XenSense-V.1: A Survey and Proposed Framework for Video Segmentation and Object Detection in Autonomous Vehicles

Mayanka Gupta[1], Ayman Amjad[2], Arjun Prabhakaran[2] , Bhanoday Kurma[2], Bhanu Prakash M[2], Dr. Kiran Agarwal Gupta[3], Dr. Chaitra Ravi[4], and Sindhoor N[5]

[1] BMS College of Engineering, Bengaluru, India
mayankag.cse@bmsce.ac.in
[2] BMS College of Engineering, Bengaluru, India
aymanamjad.cs22@bmsce.ac.in, arjunp.cs22@bmsce.ac.in,
bhanoday.cs22@bmsce.ac.in, bhanu.cs22@bmsce.ac.in
[3] Dayananda Sagar College of Engieering, Bengaluru, India
drkirangupta15@gmail.com
[4] BMS College of Engineering, Bengaluru, India
chaitrar.cse@bmsce.ac.in
[5] BMS College of Engineering, Bengaluru, India
sindhoorn.cse@bmsce.ac.in

Abstract. Despite significant progress in autonomous driving, detecting and segmenting obstacles under poor conditions remains a major chal lenge. This paper reviews deep learning models that tackle real-world difficulties like occlusion, fog, motion blur, and uneven road surfaces such as potholes and broken speed bumps—factors that heavily impact safety and detection accuracy. Over ten recent models are analyzed, including prompt-based approaches like Semantic-SAM and EPCFormer, memory-augmented ones like OOSIS and XMem, and task-specific detectors such as DR-YOLO, D-YOLO, and motion-aware YOLO variants. Each model type comes with trade-offs: prompt-based systems are flexible but depend on large vision-language datasets, while memory-based methods offer temporal consistency at the cost of increased computation. A key focus is on handling unstructured and uncertain road conditions, especially common in countries like India, where irregular infrastructure and unpredictable traffic are everyday challenges. Models trained solely on structured data often fail in these environments. To address this, the survey includes detailed comparisons and benchmark tests under difficult traffic and weather conditions. These insights inform the design of XenSense-V.1, a real-time deep learning framework using optical f low, temporal reasoning, and efficient segmentation to handle occlusion, weather issues, and complex road scenarios—particularly suited to Indian driving conditions.

Keywords: Autonomous Vehicles · Video Object Segmentation· Object Detection· Occlusion Handling· Adverse Weather Perception· Memory Augmented Deep Learning

1. INTRODUCTION

Deep learning has significantly advanced autonomous driving by enabling vehicles to perceive their surroundings at a detailed, pixel level. Techniques like video segmentation and semantic labeling form the backbone of modern perception systems. However, models that perform well in controlled settings often struggle on unpredictable, real-world roads-particularly in countries like India, where unmarked potholes, stray animals, and erratic traffic patterns are common. These challenges are further compounded by adverse weather conditions and the limitations of running complex models on edge devices in real time. This paper reviews recent deep learning methods for object detection and segmentation, highlighting their strengths and limitations in such environments. Based on these insights, we introduce XenSense-V.1, a real-time system designed to handle occlusion, motion dynamics, and unstructured road conditions with high segmentation accuracy, aiming to close the gap between lab performance and on-road reliability.

2. LITERATURE SURVEY

2.1. Object Segmentation and Semantic Labelling

Semantic segmentation and object labeling are crucial features of autonomous systems. They enable computers to perceive the world at a very high resolution, particularly in dense regions such as heavy traffic or potholed roads. The field has been revolutionized by recent advancements. Chen et al.[1]. proposed EPC Former, a transformer-based video object segmentation model. It uses text and visual cues, along with contrastive learning to better distinguish between objects. Subsequently, Li et al.[2] proposed SemanticSAM, utilizing a multiple-choice learning paradigm. On the majority of datasets, it performs moderately well when creating object and part-level segmentation masks from a single prompt. Vinoth and Sasikumar[3] combined camera and LiDAR data to address real world driving scenarios. To effectively track numerous objects in dense traffic, they used deep Q-learning in conjunction with YOLOv7 for object detection. Cheng et al.[4]. addressed the problem of occlusion using Cutie, a model that enhances the foreground-background separation based on memory for objects and a top-down query procedure in a transformer setting. This enables more efficient tracking of partially occluded objects.

Paper	Primary Task	Dataset	Metric	Score	Frames per second (FPS)
Chen et al. (2025) [EPC-Former]	Referring Video Object Segmentation	Ref-YouTube	mIoU	-VOS 56.4%	9.8
Li et al. (2023) [Semantic-SAM]	Semantic + Part Segmentation (Multi-granularity)	SA-1B, COCO, ADE20k	mIoU / 1-IoU Gain	+3.4 (1-IoU over SAM)	N/A

Vinoth & Sasikumar (2024)	Multi-object Detection + Tracking (Autonomous Driving)	Custom AV Dataset	Accuracy / MSE / Success	94.3% Accuracy, 0.06 MSE	11.2
Cheng et al. (2023) [Cutie]	Video Object Segmentation under Occlusion	MOSE (Challenging VOS)	J&F Score	74.2 (+8.7 over XMem)	22.1

Table 1: Comparison of different models for Object Segmentation and Semantic Labelling

2.2. Detection through Adverse Weather Conditions

Poor weather detection is a large problem for autonomous driving. Fog, rain, and snow all degrade visual information, resulting in missed detection and unstable tracking DR-YOLO[5] addresses fog with atmospheric scattering simulation for better feature learning and occluded object reasoning. Its specific focus on fog limits generalization to other conditions. FogGuard[6] uses a teacher-student architecture where the student is trained for fog properties through depth-aware augmentation and explicit modeling of physical principles. D-YOLO[7] uses a dual-branch architecture that achieves maximum performance in adverse condi tions through the incorporation of hazy and dehazed features and an Attention Feature Fusion (AFF) module and Dynamic Convolution. TYOLOv8[8] separates small objects from noise in the background using spatio-temporal modeling with multi-frame greyscale input and outperforms spatial baselines in windy, rainy, and foggy weather without the need for weather-specific training. These methods together mark a paradigm shift from single-fog exclusive solutions to more generalized models sensitive to temporal dynamics.

Paper	Primary Task	Dataset	Metric	Score	FPS
Zhong et al. (2024) [DR-YOLO]	Object Detection in Foggy Weather	RTTS, VF-test, VNtest	mAP	68.74%, 93.04%, 92.97%	74.8
Gharatappeh et al. (2024) [FogGuard]	Fog-Robust Object Detection	RTTS (Real Fog), RESIDE	mAP	69.43% (RTTS)	Same as YOLOv3
Chu et al. (2024) [DYOLO]	Object Detection under Diverse Weather	DAWN, Foggy Cityscapes, RTTS	mAP	72.1% (Foggy Cityscapes)	42.3
Van Lier et al. (2025) [TYOLOv8]	Spatio-Temporal Detection in Weather Conditions	Nano-VID weather (Rain, Haze, Wind)	mAP@0.25	Avg. 0.81	N/A

Table 2: Comparison of different models for Adverse Weather Object Detection

2.3. Detection of Speed and Direction of Objects Using Optical Flow

Optical flow estimates pixel motion across video frames, quantifying object speed and path. Optical flow estimation in dynamic scenes has recently been improved with deep learning. Papazoglou et al.[9] proposed a two-stage approach of motion edge detection with optical flow and segment refinement by spatio-temporal analysis, ensuring temporal continuity during sudden slowdown of objects. Liu et al.[10] proposed ODFSE, a real-time system based on FlowNet for optical flow and YOLOv2 for object detection. They estimate object speed from flow vectors in YOLO-detection regions and apply k-means clustering for object from background movement separation. Appropriate at moderate speeds, motion blur and size variation are limitations. These contributions, however, indicate the potential of optical flow for estimation of pixel dynamics at real-world speeds. XenSense-V.1 goes a step further by incorporating optical flow into its segmentation pipeline to enhance speed estimation in dense and occluded scenarios.

2.4. Detection of Partially Hidden Objects

The issue of partially obscured object detection must be addressed for autonomous driving in dynamic, unstructured environments. Conventional detection techniques often fail when objects are obscured by vehicles, people, or infrastructure. Recent studies have looked at novel strategies. Amodal segmentation models, such as A2VIS[11], use spatiotemporal signals from video to reconstruct occluded regions, improving segmentation continuity. The Bidirectional Occlusion-Guided Feature Propagation (BOFP)[12] enhances detections by correcting occlusion distorted areas by utilising forward and backward optical flow in combination with attention mechanisms.

Paper	Primary Task	Dataset	Metric	Score	FPS
Baselizadeh et al. (2025) [OOSIS]	Occlusion-Ordered Semantic Instance Segmentation	KINS (=20)	mAP@0.5:0.95 / WCS	21.6 / 77.4	N/A
Baghbaderani et al. (2024) [BOFP]	Video Semantic Segmentation	Cityscapes / VSPW	mIoU / mTC / mVC	76.5 / 70.6 / 84.5	265 ms/f
Yue et al. (2023)	Multi-Object Tracking under Full Occlusion	MOT16 / MOT17 (Occlusion Focused)	HOTA / MOTA / IDF1	HOTA: 57.7 / MOTA: 57.7 / IDF1: High	N/A

Table 3: Comparison of different models for Partially Hidden Object Detection

Generative techniques like ROODI[13] are used to implement object in painting. They predict hidden parts of objects by using 3D scene context and diffusion models. To enable robust boundary detection, CompositionalNets[14] isolates the object from its context and integrates visible components using a part-based methodology. Methods like the full occlusion-aware tracker [15], which tracks high-visibility areas using occlusion graphs and adaptive fusion, manage identity persistence even in the face of complete disappearance. Benchmark datasets such as OVIS[16], demonstrate performance problems in occlusion-heavy scenarios, highlighting the need for models that integrate spatial reasoning, memory, motion dynamics, and generative inference. Future systems need to be able to cope with visibility limitations for reliable perception in real-world driving scenarios.

2.5. Pothole and Speed Bump Detection

Road irregularities such as potholes and speed bumps are major challenges for autonomous vehicles, particularly on Indian roads. Nissimagoudar et al.[17] proposed a real-time pothole detection system based on YOLOv5 utilizing a user-devised Indian road anomaly dataset. Although efficient for bounding-box detection, it was missing shape informational details and did not analyze damage severity. To overcome small and clustered potholes, Rubin et al.[18] introduced an improved YOLO model with an Xception backbone and CBAM attention, improving detection and area estimation via image processing. Nevertheless, their approach was confined to static images without the capability to access real-time video. Both researches suggest the major limitations: dependency on bounding boxes prevents correct localization, and essential features such as depth estimation and severity measurement are still unexplored. XenSense-V1 endeavors to overcome these limitations by leveraging video-based semantic segmentation for accurate shape, depth, and severity analysis of potholes.

Paper	Primary Task	Dataset	Metric	Score	FPS
Nissimagoudar et al. (2024)	Pothole and Speed Bump Detection	Custom Indian Road Dataset	Accuracy	85% (Potholes), 83.8% (Speed Bumps)	N/A
Rubin et al. (2025)	Pothole Detection + Area Estimation	MakeML, Kaggle + Custom Kerala Dataset	mAP / IoU	mAP: 99.21%, IoU: 0.86	35.7

Table 4: Comparison of different models for Pothole and Road Defect Detection

2.6. Prompt-Based Detection

Great advancements have been achieved in prompt-based detection in the area of video segmentation, which allows both visual and textual inputs to be used for identifying and tracking objects in complex, dynamic driving scenes. A leading example is UniVS[19], a unified framework that addresses a range of segmentation tasks—from instance segmentation to referring object segmentation—all through a single architecture by framing prompts as queries. To reliably decode masks from frame to frame, UniVS introduces a new prompt crossattention mechanism (ProCA) that mixes memory based context with mean-pooled prompt features. Redefining the tasks as prompt-guided segmentation tasks, this formulation eliminates the requirement for heuristic frame matching and allows the model to generalize effectively across RefVOS, PVOS, and VOS tasks. Conversely, the DEVA framework[20] uses a decoupled, modular approach. It integrates a task-agnostic temporal propagation module that is independent of object class and a task-specific image segmentation model, such as SAM or Grounding-DINO. DEVA's bi-directional propagation mechanism integrates the advantages of image-level predictions and temporally consistent tracking even under sparse training data. This makes DEVA particularly suitable for open-world autonomous driving scenarios, where unexpected or seldom seen objects often appear.

Paper	Primary Task	Dataset	Metric	Score	FPS
Li et al. (2024) [UniVS]	Unified Video Segmentation (Prompt-based)	YouTube-VIS / Ref-DAVIS / VIPSeg	mAP / J&F / VPQ	63.8 / 66.3 / 56.0	N/A
Cheng et al. (2023) [DEVA]	Prompt-Guided Video Segmentation (Decoupled)	VIPSeg / BURST / Ref-YTVOS	VPQ / OWTA / J&F	52.2 / 70.5 / 66.0	N/A

Table 5: Comparison of different models for Prompt-Based Detection

2.7. Memory-Based Detection Using Short-Term Memory

Temporary vision loss from occlusions, glare, or sensor failures is common in real-world driving conditions. Situational awareness must be maintained during interruptions to drive safely. Memory-based methods address the limitation by accessing visual information from previous clear frames, allowing systems to decide regardless of partial visibility. An example is the Space-Time Memory Network (STM)[21], which stores object masks from previous frames and aligns them pixel-wise with the current frame and thus provides consistent segmentation in time, even with appearance change or occlusions. Building on STM, Chen et al. introduced SRNet[22], which generates auxiliary frames from neighboring frames and augments scene understanding through top-level prototype matching. Through a strengthening of segmentation stability, the method decreases visual noise and appearance-changing mistakes. These types of memory models are crucial for self-driving vehicles as they bridge perceptual gaps and maintain object detection consistent under unstable and dynamic road environments.

Paper	Primary Task	Dataset	Metric	Score	FPS
Oh et al. (2019) [STM]	Semi-supervised Video Object Segmentation	DAVIS 2017 / YouTube-VOS 2018	J&F Score	81.8 / 79.4	11.1

Chen et al. (2024) [SR-Net]	Memory-enhanced Video Object Segmentation	DAVIS 2017 / YouTube-VOS 2018	J&F Score	86.4 / 85.0	32.3

Table 6: Comparison of different models for Memory-Based Detection

The tables consolidate deep learning models for object detecting and segmenting in adverse situations and demonstrates the trade-offs between speed and performance in real-world driving conditions, critical parameters are tabulated, including FPS, accuracy, datasets, and main tasks.

3. PROPOSED MODEL PERFORMANCE PARAMETERS

This research performed preliminary runtime tests with a 1920x1080 video input resolution, which was down-scaled to 384x640 for processing, with the purpose of confirming the preliminary performance of the proposed XenSense-V.1 framework in occlusion scenarios. The model demonstrated an average runtime of 78.5 ms per frame with particular instances of 3.2 ms preprocessing, 74.7 ms for inference, and 0.6 ms post-processing. These preliminary results indicate the potential of XenSense-V.1 for real-time detection usage, especially in environments with high occlusion. In order to be deployment-ready on Indian roads and similar environment, further testing is already ongoing across multiple modules and scenarios.

Primary Task	Dataset	Metric	Score	FPS
Detection under Occlusion	Custom Indian Road Dataset	mIoU / Inference Time	78.5ms	12.7 (approx)

Table 7: Proposed model performance parameters of XenSense-V1

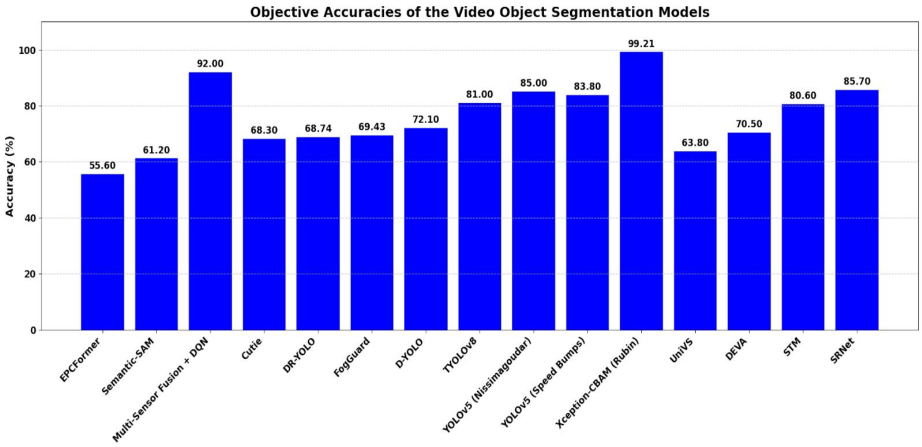

Fig.1: Accuracy comparison of surveyed Video Object Segmentation models.

4. PROBLEMS WITH CURRENT TECHNOLOGY

While deep learning has achieved strong results in semantic labeling and video segmentation, many models fail to adapt effectively to real-world autonomous driving—especially on unpredictable roads like those in India. Factors such as fog, smoke, and poor lighting conditions often obscure vision, leading to missed or inaccurate detections. Estimating object speed and trajectory is further complicated by motion blur and variable visibility. Road hazards like potholes and speed bumps also present challenges, as most detection systems rely on simple bounding boxes that lack detail on shape, depth, or severity.

In addition, many models are developed and tested in structured environments and struggle with the diversity of road types, textures, and vehicle profiles encountered in real traffic. Resource limitations on edge devices make it difficult to deploy these high-computation models in real time. Trust and interpretability remain ongoing concerns, particularly when deploying complex systems in safety-critical settings. While memory-based and prompt-guided methods show promise, more refinement is needed before they can reliably handle the full range of real-world driving conditions.

5. CONCLUSION AND FUTURE WORK

This study examined deep learning approaches for video segmentation and semantic labeling in the context of real-world autonomous driving, with a focus on challenging road conditions in India. While architectures like memory-enhanced networks and YOLO variants show potential, critical challenges remain. Effective handling of occlusions, rare object detection, and adaptation to dynamic road environments are still under active exploration.

Performance in real-world scenarios is often hindered by computational demands, speed-accuracy trade-offs, and limited interpretability—particularly when deploying on edge hardware. Bridging this gap requires models that not only retain scene context and fuse data from multiple sensors but also remain efficient enough for real-time use. Building larger, more diverse datasets and designing lightweight yet robust architectures will be central to future efforts. Addressing these issues is essential to move from benchmark success to dependable, real-world deployment in autonomous systems.

6. Declaration

We declare that the research presented in this paper, titled "XenSense-V.1: A Survey and Proposed Framework for Video Segmentation and Object Detection in Autonomous Vehicles ," is an original contribution by the authors. All sources have been properly cited. This work is submitted for academic purposes and has not been published elsewhere. The paper offers a comparative analysis of existing methods in video segmentation, adverse weather detection, optical flow-based motion estimation, and road anomaly detection, with a focus on conditions relevant to Indian roads. Any resemblance to existing works is purely coincidental, and efforts have been made to maintain academic integrity throughout.

References

1. Chen, J., Lin, J., Zhong, G., Fu, H., Nai, K., Yang, K., & Li, Z. (2025). Expression Prompt Collaboration Transformer for universal referring video object segmentation. Knowledge-Based Systems, 311, 113006.

2. Li, F., H. Zhang, P. Sun, X. Zou, S. Liu, J. Yang, C. Li, L. Zhang, and J. Gao. "Semantic-sam: Segment and recognize anything at any granularity. arXiv 2023." arXiv preprint arXiv:2307.04767.

3. Vinoth, K., and P. Sasikumar. "Multi-sensor fusion and segmentation for autonomous vehicle multi-object tracking using deep Q networks." Scientific Reports 14, no. 1 (2024): 31130.

4. Cheng, Ho Kei, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. "Putting the object back into video object segmentation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3151-3161. 2024.

5. Zhong, F., Shen, W., Yu, H., Wang, G., & Hu, J. (2024). Dehazing & Reasoning YOLO: Prior knowledge-guided network for object detection in foggy weather. Pattern Recognition, 156, 110756.

6. Gharatappeh, Soheil, Sepideh Neshatfar[1], and Salimeh Yasaei Sekeh. "FogGuard: guarding YOLO against fog using perceptual loss." In Intelligent Computing: Proceedings of the 2025 Computing Conference, Volume 3, vol. 1425, p. 135. Springer Nature, 2025.

7. Chu, Zihan. "D-YOLO a robust framework for object detection in adverse weather conditions." arXiv preprint arXiv:2403.09233 (2024).

8. Van Lier, Michel, Martin Van Leeuwen, Bastian Van Manen, Leo Kampmeijer, and Nicolas Boehrer. "Evaluation of Spatio-Temporal Small Object Detection in Real-World Adverse Weather Conditions." In Proceedings of the Winter Conference on Applications of Computer Vision, pp. 844-855. 2025.

9. Papazoglou, Anestis, and Vittorio Ferrari. "Fast object segmentation in unconstrained video." In Proceedings of the IEEE international conference on computer vision, pp. 1777-1784. 2013.

10. Liu, Kaizhan, Yunming Ye, Xutao Li, and Yan Li. "A real-time method to estimate speed of object based on object detection and optical flow calculation." In Journal of Physics: Conference Series, vol. 1004, no. 1, p. 012003. IOP Publishing, 2018.

11. Tran, Minh, Thang Pham, Winston Bounsavy, Tri Nguyen, and Ngan Le. "A2VIS: Amodal-Aware Approach to Video Instance Segmentation." Image and Vision Computing (2025): 105543.

12. Baghbaderani, Razieh Kaviani, Yuanxin Li, Shuangquan Wang, and Hairong Qi. "Temporally-consistent video semantic segmentation with bidirectional occlusion-guided

feature propagation." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 685-695. 2024.

13. Chang, Y., Dong, E., Seo, S., Kwak, N. and Yi, K.M., 2025. ROODI: Reconstructing Occluded Objects with Denoising Inpainters. arXiv preprint arXiv:2503.10256.

14. Wang, A., Sun, Y., Kortylewski, A. and Yuille, A.L., 2020. Robust object detection under occlusion with context-aware compositionalnets. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12645-12654).

15. Yue, Y., Yang, Y., Yu, Y. and Liu, H., 2023. Improving multi-object tracking by full occlusion handle and adaptive feature fusion. IET Image Processing, 17(12), pp.3423-3440.

16. Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P.H. and Bai, S., 2022. Occluded video instance segmentation: A benchmark. International Journal of Computer Vision, 130(8), pp.2022-2039.

17. Nissimagoudar, P.C., Miskin, S.R., Sali, V.N., SK, R., SK, D., HM, G., Hongal, R.S., Katwe, S.V. and CI, N., 2024. Detection of potholes and speed breaker for autonomous vehicles. Procedia Computer Science, 237, pp.675-682

18. Rubin, R., Jacob, C., Sundar, S., Stoian, G., Danciulescu, D. and Hemanth, J., 2025. Pothole Detection and Assessment on Highways Using Enhanced YOLO Algorithm With Attention Mechanisms. Advances in Civil Engineering, 2025(1), p.7911336.

19. Li, M., Li, S., Zhang, X. and Zhang, L., 2024. Univs: Unified and universal video segmentation with prompts as queries. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3227-3238).

20. Kei Cheng, H., Wug Oh, S., Price, B., Schwing, A. and Lee, J.Y., 2023. Tracking Anything with Decoupled Video Segmentation. arXiv e-prints, pp.arXiv-2309.

21. Oh, S. W., Lee, J.-Y., Xu, N., & Kim, S. J. (2019). Video Object Segmentation Using Space-Time Memory Networks. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 9226–9235.

22. Chen, Y., Zhu, W., Yang, Z.X. and Wu, E., 2024, July. Space-time Reinforcement Network for Video Object Segmentation. In 2024 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.