SEMINAR REPORT

On

Object Detection and Scene Understanding: Advance techniques for real-time video analysis

By

Atharv Kanase (TECOMP-B38)

Under the guidance of

Mrs. Avani Ray

Department of Computer Engineering

Pimpri Chinchwad College of Engineering and Research, Ravet

An Autonomous Institute — NBA Accredited (4 UG Programs) —

NAAC A++ Accredited — An ISO 21001:2018 Certified

SAVITRIBAI PHULE PUNE UNIVERSITY

(2025 - 2026)

Department of Computer Engineering

Pimpri Chinchwad College of Engineering and Research, Ravet

CERTIFICATE

This is to certify that **Atharv Kanase** from **Third Year Engineering** has successfully completed her seminar work titled **“Object Detection and Scene Understanding: Advance techniques for real-time video analysis”** at Pimpri Chinchwad College of Engineering and Research, Ravet in the partial fulfillment of the Bachelors Degree in Engineering.

Mrs. Avani Ray
Seminar Guide

Dr. Vijay A Kotkar
HOD, Computer Department

Prof. Dr. H.U. Tiwari
Director, PCCOE&R, Ravet

ACKNOWLEDGEMENT

It gives me pleasure to present a Seminar on “Object Detection and Scene Understanding: Advance techniques for real-time video analysis”. I am very much obliged to my guide **Mrs. Avani Ray**, Department of Computer Engineering, for helping me and giving proper guidance. I am very thankful to the Head of the Department **Dr. Vijay A Kotkar** and the entire staff members for their cooperation. I am also thankful to my family and friends for their support and constant encouragement towards the fulfilment of the work.

Place: Ravet, Pune **Atharv Kanase**

Date: **TECOMP-B38**

ABSTRACT

This seminar explores advanced techniques for real-time object detection and scene understanding, fundamental tasks in computer vision that enable machines to recognize, localize, and interpret multiple objects within video streams. With the rapid evolution of deep learning methods such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), real-time video analysis has become increasingly accurate and efficient, supporting critical applications like autonomous vehicles, intelligent surveillance, healthcare monitoring, and smart city systems. The seminar examines key challenges—including occlusion, variable lighting, motion blur, and computational constraints on edge devices—and evaluates state-of-the-art architectures for their speed, accuracy, and adaptability. It also highlights emerging trends such as explainable AI, zero-shot learning, and multimodal sensor fusion, emphasizing their potential to enhance the efficiency, robustness, and scalability of real-time video analytics solutions.

Keywords: Object Detection, Scene Understanding, Real-time Video Analysis, Computer Vision, Convolutional Neural Networks (CNNs), Vision Transformers (ViTs)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Introduction

The rapid growth of surveillance systems, autonomous vehicles, and smart city applications has fueled the need for efficient real-time video analysis. At the core of this challenge lies object detection and scene understanding, two interdependent tasks that enable systems to identify, localize, and interpret objects and their interactions within dynamic environments. Traditional computer vision approaches, while effective for constrained scenarios, fail to generalize in complex, real-world conditions where factors such as occlusion, illumination changes, and fast-moving objects degrade performance. The advent of deep learning, particularly convolutional neural networks (CNNs) and transformers, has revolutionized this domain, providing significant improvements in accuracy, scalability, and adaptability.

Despite these advancements, achieving robust real-time performance remains challenging due to computational complexity, the need for temporal reasoning across frames, and the requirement to balance accuracy with latency. Modern approaches integrate object detection models like YOLOv8 with scene reasoning modules and temporal analysis frameworks to provide richer insights into video streams. Such hybrid systems not only detect objects but also capture contextual relationships and temporal dynamics, enabling applications ranging from intelligent surveillance and traffic monitoring to robotics and augmented reality. This report explores existing solutions, identifies their limitations, and proposes an integrated architecture for enhancing both the efficiency and reliability of real-time video analysis.

2. LITERATURE SURVEY

Divya Nimma, Omaia Al-Omari, Rahul Pradhan, Zoirov Ulmas, RVV Krishna, Ts Yousef A Baker El-Ebiary, and Vuda Sreenivasa Rao
Object detection in real-time video surveillance using attention based transformer-YOLOv8 model [2025]
The study presents an enhanced YOLOv8 model incorporating attention-based transformers for object detection in real-time surveillance. The approach improves accuracy, robustness, and adaptability under varying lighting and occlusion conditions.(1)

Sani Abba, Ali Mohammed Bizi, Jeong-A Lee, Souley Bakouri, and Maria Liz Crespo
Real-time object detection, tracking, and monitoring framework for security surveillance systems [2024]
This work introduces a real-time framework for object detection, tracking, and monitoring in surveillance systems. It leverages deep learning-based detection with optimized tracking pipelines to enhance security monitoring efficiency and robustness in dynamic environments.(2)

Krishna Kumar, Krishan Kumar, and C. L. P. Gupta
Object Detection in Video Frames using Deep Learning [2022]
The paper applies deep learning methods for detecting objects in video frames, demonstrating how neural networks improve detection accuracy and reliability in sequential visual data.(3)

M. Koteswara Rao and P. M. Ashok Kumar
Exploring the advancements and challenges of object detection in video surveillance through deep learning: A systematic literature review and outlook [2025]
This review systematically analyzes deep learning-based object detection methods for surveillance, highlighting current progress, challenges such as real-time constraints, and future directions for intelligent monitoring systems.(4)

Paschalis Tsirtsakis, Georgios Zacharis, George S. Maraslidis, and George F. Fragulis
Deep learning for object recognition: A comprehensive review of models and algorithms [2025]
A comprehensive review of deep learning models and algorithms for object recognition,

covering CNNs, transformers, and hybrid approaches. It provides insights into their strengths, limitations, and applications across domains.(5)

Jiajun Wu

Physical scene understanding [2024]

This article focuses on physical scene understanding, examining how AI models interpret and predict real-world environments by integrating perception, reasoning, and physical interaction principles.(6)

Américo Pereira, Pedro Carvalho, and Luís Côrte-Real

A transition towards virtual representations of visual scenes [2024]

This paper explores the shift towards creating virtual representations of real-world visual scenes, emphasizing the role of computer vision, 3D reconstruction, and immersive technologies in enabling digital scene understanding and interaction.(7)

Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li

FMGS: Foundation model embedded 3D Gaussian splatting for holistic 3D scene understanding [2025]

The paper introduces FMGS, a novel approach combining foundation models with 3D Gaussian splatting to achieve comprehensive scene understanding, enabling advanced 3D perception and reconstruction.(8)

Sichao Liu, Jianjing Zhang, Robert X. Gao, Xi Vincent Wang, and Lihui Wang

Vision-language model-driven scene understanding and robotic object manipulation [2024]

This work integrates vision-language models with robotics for scene understanding and object manipulation. It demonstrates how multimodal AI improves robotic perception, planning, and interaction in unstructured environments.(9)

Table 1: **Literature survey on Object Detection and Scene Understanding**

Research Article (Author/Year)	**Objective / Proposed Work**	**Methods / Techniques**	**Datasets**	**Relevant Findings / Limitations Identified**
"Object detection in real-time video surveillance using attention based transformer-YOLOv8 model" D. Nimma et al., 2025	Real-time object detection in surveillance	Attention-based Transformer + YOLOv8	Benchmark surveillance datasets	High accuracy in challenging conditions; requires high computation
"Real-time object detection, tracking, and monitoring in security surveillance" S. Abba et al., 2024	Real-time detection, tracking, and monitoring framework	Deep learning detection + multi-object tracking	Surveillance video datasets	Achieves real-time monitoring; scalability on edge devices challenging
"Object Detection in Video Frames using Deep Learning" K. Kumar et al., 2022	Object detection in video frames	Deep learning detection models	Video frame datasets	Better accuracy than classical methods; limited benchmarking

"Exploring the Advancements and Challenges of Object Detection in Video Surveillance through Deep Learning: A Systematic Literature Review and Outlook" M. K. Rao et al., 2025	Review of deep learning in video surveillance	Systematic literature review	Literature-based	Identifies advancements and challenges; no experimental validation
"Deep learning for object recognition: A comprehensive review of models and algorithms" P. Tsirtsakis et al., 2025	Comprehensive review of recognition models	CNNs, Transformers, Hybrid DL algorithms	Literature review	Provides taxonomy of models; lacks empirical evaluation
"Physical scene understanding" J. Wu, 2024	Physical scene understanding in AI	Cognitive reasoning + physics priors	Conceptual + case study datasets	Bridges perception and reasoning; limited implementations
"A transition towards virtual representations of visual scenes" A. Pereira et al., 2024	Transition towards virtual scene representations	3D reconstruction, immersive visualization	Conceptual framework	Outlines digital twin approach; lacks real-world implementation

“FMGS: Foundation model embedded 3D Gaussian splatting for holistic 3D scene understanding” X. Zuo et al., 2025	Holistic 3D scene understanding	FMGS: 3D Gaussian splatting + foundation models	Synthetic + real 3D datasets	Enables detailed 3D reconstruction; computationally expensive
“Vision-language model-driven scene understanding and robotic object manipulation” S. Liu et al., 2024	Vision-language for robotic scene understanding	Vision-language models + robotic manipulation	Robotics datasets	Improves robotic perception; generalization remains limited

3. EXISTING SYSTEM

Recent advancements in real-time video analysis for object detection and scene understanding have evolved from conventional CNN-based models like Faster R-CNN and early YOLO variants (Kumar et al., 2022) toward more sophisticated hybrid approaches. Traditional YOLO-based frameworks (Abba et al., 2024) remain lightweight and fast for surveillance applications but face challenges with occlusion and crowded scenes. To address this, transformer-enhanced architectures such as Attention-based YOLOv8 (Nimma et al., 2025) integrate self-attention mechanisms to improve accuracy and robustness in complex environments, making them among the most effective current solutions for surveillance tasks. Complementing detection, recent work on scene understanding (Wu, 2024; Pereira et al., 2024; Zuo et al., 2025; Liu et al., 2024) leverages foundation models, 3D Gaussian splatting, and vision-language integration to move beyond bounding boxes, enabling holistic interpretation of environments and object interactions. Collectively, these approaches highlight a clear shift toward real-time, transformer-driven, and multimodal architectures that balance speed, accuracy, and contextual understanding.

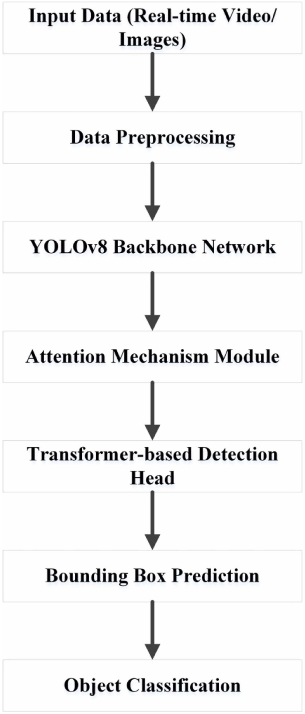

Input Data (Real-time Video/ Images)
Data Preprocessing
YOLOv8 Backbone Network
Attention Mechanism Module
Transformer-based Detection Head
Bounding Box Prediction
Object Classification

Figure 3.1: Architecture of Existing Systems

4. PROPOSED SYSTEM

4.1 Problem Statement

The central problem in real-time video analysis is the critical trade-off between speed and accuracy. While modern systems can identify objects, they often fail to do so instantly and accurately in the face of real-world complexities like occlusions, diverse lighting, and dynamic environments. This gap prevents their widespread adoption in mission-critical applications where split-second decisions are essential.

4.2 Objectives

1. Examine state-of-the-art techniques for real-time object detection and scene understanding in video analysis.
2. Compare and evaluate deep learning architectures (e.g., CNNs, ViTs) in terms of speed, accuracy, and adaptability to dynamic environments
3. Analyze key challenges such as occlusion, motion blur, variable lighting, and computational constraints.
4. Explore model optimization strategies for efficient deployment on edge and embedded systems.
5. Investigate future trends: explainable AI, zero-shot learning, and multimodal sensor fusion.

4.3 Proposed System

The proposed solution integrates real-time object detection with scene understanding and temporal reasoning to improve the accuracy and robustness of video analysis. Incoming video streams are first processed by YOLOv8 to detect and localize objects. The extracted features are then passed into two parallel modules: a Scene Reasoning Module, which uses transformer or graph-based networks to understand spatial relationships between objects, and a Temporal Module, which leverages ConvLSTMs or temporal transformers to capture motion dynamics across frames. The outputs are fused to generate annotated video feeds with alerts, enabling more intelligent monitoring that adapts to complex environments and real-time constraints.

4.3.1 YOLOv8 Backbone

YOLOv8 is a state-of-the-art object detection model that provides real-time performance while maintaining high accuracy. It uses a lightweight yet powerful convolutional backbone and optimized detection head, making it ideal for surveillance scenarios where speed and precision are critical.

4.3.2 Scene Reasoning Module

The scene reasoning module is designed to go beyond simple object detection by modeling relationships between objects in a frame. Using transformers or graph neural networks (GNNs), this module captures contextual dependencies — such as proximity, interactions, and co-occurrence — which helps the system better interpret complex scenes.

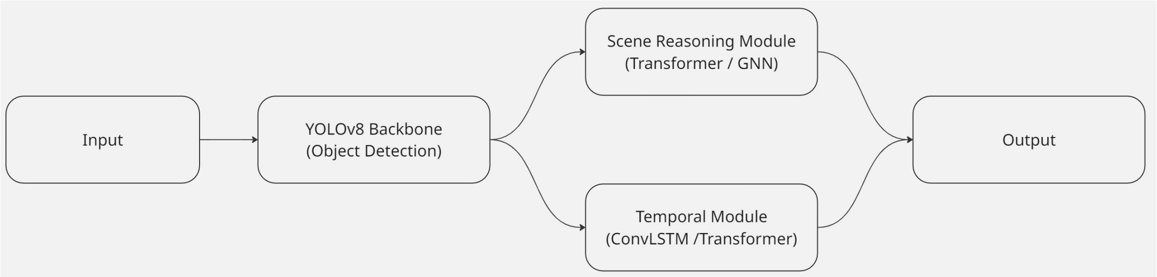

Input
YOLOv8 Backbone (Object Detection)
Scene Reasoning Module (Transformer / GNN)
Temporal Module (ConvLSTM /Transformer)
Output

Figure 4.1: Flowchart of the Proposed System

5. CONCLUSION & FUTURE SCOPE

5.1 Conclusion

Real-time video analysis through object detection and scene understanding is not just a technological advancement but a paradigm shift. We have moved from simply processing pixels to interpreting the visual world with speed and accuracy. The ability to identify objects and comprehend their relationships in dynamic environments has transcended the boundaries of research, becoming a foundational technology for a new era of intelligent systems.

5.2 Future Scope

The future of this field is focused on overcoming remaining challenges, such as improving performance on low-power edge devices and ensuring robustness in every conceivable condition. This will lead to the next generation of applications that are safer, more efficient, and more integrated into our daily lives, from fully autonomous transportation to personalized healthcare and beyond. As we continue to refine these techniques, the line between human and machine perception will continue to blur, unlocking unprecedented potential across countless industries.

REFERENCES

[1] D. Nimma, O. Al-Omari, R. Pradhan, Z. Ulmas, R. Krishna, T. Y. A. B. El-Ebiary, and V. S. Rao, "Object detection in real-time video surveillance using attention based transformer-yolov8 model," *Alexandria Engineering Journal*, vol. 118, pp. 482–495, 2025.

[2] S. Abba, A. M. Bizi, J.-A. Lee, S. Bakouri, and M. L. Crespo, "Real-time object detection, tracking, and monitoring framework for security surveillance systems," *Heliyon*, vol. 10, no. 15, 2024.

[3] K. Kumar, K. Kumar, and C. Gupta, "Object detection in video frames using deep learning," *International Journal of Computer Applications*, vol. 183, no. 51, pp. 975–8887, 2022.

[4] M. K. RAO and P. A. KUMAR23, "Exploring the advancements and challenges of object detection in video surveillance through deep learning: A systematic literature review and outlook," *Journal of Theoretical and Applied Information Technology*, vol. 103, no. 6, 2025.

[5] P. Tsirtsakis, G. Zacharis, G. S. Maraslidis, and G. F. Fragulis, "Deep learning for object recognition: A comprehensive review of models and algorithms," *International Journal of Cognitive Computing in Engineering*, 2025.

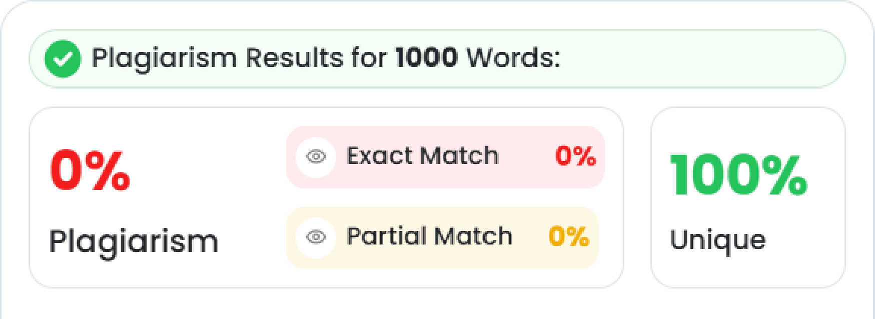
[6] J. Wu, "Physical scene understanding," *AI Magazine*, vol. 45, no. 1, pp. 156–164, 2024.

[7] A. Pereira, P. Carvalho, and L. Côrte-Real, "A transition towards virtual representations of visual scenes," *arXiv preprint arXiv:2410.07987*, 2024.

[8] X. Zuo, P. Samangouei, Y. Zhou, Y. Di, and M. Li, "Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding," *International Journal of Computer Vision*, vol. 133, no. 2, pp. 611–627, 2025.

[9] S. Liu, J. Zhang, R. X. Gao, X. V. Wang, and L. Wang, "Vision-language model-driven scene understanding and robotic object manipulation," in *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2024, pp. 21–26.

PLAGIARISM REPORT



✓ Plagiarism Results for **1000** Words:

0% Plagiarism

Exact Match 0%

Partial Match 0%

100% Unique