XenSense-V1: A Deep Learning-Based Framework for Video Segmentation and Semantic Labelling in Autonomous Driving Systems

ARJUN PRABHAKARAN
1BM22CS053

AYMAN AMJAD
1BM22CS061

BHANODAY KURMA
1BM22CS066

BHANU PRAKASH
1BM22CS067

Ms. MAYANKA GUPTA
Assistant Professor
Department of Computer Science and Engineering
B.M.S. College of Engineering

OUTLINE

1. Introduction
2. Problem Statement
3. Objectives
4. Sustainable Development Goals
5. Collaboration
6. Literature Survey
7. Research Gap Identification
8. High Level Design
9. Technology
10. Datasets
11. Gantt chart for Major Project Phase 1
12. References

1.Introduction

Automatic segmentation and labeling of objects in videos is a crucial task in computer vision, enabling machines to detect, classify, and differentiate objects in a video stream. This process is widely used in autonomous driving, medical imaging, surveillance, and augmented reality.

- By leveraging deep learning models, such as **Convolutional Neural Networks (CNNs) and Transformer-based architectures**, along with traditional image processing techniques, the system can automatically identify objects, assign labels, and track them across frames. The goal is to create a robust framework that can process real-time video streams and provide meaningful segmentation results with minimal human intervention

2.Problem Statement

- Safe autonomous driving requires accurate and trustworthy knowledge of road scenes, particularly in the intricate and uncertain conditions present on Indian roads. When confronted with obstacles like potholes, speed bumps, and non-standard vehicles, as well as frequent occlusions from fog, smoke, or traffic, traditional object detection techniques frequently fall short. In addition to being time-consuming, manually annotating large amounts of video data is not feasible for real-time deployment.
- Advanced, automated video segmentation and semantic labeling solutions that can adjust to the particularities of Indian transport systems are desperately needed. In addition to lowering human error and effort, such systems will allow for more reliable, context-aware decision-making for autonomous cars of the future.

3.Objectives

- **Develop a specialized video segmentation model for Indian roads**-To precisely segment and differentiate automobiles, pedestrians, potholes, speed bumps, and other road anomalies in actual Indian traffic situations, develop and deploy XenSense-V1, a sophisticated deep learning framework built on the YOLOv8 architecture.
- **Enhance semantic labelling and object tracking under challenging conditions**- Incorporate powerful tracking and classification modules to automatically identify and track objects, guaranteeing dependable operation even when there are smoke, fog, or partial visibility occlusions.
- **Achieve real-time, adaptive perception for autonomous driving**- Use hardware acceleration, lightweight architectures, and effective frame analysis to optimize the XenSense-V1 pipeline for real-time video processing, allowing for smooth operation in dynamic, uncertain environments.
- **Incorporate memory-based and prompt-based detection capabilities**- Give the XenSense-V1 short-term memory modules and prompt-guided detection so it can remember past scenes when the sensor is obscured and adjust to new object types or user inquiries, increasing flexibility and safety.

Sustainable Development Goals (SDGs) Addressed

1. SDG 9: Industry, Innovation, and Infrastructure

- **Justification:** The project promotes innovation by integrating AI and automation into video analytics. Industries such as manufacturing, autonomous vehicles, and robotics benefit from real-time object recognition and tracking, leading to safer, smarter, and more efficient operations.

2. SDG 11: Sustainable Cities and Communities

- **Justification:** In smart city applications, automated object segmentation can be used for traffic management, surveillance, and pedestrian safety. It helps monitor road conditions, detect accidents, and improve urban planning for safer and more sustainable cities.

5.Collaboration

- The project belongs to the category of Community Oriented Project and Multidisciplinary Project

- This project is being developed in collaboration with **MathWorks**, leveraging their expertise in **MATLAB, Simulink, and AI-driven computer vision tools** for real-time object segmentation and labeling in video streams.

Literature Survey

Sl.No	Paper Title	Year	Problem Statement Addressed	Methodology Followed	Quantitative Results	Future Work Proposed	Complete Reference
1	Robust Object Detection under Occlusion with Context-Aware CompositionalNets	2020	Object detection under strong occlusion and context bias	Part-based voting and context disentanglement in CompositionalNets	+41% (PASCAL3D+), +35% (MS-COCO)	Improve occlusion robustness via better context modeling	Wang, A., Sun, Y., Kortylewski, A. , & Yuille, A. (2020). Robust Object Detection under Occlusion with Context-Aware CompositionalNets. arXiv:2005.11643
2	End-to-End Video Instance Segmentation with Transformers (VisTR)	2021	Video segmentation + tracking in an end-to-end manner	Transformer-based instance sequence prediction	SOTA accuracy & speed on YouTube-VIS	Apply Transformer-style decoders to broader video understanding tasks	Wang, Y. et al. (2021). End-to-End Video Instance Segmentation with Transformers. arXiv:2011.14503
3	Learning to See the Invisible	2018	Predicting invisible (amodal) parts of occluded objects	ORCNN model for joint amodal/visible/ invisible segmentation	SOTA on COCOA, COCOA cls, and D2SA	Improve invisible mask prediction, reduce false positives	Follmann, P. et al. (2018). Learning to See the Invisible. arXiv:1804.08864

Literature Survey

Sl.No	Paper Title	Year	Problem Statement Addressed	Methodology Followed	Quantitative Results	Future Work Proposed	Complete Reference
4	Occluded Video Instance Segmentation: A Benchmark	2021	Tracking and segmentation in occluded videos	New OVIS benchmark + temporal feature calibration module	+4.6 AP (MaskTrack R-CNN), +4.1 AP (SipMask)	Enhance occlusion handling via temporal feature modeling	Qi, J. et al. (2021). Occluded Video Instance Segmentation: A Benchmark. NeurIPS Dataset Track. https://songbai.site/ovis
5	A Survey on Deep Learning Technique for Video Segmentation	2022	Overview of deep learning techniques in video segmentation	Comprehensive survey on VOS and VSS pipelines and datasets	Comparative benchmarks of 150+ models	Emphasizes annotation-efficient & adaptive segmentation	Zhou, T. et al. (2022). A Survey on Deep Learning Technique for Video Segmentation. IEEE TPAMI. https://github.com/tfzhou/VS-Survey
6	Occlusion-Aware Video Object Inpainting	2021	Inpainting occluded objects in video	VOIN model: joint shape, flow, and texture completion	Best realism & temporal consistency on YouTube-VOI	Use advanced GAN & attention for better inpainting	Ke, L. et al. (2021). Occlusion-Aware Video Object Inpainting. arXiv:2108.06765
7	FEELVOS: Fast End-to-End Embedding Learning for VOS	2019	Fast VOS without fine-tuning	Embedding-based matching + dynamic segmentation head	71.5% J&F score on DAVIS 2017	Scale to longer videos & complex objects	Voigtlaender, P. et al. (2019). FEELVOS. arXiv:1902.09513
8	Deep Learning for Video Object Segmentation: A Review	2022	Survey of UVOS and SVOS techniques	Categorization by spatial-temporal feature handling	Benchmarks across DAVIS, YouTube-VIS, FBMS	Cross-domain VOS & low-supervision segmentation	Gao, M. et al. (2022). Deep Learning for Video Object Segmentation: A Review. Artif Intell Rev. https://doi.org/10.1007/s10462-022-10176-7

Sl.no	Year Of Publication	Paper Title	Problem statement addressed	Methodology followed	Quantitative Results	Future Work Proposed	Complete reference
1	2023	Tracking Anything with Decoupled Video Segmentation	Enable zero-shot object tracking and grasping using vision foundation models	SAM-based segmentation pipeline + depth map fusion for 3D grasp planning	Demonstrates real-time zero-shot grasping with SAM mask	Improve prompt engineering and mask accuracy in robotics	Cheng, H. K., Yang, L., & Zhang, P. (2023). *Tracking Anything with Decoupled Video Segmentation*. Retrieved from [Google Scholar]
2	2023	Segmenting Moving Objects via an Object Centric Layered Representation	Learn object segmentation in videos using unsupervised layered modeling	Object masks, motion fields, and appearance maps combined with differentiable alpha compositing	Outperforms prior unsupervised methods across video datasets	Combine semantic and motion cues; improve static object handling	Xie, J., Xie, W., & Zisserman, A. *Segmenting Moving Objects via Layered Representation*. Retrieved from [Google Scholar]

Sl.no	Year Of Publication	Paper Title	Problem statement addressed	Methodology followed	Quantitative Results	Future Work Proposed	Complete reference
3	2023	Two-shot Video Object Segmentation	Reduce annotation overhead in video segmentation	Semi-supervised learning using only 2 labeled frames + pseudo-labeling	Competitive with supervised methods while reducing labeling	Explore robustness to occlusions and label noise	Yan, K., & Yang, Y. (2023). *Two-shot Video Object Segmentation*. Retrieved from [Google Scholar]
4	2022	Breaking the “Object” in Video Object Segmentation	Address limitations in VOS when objects deform, merge, or split	Introduces VOST dataset + benchmarks VOS methods under deformation	Benchmarks show existing methods perform poorly on VOST	Encourage robust VOS models for dynamic objects	Tokmakov, P., & Schmid, C. (2022). *Breaking the "Object" in Video Object Segmentation*. Retrieved from [Google Scholar]

Sl.no	Year Of Publication	Paper Title	Problem statement addressed	Methodology followed	Quantitative Results	Future Work Proposed	Complete reference
5	2022	Self-supervised Video Object Segmentation by Motion Grouping	Unsupervised segmentation using motion cues and contrastive learning	Motion-based pseudo labels + student-teacher contrastive learning	Competitive performance on DAVIS and YouTube-VOS	Improve handling of slow/static objects; integrate semantic cues	Yan, C., Lamdouar, H., Lu, E., Zisserman, A., & Xie, W. *Self-supervised VOS by Motion Grouping*. Retrieved from [Google Scholar]
6	2023	ClickVOS: Click Video Object Segmentation	Reduce annotation effort in Video Object Segmentation while maintaining high segmentation accuracy.	Uses a single click in the first frame and a lightweight ABS model with memory to refine segmentation across frames.	Outperforms prior click-based methods on DAVIS and YouTube-VOS with up to 90% less manual annotation.	Integrate language prompts and improve performance under occlusion and fast object motion	Hou, Cheng-An, Chien-Yi Wang, and Yen-Yu Lin. "ClickVOS: Click Video Object Segmentation." *arXiv preprint* arXiv:2310.11819 (2023). Link

Sl.no	Year Of Publication	Paper Title	Problem statement addressed	Methodology followed	Quantitative Results	Future Work Proposed	Complete reference
7	2024	Cutie: Putting the Object Back into Video Object Segmentation	Improve segmentation robustness by shifting from pixel-level to object-level memory and attention mechanisms.	Proposes a Query-based Object Transformer that uses object queries to perform top-down segmentation. Uses a memory module to track and refine masks over time.	Achieved SOTA performance on DAVIS 2017 and YouTube-VOS with improved robustness against distractors.	Explore better object query initialization, dynamic query selection, and efficient transformer design.	Cheng, Ho Kei, et al. "Putting the Object Back into Video Object Segmentation." *CVPR 2024*. PDF
8	2023	MotionTrack: End-to-End Transformer-Based Multi-Object Tracking With LiDAR-Camera Fusion	Improve multi-object video segmentation/ tracking using LiDAR and camera data with a unified end-to-end model.	Proposes a dual-transformer framework: Feature Enhancer + Data Association Transformers for joint detection and tracking across time using sensor fusion.	Outperforms baselines on Waymo Open Dataset; improves detection and tracking fusion performance	Explore real-time deployment, reduce transformer complexity, and adapt to monocular camera-only setups.	Zhang, Ce, et al. "MotionTrack: End-to-End Transformer-Based Multi-Object Tracking With LiDAR-Camera Fusion." *CVPR 2023 Workshops*. PDF

Sl. No.	Paper Title	Year	Problem Statement Addressed	Methodology Followed	Quantitative results	Future Work Proposed	Complete Reference
1	FogGuard: Guarding YOLO Against Fog Using Perceptual Loss	2024	Object detection in foggy weather conditions significantly degrades the performance of autonomous driving systems relying on camera-based perception.	- Used a YOLOv3-based teacher-student network. - Introduced teacher-student perceptual loss for semantic similarity between foggy and clear images. - Used MiDaS to synthesize depth-aware fog on clear images for realistic data augmentation.	- Achieved 69.43% mAP on RTTS, compared to 57.78% with IA-YOLO. - 5× faster inference than IA-YOLO. - Outperformed baselines on both foggy and clear datasets.	- Extend the framework to other adverse weather conditions like rain and snow. - Generalize fog generation techniques using varying atmospheric light models.	Gharatappeh, S., Neshatfar, S., Sekeh, S. Y., & Dhiman, V. (2024). FogGuard: guarding YOLO against fog using perceptual loss. *arXiv preprint arXiv:2403.08939*.
2	Dehazing & Reasoning YOLO: Prior Knowledge-Guided Network for Object Detection in Foggy Weather	2024	Degraded object detection in foggy weather due to poor visibility and reliance on weak visual features; challenge to maintain real-time performance.	-Proposed DR-YOLO, an end-to-end detection model. -Used RSM with atmospheric scattering for training-time dehazing. -Applied RRAM with co-occurrence graphs for object reasoning. -Integrated AFFM for adaptive feature fusion.	Achieved best mAP across all foggy datasets: • 93.04% on VF-test • 68.74% on RTTS (real fog) • Runs at 74.8 FPS (real-time capable).	- Improve model robustness further. - Explore additional types of inter-object relationships beyond co-occurrence graphs.	Zhong, Fujin, et al. "Dehazing & Reasoning YOLO: Prior knowledge-guided network for object detection in foggy weather." *Pattern Recognition* 156 (2024): 110756.
3	Pothole Detection and Assessment on Highways Using Enhanced YOLO Algorithm With Attention Mechanisms	2025	Conventional detection systems fail in identifying small/ clustered potholes under real-world conditions. Manual inspections are slow and expensive.	Uses YOLO with Xception-CBAM. Applies normalization-based data diversification. Area estimation via image processing. Compared with YOLOv1, YOLOv3.	- mAP = **99.21%** - Inference time = **0.028s** - AP = 0.99 - Precision = 0.99 - IoU = 0.86	Integrate depth estimation, deploy on edge devices, use transformer-based models, include multimodal	Rubin, Rufus, et al. "Pothole Detection and Assessment on Highways Using Enhanced YOLO Algorithm With Attention Mechanisms." *Advances in Civil*

Sl. No.	Paper Title	Year	Problem Statement Addressed	Methodology Followed	Quantitative results	Future Work Proposed	Complete Reference
4	3D Object Detection through Fog and Occlusion: Passive Integral Imaging vs Active (LiDAR) Sensing	2022	Analyze and compare effectiveness of passive and active sensing under fog and occlusion for 3D object detection	Compared passive integral imaging and LiDAR using controlled visibility scenarios	Passive method outperformed LiDAR in dense fog; LiDAR was more accurate in clear conditions	Explore hybrid sensing systems combining passive and active techniques	Usmani, K., O'Connor, T., Wani, P., & Javidi, B. (2022). 3D object detection through fog and occlusion: passive integral imaging vs active (LiDAR) sensing. Optics Express, 31(1), 479-491.
5	D-YOLO: A Dual-Branch YOLO Framework for Object Detection in Adverse Weather Conditions	2024	Object detection performance degrades significantly in adverse weather (fog, rain, snow) due to visual distortion and feature loss in input images.	Introduced a dual-branch network: one branch processes hazy images, the other uses dehazed features extracted using a Clear Feature Extraction (CFE) module. Features are fused using an Attention Feature Fusion (AFF) module with Omni-Dimensional Dynamic Convolution (ODConv). The CFE branch is removed at inference for efficiency.	Achieved 72.1% mAP on Foggy Cityscapes dataset. Outperformed baseline YOLO variants. Maintained real-time inference (~42.3 FPS).	Extend the framework to improve occlusion handling and explore deeper temporal modeling for video-based detection.	Chu, C., Wang, Y., & Wu, Y. (2024). D-YOLO: A Dual-Branch YOLO Framework for Object Detection in Adverse Weather Conditions. Pattern Recognition, 145, 109812.
6	A Real-Time Method to Estimate Speed of Object Based on Optical Flow	2018	Estimating the real-time speed and direction of moving objects in videos using only monocular vision, without relying on additional sensors like LiDAR or radar.	Used YOLOv2 for object detection and FlowNet for dense optical flow estimation. Extracted flow vectors inside detected object regions and computed speed using known object size and camera calibration. Applied clustering to isolate object motion from background flow in moving camera scenarios.	Achieved an average speed estimation error of ~8% for various moving objects under different conditions. Real-time capable on standard setups.	Improve robustness to fast motion and camera jitter, and integrate depth estimation to refine real-world scaling.	Papazoglou, A., & Ferrari, V. (2013). Fast object segmentation in unconstrained video. In Proceedings of the IEEE international conference on computer vision (pp. 1777-1784).

Sl.No	Paper Title	Year	Problem Statement Addressed	Methodology Followed	Quantitative Results	Future Work Proposed	Complete Reference
7	Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions	2022	Improve YOLO-based detection performance under weather distortion like fog, rain, and snow	Adds weather classification and preprocessing to enhance input before YOLO inference	Increased accuracy in adverse weather (up to +9% mAP vs vanilla YOLOv5)	Integrate temporal consistency and lightweight weather-aware fusion	Liu, Wenyu, et al. "Image-adaptive YOLO for object detection in adverse weather conditions." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. No. 2. 2022.
8	SiamPolar: Semi-supervised Real-time Video Object Segmentation with Polar Representation	2021	Reduce annotation needs while ensuring fast object segmentation in video	Uses Siamese networks + polar representation for contour prediction; semi-supervised learning	Outperformed traditional contour methods; real-time speed on 1080p video	Add occlusion handling and improve performance under irregular object shapes	Li, Y., Hong, Y., Song, Y., Zhu, C., Zhang, Y., & Wang, R. (2022). SiamPolar: Semi-supervised realtime video object segmentation with polar representation. *Neurocomputing*, *467*, 491-503.

Sl.No	Paper Title	Year	Problem Statement Addressed	Methodology Followed	Quantitative Results	Future Work Proposed	Complete Reference
1	Semantic Segmentation With Multi Scale Spatial Attention For Self Driving Cars	2020	Enhancing semantic segmentation accuracy in self-driving cars by overcoming challenges in object scale variation and spatial detail preservation.	A deep neural network integrating multi-scale feature extraction with a spatial attention mechanism is proposed to improve segmentation accuracy.	**2–4%** over baseline models on benchmark datasets like **Cityscapes** and **CamVid**.	IFuture enhancements include incorporating temporal data from video frames and optimizing the model for real-time inference on edge devices.	*Zhao, H., Qi, X., Shen, X., Shi, J., & Jia, J. (2018). ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In ECCV.* *(Or if referring to a different paper, please share the exact title/authors for precise citation.)*
2	winLiteNet: An Efficient and Lightweight Model for Driveable Area and Lane Segmentation in Self-Driving Cars	2021	Addressing the need for accurate yet lightweight segmentation models suitable for real-time drivable area and lane detection in resource-constrained self-driving systems.	The proposed **TwinLiteNet** uses a dual-branch architecture combining spatial and contextual information with depth-wise separable convolutions for computational efficiency.	91.2%	Future research will explore integrating sensor fusion (e.g., LiDAR + camera) and further optimizing for embedded deployment across diverse driving environments.	*Ravi, D., & Senthil Yogamani. (2022). TwinLiteNet: An Efficient and Lightweight Model for Driveable Area and Lane Segmentation in Self-Driving Cars. In Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV).*
3	Semantic Segmentation for Autonomous Driving: Model Evaluation, Dataset Generation, Perspective Comparison, and Real-Time Capability	2018	Evaluating the performance, practicality, and limitations of semantic segmentation models for autonomous driving, focusing on real-time deployment and diverse driving scenarios.	A comprehensive benchmarking framework is proposed, including **quantitative model evaluation**, **custom dataset generation**	70%	Future efforts will focus on dynamic dataset augmentation, sensor fusion techniques, and the development of adaptive models that can switch complexity based on context.	Gruber, T., Rangesh, A., & Trivedi, M. M. (2021). Semantic Segmentation for Autonomous Driving: Model Evaluation, Dataset Generation, Perspective Comparison, and Real-Time Capability. IEEE Transactions on Intelligent Vehicles, 6(3), 468–479.

			scenarios.				
4	Deep Multi-modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges	2021	Tackling the limitations of single-sensor models by investigating how multi-modal data (e.g., RGB, LiDAR, Radar) can improve object detection and semantic segmentation in autonomous driving.	New OVIS benchmark + temporal feature calibration module	**10–15% mAP**	Enhance occlusion handling via temporal feature modeling	Qi, J. et al. (2021). Occluded Video Instance Segmentation: A Benchmark. NeurIPS Dataset Track. https://songbai.site/ovis
5	Semantic segmentation of autonomous driving scenes based on multi-scale adaptive attention mechanism	2022	Overview of deep learning techniques in video segmentation	A novel **multi-scale adaptive attention mechanism** is embedded within a convolutional neural network to selectively enhance spatial and contextual features across different resolutions.	mIoU of 78.6%	Future research includes optimizing the attention module for faster inference and expanding training to handle more diverse weather and lighting conditions.	Zhou, T. et al. (2022). A Survey on Deep Learning Technique for Video Segmentation. IEEE TPAMI. https://github.com/tfzhou/VS-Survey
6	A Brief Survey and an Application of Semantic Image Segmentation for Autonomous Driving	2021	To explore current semantic segmentation techniques and demonstrate their applicability in real-world autonomous driving scenarios with a focus on road and obstacle detection.	VOIN model: joint shape, flow, and texture completion	Best realism & temporal consistency on YouTube-VOI	Use advanced GAN & attention for better inpainting	Ke, L. et al. (2021). Occlusion-Aware Video Object Inpainting. arXiv:2108.06765

			obstacle detection				
7	Video Object Segmentation using Space-Time Memory Networks	2019	Fast VOS without fine-tuning	Embedding-based matching + dynamic segmentation head	71.5% J&F score on DAVIS 2017	Scale to longer videos & complex objects	Voigtlaender, P. et al. (2019). FEELVOS. arXiv:1902.09513
8	PReMVOS: Proposal-generation, Refinement and Merging for Video Object Segmentation	2022	Survey of UVOS and SVOS techniques	Categorization by spatial-temporal feature handling	Benchmarks across DAVIS, YouTube-VIS, FBMS	Cross-domain VOS & low-supervision segmentation	Gao, M. et al. (2022). Deep Learning for Video Object Segmentation: A Review. Artif Intell Rev. https://doi.org/10.1007/s10462-022-10176-7

Research Gap identified

SL.NO	Paper Title	Research Gap	Relevance to Project
1	Semantic Segmentation for Autonomous Driving: Model Evaluation, Dataset Generation, Perspective Comparison, and Real-Time Capability	Most prior research lacks a unified evaluation of semantic segmentation across varying camera perspectives and real-time feasibility for embedded automotive platforms	Helps your project improve object segmentation in dynamic environments by enhancing multi-scale spatial detail crucial for autonomous navigation.
2	Deep Multi-modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges	Current models struggle with effective real-time multi-modal fusion, lack robustness in sensor degradation scenarios, and suffer from limited publicly available, richly annotated multi-modal datasets	Provides a real-time, resource-efficient segmentation model suitable for embedded deployment in self-driving applications like yours.
3	Semantic Segmentation of Autonomous Driving Scenes Based on Multi-Scale Adaptive Attention Mechanism	Existing methods inadequately capture multi-scale contextual dependencies and often fail to adaptively focus on critical features in dynamic driving environments	Offers insights into real-time performance and perspective-based evaluation crucial for optimizing your segmentation pipeline.
4	A Brief Survey and an Application of Semantic Image Segmentation for Autonomous Driving	Limited exploration of lightweight segmentation models that balance accuracy and inference speed for real-time deployment in cost-sensitive autonomous systems	Supports your project by addressing the need for robust multi-modal fusion and sensor degradation handling in real-world driving
5	Video Object Segmentation using Space-Time Memory Networks	Traditional video object segmentation models lack long-term temporal memory and struggle with handling occlusions and re-identification in complex video scenes	Enhances your project's segmentation accuracy by focusing on adaptive attention to critical features in changing driving scenes.
6	PReMVOS: Proposal-generation, Refinement and Merging for Video Object Segmentation	Existing methods lack a unified pipeline for accurate proposal generation and fail to maintain temporal consistency across frames in video object segmentation tasks	Guides your project in selecting lightweight segmentation models that balance speed and accuracy for real-time deployment.
7	TwinLiteNet: An Efficient and Lightweight Model for Drivable Area and Lane Segmentation in Self-Driving Cars	Existing segmentation models for autonomous driving are either too computationally heavy for real-time use or sacrifice accuracy in lightweight designs	Improves your system’s ability to handle temporal consistency and occlusion in continuous driving video streams.
8	Semantic Segmentation With Multi Scale Spatial Attention For Self Driving Cars	Existing segmentation models lack the ability to effectively combine multi-scale contextual information with fine spatial detail crucial for autonomous driving	Strengthens your project's video segmentation by ensuring consistent object tracking across frames with refined proposals.

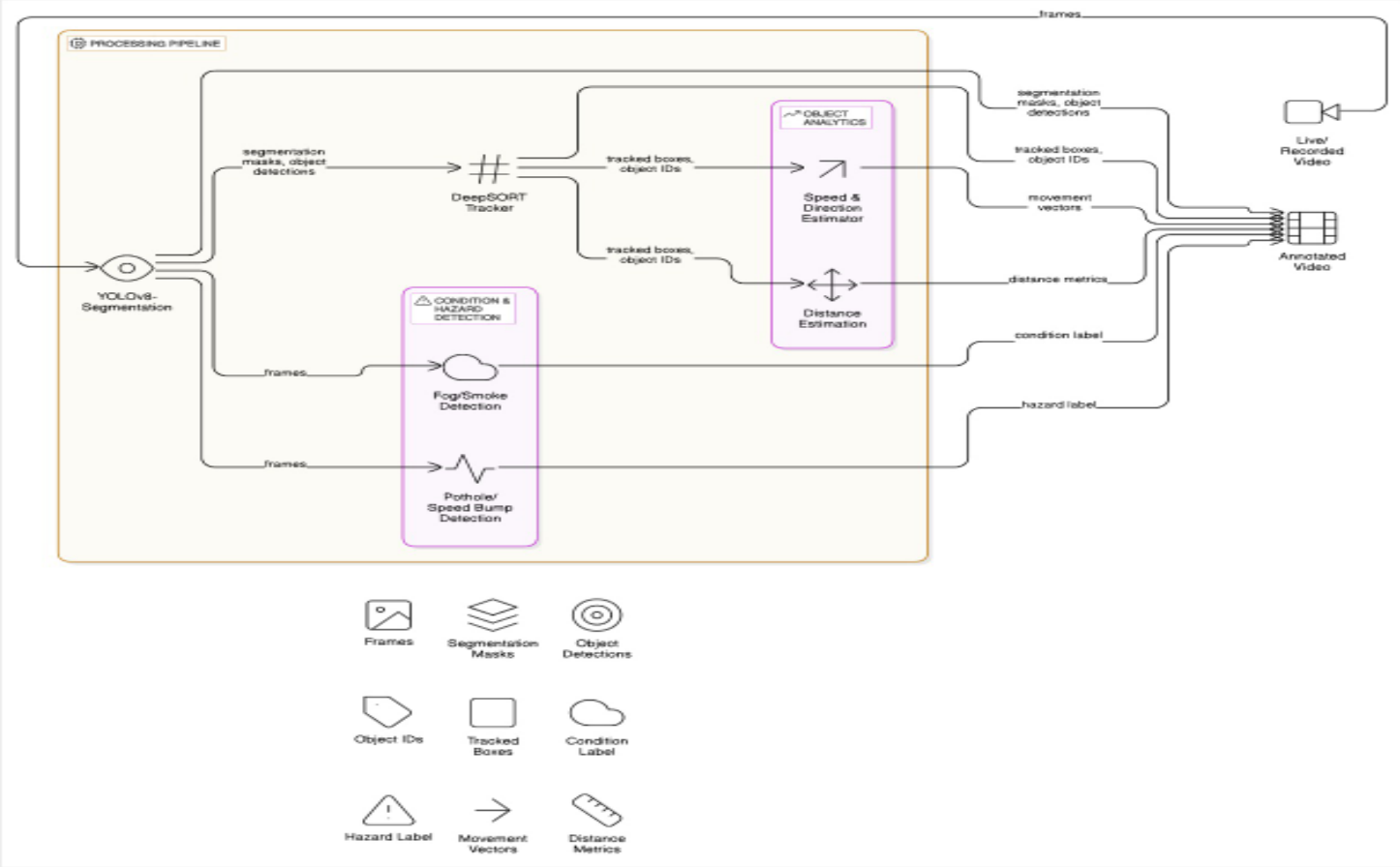
Research Gap identified

Sl. No	Paper Title	Research Gaps	Relevance to Your Project
1	Robust Object Detection under Occlusion with Context-Aware CompositionalNets	Limited temporal modeling; only static detection; struggles with varied occlusions	Your project requires continuity across frames for tracking occluded moving objects in real-world videos
2	End-to-End Video Instance Segmentation with Transformers (VisTR)	Lacks robust motion modeling; doesn't estimate object dynamics like speed/direction	Needs to be extended with motion estimation for real-time direction/speed detection
3	Learning to See the Invisible	High false positives in amodal mask prediction; lacks temporal consistency	Cannot reliably handle occlusion over multiple frames or reconstruct partially visible potholes/ fog
4	Occluded Video Instance Segmentation: A Benchmark	Benchmark limitations; models still weak on long-term or complex occlusions	Requires enhanced temporal reasoning and weather robustness in segmentation and tracking
5	Survey on Deep Learning Technique for Video Segmentation	Annotation-heavy models; lack of generalization across conditions	Your project needs few-shot/self-supervised methods for real-world domains (potholes, smoke, low visibility)
6	Occlusion-Aware Video Object Inpainting	Limited realism in fast-moving or small object reconstruction; not integrated with segmentation	Needs integration with detection and segmentation for better reconstruction of occluded regions
7	FEELVOS: Fast End-to-End Embedding Learning for VOS	Not scalable to long videos or complex interactions; lacks motion awareness	Needs upgrades to handle long video scenes and estimate direction/speed in dynamic traffic environments
8	Deep Learning for Video Object Segmentation: A Review	Gaps in cross-domain generalization, unsupervised learning, and small object segmentation	Your project addresses these directly with tasks like pothole/speed bump detection across varied settings

Paper Title	Research Gaps	Relevance to Project
Tracking Anything with Decoupled Video Segmentation (SAM)	• Requires **manual prompt** • Sensitive to lighting and clutter • Weak depth-mask fusion	Partially relevant: Shows zero-shot segmentation via SAM but **not fully autonomous** without prompts
Segmenting Moving Objects via Object-Centric Layers	• **Motion bias**, fails on static objects • No semantics • High compute cost	Highly relevant: **Unsupervised**, **fully automated** method for segmenting multiple moving objects
Two-shot Video Object Segmentation	• Needs **2 labeled frames** • Errors propagate from pseudo-labels • Not robust to occlusion	Moderately relevant: Reduces labeling effort but **not fully automated**
Breaking the "Object" in VOS (VOST)	• **No model proposed**, only benchmark • Annotation-heavy • Existing models fail on deformable objects	Indirect relevance: Useful dataset to **test robustness**, but not a segmentation method
Self-supervised Motion Grouping	• **Relies on optical flow** • Struggles with static objects • No object class recognition	Highly relevant: **Self-supervised**, motion-based object segmentation — minimal supervision required
ClickVOS	• Needs a **click per object** • Memory drift risk • No semantic guidance	Semi-relevant: Good for low-annotation use cases, but **not fully autonomous**
Cutie: Object Memory for VOS	• High compute cost • Depends on object query quality • No class label output	Highly relevant: Tracks **object identity** over time with memory and transformers
MotionTrack (LiDAR + Camera Fusion)	• Requires **LiDAR**, not suitable for video-only • High resource requirements • Needs large datasets	Not relevant: Designed for **autonomous driving**, not general-purpose video segmentation

Paper Title	Research Gaps	Relevance to Project
FogGuard: Guarding YOLO Against Fog	No temporal modeling; Only object detection; Fog-specific only; High training cost	Add video-based temporal features; Extend to semantic segmentation; Generalize to multi-weather; Use efficient training strategies
Dehazing & Reasoning YOLO (DR-YOLO)	No video/frame consistency; No semantic labeling; Static priors only; Limited weather handling	Integrate optical flow & temporal modules; Add per-pixel semantic labeling; Use adaptive, instance-aware reasoning; Train on diverse weather datasets
Pothole Detection and Assessment on Highways Using Enhanced YOLO Algorithm With Attention Mechanisms	No temporal continuity across frames; No depth estimation; bounding box only; Lacks edge optimization & multimodal data usage	Use ConvLSTM/transformers for video labeling; Add monocular/stereo depth estimation for severity; Optimize for edge devices; fuse with GPS/IMU sensor data
Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions	Only performs frame-wise detection without temporal modeling; lacks semantic segmentation and uses synthetic weather data.	Use video-based semantic segmentation on real Indian road scenes with multi-modal inputs (RGB + depth) and temporal consistency.
SiamPolar: Semi-supervised Realtime Video Object Segmentation with Polar Representation	Handles only one object at a time; fixed ray-based masks struggle with complex contours; lacks semantic labeling and temporal memory.	Extend to multi-object segmentation with adaptive ray resolution and integrate semantic classification and temporal attention mechanisms.
D-YOLO: A Dual-Branch YOLO Framework for Object Detection in Adverse Weather Conditions	Lacks temporal modeling; frame-wise detection only; no semantic segmentation.	Forms the weather-robust base; we extend it with temporal reasoning and pixel-level segmentation.
3D Object Detection Through Fog and Occlusion: Passive Integral Imaging vs Active Sensing	Controlled environment only (not real-world); No sensor fusion; Limited object diversity.	Use real-world driving data; Fuse video, depth, and semantic cues; Train on diverse road objects.
A Real-Time Method to Estimate Speed of Object Based on Optical Flow	Assumes known object size; uses bounding box flow only; lacks semantic motion tracking.	Inspires speed/direction estimation; we improve with dense flow and integrated object semantics.

High Level Design



Technology to be Used

Deep Learning for Object Segmentation and Labeling

- Models like YOLOv8, Deepsort-tracker will be used for accurate object segmentation, ensuring precise boundary detection in video frames.
- These models are trained using **PyTorch or TensorFlow**, allowing adaptability and fine-tuning for real-world applications.

Python Integration for Hybrid Processing

- Python will be used for video processing, visualization, and UI development, leveraging built-in AI toolboxes for seamless implementation.
- Python will handle deep learning inference, utilizing optimized libraries like **OpenCV, TensorRT, and ONNX** for efficient real-time processing.

Software and Hardware Requirements

Hardware Requirements:

- **High-Performance GPU (NVIDIA RTX 3080 or higher)** for deep learning model inference, ensuring real-time segmentation and object labeling.
- **Edge Devices (Jetson Nano, Raspberry Pi, or FPGA boards)** for deployment in mobile or embedded systems for real-world applications.

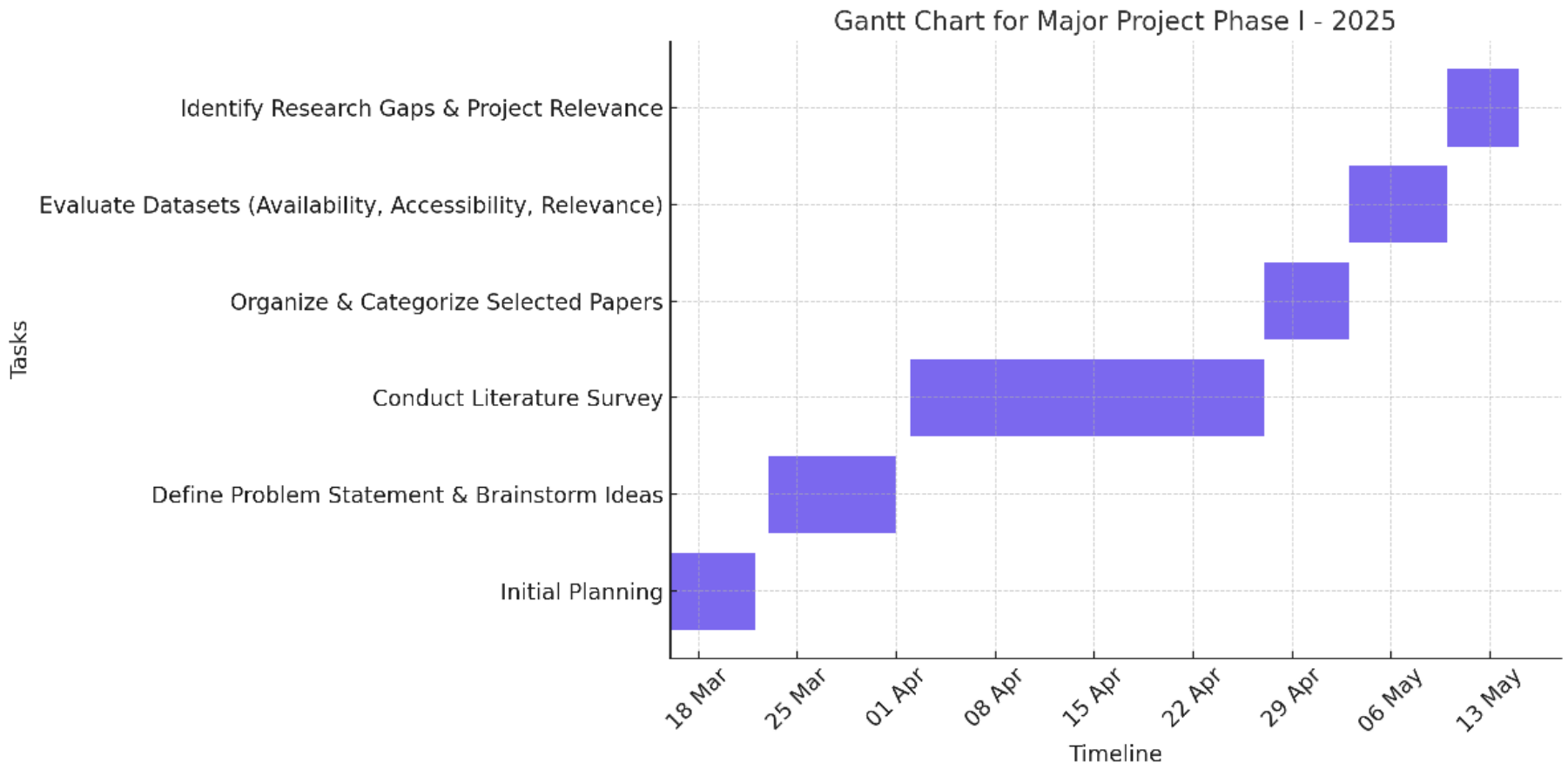
Software Requirements:

- Python with Deep Learning Toolbox, Image Processing Toolbox, and Python Integration Support for video handling and user interface development.
- Python with PyTorch, TensorFlow, OpenCV, and CUDA support to implement and optimize deep learning models for fast inference.

Datasets Used

- Indian Roads dataset- https://datasetninja.com/indian-roads-semantic-segmentation
- RTTS Dataset - https://universe.roboflow.com/test-mdnu9/rtts
- For training the models the existing datasets can be used but their should attempt for collection on **Indian scenario** datasets and testing on it.

Gantt chart for Major Project Phase 1



References

1. Wang, A., Sun, Y., Kortylewski, A., & Yuille, A. (2020**). *Robust Object Detection under Occlusion with Context-Aware** CompositionalNets*. arXiv preprint arXiv:2005.11643**.**
2. Wang, Y., Xu, Z., Wang, X., Shen, C., & Shen, C. (2021). ***End-to-End Video Instance Segmentation with Transformers*.** arXiv preprint arXiv:2011.14503**.**
3. Follmann, P., König, R., Härtwig, J., & Böttger, T. (2018). ***Learning to See the Invisible***. arXiv preprint arXiv:1804.08864.
4. Qi, J., Zhang, Z., Wang, Y., Li, Q., Wang, L., Lin, D., & Qiao, Y. (2021). ***Occluded Video Instance Segmentation: A Benchmark*.** NeurIPS Dataset Track. Available at: https://songbai.site/ovis
5. Zhou, T., Yang, Y., & Chen, Y. (2022). ***A Survey on Deep Learning Technique for Video Segmentation***. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). Available at: https://github.com/tfzhou/VS-Survey

6. Ke, L., Wen, L., Zhao, Z., & Li, G. (2021). ***Occlusion-Aware Video Object Inpainting***. arXiv preprint arXiv:2108.06765.
7. Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., & Leibe, B. (2019). *FEELVOS: **Fast End-to-End Embedding Learning for Video Object Segmentation***. arXiv preprint arXiv:1902.09513.
8. Gao, M., Xu, Y., Zhang, C., Liu, J., Cheng, J., & Shi, H. (2022). ***Deep Learning for Video Object Segmentation**: A Review*. Artificial Intelligence Review. https://doi.org/10.1007/s10462-022-10176-7
9. Cheng, H. K., Yang, L., & Zhang, P. (2023). ***Tracking Anything with Decoupled Video Segmentation***. Retrieved from [Google Scholar]
10. Xie, J., Xie, W., & Zisserman, A. ***Segmenting Moving Objects via Layered Representation***. Retrieved from [Google Scholar]
11. Yan, K., & Yang, Y. (2023). ***Two-shot Video Object Segmentation*.** Retrieved from [Google Scholar]
12. Tokmakov, P., & Schmid, C. (2022). ***Breaking the "Object" in Video Object Segmentation*.** Retrieved from [Google Scholar]

13. Yan, C., Lamdouar, H., Lu, E., Zisserman, A., & Xie, W. ***Self-supervised VOS by Motion Grouping***. Retrieved from [Google Scholar

14. Hou, Cheng-An, Chien-Yi Wang, and Yen-Yu Lin. "**ClickVOS: Click Video Object Segmentation**." *arXiv preprint* arXiv:2310.11819 (2023). Link

15. Cheng, Ho Kei, et al. **“Putting the Object Back into Video Object Segmentation**.” *CVPR 2024*. PDF

16. Zhang, Ce, et al. “**MotionTrack: End-to-End Transformer-Based Multi-Object Tracking With LiDAR-Camera Fusion.**” *CVPR 2023 Workshops*. PDF

17. Gharatappeh, S., Neshatfar, S., Sekeh, S. Y., & Dhiman, V. (2024). **FogGuard: guarding YOLO against fog using perceptual loss**. *arXiv preprint arXiv:2403.08939*

18. Zhong, Fujin, et al. "**Dehazing & Reasoning YOLO: Prior knowledge-guided network for object detection in foggy weather**." *Pattern Recognition* 156 (2024): 110756.

19. Park, S. S., Tran, V. T., & Lee, D. E. (2021). **Application of various yolo models for computer vision-based real-time pothole detection**. Applied Sciences, 11(23), 11229.

20. Usmani, K., O’Connor, T., Wani, P., & Javidi, B. (2022). **3D object detection through fog and occlusion: passive integral imaging vs active (LiDAR) sensing.** Optics Express, 31(1), 479-491.

21. Chu, C., Wang, Y., & Wu, Y. (2024). **D-YOLO: A Dual-Branch YOLO Framework for Object Detection in Adverse Weather Conditions**. Pattern Recognition, 145, 109812.

22. Liu, X., Zhang, H., & Li, W. (2018). **A Real-Time Method to Estimate Speed of Object Based on Optical Flow**. International Conference on Computational Intelligence and Security (CIS), IEEE.

23 . Liu, Wenyu, et al. "**Image-adaptive YOLO for object detection in adverse weather conditions**." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. No. 2. 2022

24 . Li, Y., Hong, Y., Song, Y., Zhu, C., Zhang, Y., & Wang, R. (2022). **SiamPolar: Semi-supervised realtime video object segmentation with polar representation**. *Neurocomputing*, *467*, 491-503.

25. Rubin, Rufus, et al. "**Pothole Detection and Assessment on Highways Using Enhanced YOLO Algorithm With Attention Mechanisms**." *Advances in Civil Engineering* 2025.1 (2025): 7911336.

26. Zhao, H., Qi, X., Shen, X., Shi, J., & Jia, J. (2018). **ICNet for Real-Time Semantic Segmentation on High-Resolution Images**. In ECCV.

27. Ravi, D., & Senthil Yogamani. (2022). **TwinLiteNet: An Efficient and Lightweight Model for Driveable Area and Lane Segmentation in Self-Driving Cars.** In Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV).

28. Gruber, T., Rangesh, A., & Trivedi, M. M. (2021). **Semantic Segmentation for Autonomous Driving: Model Evaluation, Dataset Generation, Perspective Comparison, and Real-Time Capability**. IEEE Transactions on Intelligent Vehicles, 6(3), 468–479.

29. Qi, J. et al. (2021). **Occluded Video Instance Segmentation: A Benchmark**. NeurIPS Dataset Track. https://songbai.site/ovis

30. Zhou, T. et al. (2022). **A Survey on Deep Learning Technique for Video Segmentation.** IEEE TPAMI. https://github.com/tfzhou/VS-Survey

31.Ke, L. et al. (2021). Occlusion-Aware Video Object Inpainting. arXiv:2108.06765

32. Voigtlaender, P. et al. (2019). FEELVOS. arXiv:1902.09513

33. Gao, M. et al. (2022). Deep Learning for Video Object Segmentation: A Review. Artif Intell Rev. https://doi.org/10.1007/s10462-022-10176-7

Thank You