BEST: Benchmarking and Evaluation of Segmentation and Tracking for Autonomous Vehicles in Adverse Conditions

submitted in partial fulfillment of the requirements

for the degree of

MASTER OF TECHNOLOGY

in

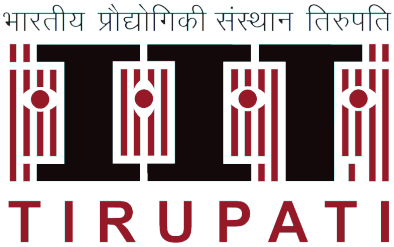
COMPUTER SCIENCE AND ENGINEERING

by

GOWLIKAR AJITESH CS24M118

Supervisor(s)

Dr. Chalavadi Vishnu



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY TIRUPATI

November 2025

DECLARATION

I declare that this written submission represents my ideas in my own words, and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented, fabricated, or falsified any idea, data, fact, or source in my submission to the best of my knowledge. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited or from whom proper permission has not been taken when needed.

Place: Tirupati
Date: 29-11-2025

Signature
GOWLIKAR AJITESH
CS24M118

BONA FIDE CERTIFICATE

This is to certify that the **BEST: Benchmarking and Evaluation of Segmentation and Tracking for Autonomous Vehicles in Adverse Conditions**, submitted by **Gowlikar Ajitesh**, to the Indian Institute of Technology, Tirupati, for the award of the degree of **Master of Technology**, is a bona fide record of the project work done by him under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Place: Tirupati
Date: 19-05-2019

Prof Dr. Chalavadi Vishnu
Guide
Assistant Professor
Department of CSE
IIT Tirupati - 517619

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my guide, **Dr. Ch. Vishnu**, for his constant support, insightful feedback, and encouragement throughout the course of this work. His guidance inspired me to explore the potential of modern foundation models, particularly the Segment Anything Model 2 (SAM2), in advancing zero-shot object segmentation.

I am also thankful to my peers and the Department of Computer Science and Engineering, IIT Tirupati, for providing a research-driven environment and the necessary resources that facilitated this study.

ABSTRACT

Adverse weather severely degrades the reliability of modern segmentation models, which often miss critical objects or generate unstable masks when fog, rain, snow, or nighttime lighting obscure visual cues. In our work, we investigate how to stabilize and enhance segmentation performance by combining two complementary models: YOLOv8 for object localization and SAM2 for prompt-guided, high-resolution segmentation. Initial experiments revealed that zero-shot SAM2 frequently overlooked visible objects and that YOLO's off-the-shelf detections were incomplete in challenging scenes. To address this, we fine-tuned YOLOv8 on the ACDC dataset to improve detection coverage across all 19 semantic classes. These refined detections were then used as high-quality prompts to fine-tune SAM2, enabling it to recover full object masks even when the underlying image structure was heavily degraded. The resulting pipeline detection guided segmentation with temporal memory achieves more reliable per-frame masks and significantly improves robustness under adverse weather. Our study highlights that targeted prompt refinement and model-aware fine-tuning can substantially extend the practical usefulness of foundation models in real-world driving conditions.

KEYWORDS: ACDC, SAM2, YOLO, zero-shot, prompts, object mask, Semantic

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ACDC	Adverse Conditions Dataset for semantic segmentation
AP	Average Precision
FPS	Frames Per Second
FT	Fine-Tuning
IoU	Intersection over Union
IITT	Indian Institute of Technology Tirupati
LR	Learning Rate
mAP	Mean Average Precision
mIoU	Mean Intersection over Union
PR	Precision–Recall
RGB	Red, Green, Blue color channels
SAM	Segment Anything Model
SAM2	Segment Anything Model 2
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
YOLO	You Only Look Once
YOLOv8	You Only Look Once, version 8

NOTATION

General Notations

$\mathbf{I}$	Input RGB image
H, W	Image height and width
$\mathbf{M}$	Predicted segmentation mask
M_i	Mask for object i
p_{ij}	Pixel value at location (i, j)
$\mathscr{R}$	Region of interest (ROI)
B	Batch size
c	Class label
$\hat{B}$	Predicted bounding box
N	Number of detected objects
σ	Sigmoid activation used in YOLO head
$\mathscr{L}_{box}$	Bounding box regression loss
$\mathscr{L}_{cls}$	Classification loss
$\mathscr{L}_{dfl}$	Distribution Focal Loss (DFL)
$\mathbf{Z}$	Output token embeddings (mask tokens)

Hybrid Model Notations

$\mathscr{D}_{YOLO}$	YOLO detections used as prompts for SAM2
$\mathscr{S}_{SAM2}$	Set of segmentation masks produced by SAM2
$\mathscr{H}$	Hybrid pipeline output (YOLO + SAM2)
P_i	Prompt corresponding to object i
M_i	SAM2 mask conditioned on YOLO prompt P_i

Evaluation Metrics

IoU	Intersection over Union
mIoU	Mean Intersection over Union
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
mAP_{50}	Mean Average Precision at IoU threshold 0.50
$\text{mAP}_{50:95}$	Mean Average Precision averaged from 0.50 to 0.95
TP, FP, FN	True Positives, False Positives, False Negatives

Miscellaneous Symbols

θ	Model parameters
η	Learning rate
∇	Gradient operator
$\mathscr{L}$	Total training loss
$\mathbb{R}$	Real number space

CHAPTER 1

INTRODUCTION

Semantic segmentation in autonomous vehicles requires models that can reliably identify objects across highly variable real-world conditions. While recent foundation models such as SAM2 [2] have shown strong performance in both image and video segmentation, their behavior in challenging environments—especially under adverse weather conditions—remains underexplored. Autonomous vehicles and safety-critical systems demand consistent object recognition despite fog, rain, snow, night-time illumination changes, and visual degradation. This project investigates whether SAM2 can maintain segmentation quality under such conditions and explores ways to enhance its robustness.

1.1 Motivation

Although SAM2 offers impressive generalization capabilities, our experiments revealed several critical limitations when deployed on scenes with low visibility or clutter:

Forgetting Previously Seen Objects: SAM2 occasionally loses track of objects across frames, especially when visibility drops or objects undergo partial occlusion.

Confusion Between Similar-Looking Objects: In visually degraded conditions (dense fog, night scenes), SAM2 often merges or confuses similar classes—e.g., pedestrians vs. signboards, or vehicles in low contrast.

Poor Segmentation Under Adverse Weather: SAM2's pre-training on mostly clean and well-lit imagery makes it struggle with scenes from datasets such as ACDC, where fog, rain streaks, snow, and night illumination drastically distort object appearance.

From these observations, we identified multiple potential problem statements: improving long-term tracking, reducing class confusion, and enhancing segmentation robustness. After evaluating all these directions, we decided to focus primarily on segmentation in adverse weather conditions, as this is the most fundamental failure point and directly impacts safety in autonomous navigation.

1.2 Problem Statement

Despite its large-scale training and architectural strengths, SAM2 fails to produce reliable segmentation under adverse weather. Visual degradations cause:

- Boundary distortion, where object edges become unclear and masks appear coarse or incorrect
- Missed detections, especially for small objects like pedestrians, poles, and distant vehicles
- Color confusion, where fog/snow alters pixel intensities, leading SAM2 to misclassify or ignore objects
- Loss of temporal consistency, where objects segmented correctly in one frame are lost in the next

These issues indicate that SAM2's learned features are not sufficiently robust to the domain shift introduced by fog, rain, night-time, and other adverse conditions. Since autonomous driving must operate in all weather scenarios, improving SAM2's resilience is essential.

1.3 Objectives

Our primary objective is to enhance SAM2's segmentation performance in adverse weather conditions and explore strategies that improve its adaptability to degraded inputs. Direct fine-tuning of SAM2 is non-trivial due to:

- extremely large model size
- memory-dependent architecture
- complex prompt-based and mask-token mechanisms
- limited ground-truth masks for fine-tuning in adverse scenarios

Therefore, our approach focuses on designing a practical, efficient enhancement strategy that improves SAM2 without requiring full-scale fine-tuning.

Additional objectives include:

Segment as many objects as possible under adverse scenes

Increase mask quality and boundary accuracy

Preserve consistency across frames (if extended to videos)

Build a modular pipeline scalable to future fine-tuned or hybrid models

Lay groundwork for a model capable of long-term robust tracking

1.4 Scope of the Work

The scope is defined in multiple stages:

- Stage 1 – Image-Level Evaluation and Enhancement (Current Work): Focus on image segmentation performance under fog, rain, snow, and night conditions using the ACDC dataset. If the improved method produces clearer, more accurate masks, it confirms the feasibility of upgrading SAM2 for harsh environments.
- Stage 2 – Extension to Videos (Planned): If image-level improvements succeed, the methodology can be adapted to videos by utilizing SAM2's memory components. This would allow:
  - long-term object tracking
  - occlusion recovery
  - consistent mask propagation
  - improved temporal stability under varying weather
- Stage 3 – Toward a Fully Robust Tracking System (Future Scope)
  - A highly stable SAM2-based model that:
  - works in any weather condition
  - can track any object with minimal failure
  - maintains identity consistently until the end of the sequence

Such a system would be valuable not only for autonomous driving but also for military reconnaissance, border surveillance, and drone-based monitoring, where robust segmentation in adverse conditions is critical.

CHAPTER 2

Literature Review

The Semantic segmentation and video object tracking have evolved rapidly with the introduction of large-scale foundation models and memory-driven architectures. SAM2 represents the latest advancement in this direction by unifying image and video segmentation within a single framework. Its design integrates an image encoder, prompt encoder, and a dedicated memory encoder that stores compact object representations over time. The introduction of memory attention and a long-term memory bank enables SAM2 to maintain object consistency across challenging frames containing motion, occlusion, or appearance changes, making it highly suitable for real-world scenarios.

Another foundational approach in this space is the STM (Space–Time Memory) network, which introduced the concept of a key–value memory structure for video segmentation. STM allows the model to reference past frames through attention-based retrieval, enabling reliable tracking under heavy occlusion or rapid object deformation. STM laid the conceptual groundwork for subsequent memory-based architectures such as XMem, AOT, and SAM2.

To evaluate segmentation performance in challenging environments, datasets designed for autonomous driving play a crucial role. The ACDC [3] dataset focuses specifically on adverse weather conditions including fog, rain, snow, and night-time scenes. It provides high-quality, pixel-level semantic annotations from real-world settings, making it an ideal benchmark for testing the robustness of segmentation models under visibility degradation—an essential requirement for autonomous vehicle perception.

For semantic segmentation baselines, transformer-based models such as Mask2Former [1] and SegFormer have become widely adopted due to their strong performance and architectural efficiency. These models leverage global self-attention and multi-scale feature processing to deliver competitive results across diverse conditions. They serve as important comparison points for analyzing the benefits of memory-driven approaches like SAM2, particularly in settings where temporal context or long-term consistency is required.

CHAPTER 3

Proposed Methodology

The overall architecture of the proposed framework is shown in **??**. The YOLOv8 is used to process points or boxes to adapt SAM prompt input. In SAM2, the decoder never uses a raw image embedding; it receives a version already conditioned on past predictions and any prompted frames, which may even come from later time steps. After each frame is segmented, the memory encoder stores a compact representation in the memory bank. When a new frame arrives, memory attention fuses its fresh embedding with these stored memories before passing the result to the mask decoder for the final prediction.

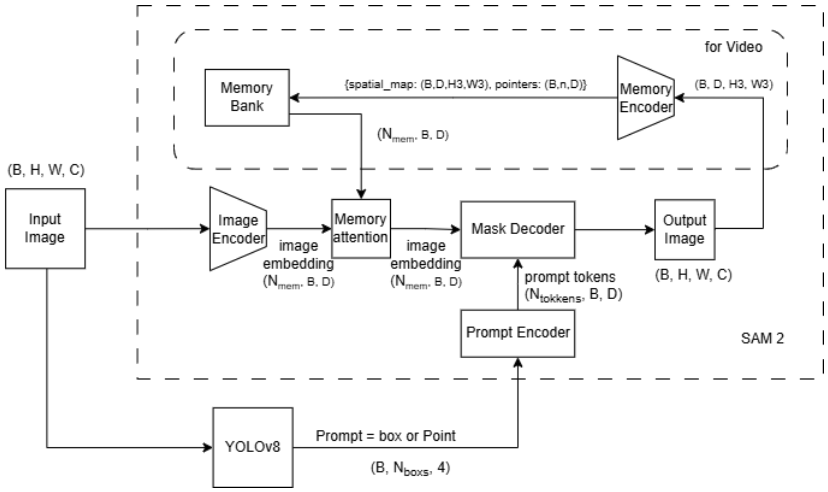


Figure 3.1: our proposed YOLOv8 + SAM2 architecture

3.1 YOLOv8 — Detection and Prompt Source

The YOLOv8 module first localizes objects, effectively deciding where the subsequent SAM2 model should direct its attention for segmentation. While its architecture, with its efficient Backbone, feature-refining Neck, and anchor-free Head, builds convincingly upon the YOLO legacy to balance speed with robust accuracy, its true role in our pipeline is more nuanced than mere detection. We use it to generate simple bounding boxes from frames resized with preserved aspect ratio but these detections aren't our final product. Instead, these bounding boxes are used as prompts for SAM2. This is crucial because SAM2, is fundamentally a prompt-driven

segmenter. without clear spatial guidance, it can easily be distracted by visual noise like fog artifacts or misleading shadows in adverse weather.

3.2 Image Encoder

Our system uses the SAM2.1 image encoder as its primary visual backbone. It relies on a hierarchical transformer paired with an FPN to extract features at multiple resolutions. We fuse the stride-16 and stride-32 stages to form the main image embedding, since these deeper features capture stable semantic structure—something especially useful when fog, rain, or nighttime lighting hide fine details. The higher-resolution stride-4 and stride-8 features are handled differently: rather than sending them through the memory pathway, which tends to wash out detail, we feed them directly into the decoder's upsampling layers. This skip-connection path helps recover crisp boundaries around thin structures such as poles, pedestrians, or road edges. For positional encoding, we use windowed absolute embeddings and interpolate global positional information across windows, avoiding the complexity of relative positional encoding. The encoder scales across model sizes (T, S, B+, L) and employs global attention in only a selected subset of layers to keep the compute manageable.

3.3 Memory Attention

The memory attention module integrates past information into the current frame's processing. It uses sinusoidal absolute positional embeddings along with 2D spatial RoPE in both self-attention and cross-attention layers, giving the model a stable sense of spatial layout across frames. We intentionally exclude the object-pointer tokens from RoPE because they do not correspond to any fixed spatial position. The default configuration uses four memory-attention layers, which is sufficient to maintain object identity without overwhelming the system. This component is responsible for aligning the current image features with the stored memory features, allowing the model to continue tracking objects even when they reappear after partial occlusion or poor visibility.

3.4 Prompt Encoder and Mask Decoder

The prompt encoder follows the original SAM design, turning user-provided or detector-generated prompts into tokens that guide the segmentation process. The mask decoder, however, incorporates several additions specific to our video setting. We repurpose the output mask token as the object-pointer token for that frame and store it in the memory bank. To handle occlusions, we introduce an additional occlusion-prediction token; an MLP head evaluates whether the object is visible or hidden in the current frame. If the model predicts an occlusion event, we attach a learned occlusion embedding to that frame's memory entry. The decoder also benefits from the high-resolution stride-4 and stride-8 encoder features, which are injected during upsampling and help refine object boundaries. SAM2's multi-mask behavior is retained: when ambiguity arises, the decoder produces multiple candidate masks and selects the one with the highest IoU score for propagation.

3.5 Memory Encoder and Memory Bank

The memory encoder does not rely on an additional image backbone; instead, it reuses the embeddings already produced by the main image encoder. These are combined with the predicted mask to form memory tokens that capture both appearance and geometry. Each memory token is projected to 64 dimensions before being stored in the memory bank, keeping the memory operations lightweight. The 256-dimensional object-pointer token is split into four 64-dimensional tokens when interacting with the memory bank, making cross-attention efficient. When multiple objects appear in a video, the image encoder is shared across all of them, but each object maintains its own memory bank and decoder. This design maintains object-specific consistency without duplicating expensive visual processing.

Fine tuning

Our initial experiments made the limitations of zero-shot models very clear. YOLOv8 performed reasonably well out of the box, but its detections were far from complete—especially in fog and nighttime scenes, where several objects were either missed or incorrectly localized. SAM2 struggled even more. While it could segment many structures in the frame, it consistently

failed on some of the very objects that were visually obvious to humans. These gaps were not random: whenever YOLO failed to propose a bounding box, SAM2 had no useful prompt and therefore ignored the object entirely. This pushed us toward a more deliberate refinement process. We first fine-tuned YOLOv8 on the ACDC dataset so that it learned to detect a richer and more weather-aware set of objects across all 19 classes. Once the detector became more reliable, its improved bounding boxes served as higher-quality prompts for SAM2. This enabled the second stage of our pipeline: fine-tuning SAM2 using the refined YOLO outputs and the ground-truth semantic masks. By training SAM2 on object-specific, box-guided prompts rather than relying on its purely zero-shot behavior, we aimed to teach the model how to recover full, high-resolution masks even when the underlying scene was degraded. In effect, fine-tuning YOLO enhanced the prompting signal, and fine-tuning SAM2 allowed the segmentation model to adapt to ACDC's unique distortions—closing the loop between detection and segmentation in a way neither model could achieve alone.

CHAPTER 4

Experimentation and Results

This section presents the experimental evaluation of YOLOv8, SAM2, and our proposed hybrid YOLO+SAM2 pipeline under adverse weather conditions. Since fog severely degrades visibility, we use a representative fog-frame from the ACDC dataset to illustrate performance differences across models.

4.1 Quantitative Results

Table 4.1: Performance comparison of YOLOv8 before and after fine-tuning on fog weather images. Fine-tuning significantly improves detection quality across all metrics.

Metric	Before Fine-Tuning	After Fine-Tuning	Improvement
Precision	0.41	0.75	+0.34
Recall	0.21	0.49	+0.28
mAP@50	0.22	0.54	+0.32
mAP@50-95	0.10	0.34	+0.24

4.2 Qualitative Results

Table 4.2: Qualitative assessment of segmentation performance of SAM2 vs. the proposed YOLO+SAM2 hybrid model.

Model	Large Object Recall	Small Object Recall	Segmentation Quality
SAM2 Only	High	Low	Misses poles, trees, and distant objects.
YOLO+SAM2 (Before FT)	High	Medium	Uses YOLO but still limited by misses.
YOLO+SAM2 (After FT)	High	High	Clearer and perfect segmentation.

4.3 Segmentation Results

The segmentation results on fog weather conditions on different models and our proposed models.

4.3.1 Input Image

Figure 4.1: Raw input image under fog weather conditions.

4.3.2 YOLOv8 Prediction (Before Fine-Tuning)

Figure 4.2: YOLOv8 detection results before fine-tuning.

4.3.3 SAM2 Segmentation Output

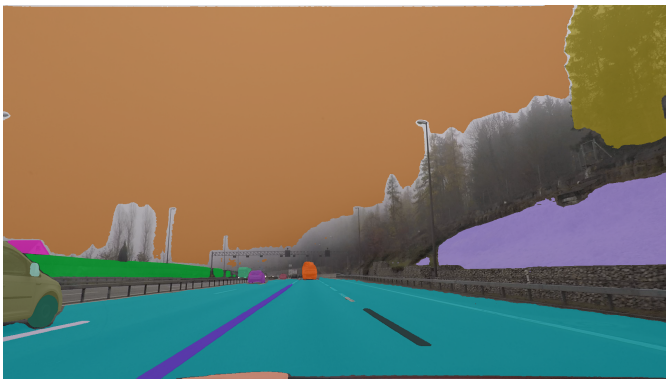


Figure 4.3: SAM2 segmentation on fog conditions (without prompts).

4.3.4 Our Hybrid Model (YOLO + SAM2)

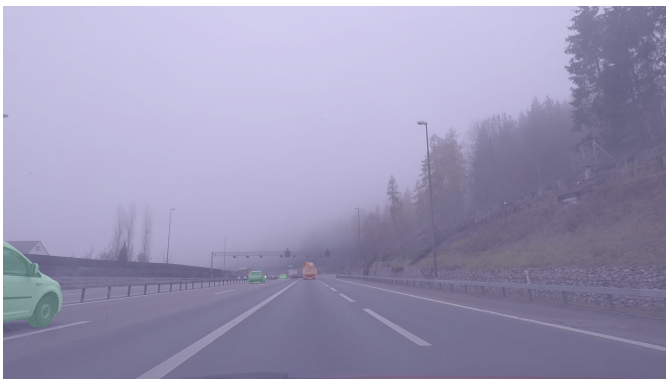


Figure 4.4: Hybrid YOLO+SAM2 output before YOLO fine-tuning.

Observation: Under fog, YOLOv8 detects mostly cars while missing pedestrians, poles, and distant objects. SAM2 segments broad regions (road, cars) but fails to segment trees, poles, and small structures. The hybrid model improves object recall slightly, but ultimately inherits YOLO's missed detections.

4.4 Fine-Tuning YOLOv8

To address poor detection quality, we fine-tuned YOLOv8 by unfreezing the last 8 layers while keeping the remaining layers frozen. This allows the model to adapt to fog-specific visual cues without overfitting.

4.4.1 Training and Validation Visualization

The below are visualization of training and validation.

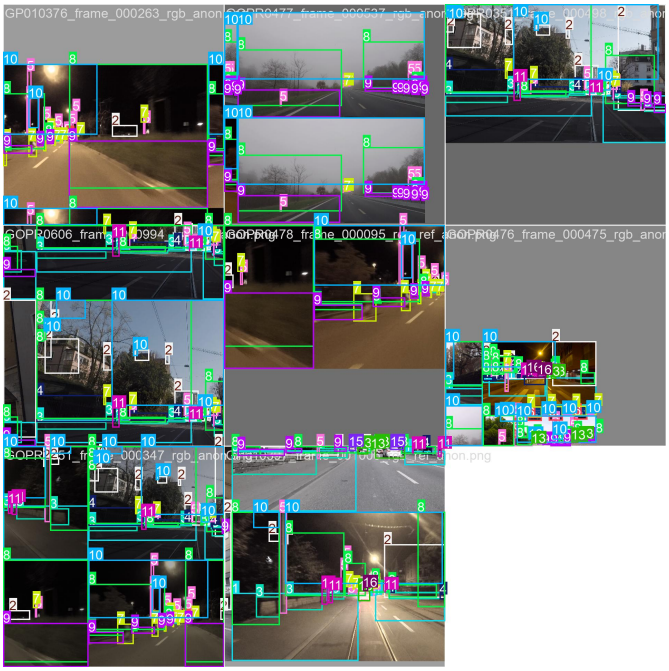


Figure 4.5: Sample training batch during YOLOv8 fine-tuning.

Figure 4.6: YOLOv8 predictions on validation samples.

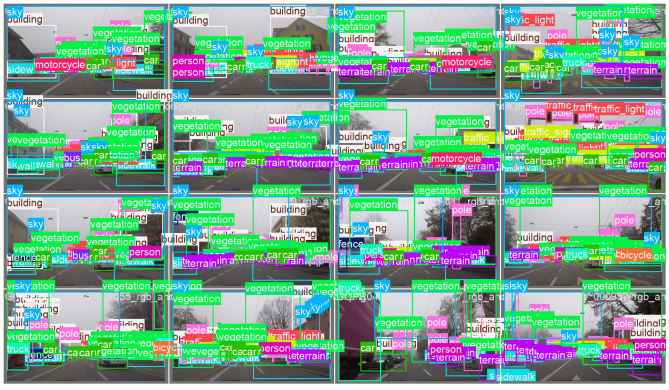


Figure 4.7: Ground-truth labels for validation samples.

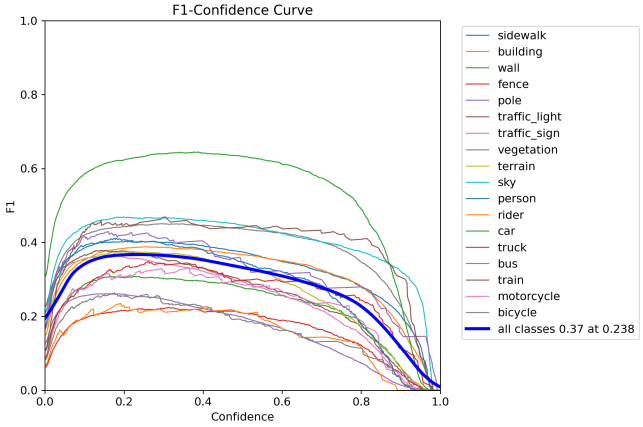


Figure 4.8: Box F1-score curve for fine-tuned YOLOv8.

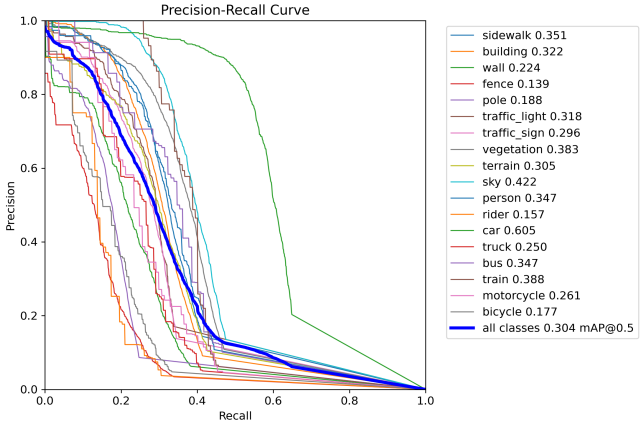


Figure 4.9: Precision–Recall curve for object detection.

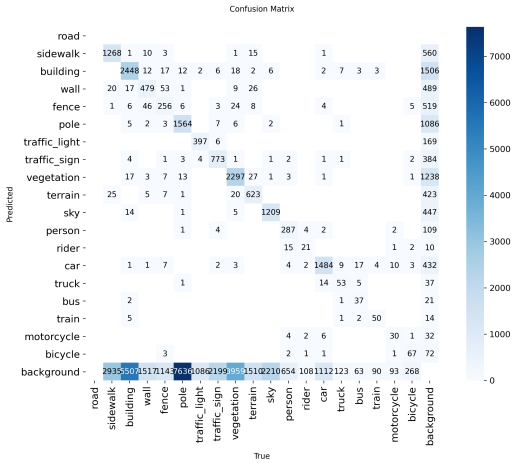


Figure 4.10: Confusion matrix (unnormalized).

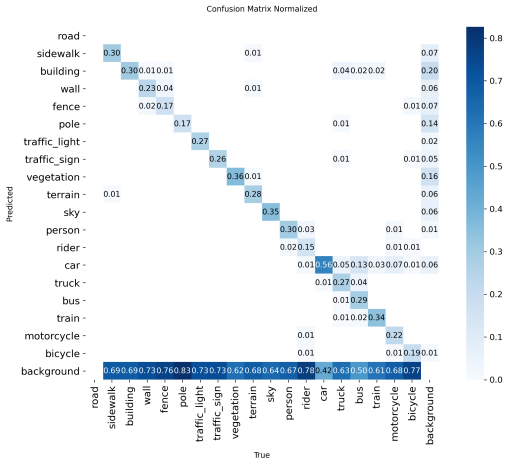


Figure 4.11: Confusion matrix (normalized).

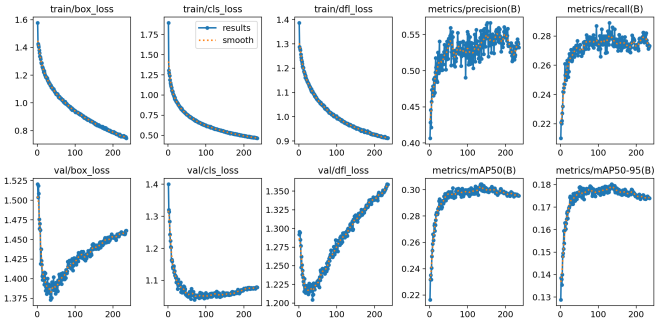


Figure 4.12: Training curves for fine-tuned YOLOv8.

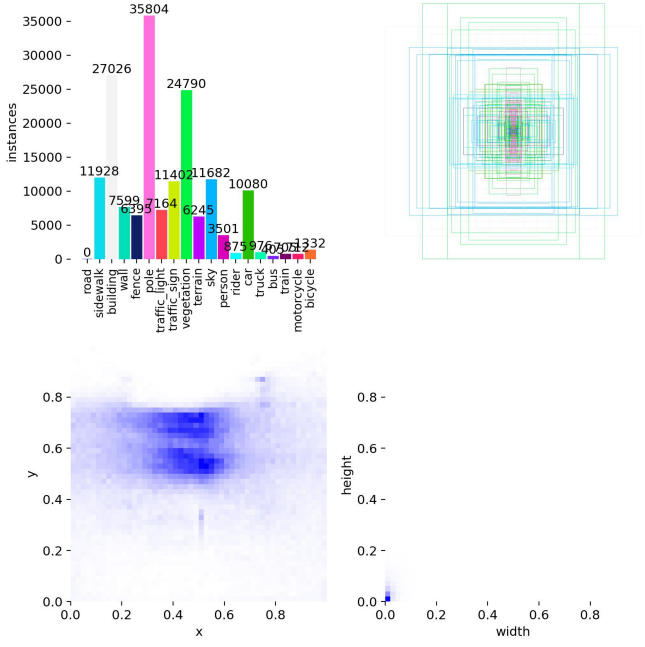


Figure 4.13: Class frequency distribution in the training dataset.

Figure 4.14: YOLOv8 output after fine-tuning on fog data.

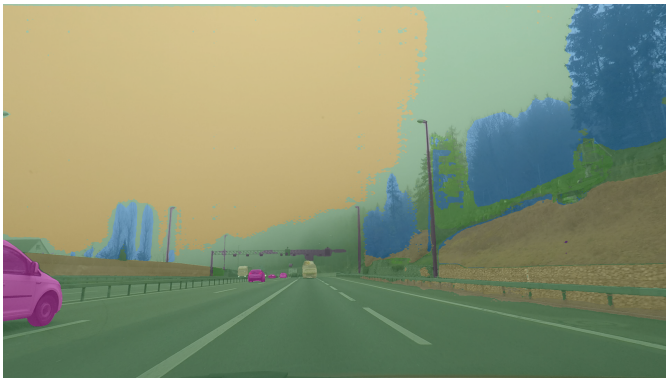


Figure 4.15: Hybrid YOLO+SAM2 segmentation after YOLO fine-tuning.

4.5 Post Fine-Tuning Results

4.5.1 YOLOv8 After Fine-Tuning

4.5.2 Hybrid Model After Fine-Tuning

Final Observation: After fine-tuning, YOLOv8 detects significantly more objects under fog, including poles, pedestrians, and distant vehicles. As a result, the hybrid YOLO+SAM2 model also produces much richer and more accurate segmentation masks, particularly for small and fog-obscured objects.

CHAPTER 5

Conclusion and Future works

Looking ahead, this project opens up more questions than it answers. We've seen that SAM2 behaves almost like two different models depending on the weather: confident and clean in normal daylight, and then surprisingly fragile once fog, rain, or nighttime glare enter the frame. After fine-tuning, the model becomes noticeably better with adverse-weather images, but only for images. The moment we start thinking about videos—continuous motion, objects that disappear and reappear, headlight flares, raindrops sticking to the camera lens—the weaknesses show up again.

One obvious direction is to scale our current system from single-frame inference to truly stable video-level tracking. SAM2's Video Predictor is promising, but it still has a tendency to forget what it was following a few frames earlier, especially if the object gets occluded or rotates out of the camera's field for a moment. There are also those awkward moments where two similar-looking cars pass close together and SAM2 mixes up their identities, handing the mask of one object to another as if nothing happened. It doesn't completely break the system, but it's enough to make the tracking look unreliable.

Another area worth exploring is multi-object handling. At the moment, the model manages a handful of vehicles reasonably well, but the cracks start to show in busier urban scenes—think a four-way junction on a rainy evening where headlights overlap and motion blur becomes unavoidable. In such cases, the model simply loses track of some objects or merges them accidentally. Improving this might require integrating a dedicated multi-object tracker, or experimenting with transformer-based temporal attention modules that help the model "remember" better across frames.

In short, while the current system works reasonably well for controlled experiments, a stronger, more resilient version would require a mix of better tracking, richer data, improved temporal modeling, and possibly multi-modal sensing. Each of these directions feels like a natural extension of what we've already built, and tackling them could bring the system much closer to real-world reliability.

REFERENCES

[1] **B. Cheng**, **I. Misra**, **A. G. Schwing**, **A. Kirillov**, and **R. Girdhar**, Masked-attention mask transformer for universal image segmentation. 2022.

[2] **N. Ravi**, **V. Gabeur**, **Y.-T. Hu**, *et al.* (2024). Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*. URL https://arxiv.org/abs/2408.00714.

[3] **C. Sakaridis**, **D. Dai**, and **L. Van Gool**, ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.