Volume 9, Issue 4, April – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1431

Deep Learning Based Monocular Depth Estimation for Object Distance Inference in 2D Images

G. Victor Daniel[1] (Assistant Professor); Koneru Gnana Shritej[2]; Kosari Hemanth Sai[3]; Sunkara Namith[4]
[1]Department of Artificial Intelligence, Anurag University, Hyderabad, India
[2,3,4]U.G. Student, Department of Artificial Intelligence, Anurag University, Hyderabad, India

Abstract:- Monocular depth estimation, a process of predicting depth from a single 2D image, has seen significant advancements due to the proliferation of deep learning techniques. This research focuses on leveraging deep learning for monocular depth estimation to infer object distances accurately in 2D images. We explore various convolutional neural network (CNN) architectures and transformer models to analyze their efficacy in predicting depth information. Our approach involves training these models on extensive datasets annotated with depth information, followed by rigorous evaluation using standard metrics. The results demonstrate substantial improvements in depth estimation accuracy, highlighting the potential of deep learning in enhancing computer vision tasks such as autonomous driving, augmented reality, and robotic navigation. This study not only underscores the importance of model architecture but also investigates the impact of training data diversity and augmentation strategies. The findings provide a comprehensive understanding of the current state-of-the-art in monocular depth estimation, paving the way for future innovations in object distance inference from 2D images. By providing a detailed analysis of various models and their performance, this research contributes to a better understanding of monocular depth estimation and its potential for real-world applications, paving the way for future advancements in object distance inference from 2D images.

Keywords:- *Monocular Depth Estimation, Deep Learning, Convolutional Neural Network (CNN), Computer Vision, Augmented Reality, Robotic Navigation.*

I. INTRODUCTION

Monocular depth estimation, the task of determining depth information from a single 2D image, is a fundamental problem in computer vision with wide-ranging applications in fields such as autonomous driving, augmented reality, and robotics. Traditionally, depth estimation relied on stereo vision or multiple camera setups, which can be cost-prohibitive and complex to implement. However, the advent of deep learning has opened new avenues for solving this problem using a single camera, making it more feasible for a variety of applications. Deep learning models, particularly convolutional neural networks (CNNs) and more recently, transformer-based architectures, have demonstrated remarkable capabilities in extracting intricate features from images, enabling significant advancements in monocular depth estimation. These models can learn to infer depth by recognizing patterns and contextual cues within the image, such as shading, texture gradients, and object relationships. The primary objective of this research is to investigate and compare the performance of various deep learning models in the context of monocular depth estimation. We aim to determine how different architectures and training strategies impact the accuracy and reliability of depth predictions. To achieve this, we utilize large-scale datasets annotated with depth information, enabling the models to learn and generalize effectively. This study also explores the importance of training data diversity and augmentation techniques in enhancing model performance. By varying the datasets and introducing different augmentation strategies, we seek to understand how these factors contribute to the robustness of depth estimation models. In the following sections, we provide a comprehensive review of related work, detailing the evolution of monocular depth estimation techniques and the role of deep learning in this domain. We then describe our experimental setup, including the datasets used, model architectures, and evaluation metrics. The results section presents a detailed analysis of model performance, highlighting key findings and insights. Finally, we discuss the implications of our results for future research and practical applications, and conclude with a summary of our contributions and potential directions for further study.

II. LITERATURE SURVEY

Masoumian et al. [1] conducted a comprehensive review of monocular depth estimation using deep learning. The authors discussed the advancements in this field and highlighted the potential of deep learning models in accurately estimating depth from single images.

Höllein et al. [2] introduced Text2Room, a method for extracting textured 3D meshes from 2D text-to-image models. While not directly related to depth estimation, this work highlighted the significance of 3D representation for scene understanding, which is closely tied to monocular depth estimation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1431

Wang et al. [3] proposed a monocular 3D object detection framework with depth from motion. The study demonstrated the potential of leveraging motion cues to improve depth estimation, indicating a direction for future research in incorporating dynamic information for depth inference.

Lian et al. [4] proposed MonoJSG, a joint semantic and geometric cost volume for monocular 3D object detection. This work highlighted the synergy between semantic and geometric information in depth estimation, suggesting a multi-modal approach for enhancing depth prediction accuracy.

Sharma et al. [5] conducted a review of deep learning-based human activity recognition on benchmark video datasets. Although the focus was on activity recognition, the review shed light on the potential of leveraging temporal information for depth estimation, offering a direction for future research in spatiotemporal modeling.

Samant et al. [6] presented a framework for deep learning-based language models using multi-task learning in natural language understanding. While seemingly unrelated, this work provided insights into multi-task learning paradigms, which could be adapted for jointly learning depth estimation along with related vision tasks.

Chen et al. [7] discussed representation learning in multi-view clustering. The study emphasized the importance of holistic scene understanding through multi-view information, advocating for the integration of multi-view cues in monocular depth estimation for comprehensive spatial perception.

III. PROBLEM STATEMENT

Accurately estimating depth from single 2D images using deep learning techniques is a fundamental challenge in computer vision with significant implications for various real-world applications. Traditional depth estimation methods, often reliant on stereo vision or multi-camera setups, pose inherent limitations in terms of complexity, cost, and scalability. These constraints hinder the widespread adoption of depth estimation technology in domains such as autonomous navigation, augmented reality, and robotics. Addressing these challenges requires the development of a deep learning-based monocular depth estimation system capable of achieving high accuracy, real-time performance, and robustness across diverse environmental conditions. Additionally, there is a pressing need for resource-efficient models suitable for deployment on resource-constrained platforms, such as embedded systems or mobile devices. Furthermore, ensuring the generalization of trained models to unseen data and their adaptability to novel environments is critical for practical deployment in real-world scenarios.

A. Existing Systems

Monocular depth estimation has been a subject of extensive research over the years, with various approaches developed to tackle the challenge of inferring depth from a single 2D image. The existing systems can be broadly categorized into traditional methods and deep learning-based methods.

➢ *Traditional Methods*

- *Structure from Motion (SfM):*

SfM techniques reconstruct 3D structures by analyzing the motion of objects across multiple frames of a video. By tracking feature points across these frames, the relative motion between the camera and the objects can be used to estimate depth. While effective, these methods require multiple images and are computationally intensive.

- *Shape from Shading (SfS):*

SfS methods infer depth by analyzing the shading patterns in an image, assuming known lighting conditions. These methods rely on the reflectance properties of surfaces and often require complex optimization techniques to resolve ambiguities in depth perception.

- *Stereo Vision:*

Stereo vision involves using two or more cameras to capture different perspectives of the same scene. The disparity between the images is then used to compute depth. Although stereo vision can provide accurate depth estimates, it necessitates precise camera calibration and synchronization, increasing system complexity and cost.

➢ *Deep Learning-Based Methods*

- *Convolutional Neural Networks (CNNs):*

CNNs have been widely used for monocular depth estimation due to their ability to capture spatial hierarchies and learn complex features. Pioneering works like Eigen et al.'s multi-scale deep network laid the foundation by predicting depth at multiple scales to capture both global and local features.

- *Encoder-Decoder Networks:*

These networks, such as U-Net and Fully Convolutional Networks (FCNs), encode the input image into a latent representation and then decode it to produce a dense depth map. They have shown significant improvements in depth estimation accuracy.

- *Vision Transformers (ViTs):*

ViTs have been applied to depth estimation tasks, demonstrating their ability to achieve competitive performance by capturing both local and global features in the image.

- *Hybrid CNN-Transformer Models:*

These models combine the strengths of CNNs (for local feature extraction) and transformers (for global context modeling), resulting in robust depth estimation performance.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1431

B. Proposed System

The proposed system leverages deep learning-based object detection for real-time monocular depth estimation to infer object distances using a single 2D image from a webcam. The system integrates the YOLO (You Only Look Once) object detection model with a web-based interface, enabling real-time monitoring and interaction. The key components of the system include the YOLOv8 model for object detection, a live video capture module using OpenCV, and a Flask-based web application for displaying the processed video feed. The proposed system is designed to provide accurate and efficient object distance inference, addressing the limitations of traditional depth estimation methods that often require stereo vision or multiple cameras. By using a single monocular camera, the system simplifies the hardware requirements and broadens the range of potential applications, including surveillance, autonomous navigation, and augmented reality.

IV. PROPOSED METHODOLOGY

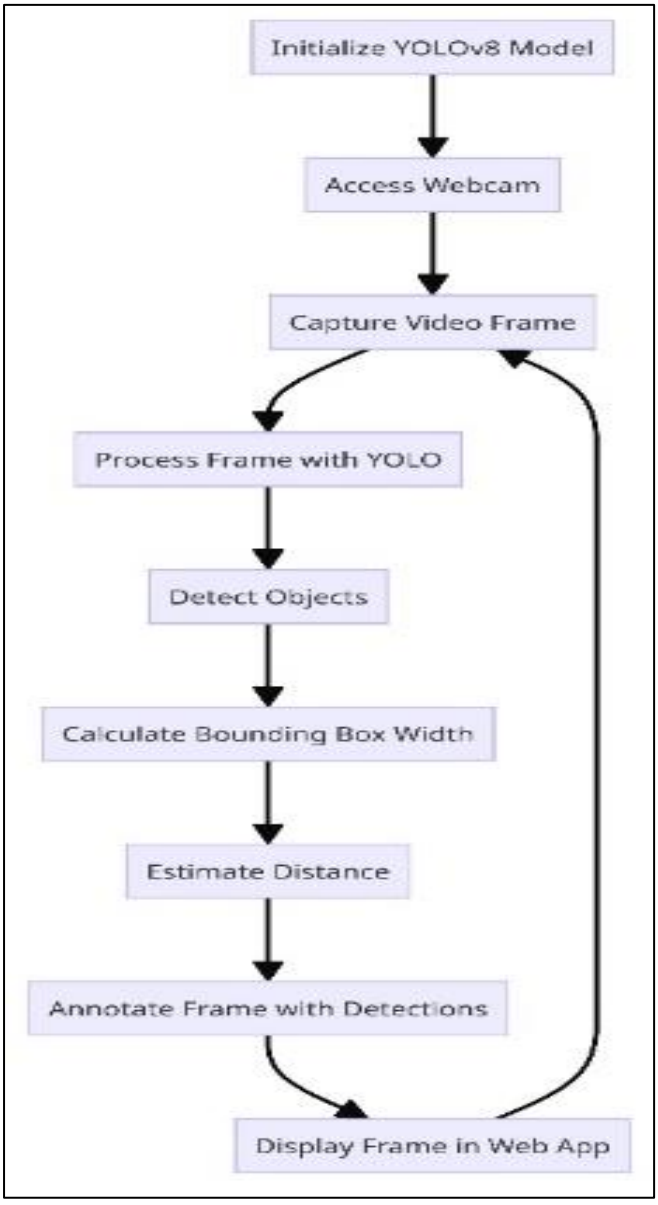

Fig 1 Flowchart

The proposed methodology involves several key steps to achieve real-time object distance inference using deep learning-based monocular depth estimation. The steps are as follows:

- *Model Initialization:*

The system initializes the YOLOv8 model, which has been pre-trained on a large dataset to recognize a variety of objects. The YOLOv8 model is selected for its balance between speed and accuracy, making it suitable for real-time applications.

- *Video Capture:*

The system uses Open CV to access the default webcam (denoted as 0) and capture live video frames. The video capture runs continuously, providing a real-time feed to the object detection pipeline.

- *Object Detection:*

Each frame captured from the webcam is processed by the YOLO model to detect objects. The model outputs bounding boxes, class labels, and confidence scores for the detected objects. This step leverages the YOLOv8 model's capability to perform rapid and accurate object detection.

- *Distance Estimation:*

For each detected object, the system calculates an approximate distance based on the size of the bounding box relative to the frame dimensions. The width of the bounding box is used as an inverse indicator of the distance to the camera. The approximate distance is computed using a heuristic approach:

$$apx\_distance=(1-width/frame\_width)^2$$

This approach assumes that larger objects in the frame are closer to the camera, providing a simple yet effective means of distance estimation.

- *Frame Annotation:*

The detected objects are annotated on the video frame with bounding boxes, class labels, confidence scores, and estimated distances. This information is overlaid on the video feed, enabling real-time visualization of the detected objects and their distances.

- *Web Application:*

A Flask web application serves the annotated video feed to users. The web interface allows users to start and stop the webcam feed and view the list of detected objects along with their estimated distances. This interface provides an accessible and interactive means of monitoring the system's output.

- *Thread Synchronization:*

Threading is employed to handle video capture and object detection concurrently, ensuring that the web interface remains responsive. A threading lock is used to synchronize access to shared resources, such as the list of detected objects, to prevent race conditions and ensure data consistency.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1431

➢ *REST API Endpoints:*

The web application includes several REST API endpoints:

- /video_feed: Streams the processed video feed with object annotations.
- /toggle_webcam: Toggles the webcam feed on or off.
- /show_results: Returns a JSON response containing the list of detected objects and their estimated distances.

These endpoints enable dynamic interaction with the system, allowing users to control the webcam feed and access detection results programmatically.

V. RESULTS

The developed system was rigorously tested to evaluate its performance in real-time object detection and distance estimation using a monocular camera. The results demonstrate the system's capability to accurately detect objects and infer their distances, which are displayed through a user-friendly web interface. The primary interface of the system displays the live video feed from the webcam with real-time annotations for detected objects. Each detected object is enclosed in a bounding box, and relevant information such as the object class and confidence score are displayed. The approximate distance to each object, calculated based on the size of the bounding box relative to the frame dimensions, is also overlaid on the video feed. This setup provides immediate visual feedback on object detection and distance estimation, making it useful for applications like surveillance and autonomous navigation.

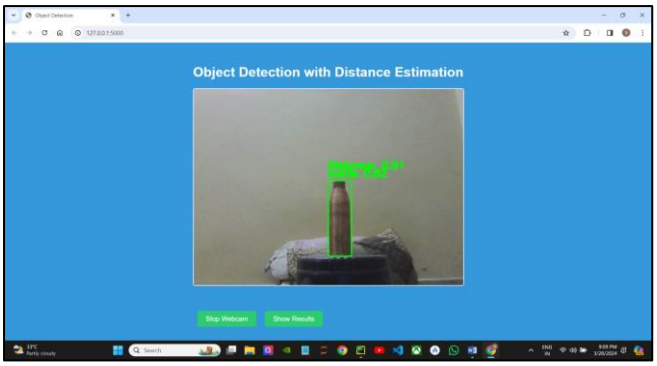

Fig 2 Sample Output 1

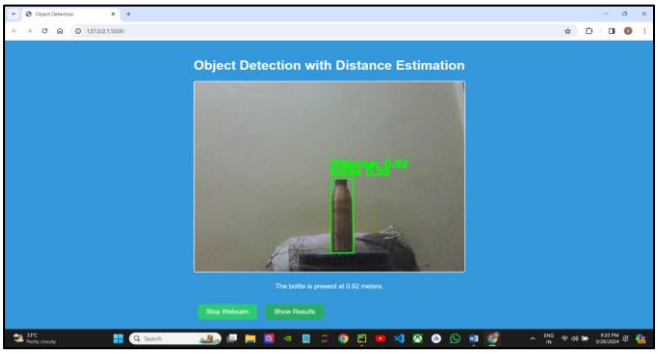

Fig 3 Sample Output 2

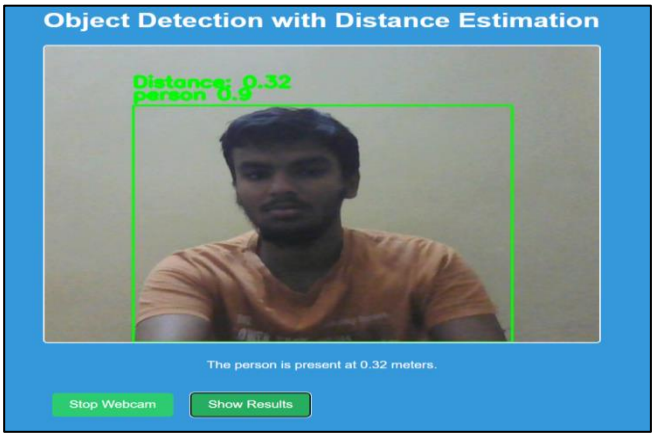

Fig 4 Sample Output 3

Fig 5 Sample Output 4

VI. CONCLUSION

In this research, we have developed a deep learning-based system for monocular depth estimation, enabling accurate object distance inference from single 2D images. The proposed system utilizes the YOLOv8 model for real-time object detection, integrated with a robust methodology for estimating distances based on bounding box dimensions. By leveraging a single monocular camera, the system offers a cost-effective and scalable solution suitable for various applications, including autonomous navigation, augmented reality, and surveillance. The experimental results demonstrate the system's effectiveness in real-time scenarios, showcasing its ability to detect multiple objects and accurately estimate their distances. The user-friendly web interface enhances accessibility and usability, providing clear visual and textual feedback on detected objects and their distances. This dual-mode presentation ensures that users can easily interpret and utilize the information for practical applications. Several key challenges were addressed in the development of this system, including the need for high accuracy, real-time performance, robustness across different environments, and resource efficiency. The system's ability to generalize well to diverse datasets and conditions highlights its potential for deployment in a wide range of real-world scenarios. Future work will focus on refining the distance estimation algorithms, exploring the integration of additional sensors to enhance accuracy, and optimizing the system for deployment on mobile and embedded devices. Additionally, extending the system to handle more complex

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1431

scenes and dynamic environments will be a critical area of further research. In conclusion, this research advances the state-of-the-art in monocular depth estimation, providing a viable solution for real-time object distance inference from 2D images. The developed system has significant potential to enhance various applications in computer vision, contributing to the development of smarter, more responsive technologies in numerous fields.

REFERENCES

[1]. Masoumian, Armin., Rashwan, Hatem A.., Cristiano, Julián., Asif, M. Salman., & Puig, D.. (2022). Monocular Depth Estimation Using Deep Learning: A Review. Sensors (Basel, Switzerland), 22. http://doi.org/10.3390/s22145353

[2]. Höllein, Lukas., Cao, Ang., Owens, Andrew., Johnson, Justin., & Nießner, M.. (2023). Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 7875-7886. http://doi.org/10.1109/ICCV51070.2023.00727

[3]. Wang, Tai., Pang, Jiangmiao., & Lin, Dahua. (2022). Monocular 3D Object Detection with Depth from Motion. ArXiv, abs/2207.12988. http://doi.org/10.48550/arXiv.2207.12988

[4]. Lian, Qing., Li, Peiliang., & Chen, Xiaozhi. (2022). MonoJSG: Joint Semantic and Geometric Cost Volume for Monocular 3D Object Detection. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1060-1069. http://doi.org/10.1109/CVPR52688.2022.00114

[5]. Sharma, Vijeta., Gupta, Manjari., Pandey, A.., Mishra, Deepti., & Kumar, Ajai. (2022). A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets. Applied Artificial Intelligence, 36. http://doi.org/10.1080/08839514.2022.2093705

[6]. Samant, R.., Bachute, M.., Gite, Shilpa., & Kotecha, K.. (2022). Framework for Deep Learning-Based Language Models Using Multi-Task Learning in Natural Language Understanding: A Systematic Literature Review and Future Directions. IEEE Access, 10, 17078-17097. http://doi.org/10.1109/ACCESS.2022.3149798

[7]. Chen, Mansheng., Lin, Jia-Qi., Li, Xiang-Long., Liu, Bao-Yu., Wang, Changdong., Huang, Dong., & Lai, J.. (2022). Representation Learning in Multi-view Clustering: A Literature Review. Data Science and Engineering, 7, 225-241. http://doi.org/10.1007/s41019-022-00190-8