CAP-6411 Computer Vision Systems Assignment 3- SAM2

Karthik Ramasamy
University of Central Florida
ka234388@ucf.edu

1 Introduction

The core requirement is to implement the Segment Anything Model 2 (SAM2) without any manual intervention, such as clicking points or drawing boxes. I first establish a simple **baseline** algorithm that uses a zero-shot object detector to provide textual prompts and then introduce an **improved** algorithm with enhanced techniques designed to outperform this baseline. Performance is quantitatively measured and compared using the Dice similarity coefficient.

2 Dataset and Technique used

The Cambridge-Driving Labeled Video Database (CamVid) validation set.

Baseline Pipeline: I have used OWL-ViT v2 to generate bounding boxes from simple, single-word text prompts, which are then passed directly to SAM2 to create segmentation masks.

- OWL-ViT (OWLv2) → boxes → SAM2 → masks
- Detector: HuggingFace OWL-ViT zero-shot object detector.
- Takes an image + text prompts (e.g., ”person”, ”car”, ”bus”).
- SAM2 uses the bounding boxes from OWL-ViT as input “prompts” and produces pixel-level masks for each region.

Improved Pipeline: I have used yolov8 along with the OWLv2 model:

- OWL-ViT predicts boxes from text prompts (zero-shot).
- YOLOv8 predicts boxes using a pretrained detector.
- Then i merge the boxes from both models before passing them to SAM2 for mask generation.

  OWL-ViT + YOLOv8 → merged boxes → SAM2 → final masks

3 Insights

3.1 Baseline vs Improved Pipeline

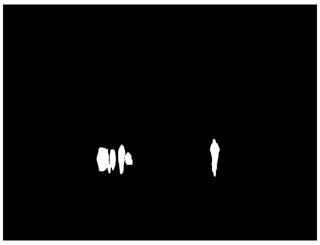
Baseline Pipeline:

- **Simple Prompts:** Uses basic, single-word text prompts.
- **Finds Objects:** The OWL-ViT model finds the general location of objects based on these simple prompts.
- **Direct Segmentation:** SAM2 takes the exact bounding boxes from OWL-ViT and creates a mask.

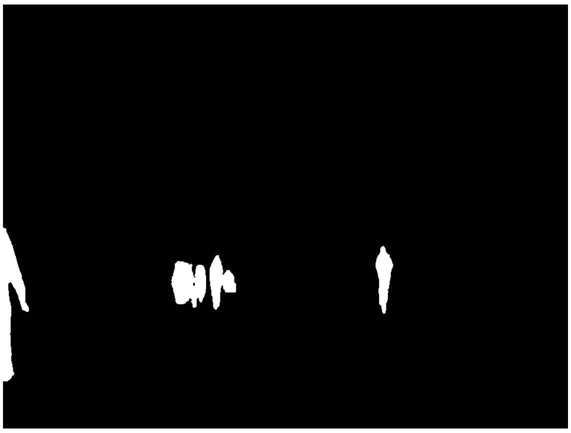
Improved Pipeline:

- **Enhanced Prompts:** Uses a smarter, more descriptive list of prompts to detect a wider variety of objects.
- **Tuned Thresholds:** Applies carefully chosen confidence scores for each class (stricter for vehicles, more lenient for people) to balance object detection and reduce errors.
- **Box Jittering:** Algorithmically generates slightly larger or adjusted versions of detected bounding boxes to give SAM2 more context, helping it capture entire objects.
- **Mask Refinement:** After masks are created, a final clean-up step removes small noisy regions and fills holes to produce smoother, more accurate masks.

4 Results and Comparisons



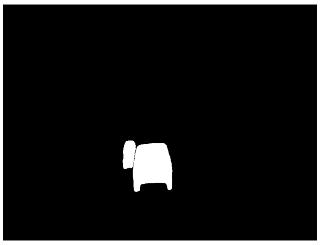
PEOPLE BASELINE PIPELINE MASK



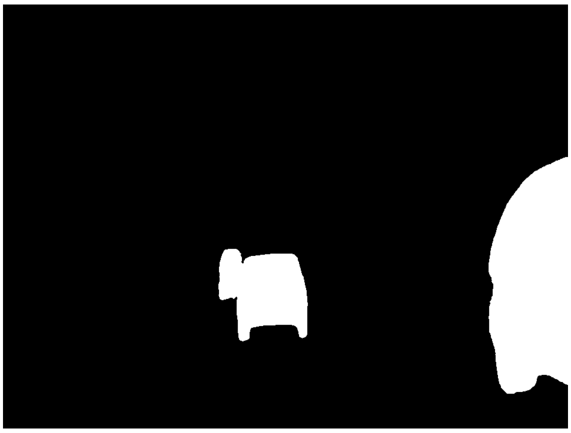
PEOPLE IMPROVED PIPELINE MASK

The figure compares the people masks generated by the baseline and improved pipelines. In the baseline output, only a few small fragments of people are detected, resulting in incomplete and unreliable segmentation.

By contrast, the improved pipeline produces larger and more consistent masks that clearly capture multiple pedestrians in the scene. This demonstrates that the proposed enhancements significantly improve the recall and coverage of people segmentation, addressing one of the main weaknesses of the baseline approach.



VEHICLE BASELINE MASK



VEHICLE IMPROVED MASK

Here the above vehicle masks illustrate how the baseline and improved pipelines perform on cars in the scene. In the baseline output, only the central vehicle is segmented, while part of the nearby car is missed.

The improved pipeline not only preserves the segmentation of the main vehicle but also extends coverage to include the additional car on the right side, providing a more complete representation of the scene. This shows that the improved method enhances recall for vehicles, although the baseline was already fairly strong.

4.1 Comparision between Dice Scores for Person and Vehicle

4.1.1 Overall Performance Summary

Table 1 shows the mean Dice scores for the entire validation set, comparing the baseline and improved pipelines.

Explanation:

- **Baseline Mean Dice:** The average Dice score using the baseline pipeline (OWL-ViT → SAM2). For people, the baseline achieved 0.6071 and for vehicles, 0.5625.

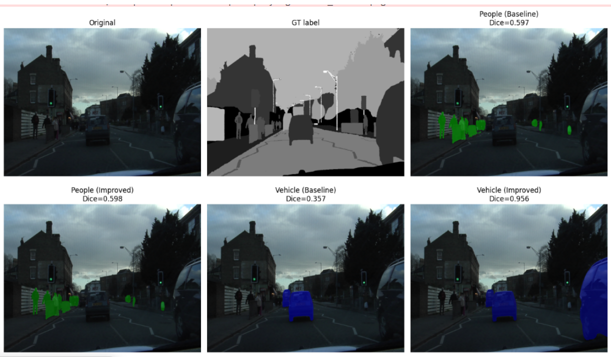
Category	Baseline MD	Improved MD	M Gain
People	0.6071	0.6107	**0.0037**
Vehicle	0.5625	0.7022	**0.1397**

Table 1: Mean Dice scores for baseline and improved pipelines across the validation set.

- **Improved Mean Dice:** Average Dice score using the improved pipeline (OWL-ViT + YOLOv8 ensemble + SAM2 + refinement). The score slightly improved for people (0.6107) and significantly for vehicles (0.7022).
- **Mean Gain:** Difference between improved and baseline Dice scores, showing the benefit of the improved pipeline. The gain is marginal for people (+0.0037) and substantial for vehicles (+0.1397).

Conclusion:

- In the code execution people segmentation shows minor improvement, indicating imporved pipeline and for vehicle segmentation benefits greatly from the improved pipeline, with ensemble detection and mask refinement boosting performance significantly.
- Overall, the improved pipeline demonstrates the effectiveness of combining multiple detectors and refining masks, particularly for challenging object classes.



EVALUATION RESULT OF SAMPLE IMAGE