Real-Time Object Detection & Tracking

Using YOLOv11 and SAM2 for Robust, Accurate, and Scalable Object Detection

Team Name: Vision_Titan
College Name: IIIT Jabalpur

Problem Statement

Develop an object detection model that can accurately detect, classify, and localize multiple objects within an image or video frame.

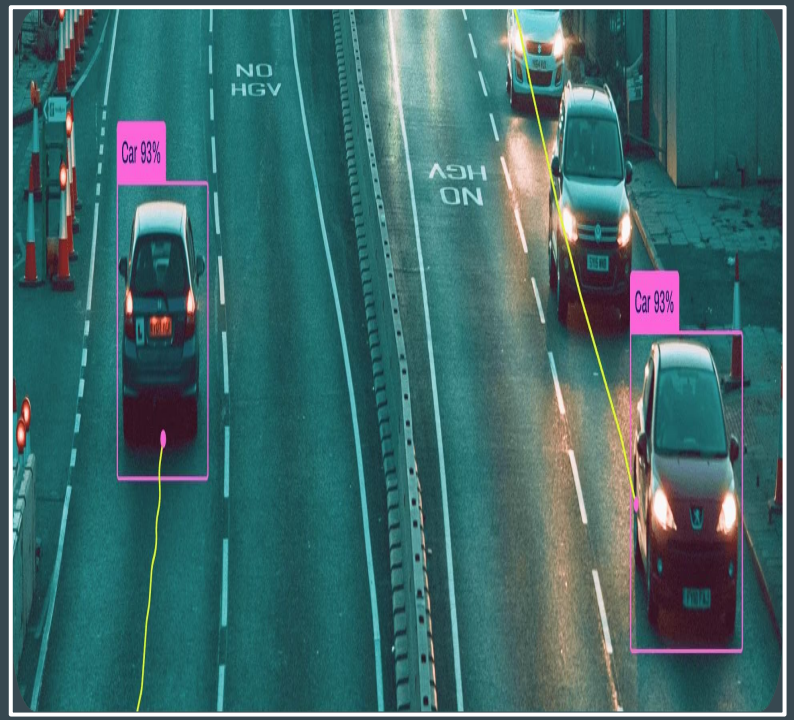
Key Requirements:

- High precision: Minimize false positives and false negatives.
- Real-time processing: Enable practical deployment in real-world applications with minimal latency.
- Environment versatility: Handle varying lighting, object angles, and occlusions.
- Tracking capability: Track objects across multiple frames in videos.

Real-World Applications:

- Surveillance systems.
- Autonomous vehicles (e.g., self-driving cars).
- Retail automation (e.g., inventory tracking).

Introduction to Object Detection



1

Core Computer Vision Task

Object detection aims to identify and localize objects within an image or video.

2

Essential for Many Applications

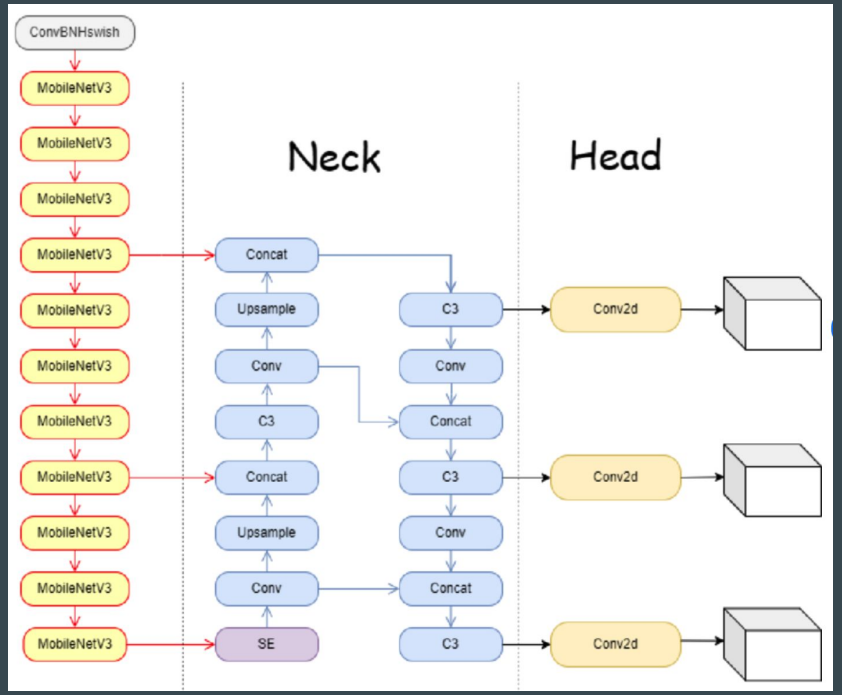
It's a fundamental component in various applications, including autonomous driving, robotics, and security.

3

Types of Object Detectors

There are two main categories: two-stage detectors (e.g., R-CNN) and one-stage detectors (e.g., YOLO).

Solution Approach



Model Selection:

- Use YOLOv11 (You Only Look Once) for real-time object detection.
- Incorporate SAM2 (Segment Anything Model) for high-precision segmentation and object localization.

Key Features:

- YOLOv11: Fast, accurate detection with strong generalization to new environments.
- SAM2: Provides pixel-level segmentation for better localization in challenging scenarios.

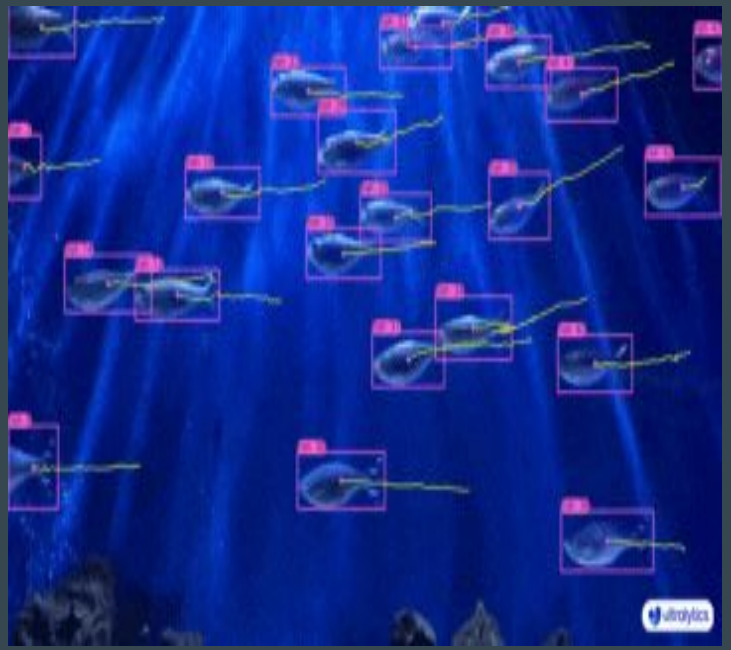
Tracking:

Integrate object tracking algorithms to maintain object IDs across multiple frames for dynamic video input.

Methodology

1. **Data Preparation**:
    - Load input images or video frames for object detection.
    - Preprocess images for YOLO11 and SAM2 input formats.
2. **Object Detection and Tracking**:
    - Use YOLO11 to detect objects in the input images or video frames.
    - Extract bounding boxes, class labels, and confidence scores for detected objects.
    - Track objects across frames using BOT_sort, which efficiently handles multiple objects and occlusions. BOT_sort also retains object identity across frames, even if an object temporarily leaves and re-enters the frame.
3. **Object Segmentation**:
    - Use SAM2 to segment detected objects and refine the boundaries for precise localization.
    - Generate pixel-wise masks for each object to separate them from the background.
    - Combine segmentation results with detection for accurate object localization.
4. **Integrating YOLO11 and SAM2:**
    - Combine the detection results from YOLO11 with the segmentation results from SAM2 to achieve accurate object detection and localization.
    - Use the tracking information to maintain consistency across frames and track objects effectively.
    - Align annotation formats between YOLO11 and SAM2 by converting SAM2 output to YOLO11-compatible format for final output.

Overview of YOLO (You Only Look Once)



1. **Single-Stage Detector**
   YOLO processes the entire image in a single forward pass to make predictions, making it very fast.

2. **Grid-Based Approach**
   The image is divided into a grid, and each grid cell is responsible for predicting objects.

3. **Bounding Boxes and Confidence Scores**
   YOLO predicts bounding boxes around objects and confidence scores representing the likelihood of a detection.

Why YOLOv11?

Real-Time Detection:	YOLOv11 offers superior speed while maintaining high accuracy, making it ideal for real-time applications in dynamic environments.
High Precision	YOLOv11 minimizes both false positives and false negatives with better performance on small and overlapping objects compared to other models.
End to End Architecture	It uses a single neural network to predict class probabilities, bounding box coordinates, and confidence scores simultaneously, resulting in faster processing.
Flexibility and Scalability:	YOLOv11 supports detection across multiple object classes and varied environments (lighting, occlusions, etc.).
Key Advantage:	Faster than traditional CNN-based methods (like Faster R-CNN), suitable for real-time applications like surveillance and autonomous driving

Segmentation-Aided YOLO (SAM2)

1

Improving YOLO's Accuracy

SAM2 leverages semantic segmentation to provide additional information about the objects.

2

Segmentation and Object Detection

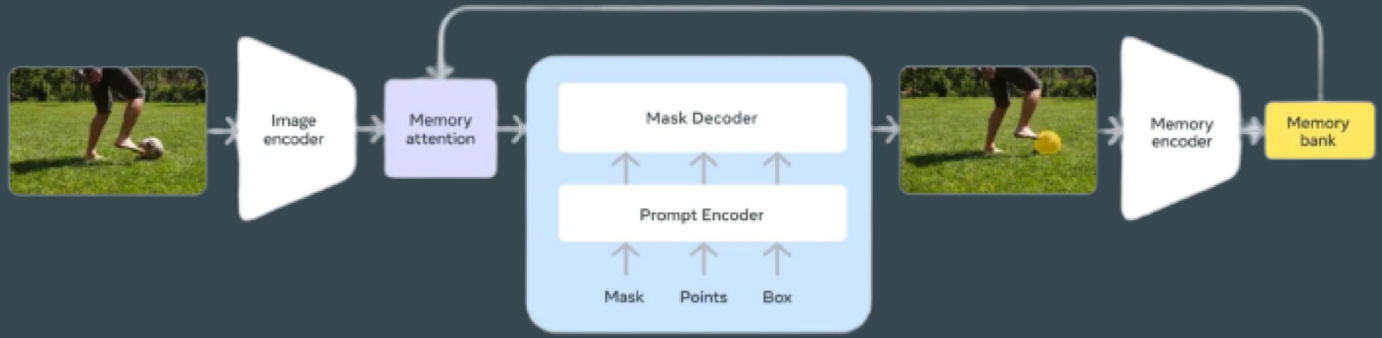
Segmentation provides a more precise outline of the object's shape, leading to more accurate bounding boxes.

3

Integration with YOLO

SAM2 integrates semantic segmentation into the YOLO framework to enhance its performance.

Why SAM2 ?



- Superior Segmentation:
  SAM2 can segment objects in complex scenes with pixel-perfect precision, useful for scenarios with occlusions or highly cluttered environments.
- Zero-Shot Segmentation:
  SAM2 can detect and segment novel object categories without the need for additional training, adding robustness in unpredictable environments.
- Enhanced Localization:
  SAM2's segmentation capabilities allow for precise localization of objects, especially when objects overlap or are partially obscured.
- Integration with YOLOv11:
  By combining SAM2's segmentation with YOLOv11's detection, we achieve greater accuracy in multi-object environments.

Key Features and Innovations of SAM2

Semantic Segmentation Module

The segmentation module generates pixel-level object masks, providing richer information about the object's shape.

Enhanced Bounding Box Refinement

SAM2 refines the predicted bounding boxes using the segmentation masks, resulting in tighter and more accurate detections.

Adaptive Feature Fusion

The system combines features from both the segmentation module and the YOLO network for improved performance.

Results

This solution effectively detects and segments objects in diverse scenarios, making it suitable for applications like:

- Surveillance: Identifying and tracking people or objects in security footage.
- Autonomous Driving: Detecting vehicles, pedestrians, and obstacles in real-time.
- Retail Automation: Identifying products on shelves or tracking customer movements.




Challenges Faced

- Real-Time Constraints:
  Balancing high accuracy with low latency in real-time applications such as video feeds from drones or surveillance cameras.
- Object Occlusion and Overlap:
  Detecting and localizing objects that are partially hidden or stacked can affect detection accuracy. SAM2 helps mitigate this, but it's still a challenge in highly dynamic scenes.
- Lighting and Environmental Variability:
  Handling objects in challenging conditions such as low light, shadows, or glare. Model tuning and data augmentation help address this challenge.
- Resource Constraints:
  Processing high-resolution images or videos in real-time requires significant computational power, especially for high-precision models like SAM2.
- Multi-Object Detection in Crowded Environments:
  Efficiently tracking and distinguishing between multiple objects when they are close together or overlap remains a challenging problem.

Future Scope

- Improving Model Efficiency:
  Future developments could include optimizing YOLOv11 and SAM2 for lower computational overhead while maintaining accuracy, enabling deployment on edge devices or mobile platforms.
- Multimodal Inputs:
  Integrating thermal or infrared data to improve object detection under challenging lighting conditions (e.g., at night or in fog).
- Cross-Domain Generalization:
  Enhancing the model's ability to generalize across different domains (e.g., urban surveillance vs. rural areas) without requiring extensive retraining.
- Advanced Tracking Capabilities:
  Developing more robust multi-object tracking (MOT) algorithms to handle fast-moving objects, especially in autonomous driving or drone navigation.
- Human-Robot Interaction (HRI):
  Integrating object detection with robotics for autonomous manipulation tasks, such as warehouse automation or robot-assisted surgery.

Real-World Applications and Use Cases

Autonomous Driving

Identifying objects in real-time is essential for autonomous vehicle navigation and safety.

Robotics

Object detection enables robots to perceive their environment and interact with objects intelligently.

Security

Security systems use object detection for surveillance, intrusion detection, and access control.

Healthcare

Medical imaging analysis and patient monitoring benefit from accurate and fast object detection.

Conclusion and Future Directions

SAM2 represents a significant advancement in real-time object detection, offering improved accuracy and efficiency. Future research focuses on further enhancing performance, addressing edge-case scenarios, and exploring new applications.

The Team

Shivansh Fulper	Dhruv Parmar	Ashwin Kothawade	Prasad Ayush
Team Leader	Team Member 1	Team Member 2	Team Member 3