PROXI-NAV: Proximity-Aware Navigation with Audio-Visual Sensing Intelligence

Aswinkumar V
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, India.
aswinkumar.v9805@gmail.com

Akshita Jawahar
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, India.
akshita1009@gmail.com

Sreya Mynampati
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, India.
sreyamynampati@gmail.com

Sajidha S A*
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, India.
sajidha.sa@vit.ac.in

Sumaiya Thaseen
School of Computing Science and Informatics
De Montfort University, Dubai Internet City, Dubai, UAE
sumaiyathaseenikram@ieee.org

Abstract

Pedestrians with a visual disability can experience continuous difficulties when navigating the ever-changing and demanding pedestrian ecosystem. Mobility assistance tools typically provide little to no situational awareness or context to help such pedestrians. Although systems that provide assistance using information from eyes have shown promise in helping these pedestrians navigate; however, they primarily utilize heuristic fusion from multiple sensory perceptions, which results in context-dependent and brittle navigation techniques. This paper introduces **PROXI-NAV**, a learning-based navigation assisting framework that employs a dense spatial risk map to model navigation safety using monocular vision. Using this spatial risk map, a *RiskCNN* network combining multiple complementary perceptual predictors is created to assist visually disabled individuals in making safe navigation decisions. Pseudo risk labels created using perceptual signals during weakly supervised training allow for the creation of these types of networks without the cost of manually labeling risks. RiskCNN eliminates the reliance on heuristic fusion during inference; therefore, it can produce smooth and understandable risk heatmaps that support navigation decision making. The entire system is implemented using an NVIDIA Jetson Nano edge computing device and provides real-time audio assistance for use in wearable assistive devices. The results of the tests show valid and consistent navigation cue feedback for visually disabled individuals when tested in real-life pedestrian environments.

Index Terms

Assistive navigation, object detection, depth-aware perception, monocular vision, navigation risk modeling, weakly supervised learning, edge AI, real-time audio guidance.

I. Introduction

Individuals who use convectional mobility aids such as white canes and guide dogs to aid them in avoiding obstacles at close range rely mostly on their own experience in how to navigate and expect obstacles to move in approach with the person using the cane. However, these devices do not provide any assistance in avoiding or predicting hazards beyond the range of the device used. Therefore, the ability for individuals living with visual impairments to independently navigate safely in outdoor pedestrian environments where there are dynamic obstacles, irregularly shaped surfaces, or limited space is important to their overall mobility and independence.

Advances in embedded computing and computer vision have created more interest in assistive navigation systems that utilize vision. Using lightweight cameras alongside object detection, semantic segmentation, and frequent monocular depth, these systems allow users to perceive a more robust representation of their surrounding environment through a wearable platform. However, even though there has been significant technological progress made, many current systems are still limited by the fact that they utilize heuristic rules or threshold-based logic as methods to convert the perceptual outputs into navigation guidance. Due to this reliance on heuristic rules and thresholds as the basis for navigation guidance, this approach has many disadvantages. Sensitivity to environmental variability, the need for extensive manual calibration across different environments, and the tendency to provide unreliable or non-consistent navigation guidance within cluttered or unfamiliar scenes are all limitations associated with this method of operation.

A primary limitation of this approach is that it doesn't very clearly represent safety in navigation. Although the perception module contains information about *what* is present in the environment, it doesn't directly encode *where* it is safe to walk. Effective navigation for pedestrians relies upon the interplay of many different cues, including the proximity of obstacles, the continuity of surfaces, and the local geometric context, as well as the way those cues interact geometrically, all of which cannot be modelled effectively through fixed analytical rules. In addition, using fixed analytical rules does not account for how pedestrians adaptively use these cues in real time.

The limitation of assistive navigation has been addressed in this work by viewing this issue through the lens of the learning of a spatial risk, as opposed to directly mapping the perceptual information to a navigation command. In the proposed framework, the dense spatial risk field (which shows how safe the user is at any given point in his/her field of view to travel to) becomes the intermediate representation that decouples the perceptual domain from the decision-making domain and allows for a better understanding of the user's decision-making process based on current context. By learning this representation from experience or data, an assistive navigation system is able to learn nonlinear spatial relationships that can be very difficult to write down with human-created decision-making policies.

The lack of clearly defined ground truth labels for navigation risk makes it difficult to train models to properly represent them. The solution we propose is a weakly supervised approach in which surrogate navigation risk labels are created from perceptual cues. Therefore, this allows structured supervision of the model while also allowing the model to learn on its own based on spatial regularities and contextual dependencies. Thus, our method permits scalable training beyond initial formulation simultaneously without the need to rely on subjective manual annotation.

The resulting system provides wearable assistance to those in need by optimizing each component to run on an embedded edge platform. Because of this, monocular video streams can be processed continuously, allowing for low-latency generation of auditory navigation cues. With this combined approach of learning-based risk modelling to detect hazards and real-time sensor information, this framework will make it easier and safer for pedestrians to navigate through their everyday lives.

II. Related Work

Much effort has been devoted to vision-based assistive navigation to promote independent mobility of the visually impaired. Early works have mostly integrated camera-based perception with audio feedback to improve environmental awareness. Li et al. [1] introduced a vision-based indoor navigation aid that coupled object detection with audio cues, showing improved situational awareness, but with dependence on predefined navigation logics. More recently, advances on monocular depth estimation enabled richer geometric reasoning from a single camera. Ranftl et al. [2]• showed robust generalization of depth across diverse environments.

Learning-based navigation methods have also been pursued to transcend the shortcomings of rule-based systems. Lu et al. [3] deployed deep reinforcement learning for assistive navigation in robotic settings; indeed, adaptive behavior ensued at the price of burdensome training requirements. Phone-based navigation assistants like DeepNAVI [4] further illustrated the feasibility of deep learning for assistive guidance but kept coupling of navigation decisions tight with handcrafted heuristics. Complete state-of-the-art surveys [5] detail the fact that heuristic fusion and manual tuning continue to be major bottlenecks in attaining high real-world robustness performance.

A way which emerged as promising for such cases is weakly supervised learning. Chang et al. [6] proved the applicability of weakly supervised learning for such issues as navigation perception by learning meaningful representations in such conditions. More recently, risk-aware learning techniques have been explored for such purposes as learning traversal representations. Cai et al. [7] proved the applicability of evidential traversal learning for autonomous agents.

The need for embedded deployment has been raised in various assistive systems. Embedded cognitive assistance systems using deep learning have been proposed for real-time feasibility even on low-power systems [8], [9]. Along with the development of modern methods of object detection, real-time detection has been gaining application in assistive systems even based on wearables or edge computing. Applications of YOLOv8 have been demonstrated by Jeong et al. [10] for outdoor assistive systems and by Saranya Arulselvarani [11] for assistive systems in indoors. Performance analysis of real-time detection systems also emphasis various trade-offs regarding accuracy, complexity, and latency [12].

The application of monocular depth estimation has also been included in the design of embedded assistive systems to enhance the perception of proximity. The work by Anjom et al. [13] included depth estimation in a real-time notification system for the

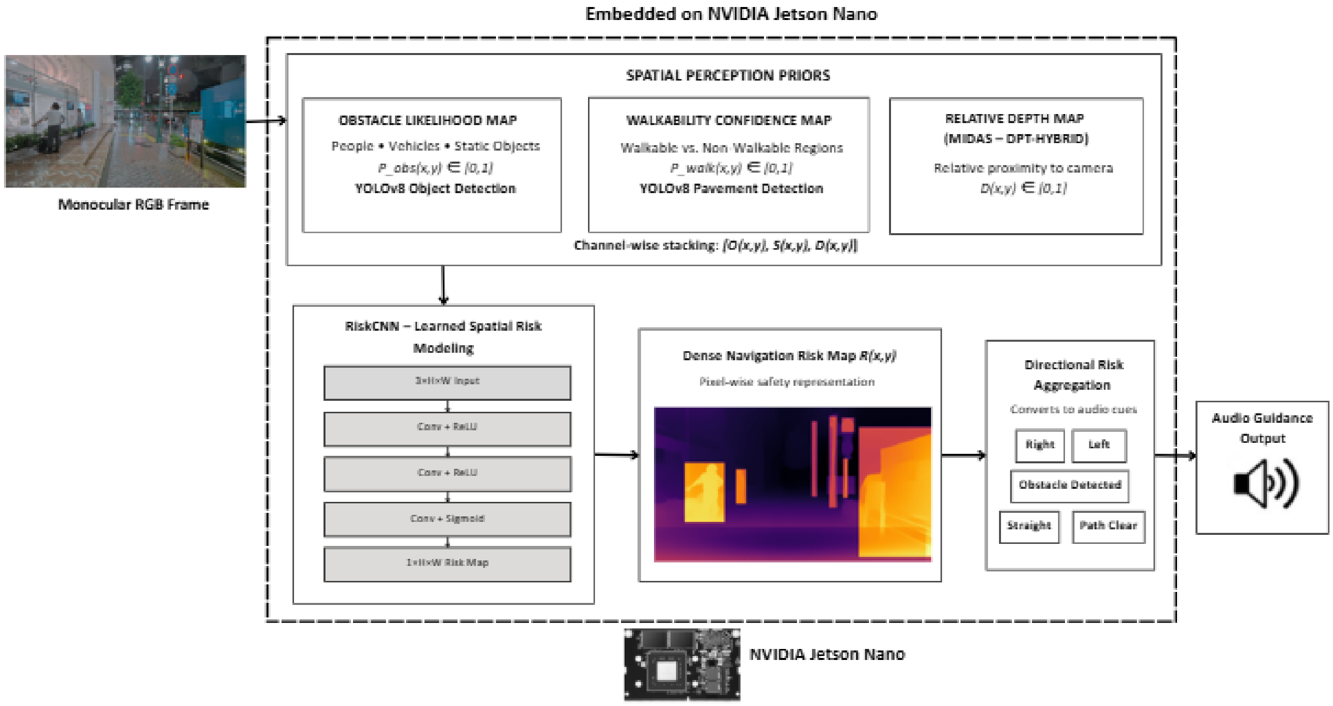


Fig. 1. System Architecture diagram.

visually impaired, and recognition systems on smartphones are continuously being enhanced for increased access in everyday settings [14]. However, the translation of the perceivability data into navigation directions in the majority of the existing systems does not define navigation safety.

Differing from previous methods, the proposed PROXI-NAV framework translates assistive navigation to a density risk representation learning problem. By weaving together the threads of object detection, depth perception, and walkability under the umbrella of weakly supervised learning, the proposed solution separates the concerns of perception and navigation reasoning and allows for the opportunity to reason proximity-related guidance on the embedded edge platform.

III. Architecture

The proposed architecture is intended to offer real-time assistive navigation capabilities based on the integration of parallel perceptual processing with the inference of spatial risk information through learning. The architecture enforces the perception–risk–decision framework. Within this framework, the interpretation of semantics and geometry in the scene and the interpretation of navigation are decoupled. This is depicted in Fig. 1.

A. System Overview

Given an input monocular RGB frame

$$\mathbf{I} \in \mathbb{R}^{H \times W \times 3}, \tag{1}$$

the system is able to extract a condensed set of spatial feature representations concerning the presence of obstacles, the traversability of the surfaces, and geometric closeness using parallel feature computation through specific perception modules that are then combined using a light-weight fully convolutional neural network, termed as *RiskCNN*.

The architecture is consciously designed to work on the spatial feature maps, as opposed to the output of the full model, which helps to reduce the amount of data that has to be transferred and the amount of computations that have to be performed.

B. Parallel Perceptual Feature Extraction

The semantic perception is achieved through a YOLOv8 model featuring a multi-head structure that accomplishes both obstacle detection and walkable surface estimation in a shared backbone network without having the need to run multiple standalone networks independently on the same system or environment concerning related perception perception processes.

The joint backbone yields generalized spatial features, which are concurrently processed by task-specific heads. Instead of revealing detailed detection metadata or segmentation results, the perception module directly outputs compact spatial confidence maps, which are represented by

- An *obstacle likelihood map* encoding the confidence-weighted presence of obstacles projected onto the image plane.
- A *walkability confidence map* representing traversable regions with softened spatial boundaries.

In parallel, geometric proximity is estimated using a lightweight monocular depth estimation network optimized for embedded inference. The depth estimator outputs a dense relative depth map aligned with the RGB frame that is normalized to form a proximity-aware spatial feature map. All perceptual modules run in parallel on the same input frame, so semantic and geometric features become available without introducing sequential dependencies.

C. Spatial Risk Inference Module

The obstacle likelihood map $\mathbf{O}$, walkability confidence map $\mathbf{S}$, and relative depth map $\mathbf{D}$ are stacked channel-wise to form the RiskCNN input:

$$\mathbf{X} = [\mathbf{O}, \mathbf{S}, \mathbf{D}] \in \mathbb{R}^{3\times H\times W}. \tag{2}$$

RiskCNN is implemented as a small fully convolutional network consisting of successive 3×3 convolutional layers with nonlinear activations, followed by a 1×1 convolution and sigmoid output. The network keeps the spatial resolution throughout and outputs a dense navigation risk map.

$$\mathbf{R} = f_\theta(\mathbf{X}), \tag{3}$$

where $\mathbf{R} \in [0, 1]^{H\times W}$.

Hence, by operating exclusively on spatial feature maps, RiskCNN captures local context and nonlinear interactions between semantic and geometric cues within a computationally efficient manner.

D. Decision Interface and Embedded Deployment

The risk map can be spatially aggregated for analyzing feasible directions for navigation, and this result is then provided as feedback to the user through auditory signals.

The entire pipeline runs on the NVIDIA Jetson Nano edge platform. The capabilities provided by perceptual parallel processing, the usage of shared feature extraction, as well as the compact architecture in the RiskCNN model make it possible to provide real-time functionality in assistive navigation systems operating outdoors.

IV. Methodology

In this section, we will discuss the learning formulation and inference procedure that enable us to model our proposed risk-based reasoning framework for navigation. We will also discuss our approach in learning risk for navigation from prior perceptions.

A. Learning Objective

The objective is to discover a mapping from feature maps to a densely defined spatial risk space, representing a relative safety over space for movement. The risk during navigation is a continuous probabilistic space permitting smooth reasoning over space and decision-making with awareness of proximity.

B. Weakly Supervised Risk Construction

As accurate annotations for risk in ground truth for navigation are not feasible, a weakly supervised learning model is used instead. Pseudo-risk annotations are generated by weight combination of perceptual priors:

$$\hat{\mathbf{R}} = \alpha\mathbf{O} + \beta\mathbf{D} + \gamma(1 - \mathbf{S}), \tag{4}$$

where α, β, and γ control the influence of obstacle presence, geometric proximity, and non-walkable regions.

The pseudo-risk map is clipped to the interval $[0, 1]$ and spatially smoothed to encourage coherence and reduce high-frequency noise.

C. RiskCNN Training Procedure

RiskCNN is trained to regress the pseudo-risk map from stacked perceptual inputs. Given an input $\mathbf{X}$ and pseudo-label $\hat{\mathbf{R}}$, the network predicts $\mathbf{R} = f_\theta(\mathbf{X})$.

RiskCNN is trained using a pixel-wise binary cross-entropy loss, treating navigation risk as a dense probabilistic field. Given the predicted risk map $\mathbf{R}$ and the pseudo-risk supervision $\hat{\mathbf{R}}$, the training objective is defined as

$$\mathcal{L}_{\text{risk}} = -\frac{1}{HW} \sum_{x,y} \Big[\hat{\mathbf{R}}_{xy} \log(\mathbf{R}_{xy}) + (1 - \hat{\mathbf{R}}_{xy}) \log(1 - \mathbf{R}_{xy}) \Big] \tag{5}$$

Optimization is carried out using the Adam optimizer, while all perception modules remain frozen to ensure that learning is confined to the risk modeling component.

D. Inference and Direction Selection

At inference time, perceptual feature maps are computed and passed to the trained RiskCNN to obtain the navigation risk map $\mathbf{R}$. The image plane is partitioned into directional regions corresponding to feasible motion directions. For each region Ω_k, the average risk is computed as

$$\rho_k = \frac{1}{|\Omega_k|} \sum_{(x,y)\in\Omega_k} \mathbf{R}(x, y), \tag{6}$$

and the direction associated with the minimum aggregated risk is selected.

E. Temporal Stabilization and Audio Feedback

In order to get smooth guidance, directional decisions are temporally smoothed using a sliding window over predictions. The final decision regarding navigation is provided to the user in real time, using auditory feedback.

V. Experimental Setup

The proposed model is tested on publicly available datasets for outdoor route-following tasks involving sidewalks, pavements, pedestrians, and urban obstacles, under various lighting conditions and levels of complexity and illumination. The Walk On The Road (WOTR), obstacle detection part, is trained on the WOTR dataset, and the pavement dataset is used for walkability score support, which has marked pavement areas including sidewalks and curbs. This dataset is all resized into the same fixed resolution $H \times W$, and is separated into training (70%), validation (15%), and testing (15%) subsets.

The training process is conducted through a cloud-enabled GPU computing setup that comes equipped with NVIDIA Tesla T4 (16-GB VRAM). The testing and execution phases are done using two different systems. The first one is a laptop configuration that relies on the AMD Ryzen 7 8845HS processor along with integrated graphics. The second one is the NVIDIA Jetson Nano, which is equipped with the quad-core ARM Cortex-A57 processor and the 128-core Maxwell GPU.

There are three dimensions of system performance being evaluated: the quality of perception, real-time capability, and navigation reasoning. Obstacle detection is evaluated on Perception Quality using mean Average Precision, mAP@0.5. Realtime capability is quantified with average perframe latency and frames per second. Qualitative and stability-based evaluations are conducted on Navigation Reasoning to determine how well the learned risk representation aids in making safe motion decisions. To motivate the application of learned spatial risk inference, consider the various alternative strategies to model navigation risks in Table I. Learned risk modeling allows for spatial context reasoning and corridor-level safety assessments that are not possible with fixed formulations.

TABLE I
COMPARISON OF NAVIGATION RISK MODELING STRATEGIES.

Strategy	**Spatial**	**Proximity**	**Corridors**	**Adaptability**
Rule-Based Logic	No	Limited	No	Low
Fixed Risk Fusion	No	Yes	Limited	Low
RiskCNN (Ours)	Yes	Yes	Yes	High

VI. RESULTS AND DISCUSSION

This section provides quantitative and qualitative results to assess perception reliability, real-time performance, and the effectiveness of learned spatial risk modeling for navigation.

A. Obstacle and Pavement Detection Performance

Table II reports obstacle detection performance on the WOTR dataset. The proposed multi-head perception model is reaching detection accuracy on par with established single-task detectors while enabling simultaneous inference of multiple perceptual cues. This confirms that sharing features within a unified backbone preserves detection reliability while allowing for subsequent efficient processing.

B. Walkable Area Detection Performance

Table III illustrates the performance of the proposed multi-head perception model on both curb and pavement datasets for walkable area detection. It is evaluated against established semantic segmentation and instance-aware baselines using standard metrics.

The multitask head model sets walkability detection accuracy comparable to specialized single-task segmentation networks sharing a unified backbone with the obstacle detection task. This shows that the shared feature representation captured enough surface-level semantic cue without task interference during joint inferences.

TABLE II
OBSTACLE DETECTION PERFORMANCE COMPARISON.

Model	**mAP (%)**	**F1 (%)**	**Recall (%)**
YOLOv8	86.1	83.9	85.2
Faster R-CNN	87.5	85.1	86.7
DETR	84.8	81.7	83.5
Multi-Head (Ours)	85.2	83.0	84.3

Even though it does not surpass the absolute peak accuracy for all single-task benchmarks, the proposed model of multihead perception always provides a level of competitiveness in both obstacle detection and walkable area detection tasks. This should come as no surprise, considering the purpose of the system.

TABLE III
WALKABLE AREA DETECTION PERFORMANCE COMPARISON.

Model	**mIoU (%)**	**F1 (%)**	**Recall (%)**
U-Net	85.8	82.0	84.7
DeepLabv3+	86.7	83.5	85.8
Mask R-CNN	87.2	84.1	86.3
Multi-Head (Ours)	85.9	82.4	84.6

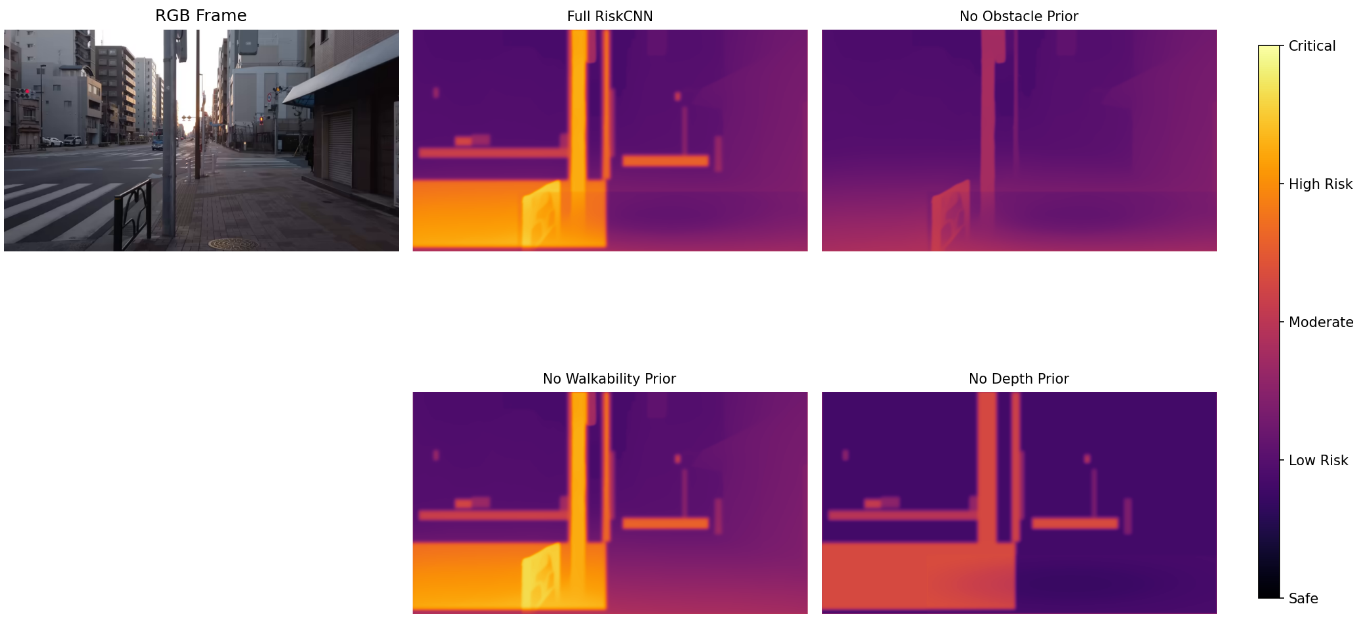


Fig. 2. Ablation Study

Through joint usage of a feature extraction backbone, there is elimination of redundant computations and simultaneous processing for multiple perception cues. The minor degradation in accuracy for specific tasks is compensated by the substantial decrease in task latency and computational cost. The findings suggest that the design brings forth a good balance between perceptual robustness and efficiency and can be applied in risk reasoning and real-time assistive navigation.

C. Real-Time Inference Performance

Table IV compares embedded inference latency on the NVIDIA Jetson Nano. The multi-head perception model significantly reduces per-frame latency compared to sequential execution of independent detection and segmentation models. By avoiding redundant backbone computation and enabling parallel feature extraction, the proposed pipeline achieves real-time operation suitable for wearable navigation assistance.

TABLE IV
REAL-TIME PERFORMANCE ON JETSON NANO.

Setup	**Latency (ms)**	**Std (ms)**	**Max (ms)**	**FPS**
Sequential Models	195	15	245	5
Multi-Head (Ours)	130	8	165	8

D. Ablation Analysis of Risk Modeling

An ablation study was performed in order to examine the contribution of specific learned and prior perceptions as well as their learned combination. Simply basing decision upon obstacle likelihood provides disparate risk responses. Incorporation of depth significantly enhances proximity perception but remains noisy. Overall, the full risk model provided significantly more stable and coherent spatial risk responses and superior navigation decisions. Table V provides ablation study of our risk modeling approach. Removing either learned combination or basing risk reasoning purely upon specific prior perceptions compromises spatial and temporal stability of risk responses. Overall, our full risk model provides superior decision making(Fig 2).

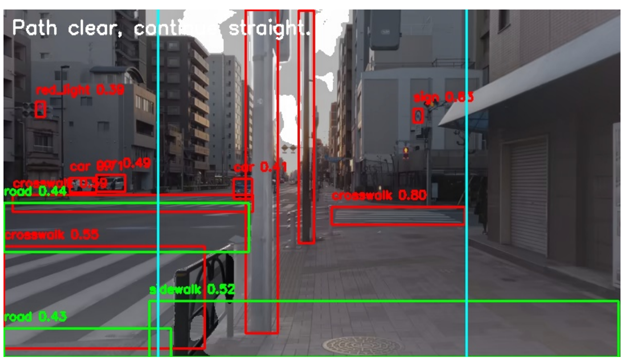
TABLE V
ABLATION STUDY OF NAVIGATION RISK MODELING COMPONENTS.

Configuration	**Coherence**	**Stability**	**Reliability**
Obstacle Only	Low	Low	Low
Obstacle + Depth	Moderate	Low	Moderate
Handcrafted Fusion	Moderate	Moderate	Moderate
RiskCNN (Ours)	High	High	High

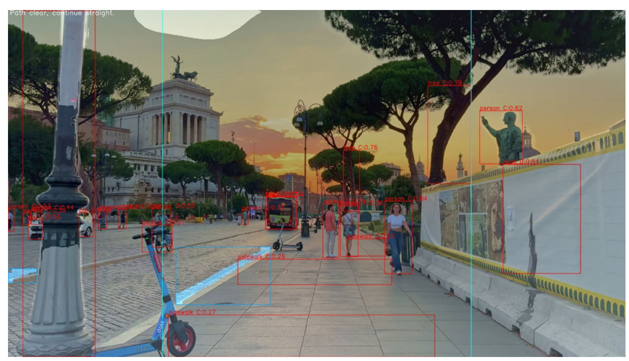
E. Effectiveness of Learned Risk Modeling

The results in Fig. 3 show the system performance in navigating various outdoor conditions. The modules detect the obstacles and the possible routes to walk through, so the RiskCNN combines all the elements to generate a risk map that helps it navigate.

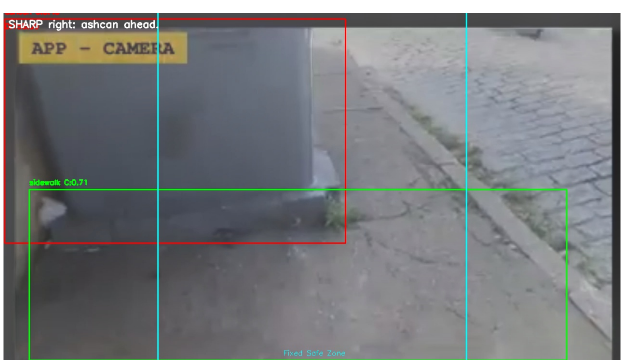
Compared to direct rule-based reasoning, the risk model learned has smoother spatial distributions for risk and is less affected by localized noise in perception. The natural weighting of obstacles outside the proximity range, combined with less emphasis on non-walkable areas and near-field hazards, provides more robust weighting for navigating decisions.



(a) Curb Based Path Guidance: Continue Straight



(b) Object too far to interfere: Distance calculated by depth map



(c) Obstacle Detection with interference: Sharp right

(d) Obstacle (Person) outside safe zone

Fig. 3. Inference results with depth overlay and instructions.

F. Discussion

The effectiveness of the proposed system has been verified, and it has been found that the system allows for the perception performance to be preserved along with the realization of a real-time embedded system. The output of the model has been shown in Fig 3. The combination of RiskCNN brings a structured spatial reasoning component that translates the perceptual information into risk concepts for navigation. This decoupling of navigation and risk reasoning allows for safe navigation in the outdoor world while maintaining a low latency level.

VII. Conclusions and Future Works

In this paper, a novel monocular vision-based assistive navigation system PROXI-NAV, which learns a dense representation of risk for navigation through weak supervision, was presented. The approach removes the constraints imposed on navigation systems by traditional task-agnostic representation-based pipelines with a lightweight CNN for a task-related risk field. Results show that the presence of obstacles, walkability confidence, and proximity have been effectively incorporated in the risk representation by the network. Ablation studies have also been able to show the need to fuse risk factors in the model in order to predict navigation risk effectively.

Future research will investigate a range of extensions. Deploying recurrent or attention models could improve the robustness of risk predictions in time-varying environments. Sharing nets could also enhance efficiency. Moreover, learning a navigation policy based upon demonstrations from experts could be a highly effective replacement for our heuristic directional aggregation strategy. Last but not least, experiments conducted with visually impaired subjects will be essential for the verification of our hypotheses.

VIII. Funding Sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] B. Li, J. P. Muñoz, X. Rong, Q. Chen, J. Xiao, Y. Tian, A. Arditi, and M. A. Yousuf, "Vision-Based Mobile Indoor Assistive Navigation Aid for Blind People," *IEEE Transactions on Mobile Computing*, vol. 18, no. 3, pp. 702–714, 2019.

[2] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1623–1637, 2022.

[3] C.-L. Lu, Z.-Y. Liu, J.-T. Huang, C.-I. Huang, B.-H. Wang, Y. Chen, N.-H. Wu, H.-C. Wang, L. Giarré, and P.-Y. Kuo, "Assistive Navigation Using Deep Reinforcement Learning Guiding Robot With UWB/Voice Beacons and Semantic Feedbacks for Blind and Visually Impaired People," *Frontiers in Robotics and AI*, vol. 8, 2021.

[4] B. Kuriakose, R. Shrestha, and F. E. Sandnes, "DeepNAVI: A Deep Learning Based Smartphone Navigation Assistant for People with Visual Impairments," *Expert Systems with Applications*, vol. 212, Art. no. 118720, 2023.

[5] M. H. Abidi, A. N. Siddiquee, H. Alkhalefah, and V. Srivastava, "A Comprehensive Review of Navigation Systems for Visually Impaired Individuals," *Heliyon*, vol. 10, no. 11, 2024.

[6] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, and M.-H. Yang, "Weakly-Supervised Semantic Segmentation via Sub-Category Exploration," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] X. Cai, S. Ancha, L. Sharma, P. R. Osteen, B. Bucher, S. Phillips, J. Wang, M. Everett, N. Roy, and J. P. How, "EVORA: Deep Evidential Traversability Learning for Risk-Aware Off-Road Autonomy," *arXiv preprint* arXiv:2311.06234, 2023.

[8] H.-H. Ngo, H. L. Le, and F.-C. Lin, "Deep-Learning-Based Cognitive Assistance Embedded Systems for People with Visual Impairment," *Applied Sciences*, vol. 15, no. 11, Art. no. 5887, 2025.

[9] R. Bharadwaj, B. Bachewar, S. Barsude, H. Badagandi, and S. Darade, "VisionTool: A Deep Learning Embedded Device for Blind People," *Journal of Information Technology Education: Innovations in Practice*, vol. 24, 2025.

[10] I.-K. Jeong, K. Kim, J. Jung, and J. Cho, "YOLOv8-Based XR Smart Glasses Mobility Assistive System for Aiding Outdoor Walking of Visually Impaired Individuals," *Electronics*, vol. 14, no. 3, Art. no. 425, 2025.

[11] M. Saranya and S. Arulselvarani, "Real-Time Obstacle Detection Using YOLOv8 for Assistive Navigation," *Indian Journal of Science and Technology*, vol. 18, no. 25, pp. 2009–2022, 2025.

[12] A. Pratap, S. Kumar, and S. Chakravarty, "Adaptive Object Detection for Indoor Navigation Assistance: A Performance Evaluation of Real-Time Algorithms," *arXiv preprint* arXiv:2501.18444, 2025.

[13] J. Anjom, R. I. Chowdhury, T. Hasan, and M. I. A. Hossain, "An Embedded Real-Time Object Alert System for Visually Impaired: A Monocular Depth Estimation Based Approach," *arXiv preprint* arXiv:2507.08165, 2025.

[14] F. A. Rahman, F. A. Pusparani, W. L. Yeoh, and O. Fukuda, "Smartphone-Based Seamless Scene and Object Recognition for Visually Impaired Persons," *Information*, vol. 16, no. 9, Art. no. 808, 2025.