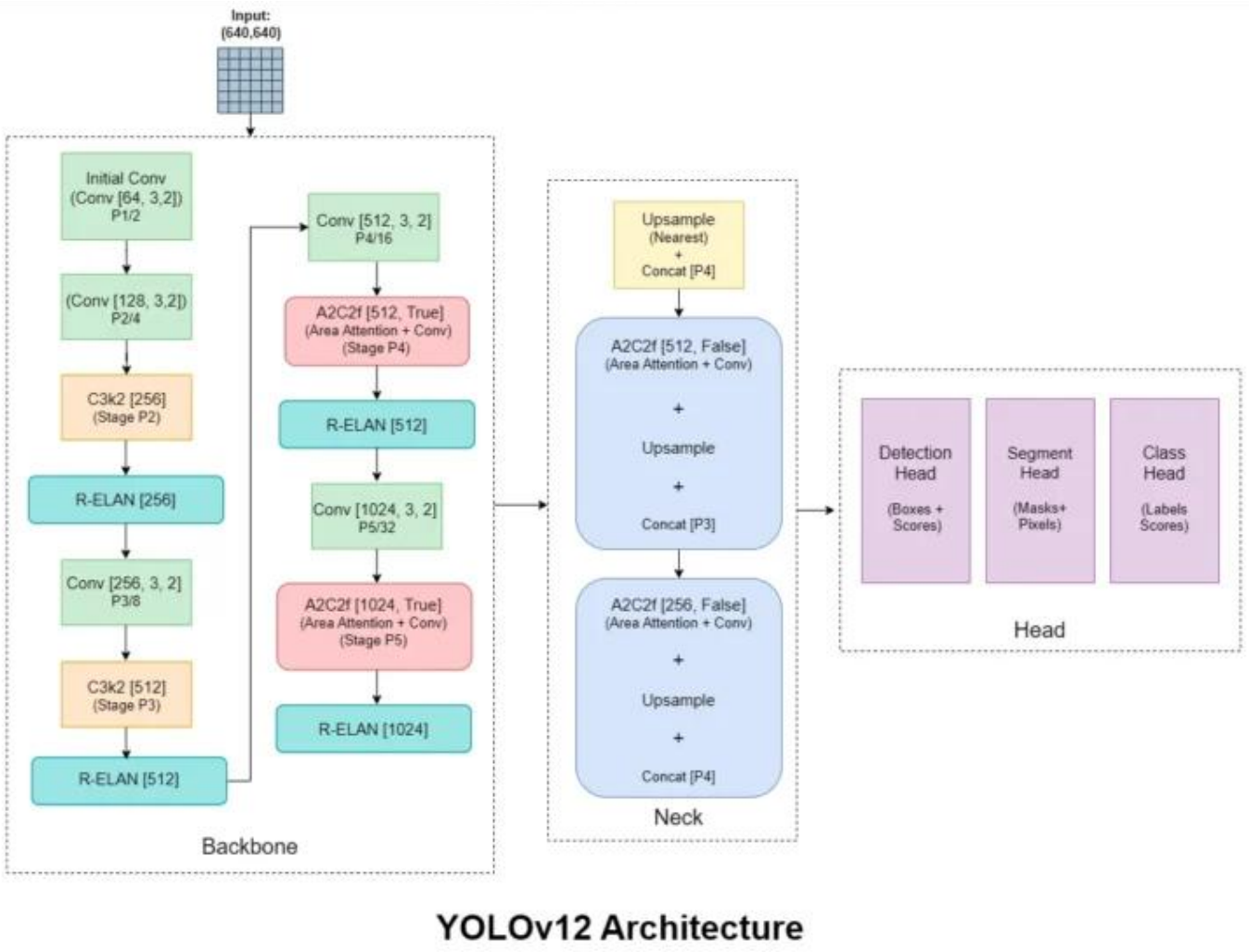


One-Stage Detector

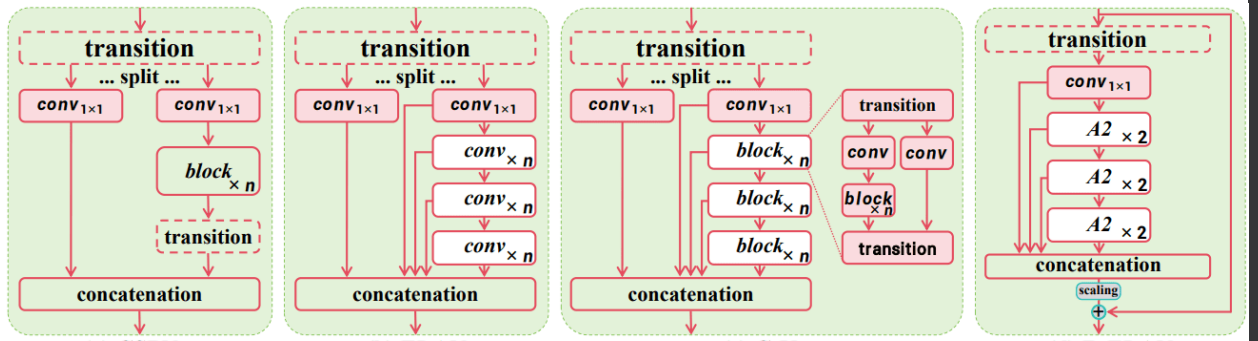
YOLOv12

- YOLOv12 is the latest iteration of the **YOLO** series, bringing enhanced **accuracy, speed, and efficiency** in real-time object detection.

- **Key Features:**
  - **Architecture:**
    - Optimized single-stage detector with an attention-first design for richer contextual understanding.
    - Incorporates efficient attention mechanisms (e.g., FlashAttention) for speed without sacrificing accuracy.
  - **Backbone Network:**
    - Hybrid CNN-Transformer backbone for combining local detail extraction with global context.
    - Uses Residual Efficient Layer Aggregation (R-ELAN) for improved multi-scale feature fusion.
  - **Frame Rate:**
    - State-of-the-art FPS on modern GPUs, with higher mAP than YOLOv11.
    - Example benchmarks on NVIDIA A100 GPU:
      - **YOLOv12-N:** ~40.6% mAP at ~1.64 ms latency.
      - Larger models achieve higher accuracy with slightly reduced FPS.
  - **Scalability:**
    - Available in multiple sizes (N, S, M, L, X) to balance speed and accuracy.
    - Optimized variants for edge and embedded systems.



YOLOv12 Architecture

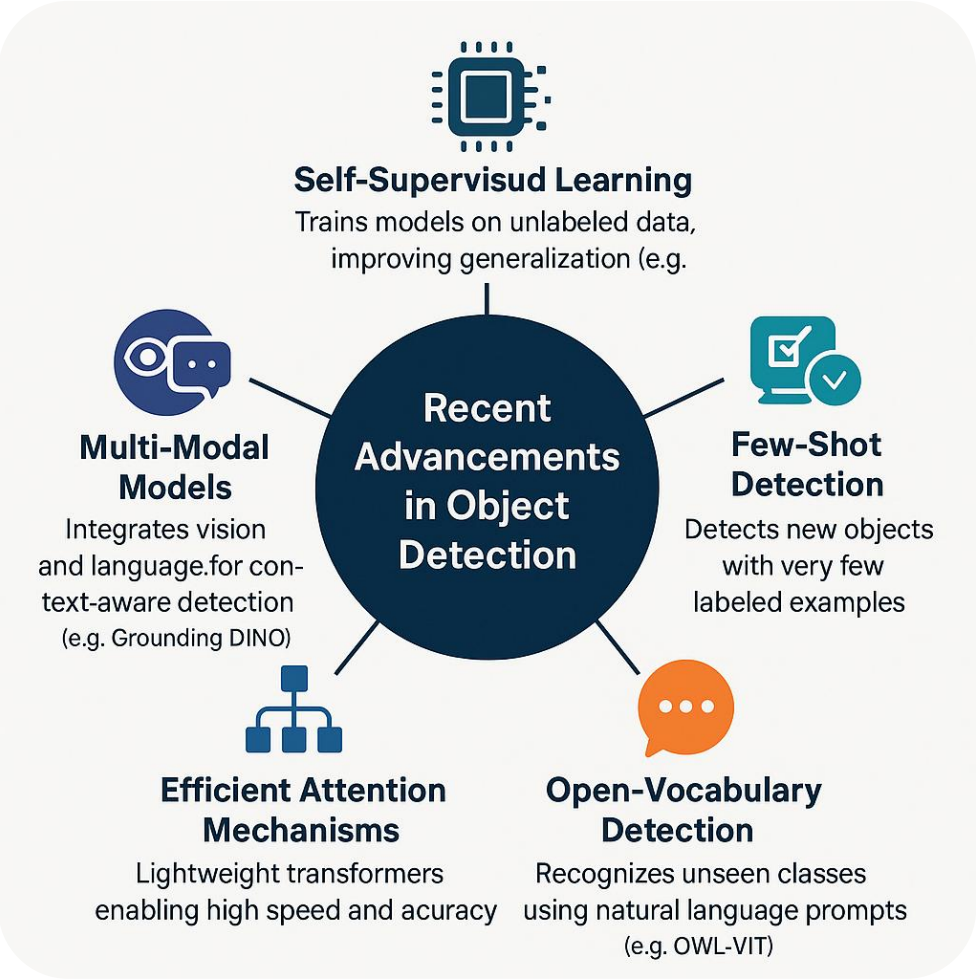


State-of-the-Art Models (2025)

- **YOLOv12:**
  - Next-gen YOLO, improved small-object detection, hybrid CNN-Transformer backbone, efficient attention, state-of-the-art FPS & accuracy, optimized for edge devices.
- **RT-DETR:**
  - Real-time DEtection TRansformer, end-to-end detection without NMS, high accuracy in real-time.
- **Grounding DINO:**
  - Zero-shot detection using natural language prompts, large-scale pretraining.
- **DINOv2:**
  - Self-supervised vision transformer with superior feature generalization for detection tasks.
- **OWL-ViT:**
  - Open-vocabulary detection, few-shot adaptability.
- **Detectron2:**
  - Modular framework for high-performance detection & segmentation.

Recent Advancements in Object Detection (2025)

- **Self-Supervised Learning:** Trains models on unlabeled data, improving generalization (e.g., DINOv2).
- **Few-Shot Detection:** Detects new objects with very few labeled examples.
- **Efficient Attention Mechanisms:** Lightweight transformers enabling high speed and accuracy.
- **Open-Vocabulary Detection:** Recognizes unseen classes using natural language prompts (e.g., OWL-ViT).
- **Multi-Modal Models:** Integrates vision and language for context-aware detection (e.g., Grounding DINO).

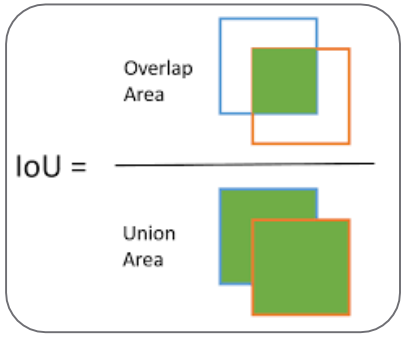


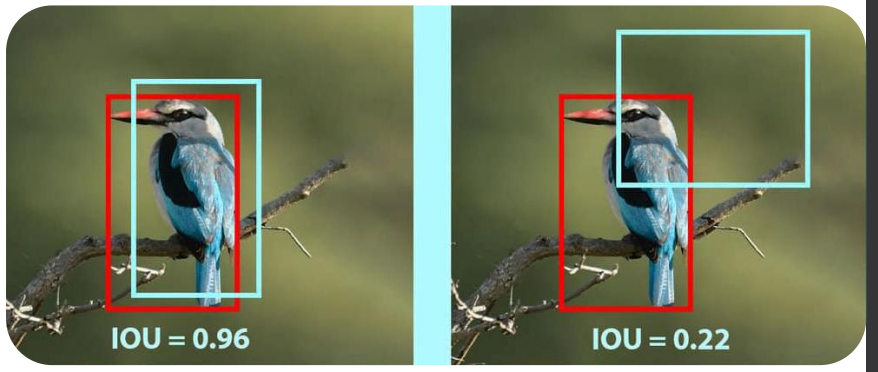
Performance Metrics

1. **Intersection over Union (IoU):**
   - Measures overlap between predicted and ground-truth boxes:
   - **Formula:**

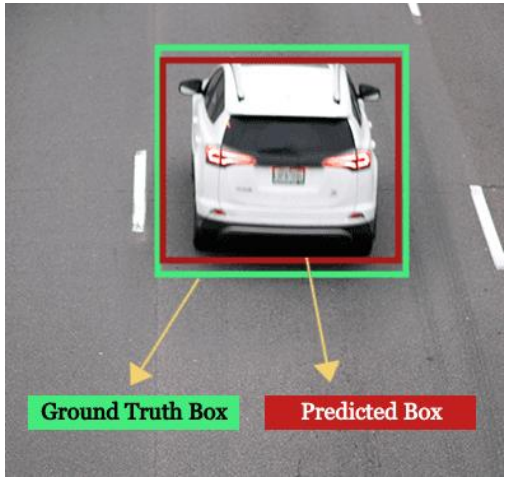
$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$

   - It's value between 0 and 1.
     - 1.0 = perfect match
     - 0.0 = no overlap.





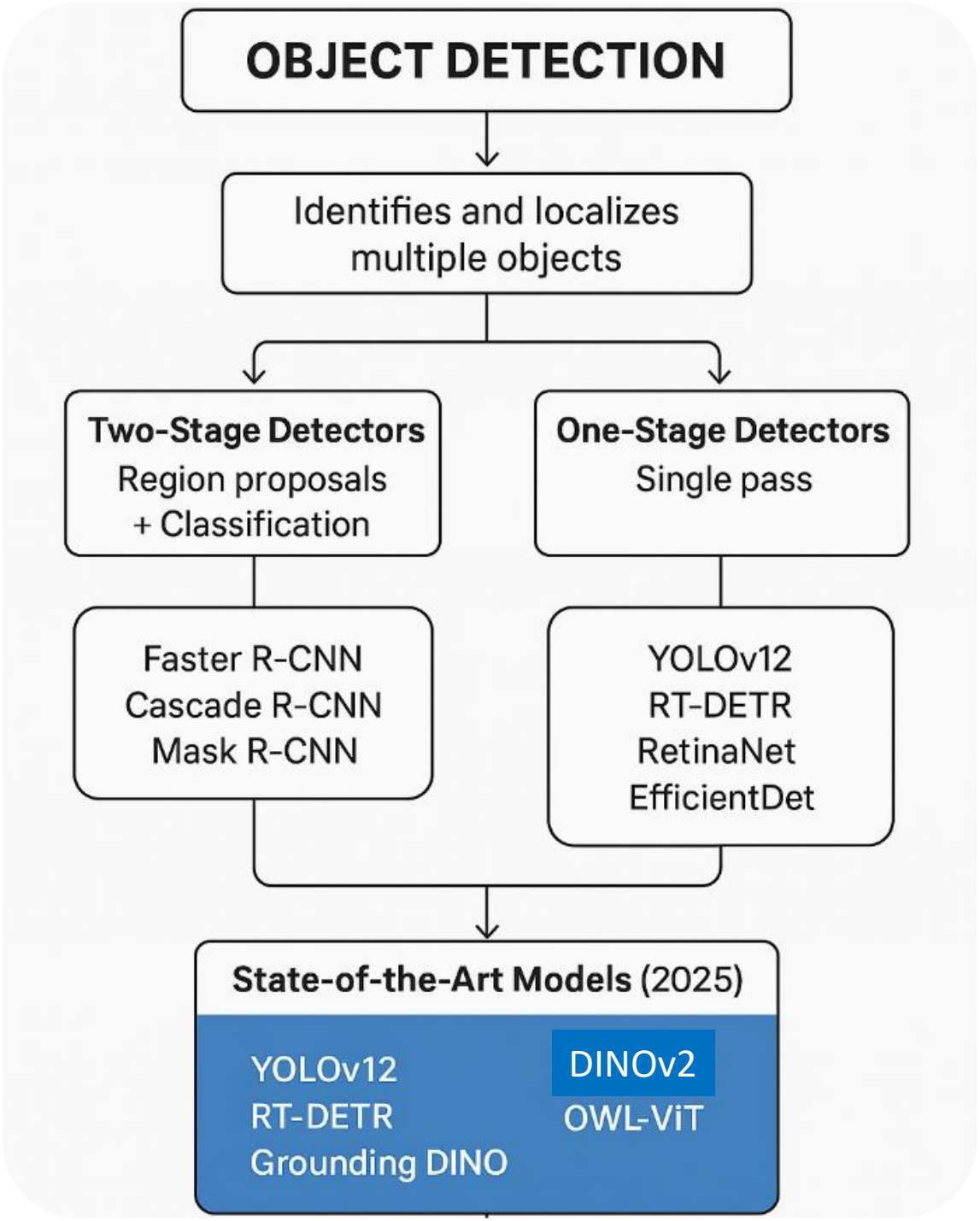
2. **Mean Average Precision (mAP):**
   - Combines classification & localization accuracy.
   - Average precision is computed over multiple IoU thresholds (e.g., 0.5:0.95).
   - Higher mAP = better detection performance.
   - **Example:**
     - COCO benchmark reports mAP across IoU thresholds from 0.5 to 0.95.



$$mAP = \frac{1}{k}\sum_{i}^{k} AP_i$$

Real-World Benchmarks (2025)

Model	mAP@0.5:0.95 (COCO)	FPS (A100)	Key Strength
YOLOv12x	57.8	80	Real-time, high accuracy
RT-DETR	56.2	65	Transformer-based speed
DINOv2	58.1	45	Open-vocabulary detection


OBJECT DETECTION
Identifies and localizes multiple objects
Two-Stage Detectors
Region proposals + Classification
One-Stage Detectors
Single pass
Faster R-CNN
Cascade R-CNN
Mask R-CNN
YOLOv12
RT-DETR
RetinaNet
EfficientDet
State-of-the-Art Models (2025)
YOLOv12
RT-DETR
Grounding DINO
DINOv2
OWL-ViT

Segmentation

Image Segmentation

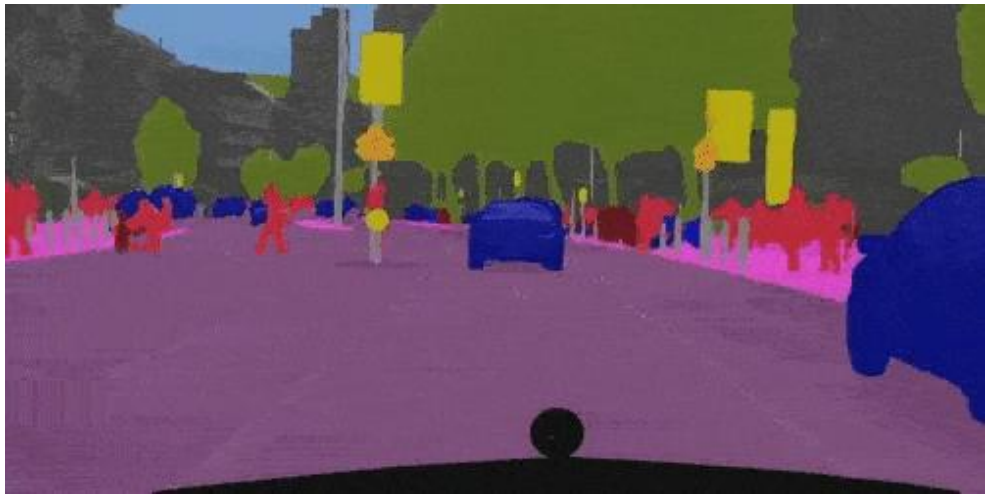
- Partitioning an image into distinct regions or segments.
- **Objective:** Precise localization and identification of objects at the pixel level.

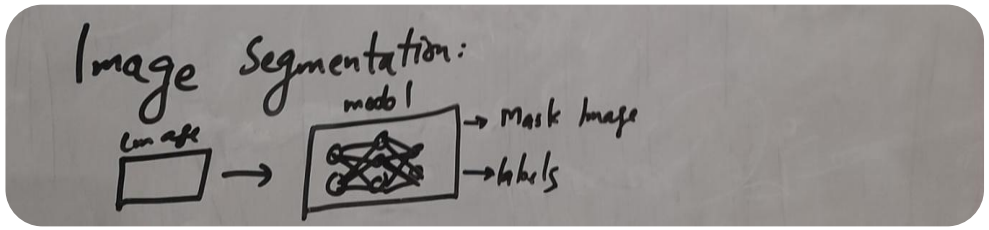
Applications:

- Medical Imaging (e.g., tumor detection).
- Autonomous Vehicles (e.g., road scene understanding)

Output:

- **Pixel-wise mask** identifying object boundaries and regions.





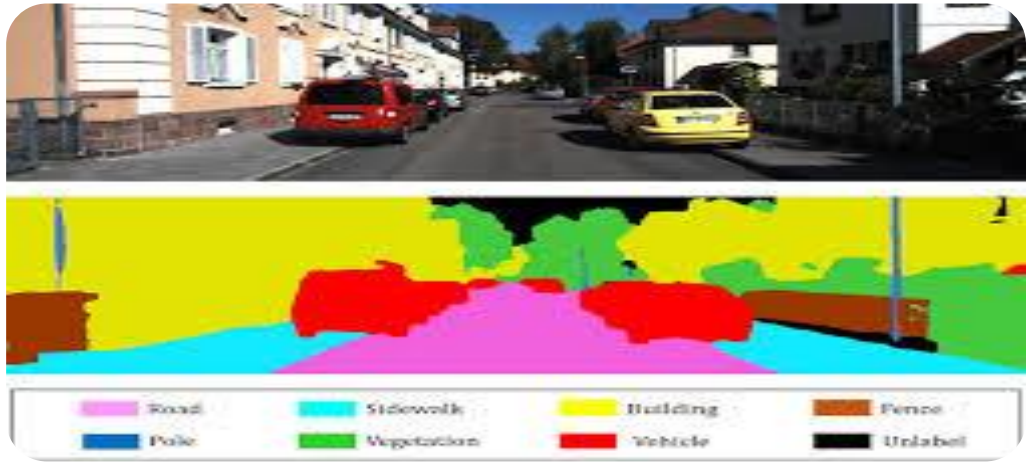
Types of Image Segmentation

1. Semantic Segmentation

- Assigns a class label to every pixel in an image.
- Segments the image into regions corresponding to different semantic categories (e.g., person, car, tree).
- Focuses on understanding the overall scene layout and context.

- **Applications:**
  - Urban planning (e.g., land cover mapping).
  - Precision Agriculture (e.g., crop yield estimation).

- **Models:** U-Net, DeepLab, FCN, SegNet



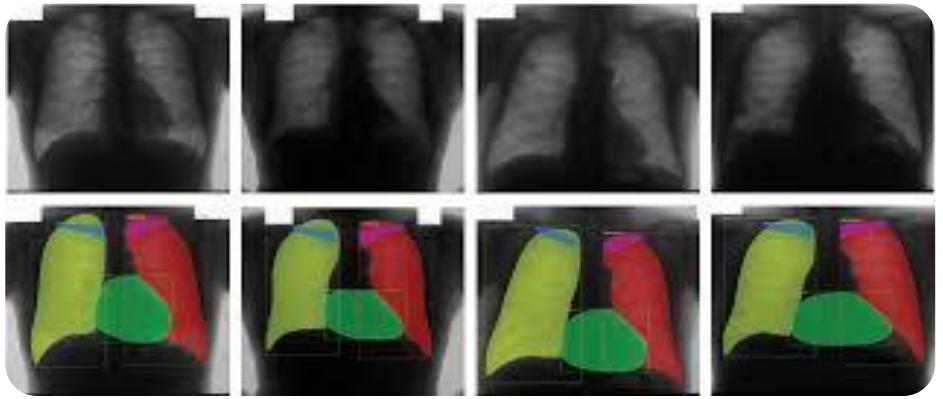
Types of Image Segmentation

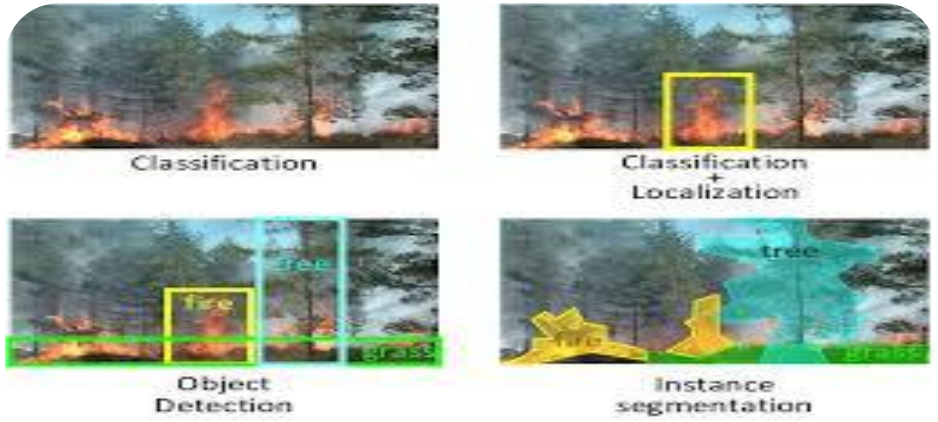
2. Instance Segmentation

- Identifies and distinguishes individual object instances.
- Provides pixel-level masks for each object instance in an image.
- Differentiates between multiple objects of the same class, even if they overlap.
- Widely used in AI and robotics competitions, such as the DARPA Robotics Challenge and RoboCup, where robots perform complex tasks in dynamic environments.

- **Application**
  - **Medical Imaging** (e.g., organ segmentation).
  - **Autonomous Driving** (e.g., pedestrian detection).
  - **Surveillance and Security**

- **Models:** Mask R-CNN, YOLO-based instance segmentation.





Types of Image Segmentation

3. Panoptic Segmentation

- Combines semantic and instance segmentation.
- Assign class labels to every pixel and provide instance masks.

- **Assigning Class Labels to Every Pixel**:
  - Like semantic segmentation, each pixel is labeled with a category (e.g., person, car, road).
- **Providing Instance Masks for Objects**:
  - In addition to class labels, panoptic segmentation gives unique instance masks for each object, allowing precise delineation and identification.
- **Applications:**
  - Scene understanding (e.g., urban environments).
  - Robotics (e.g., object manipulation)
- **Models:** Panoptic FPN, Mask2Former, OneFormer, K-Net, UPSNet, DETR-Panoptic

Input Image

Image Segmentation

Semantic Segmentation

Output

- Label of each category
- Pixel values of each category

Instance Segmentation

Output

- Pixel-wise mask for each object instance
- Label of each category

Panoptic Segmentation

Output

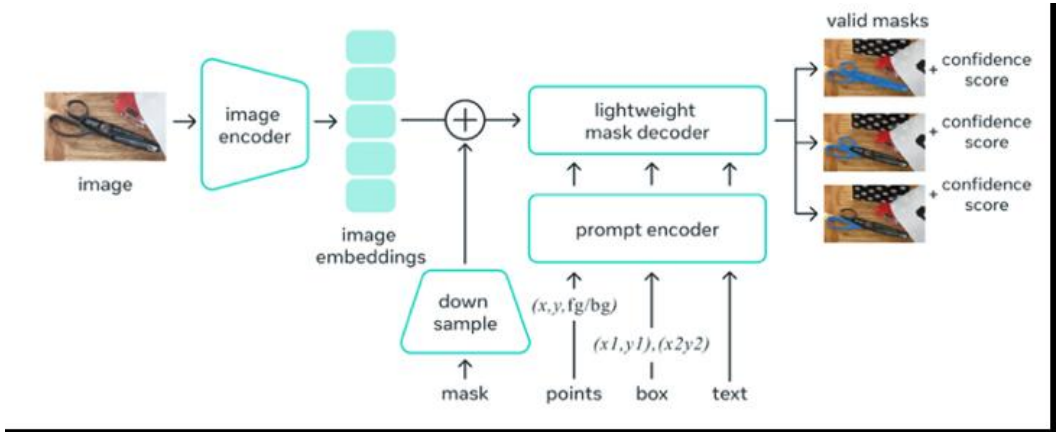
- Label of each category
- Pixel-wise mask for each object instance

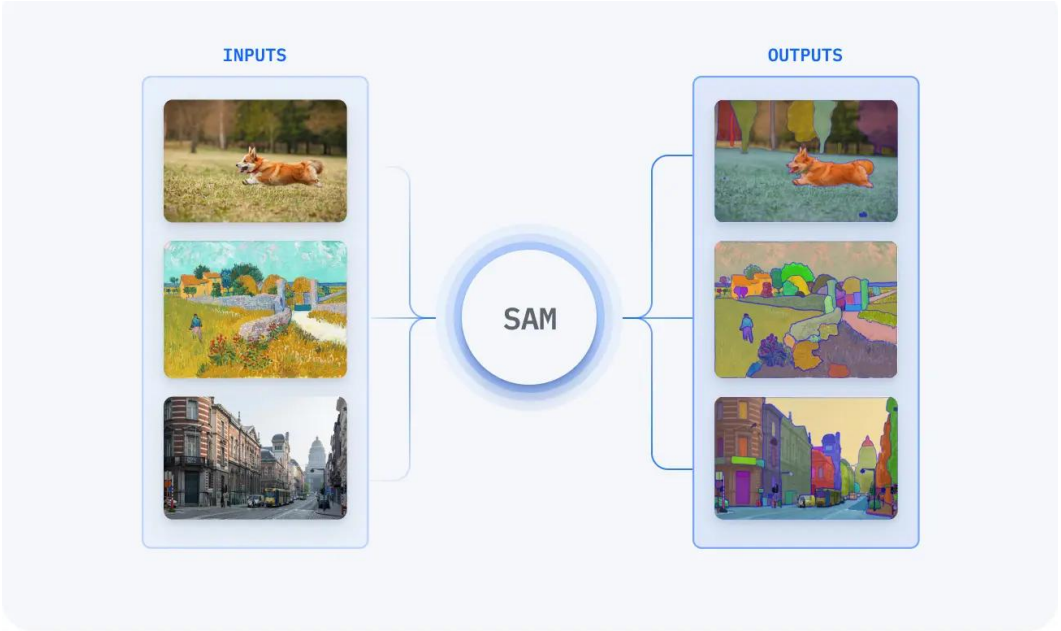
YOLO-based Instance Segmentation

- **YOLO:**
  - A fast, single-stage object detector.
  - YOLO is a popular object detection algorithm known for its speed and efficiency.
  - Adapted to perform instance segmentation by predicting segmentation masks along with bounding boxes.
- **Example:**
  - YOLOv8-seg
  - YOLACT / YOLACT++
- **Strength:**
  - Real-time performance.
  - Efficient and lightweight
- **Trade-off:**
  - May struggle with overlapping objects or fine boundaries compared to Mask R-CNN.
- **Applications:**
  - Drones & robotics
  - Surveillance
  - Edge deployment

Segment Anything (SAM) – Foundation Model

- Developed by Meta AI (SAM-1 in 2023, SAM-2 in 2024).
- This model can identify the precise location of either specific objects in an image or every object in an image.
- Performs segmentation without task-specific training (zero-shot).
- SAM's game-changing impact lies in its zero-shot inference capabilities.
- This means that SAM can accurately segment images without prior specific training, a task that traditionally requires tailored models.
- **Strength**: Zero-shot, versatile, scalable to large datasets.
- **Use Case**: Annotation, interactive labeling, general-purpose segmentation.





Segmentation Advancements (2025)

- **Zero-Shot Generalization**:
  - Models like SAM (Segment Anything Model) can handle unseen objects without additional training.
- **Foundation Models for Segmentation**:
  - Large-scale pre-trained models that adapt quickly to multiple domains.
- **Real-Time Edge Segmentation:**
  - Lightweight architectures enabling segmentation on IoT and embedded devices.
- **Multimodal Segmentation**:
  - Combining vision, text, and depth data for richer scene understanding.
- **Explainable Segmentation:**
  - Tools that visualize how models make segmentation decisions to build trust.

Summary & Future Outlook

- **Key Takeaways**
  - Classification → Detection → Segmentation
  - **Image Classification:** Understanding "what" is in an image
  - **Object Detection:** Finding "what" and "where" objects are located
  - **Segmentation:** Precise pixel-level understanding of scenes

- **Evolution of Architectures**
  - From Two-Stage Detectors (R-CNN family) to One-Stage Detectors (YOLO series)
  - YOLOv1 to YOLOv12: Continuous improvement in speed and accuracy
  - Foundation Models like Segment Anything (SAM) revolutionizing segmentation

Thank You