International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0349

e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2423-2429

# Enhancing Traffic Scene and Understanding Through Image Captioning and Audio

*Sejal Pawar[1], Shruti Mulay[2], Jivani Suryawanshi[3], Vaishnavi Walgude[4], Prof. K. V. Patil[5]*

*[1,2,3,4]Student, Department of Information Technology, Bharati Vidyapeeth's College of Engineering for Women, Pune, India.*

*[5]Professor, Department of Information Technology, Bharati Vidyapeeth's College of Engineering for Women, Pune, India.*

***Emails:*** *sejalmegha.1412@gmail.com[1], Shrutimulay06@gmail.com[2], Jivanisuryawanshi9@gmail.com[3], vaishnaviwalgude@gmail.com[4], Kamlesh.patil@bharatividyapeeth.edu[5]*

## Abstract

*Navigating significant collections of traffic images on the net provides a tremendous task, mainly for users in search of particular facts. Many images lack captions, making it difficult to locate applicable content. Our undertaking addresses this difficulty by way of developing an automated labelling service that generates object-based totally description and gives auditory cues about their distances, the usage of a combination of computer vision and audio description techniques. With this, automation has become the need of the hour. The use of automation in motor vehicles is one similar area that's getting further and further significance and recognition around the world. Our technique leverages state-of-the-art object detection strategies, specially the YOLO (You Only Look Once) version, to identify and label items within traffic photos. By expertise the content of each image, our service generates labels for items detected, and presents additional information concerning their distances from the viewer. To enhance accessibility, our service includes a Text-to-Speech (TTS) engine for sounding audio description. This function caters to users with visible impairments and people who decide on auditory facts. Our studies pursuits to automate the method of annotating site visitor's images, lowering the reliance on human intervention, especially for massive databases. We make use of a deep neural network structure, which include the YOLO version for object detection, and extra additives for distance estimation. This functionality is designed to accommodate individuals with visual impairments as well as those who favor auditory cues. By bridging the distance between image content and textual/audio descriptions, our device offers a promising solution for correctly accessing records within traffic scenes.*

***Keywords:*** *Convolution Neural Networks (CNN), You Only Look Once (YOLO), ADAS – Advance Driver Assistance System, TTS – Text-To-Speech.*

## 1. Introduction

The internet hosts an overwhelming number of images, many of which are uncategorized and lack captions, making it challenging for users to find specific information efficiently. This issue is particularly significant for applications. Developing an automatic captioning system for traffic images, augmented with audio descriptions, could greatly improve this situation by providing accessible and descriptive information about traffic. Recent advancements in computer vision and natural language processing (NLP) provide an opportunity to address these challenges more effectively. By leveraging deep learning techniques, we can develop more sophisticated models capable of understanding and describing the intricate details of traffic scenes. One such advancement is the You Only Look Once (YOLO) algorithm, which offers real-time object detection. [1] YOLO's efficiency and accuracy in detecting multiple objects within an image make it an ideal candidate for analyzing traffic scenes. [2] We integrate YOLO for object detection with advanced captioning techniques to create a comprehensive system for traffic scene understanding. YOLO will be used to detect and identify various traffic elements, such as vehicles, pedestrians, traffic signs, and road conditions, in real-time. Our topic endeavors to contribute substantial fresh insights by presenting a pioneering approach to autonomously produce captions and audio descriptions for traffic scenes. By harnessing cutting-edge deep learning methodologies, we aspire to comprehend the essence of traffic images/video and articulate their semantic details in a natural language format. [3] This methodology not only expedites the annotation process for traffic images but also enriches accessibility for individuals with visual impairments. The key challenges we face include accurately interpreting the complex visual information present in traffic scenes, generating textual descriptions, and integrating audio synthesis for enhanced accessibility. [4] Overcoming these challenges requires a multidisciplinary approach that combines expertise from computer vision, natural language processing, and audio engineering. [5]

## 2. Literature Survey

AI based Driver Assistant system - 04 | Apr 2020, IEEE Xplore: This system focuses mainly on traffic sign detection and recognition (TSDR) which is vital for advanced driver assistance systems and autonomous vehicles, aimed at enhancing road safety. [6] Early methods relied on color segmentation and shape analysis but struggled with varying lighting and complex backgrounds. Machine learning introduced feature-based techniques like HOG and SVM, which, while better, required manual feature extraction. The advent of deep learning, particularly Convolutional Neural Networks (CNNs), significantly improved accuracy with models like R-CNN, Fast R-CNN, and Faster R-CNN, although they were computationally intensive. Single-stage detectors like SSD and YOLO addressed real-time requirements but often at the cost of some accuracy. TSDR still faces challenges such as small object detection, environmental variations, and false positives. Recent advancements include attention mechanisms, transformer-based models, and hybrid approaches combining traditional and deep learning methods. [7] Integrating voice feedback through text-to-speech enhances driver assistance, and future research aims at leveraging edge computing, advanced data augmentation, and multi-modal systems to further improve TSDR systems' speed, accuracy, and robustness. AI-Based Autonomous Driving Assistance System - 2021, IEEE: In this article it presents the rapid advancement in modern technology has significantly influenced the automotive industry, leading to a surge in the development of self-driving cars and advanced driver assistance systems (ADAS). The methods often struggled with varying environmental conditions and complex backgrounds. The introduction of machine learning brought feature-based methods like Histogram of Oriented Gradients (HOG) combined with Support Vector Machines (SVM), which offered improved accuracy but required extensive manual feature extraction. [8] The emergence of deep learning revolutionized the field, with Convolutional Neural Networks (CNNs) such as R-CNN, Fast R-CNN, and Faster R-CNN significantly enhancing detection accuracy but at the cost of computational efficiency. Single-shot detectors like SSD and YOLO addressed real-time processing needs but sometimes sacrificed precision. Current research is tackling challenges such as the detection of small objects, handling diverse environmental conditions, and reducing false positives through advanced techniques like attention mechanisms and transformer-based models. Moreover, integrating voice feedback via text-to-speech systems enhances driver interaction and safety. Object Detection Learning Techniques for Autonomous Vehicle Applications - 2019: In this paper, we study that the critical domain of autonomous vehicles, which represents a transformative innovation in intelligent transportation. [9] Traditional sensors like LIDAR, while effective, often falter in adverse weather, necessitating robust vision-based learning techniques. The advent of deep learning has revolutionized this domain, with models like "You Only Look Once" (YOLO) and the Single Shot Multibox Detector (SSD) demonstrating superior accuracy and speed. [10] These models leverage convolutional neural networks (CNNs) to process visual data efficiently, enabling real-time detection crucial for autonomous driving. Comparative studies reveal that while SVMs lack the necessary performance for dynamic driving environments, YOLO and SSD excel, providing the rapid, accurate object recognition needed to make swift driving decisions. This research highlights the importance of continuing to refine these deep learning models and integrating them seamlessly with other vehicle systems to achieve the high safety standards required for autonomous vehicles. Caption Generation from Road Images for Traffic Scene Construction - 2020: In this paper image captioning has advanced significantly with deep learning, particularly through CNNs and RNNs like LSTMs. Early methods struggled with accuracy, but models such as Show and Tell and Show, Attend and Tell improved coherence and relevance by using the encoder-decoder framework. The encoder (CNN) extracts visual features, and the decoder (RNN) generates text, with attention mechanisms enhancing performance by focusing on relevant image parts. In traffic scenes, combining object detection (e.g., Faster R-CNN) with segmentation models (e.g., Mask R-CNN) has led to precise identification of static and dynamic elements. Recent approaches use graph-based methods and transformers to capture complex interactions. Benchmark datasets like TSD-max and COCO standardize evaluations, driving further progress. These advances highlight the potential of image captioning networks in autonomous vehicle testing and smart city infrastructure through accurate traffic scene modeling. The Traffic Scene Understanding and Prediction Based on Image Captioning - 2021, IEEE Access: Recent advancements in image captioning and traffic scene understanding have significantly benefited from deep learning techniques, particularly the integration of convolutional neural networks (CNNs) for feature extraction and Long Short Term Memory (LSTM) networks for sequence generation. In traffic scene understanding, combining object detection (e.g., Faster R-CNN) and semantic segmentation (e.g., Mask R-CNN) with LSTM-based captioning improves the recognition and description of both static and dynamic traffic elements. Recent innovations include graph-based methods and transformers to better capture relationships between traffic components. Benchmark datasets such as TSD-max and COCO standardize evaluations and drive progress. These advanced models not only generate detailed scene descriptions but also offer potential in creating natural language driving strategies, contributing to autonomous vehicle development and intelligent transportation systems. [11]

## 3. Overview of The System

Data Collection: Gather a dataset of traffic scenes, annotating traffic objects with bounding boxes and captions. Utilize public datasets like COCO/Flicker or record your own scenes.

### 3.1. Data Pre-processing

Trim footage to remove irrelevant frames and enhance image quality through techniques like resizing and normalization. Model Training: Choose YOLOv3 for object detection and a captioning model for generating captions. Split the dataset for training and testing, then train both models on annotated data.

### 3.2. Integration with TTS Engine

Develop a mechanism to trigger voice feedback based on detected objects. Utilize captions generated by the model for voice feedback.

## 4. Proposed System

To enhance the understanding of traffic scenes using image captioning and audio descriptions, we can employ YOLO (You Only Look Once) for object detection and TTS (Text-to-Speech) for generating audio descriptions. Figure 1 refers to the flow chart of the Proposed system.

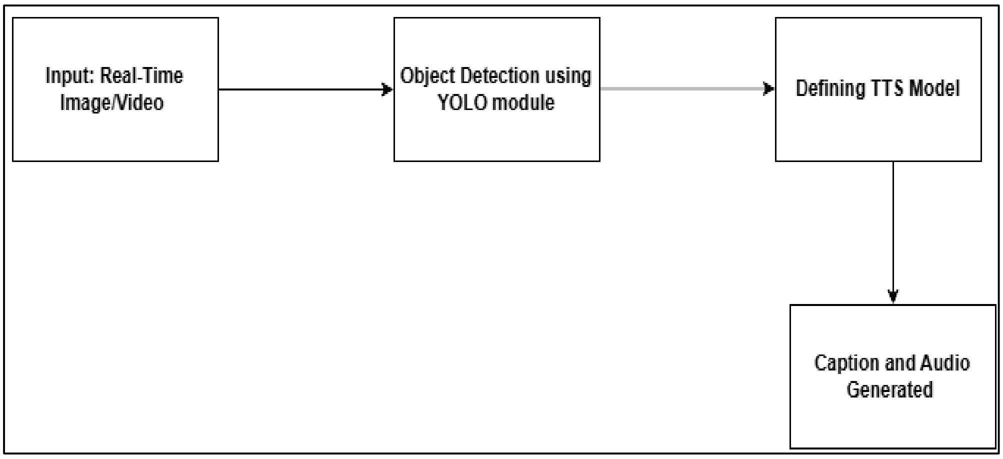
Input: Real-Time Image/Video
Object Detection using YOLO module
Defining TTS Model
Caption and Audio Generated

**Figure 1** Flow Diagram

### 4.1. Input Layer: Real-Time Image/Video Capture

The architecture begins with the input layer, where cameras strategically placed at various locations, such as traffic intersections or on moving vehicles, capture real-time images or videos of traffic scenes. These cameras continuously feed live visual data into the system, ensuring that the most current traffic conditions are analyzed. This real-time input is crucial for applications that rely on up-to-date information, such as traffic management and autonomous vehicle navigation.

### 4.2. Object Detection Layer: YOLO

The core of the object detection process is handled by the YOLO algorithm. YOLO is renowned for its ability to process images in real-time with high accuracy and speed. Unlike traditional methods that repurpose classifiers to perform detection, YOLO applies a single neural network to the full image. It divides the image into a grid and directly predicts bounding boxes and class probabilities for each grid cell. This approach allows YOLO to identify and locate various objects, such as cars, pedestrians, and traffic lights, quickly and accurately. The output of this layer is a set of detected objects, each with its bounding box coordinates and class labels, which form the foundational data for further analysis. Once the objects in the traffic scene are detected by YOLO, the next step is to generate textual descriptions of the scene. This approach ensures that the generated captions are straightforward and directly reflect the objects identified by YOLO.

### 4.3. Text-To-Speech (TTS) Layer

TTS synthesizes natural-sounding speech from the textual descriptions. The audio output provides real-time information about the traffic scene. The TTS layer is focused on providing critical audio warnings rather than general audio descriptions.

### 4.4. Output Layer

The final layer of the architecture is the output layer, which presents the generated textual and audio description to the end-users. The textual output can be displayed on screens or as overlays on the video feed, providing a visual summary of the traffic scene. The audio output, generated by the TTS engine, is played through speakers, offering real-time audio alerts and descriptions. This dual-modality output ensures that the information is accessible to a wide range of users.

## 5. Algorithms

### 5.1. YOLO Algorithm

The YOLO (You Only Look Once) algorithm is a popular and powerful deep learning model used for object detection. It was developed by Joseph Redmon and Ali Farhadi. YOLO stands out for its speed and accuracy, allowing real-time object detection in various applications. How YOLOv3 Works, Input Image: The input image is divided into a grid. Each grid cell is responsible for detecting objects whose centers fall within the cell. Feature Extraction: The Darknet-53 backbone processes the input image to extract features. These features are then used to predict bounding boxes and class probabilities. Scale Predictions: YOLOv3 predicts bounding boxes at three different scales: First Scale: Uses the features from the final convolutional layer. Second Scale: Uses features from an intermediate layer and up samples them to match the resolution of the next layer. Third Scale: Uses features from an even earlier layer, again up sampling to match the resolution. Anchor Boxes: Anchor boxes are predefined shapes that the network uses to predict the bounding boxes. YOLOv3 uses k-means clustering on the training dataset to determine the dimensions of these anchor boxes. Objectness Score: Each bounding box has an objectness score that indicates the likelihood of an object being present within the box. Non-Maximum Suppression (NMS): YOLOv3 applies NMS to remove duplicate boxes and keep only the ones with the highest confidence scores.

### 5.1.1. YOLOv3 Architecture

The architecture of YOLOv3 can be broken down into several stages: Convolutional Layers: These layers extract features from the input image using the Darknet-53 backbone. Residual Blocks: YOLOv3 uses residual connections (shortcuts) to improve gradient flow and allow for deeper networks. Detection Layers: The network has several detection layers that predict bounding boxes at different scales.

## 6. UML Diagram

### 5.2. TTS Algorithm

Integrating Text-to-Speech (TTS) technology with image captioning provides a means to convert text into natural language speech, enhancing accessibility for users. This integration enables individuals to receive audio descriptions of images, facilitating comprehension and accessibility. The audio output provides real-time information about the traffic scene. The primary function of the TTS layer is to convert the calculated distances between the detected objects and the system into audio warnings if the objects are too close.

- Distance Calculation: The system calculates the distance between the detected objects and the camera using methods such as the pinhole camera model. This involves determining the real-world distance based on the size and position of the objects in the image.
- Threshold Check: The system compares the calculated distances with predefined safety thresholds. If any object is detected within a dangerous proximity, the system generates a warning message.
- Audio Output: The TTS model converts these warning messages into audio.

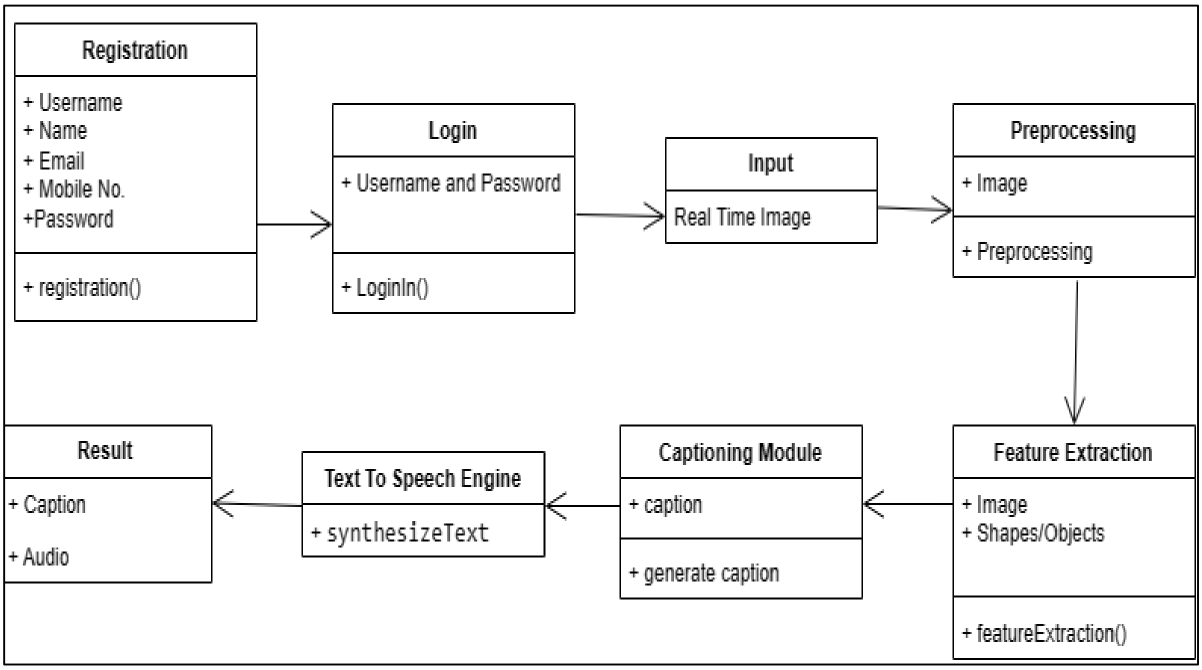

**Figure 2** Class Diagram

A class diagram for the system focused on generating caption and audio as output for real time image captured, which represents the various classes or entities within the system and their relationships. It provides structural overview of the system's components and how they interact. Figure 2 refers to the Class Diagram of the system and Figure 3 explains the Activity diagram of the system.

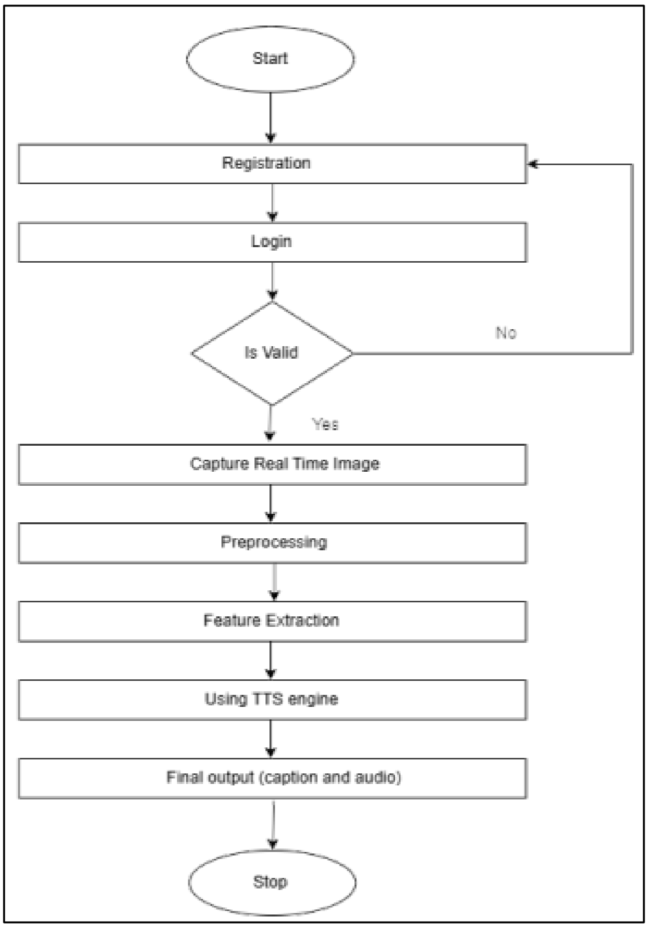

**Figure 3** **Activity Diagram**

## 7. Future Scope

The future scope of this project encompasses several avenues for further development and enhancement: Advanced Object Detection Techniques: Explore newer object detection algorithms beyond YOLOv3, such as YOLOv4 or EfficientDet, to improve accuracy and efficiency in detecting traffic objects. Semantic Segmentation: Incorporate semantic segmentation models to provide more detailed understanding of traffic scenes, enabling finer-grained analysis and captioning. Contextual Understanding: Enhance captioning models with contextual understanding to generate more descriptive and contextually relevant captions, taking into account scene dynamics and environmental factors. Multimodal Integration: Integrate additional modalities such as lidar data or radar signals to complement visual information, enhancing object detection and scene understanding in challenging conditions like low visibility or adverse weather. Adaptive Learning: Implement adaptive learning techniques to continuously improve the system's performance over time, learning from user interactions and feedback to refine object detection, captioning, and voice feedback.

## 8. Result

The results are shown in Figure 4 and Figure 5

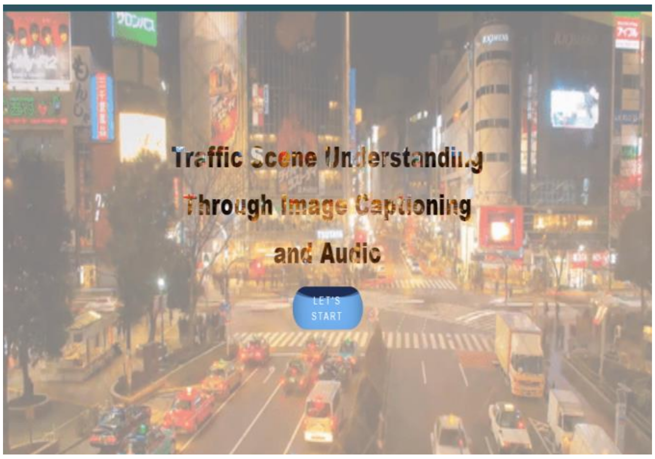

**Figure 4** **Result Page 1**

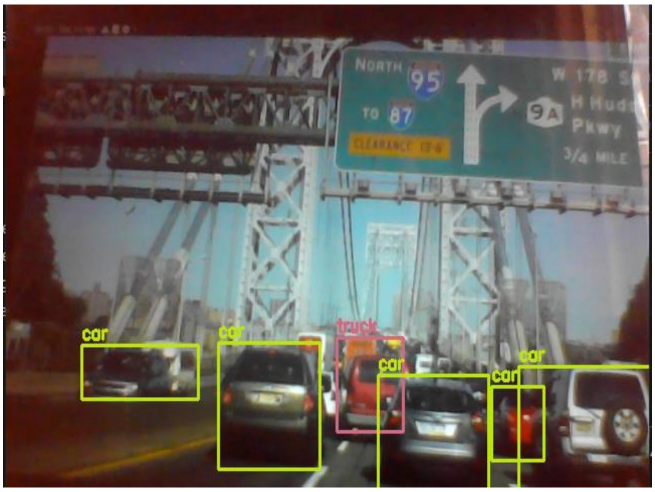

**Figure 5** **Result Page 2**

## Conclusion

This project provides a comprehensive overview of object detection methodologies within the realm of traffic scene understanding and recognition, showcasing the evolution from conventional multi-stage convolutional neural network architectures to the more efficient and accurate YOLO v3. By leveraging a single neural network to analyze entire images, YOLO significantly streamlines the object detection process, enabling real-time capabilities. Moreover, the project's outcome extends beyond mere detection by providing captions and audio feedback generated from detected traffic signs and objects. However, challenges persist, particularly in synchronizing voice feedback with detected objects. Despite these challenges, the advancements in object detection, coupled with the provision of captioning and audio feedback, hold tremendous promise for enhancing traffic management systems and elevating road safety standards.

## References

[1]. Chuan Wu, Yaochen Li, Ling Li, Le Wang, Yuehu Liu, “Caption Generation from Road Images for Traffic Scene Construction”, 19 Oct 2020, IEEE.

[2]. WEI LI 1,2, ZHAOWEI QU 1, HAIYU SONG 1,2, PENGJIE WANG 2, AND BO XUE2, “The Traffic Scene Understanding and Prediction Based on Image Captioning”, 2021, IEEE Access.

[3]. Mehdi Masmoudi, Hakim Ghazzai, Mounir Frikha, Yehia Massoud, "Object Detection Learning Techniques for Autonomous Vehicle Applications", 01 Sep 2019, IEEE.

[4]. P. Aishwarya Naidu1 *, Satvik Vats2, Gehna Anand3, Nalina V.4, “A Deep Learning Model for Image Caption Generation’. International Journal of Computer Sciences and Engineering on 6th June 2020.

[5]. V.Geetha, C K Gomathy, T. Harshitha, P. Vijay Nagendra Varma, “A Traffic Prediction for Intelligent Transportation System using Machine Learning,” in International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-10 Issue-4, April 2021.

[6]. Karthik N, Sudhir Shenai,"Enabling Object Detection and Distance Calculation in AI based Autonomous Driving System", IEEE, 13 December 2022.

[7]. Gauri Rao, Bhavya Divecha, Jenish Joshi, Anishk Jaiswal, "Traffic Light Management System Using Image Processing", July 2022, IJIRT Volume 9 Issue 2.

[8]. Xiaoyuan Liang, Xusheng Du, Guiling Wang, Zhu Han Fellow, "Deep Reinforcement Learning for Traffic LightControl in Vehicular Networks", 29 Mar 2018, IEEE.

[9]. Manisha M. Patil,"Experiment based on Deep Learning: Image Caption Generator",2021 IJCRT Volume 9, Issue 12 December 2021.

[10]. Jasmita Khatal, Prajkta Jadhav, Shraddha Parab,"Real Time Image Captioning and Voice Synthesis using Neural Network",International Research Journal of Engineering and Technology (IRJET),01 Jan 2021.

[11]. Yuki Mori; Tsubasa Hirakawa; Takayoshi Yamashita; Hironobu Fujiyoshi,"Image Captioning for Near-Future Events from Vehicle Camera Images and Motion Information", IEEE Intelligent Vehicles Symposium (IV), 1 Nov 2021.