Indoor Navigation for the Blind: Real-Time Object Detection and Mapping

Abhishek Bansal
IIITD, Delhi
abhishek22021@iiitd.ac.in

Dhruv Sharma
IIITD, Delhi
dhruv22170@iiitd.ac.in

Pratyush Gupta
IIITD, Delhi
pratyush22375@iiitd.ac.in

Vinayak Agrawal
IIITD, Delhi
vinayak22574@iiitd.ac.in

1. Problem Statement

Monocular depth estimation models are often too computationally intensive for low-cost edge devices, leading to unsafe delays in obstacle detection for blind navigation. While lightweight architectures and compression reduce inference time, they typically sacrifice depth precision in cluttered indoor environments. This project optimizes depth estimation pipelines for reliable, real-time performance on resource-constrained hardware by co-designing efficient model architectures, hardware-aware optimizations, and robust navigation logic. The resulting system, validated in dynamic indoor settings, processes a single RGB camera feed (e.g., from a smartphone) to produce depth maps integrated with obstacle detection, door recognition, and path planning cues.

2. User Interface Description

The application's user interface is composed of two primary screens that emphasize simplicity, accessibility, and clarity. The first screen presents a full-size background image accompanied by a prominent "Start Navigating" prompt, serving as a clear call-to-action for users. A large, centrally placed "Start" button, positioned near the bottom of the screen, allows users to easily initiate the navigation or object detection process. Upon activation, the interface transitions to a second screen where a live camera feed dominates the display area, capturing real-time visual data. Below this feed, automatically generated text provides detailed descriptions of objects or obstacles within view, ensuring that visually impaired users receive immediate, comprehensible feedback. Additionally, a conspicuous "TTS" (text-to-speech) button is available, enabling users to have the displayed text read aloud on demand, thereby reducing reliance on visual cues and enhancing accessibility.

Furthermore, it is planned that after the interim submission, the application will be developed as a responsive website using NextJS. This web-based implementation will integrate the same two-screen design, allowing users to access real-time navigation assistance via their browsers on various devices, while leveraging NextJS to ensure fast performance and an optimal user experience.

3. Literature Review

Navigating indoor environments for visually impaired individuals is complex due to the need for real-time obstacle detection, precise spatial mapping, and reliable depth estimation. Prior work has addressed these challenges through RGB-D scene understanding, deep learning-based positioning, and advanced depth prediction methods, all of which inform our approach.

Silberman et al. [12] introduced the NYU Depth V2 dataset, a comprehensive resource that supports indoor scene segmentation by analyzing RGB-D data and inferring spatial relationships among objects. Their work emphasizes how depth cues are crucial for identifying structural elements—walls, floors, and obstacles—that are essential for safe navigation. The dataset's diversity aids in training models that generalize well across varied indoor environments.

Mahida et al. [6] explored deep learning-based positioning techniques specifically designed for indoor navigation assistance. By integrating multiple sensor inputs, they achieved refined, real-time position estimation. This multi-sensor fusion strategy underscores the importance of combining visual data with other sensor modalities to provide adaptive and precise navigational support.

Sathyamoorthy et al. [11] developed the DenseCAvoid framework for real-time obstacle avoidance in dynamic environments. Although primarily aimed at human-robot interaction, their approach—employing anticipatory behaviors and trajectory adjustments—can be adapted to enhance obstacle detection and avoidance in crowded indoor settings, ensuring the system remains responsive to rapid changes.

Eigen et al. [2] proposed a multi-scale deep network for predicting depth maps from single RGB images. This method is particularly relevant in scenarios where traditional depth sensors are unavailable or unreliable. By inferring depth directly from visual cues, their approach provides a robust alternative that, when combined with other sensor inputs, enhances the accuracy and reliability of obstacle detection and spatial mapping.

4. Datasets, Evaluation Metrics, and System Analysis

4.1. Dataset Description and Exploratory Data Analysis (EDA)

The NYU Depth V2 dataset serves as a critical benchmark for indoor scene understanding, providing high-resolution RGB images and precisely calibrated depth maps captured via a Microsoft Kinect sensor.

Data Composition and Modalities

- **RGB and Depth Images:** Detailed images with corresponding depth maps capture the spatial layout of indoor scenes, enabling accurate obstacle detection and navigation mapping.
- **Monocular Depth Estimation:** Leverages single-channel depth estimation techniques for scenarios with limited sensor data.
- **Annotations:** Semantic segmentation labels and ground truth depth data support rigorous model training and evaluation.

Relevance to Indoor Navigation for the Blind

The dataset addresses navigation challenges by:

- **Spatial Mapping:** Generating detailed, real-time indoor navigational maps.
- **Obstacle Detection:** Fusing RGB and depth information to detect obstacles in complex environments.
- **Environment Diversity:** Providing images across various indoor settings to ensure system robustness.

EDA Insights

Key observations include:

- **Data Variability:** Challenges from lighting variations and image quality disparities.
- **Structural Limitations:** Inherent class imbalance and domain shifts across indoor scenes.

4.2. Evaluation Metrics

The performance of our models was assessed using both standard and task-specific metrics:

Depth Estimation Metrics

- *Absolute Relative Error (AbsRel)*
- *Root Mean Squared Error (RMSE)*
- *RMSE_log* (logarithmic scale)
- *Delta Accuracies* with thresholds: $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$

These metrics allow us to evaluate the accuracy of both paired depth maps and **monocular depth estimation** outputs.

4.3. System Architecture and Model Compression

Object Detection with YOLOv8-nano

To achieve real-time performance with limited computational resources, we employed the **YOLOv8-nano** model from the `ultralytics` package. Key advantages include:

- **Model Compression:** The nano version significantly reduces the model size and inference time, which is crucial for deployment in low-resource settings.
- **Detection Accuracy:** Despite its compactness, YOLOv8-nano maintains competitive accuracy in detecting obstacles, which is essential for safe navigation.

Depth Estimation Models

For depth estimation, initial attempts with FastDepth encountered repository issues. We therefore adopted the **DPT_Hybrid** model from the MiDaS repository. Both **MiDaS_small** and **DPT_Hybrid** models were evaluated, with a focus on integrating **monocular depth estimation** The models are benchmarked using the aforementioned depth estimation metrics.

4.4. Analysis of Results and Sample Outputs

Quantitative Analysis

The table below summarizes the average depth metrics for the evaluated models: Reference Table 1

Qualitative Analysis and Visualizations

Extended visualizations were generated to illustrate model performance:

- **Object Detection:** Sample outputs (Figures 1–3) display detection results using YOLOv8-nano, with bounding boxes and confidence scores overlaid on the RGB images.
- **Depth Estimation:** Figures 4 and 5 show sample depth maps produced by MiDaS and DPT_Hybrid, respectively. These include results from both paired depth estimation and **monocular depth estimation** methods.

4.5. Compute Requirements and Resource Analysis

Efficient resource management is essential for real-time assistive applications:

- **Computing and Storage:** Training and evaluation were performed on GPU systems with moderate storage needs (Google Colab T4) for the NYU Depth V2 dataset.
- **RAM and Energy:** Batch processing and models like YOLOv8-nano minimize memory usage

Average Inference Time Plots

To illustrate performance, we included plots comparing the average inference times for the object detection and depth estimation models. (Figures 6 and 7)

5. Individual Contributions

All team members contributed equally to the project Their individual contributions are detailed below.

- **Abhishek Bansal**
  - **Work done till interim submission:** Developed the user interface, described and analyzed the NYU dataset, and contributed to report writing.
  - **Tentative future work:** Complete the user interface part fully.
- **Dhruv Sharma**
  - **Work done till interim submission:** Researched, found, and analyzed the NYU dataset, and helped in report writing.
  - **Tentative future work:** Perform benchmarking on the SUNRGB dataset and analyze potential models and techniques.
- **Pratyush Gupta**
  - **Work done till interim submission:** Conducted initial benchmarking on the NYU dataset, analyzed and identified the SUNRGB dataset, and contributed to report writing.
  - **Tentative future work:** Assist in analyzing data handling strategies (preprocessing and post processing) and in hyperparameter finetuning.
- **Vinayak Agrawal**
  - **Work done till interim submission:** Analyzed related work through research papers and contributed ideas which future helped in developing our approach
  - **Tentative future work:** Implement the Text-to-Speech feature and help in running the complex models that can handle 3D data.

6. Future Work

Building on our initial benchmarking with the NYU Depth V2 dataset, we now plan to use the comprehensive SUN RGB-D dataset [13] for a robust indoor navigation system for the blind. SUN RGB-D's 10,335 real RGB-D images and rich 2D/3D annotations provide a superior platform for developing models that achieve total scene understanding.

End-to-End Pipeline Development

Our objective is to integrate the following modules into a unified real-time system:

- **Object Detection:** Fine-tune SOTA models such as YOLOv5 [4] and Faster R-CNN [10] on SUN RGB-D, incorporating RGB and depth fusion for improved detection in cluttered indoor scenes.
- **Semantic Segmentation:** Use DeepLabv3 [1] and Fully Convolutional Networks (FCNs) [5] to produce precise pixel-wise labels, benchmarked via mean Intersection over Union (mIoU) and pixel accuracy.
- **Monocular Depth Estimation:** Integrate models like MiDaS [9] or Monodepth2 [3] to predict depth from RGB images, ensuring robust performance even with noisy sensor data.
- **3D Mapping and SLAM:** Incorporate a real-time SLAM module (e.g., ORB-SLAM2 [7]) and explore learning-based methods (e.g., VoteNet [8]) to continuously reconstruct indoor environments.
- **Additional Tasks:** Extend the pipeline to include 3D object orientation, room layout estimation, and other scene understanding tasks, all of which contribute to accurate obstacle detection and spatial awareness.
- **Multi-Modal Fusion and Feedback:** Design fusion layers to merge RGB, depth, and semantic features, generating a comprehensive spatial representation that supports auditory real-time feedback as mentioned in the user design section.

Benchmarking and Evaluation

- **Detection and Segmentation:** Evaluated via metrics such as mean Average Precision (mAP) and mIoU.
- **Depth Estimation:** Measured with error metrics such as Root Mean Squared Error (RMSE) and absolute relative difference.
- **Mapping Accuracy:** Assessed via trajectory error and 3D reconstruction accuracy.
- **Real-Time Efficiency:** Monitored via frames per second (FPS) and inference latency to ensure our system meets real-time constraints on platforms such as Google Colab with a T4 GPU.

Impact on Indoor Navigation for the Blind

The fully integrated pipeline will enhance indoor navigation by providing accurate obstacle detection and spatial mapping for safe path planning, utilizing monocular depth estimation to compensate for sensor noise, delivering comprehensive scene understanding through object orientation and room layout estimation, and enabling auditory feedback to assist visually impaired users.
The below demonstration video outlines the capabilities of the SUNRGB dataset: https://www.youtube.com/watch?time_continue=117&v=fOQdC7aeIr8.

Metric	MiDaS_small	DPT_Hybrid
AbsRel	1.2269	1.2840
RMSE_log	1.3655	1.1121
$\delta < 1.25$	0.1862	0.1847
$\delta < 1.25^2$	0.3441	0.3450
$\delta < 1.25^3$	0.4794	0.4788

Table 1. Average Depth Metrics for the Evaluated Models.

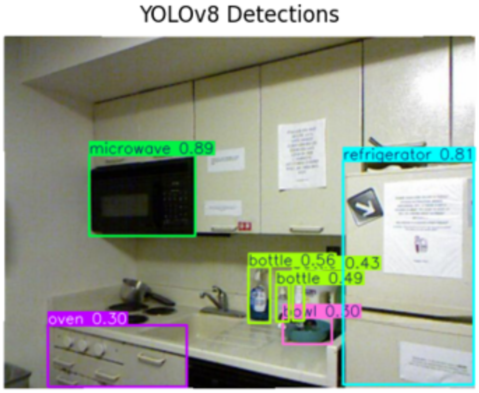

Figure 1. Sample object detection output using YOLOv8-nano.

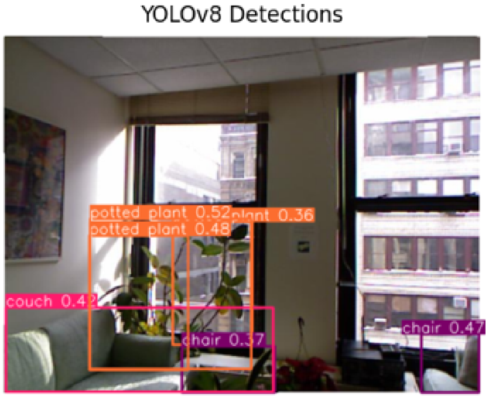

Figure 2. Additional sample detection output from YOLOv8-nano.

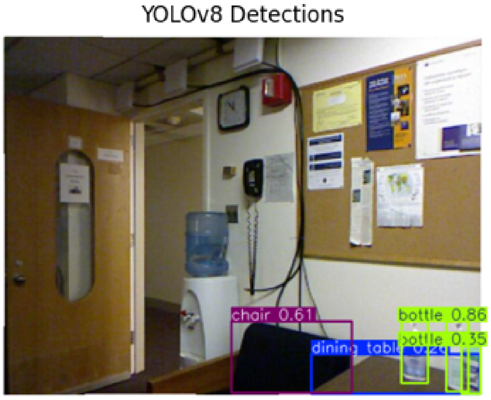

Figure 3. Another detection output using YOLOv8-nano.

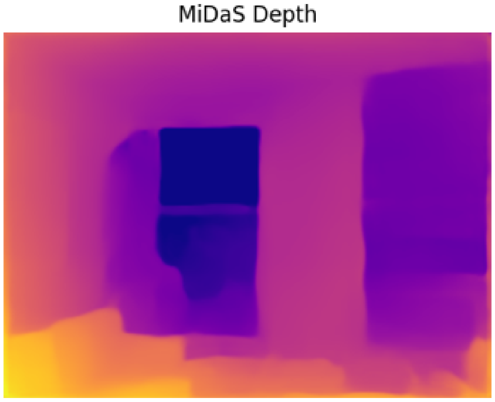

Figure 4. Sample depth map predicted by MiDaS.

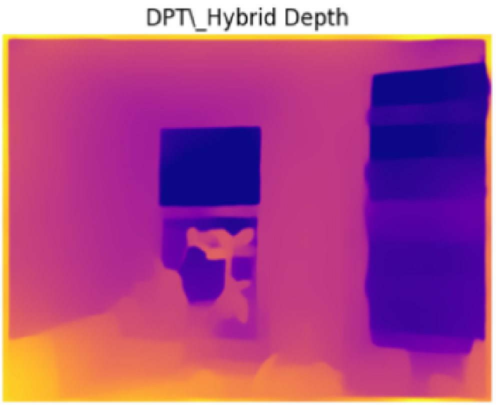

Figure 5. Sample depth map predicted by DPT_Hybrid.

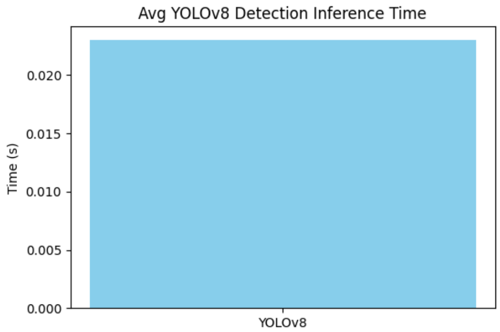

Figure 6. Avg YOLOV8 Detection time

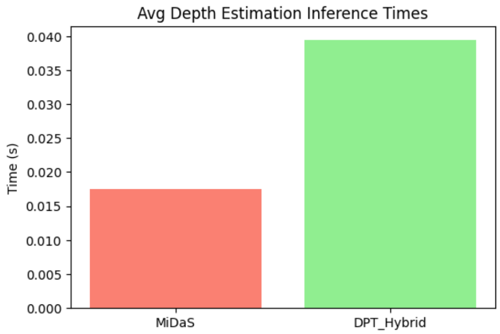

Figure 7. Avg depth estimation inference time

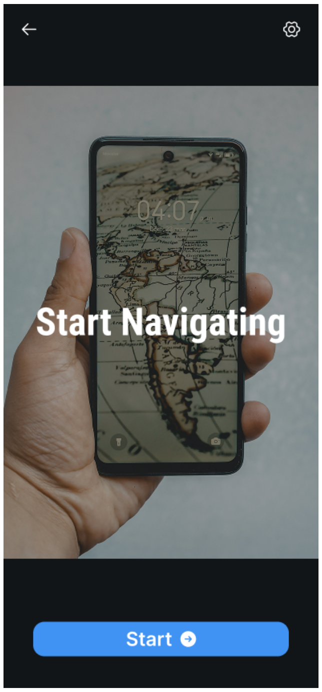

Figure 8. Landing page of the app

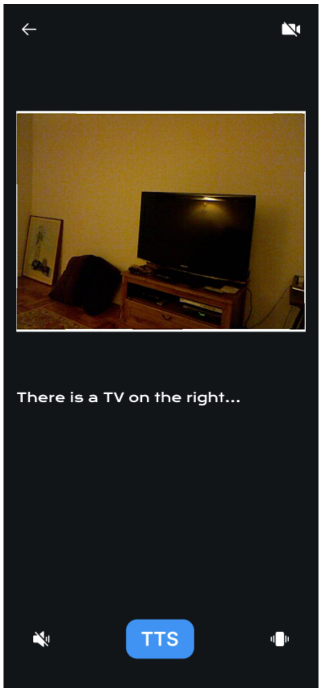
There is a TV on the right...

Figure 9. Main page that takes the video feed

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Rethinking atrous con-

volution for semantic image segmentation. In *arXiv preprint arXiv:1706.05587*, 2017. 3

[2] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, 2014. 2

[3] Clément Godard, Oisin Mac Aodha, Mike Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[4] Glenn Jocher. Ultralytics yolov5. `https://github.com/ultralytics/yolov5`, 2020. 3

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3

[6] Payal Mahida, Seyed Shahrestani, and Hon Cheung. Deep learning-based positioning of visually impaired people in indoor environments. In *Proceedings of CVPR*, 2025. 1

[7] Raúl Mur-Artal and Juan D. Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. In *IEEE Transactions on Robotics*, 2015. 3

[8] Charles R. Qi, Yang Liu, Cheng Wu, Hao Su, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[9] Rupert Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 3

[11] Adarsh Jagan Sathyamoorthy, Jing Liang, Utsav Patel, Tianrui Guan, Rohan Chandra, and Dinesh Manocha. Densecavoid: Real-time navigation in dense crowds using anticipatory behaviors. In *Proceedings of CVPR*, 2025. 1

[12] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, 2012. 1

[13] Scott Song, Jean Ponce, et al. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015. 3