REINFORCEMENT LEARNING WITH SYNTHETIC NAVIGATION DATA ALLOWS SAFE NAVIGATION IN BLIND DIGITAL TWINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Limited access to dedicated navigation data in visually impaired individuals is a significant bottleneck in the development of AI-driven assistive devices. To address this, we have developped a virtual environment designed to extract various human-like navigation data from procedurally generated labyrinths. Using reinforcement learning and semantic segmentation, we trained a convolutional neural network to perform obstacle avoidance from synthetic data. Our model outperformed state-of-the-art backbones including DINOv2-B in safe pathway identification in real world. In conclusion, despite being trained only on synthetic data, our model successfully extracted features compatible with safe navigation in real-world settings, opening new avenues for visually impaired.

1 INTRODUCTION

Globally, millions of people live with blindness, which is associated to severe restrictions in mobility and as well as a significant increased risk of fall-related injuries (Wood et al., 2011; Brunes & Heir, 2021; Singh & Maurya, 2022). The recent development of retinal and cortical prostheses has not achieved vision restoration in blinds yet. As an alternative, Sensory Substitution Devices (SSDs) hold significant promise for aiding visually impaired individuals by converting environmental sensor data into tactile or auditory stimuli (Bachy Rita, 1972; Jicol et al., 2020). However, current SSDs face limitations in conveying complex visual scenes through skin or ears (Elli et al., 2014), primarily due to the narrow bandwidth of these sensory channels, which can lead to cognitive burden (De Jong, 2010). Indeed, artificial intelligence (AI), and in particular deep learning, enables the extraction of relevant information to be conveyed to the blind.

Contrary to autonomous driving (Chen & Krähenbühl, 2022; Toromanoff et al., 2019) and robotics (Shah & Levine, 2022; Kruse et al., 2013), SSDs have not benefited from the availability of navigation datasets. As a result, current applications of deep learning for effective navigation aids predominantly rely on general-purpose datasets for object recognition and classification (Scalvini et al., 2023; Mukhiddinov & Cho, 2021; Kim et al., 2023; Kerdegari et al., 2016), semantic segmentation (Tapu et al., 2017; Zheng & Weng, 2016), or depth estimation (Bai et al., 2017; Sharma et al., 2016; Asiedu Asante & Imamura, 2023). While these applications can extract high-level features that improve environmental understanding for the visually impaired, they do not convey information about navigation decisions. A few studies have used navigation-specific data to provide guidance (Zheng & Weng, 2016; Kerdegari et al., 2016), but the amount of data was limited to a few hundred samples due to the resource-intensive nature of the collection process.

The aim of this study is twofold: (i) To address the unavailability of human navigation data, we propose a generic method for training AI systems specifically designed for the blind. Our approach relies on synthetic semantic segmentation maps to optimize SSD outputs in virtual environments, enabling straightforward real-world transferability without requiring advanced domain adaptation (Xu et al., 2022; Zhu et al., 2023). (ii)

We evaluate the efficacy of this method by first training an AI-based system for obstacle avoidance in virtual environments and then demonstrating its ability to enable safe navigation using low-dimensional navigation cues in both virtual and real-world settings. To support this, we introduce NavIndoor, a new virtual environment designed for the automatic generation of synthetic human-like navigation data. NavIndoor leverages procedural generation to create randomized, obstacle-filled mazes with structure similar to real-world indoor environments, allowing for the simulation of various navigation scenarios in an efficient, safe, and scalable manner.

In Section 3, we describe our method for optimisation of AI-based SSDs within virtual environments and deployment in real-world setting. In Section 4, we present the newly proposed virtual environment and demonstrate that a compact CNN allows safe navigation in it. Finally, in Section 5, we illustrate the model's proficiency in real-world settings by estimating the forward navigation boundary in the Active Vision Dataset and showing that our model outperforms standard backbones in safe pathway identification (AUC 0.92). Furthermore, we show that incorporating random morphological operators around obstacles during training in virtual environments improves generalization to real-world data. The scalable and flexible nature of our method, combined with its potential for generalization to various navigation tasks, underscores its promise in enhancing feature extraction for future sensory substitution devices.

We summarize our contributions as follows:

- We identify a significant gap in the literature regarding the use of navigation data to improve Sensory Substitution Systems.
- We release NavIndoor, an open-source software for the computationally efficient generation of procedurally generated, obstacle-filled environments, enabling seamless integration with AI systems. NavIndoor facilitates the efficient creation of various large-scale, human-like navigation datasets.
- We show that synthetic data enables the extraction of low-dimensional features for navigation by individuals with visual impairments.
- We demonstrate that applying basic morphological operators to synthetic semantic segmentation maps enhances performance in real-world conditions after training.

2 Related Works

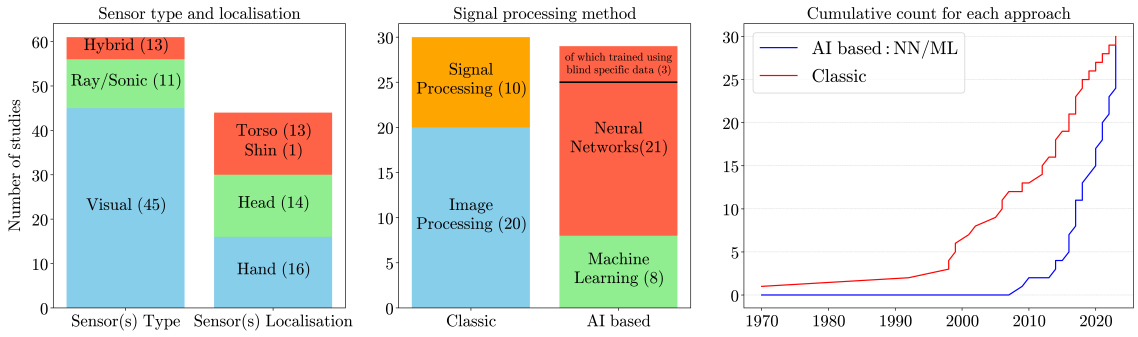

Figure 1: Overview of Sensory Substitution Devices Publications. (Left) Sensor type and their body localisation. (Middle) Processing methods used in the studies for classic and AI-based approaches. (Right) Cumulative count of publications over the years for machine learning (ML) or neural networks (NN) and classic approaches.

SENSORY SUBSTITUTION DEVICES AND AI FOR THE VISUALLY IMPAIRED

We reviewed the literature on sensory substitution devices and electronic travel aids released from 1970. The literature review was performed in PubMed and completed with research using arXiv, Elicit, and conventional search engines. Papers proposing new systems for sensory substitution were included, resulting in 61 studies. Detailed methodology and results are presented in Appendix A.

Various SSDs have been proposed to improve navigation for the blind. Traditional visual SSDs often utilized image mapping through haptic feedback (Bach-y Rita et al., 1998; Kajimoto et al., 2004; 2006; DANILOV & TYLER, 2005) or audio representations derived from images (Auvray et al., 2007; Meijer, 1992a). However, these approaches demand substantial attentional resources (Lee, 2019; Theurel et al., 2013), leading to cognitive overload (De Jong, 2010), and limiting their application to controlled environments (Elli et al., 2014).

Consequently, image and signal processing have been used to improve extraction of salient features using infrared, LiDAR, ultrasonic and mostly (50/61, 82%) visual sensors.

The use of machine learning and neural networks were introduced in 2009 and changed drastically the design of new devices by involving 58% (28/48) studies since then. Applications of neural networks for sensory substitution devices include image segmentation, object recognition, or classification (Busaeed et al., 2022; Scalvini et al., 2023; Asiedu Asante & Imamura, 2023; Afif et al., 2020; Sulaman et al., 2023; Bhatlawande et al., 2022; Chaudhary & Dr. PrabhatVerma, 2023; Mukhiddinov & Cho, 2021), object tracking (Tapu et al., 2017), speech understanding (Bai et al., 2017), image captioning (Ganesan et al., 2022; Kavitha et al., 2023), optimization of auditory representation of images with GANs (Kim et al., 2021; 2023; Hu et al., 2019; Port et al., 2021), and best action prediction (Zheng & Weng, 2016; Kerdegari et al., 2016).

AI-BASED DEVICES FOR OBSTACLE AVOIDANCE

Out of 20 neural-network SSD studies, 9 were aimed at obstacle avoidance tasks, which is critical for safe navigation. Limitations with such approaches mainly rely on computational and energetic cost, because such systems often require substantial hardware resources to perform multiple scene understanding tasks in parallel and in real-time (Mahendran et al., 2021). Also, complex operations such as depth estimation may require expensive or heavier sensors, such as stereo camera (Caraiman et al., 2017; Asiedu Asante & Imamura, 2023).

On the other hand, 2 studies (Zheng & Weng, 2016; Kerdegari et al., 2016) proposed to estimate directly the best possible action for the blind, but such tasks require extensive and costly acquisition of human navigation data. In (Zheng & Weng, 2016), authors collected 4109 tuples of GPS/visual sensor information labelled with the best possible action (forward/left/right/stop) predicted with a deep neural network. In (Kerdegari et al., 2016), authors collected 4051 samples comprising an ultrasonic measurement coupled with a performed action (forward/left/right) predicted with a multilayer perceptron.

DOMAIN ADAPTATION FROM SIMULATION TO REAL ENVIRONMENTS

Domain shift is a primary concern when deploying deep learning models trained in simulations into the real-world. Approaches to enable the transferability of AI systems from virtual to real-world environments have primarily focused on achieving photorealism by scanning real 3D scenes, as highlighted in (Xia et al., 2018). This strategy allows AI systems to utilize material textures for executing complex tasks and to leverage detailed 3D scene representations stored in external memory. Such systems have primarily benefited robotics (Hirose et al., 2019; Kang et al., 2019), where they enable fully autonomous agents to undertake complex scene understanding tasks.

Besides, SSDs are designed to prioritize the transmission of low-dimensional features that can be computed in real-time and easily interpreted by humans for navigation, and thus do not necessarily need to store a 3D rich representation of the surrounding environments for complex autonomous navigation tasks. Indeed, on the other hand, semantic segmentation has been proposed as domain-agnostic features to allow better generalization for robotics navigation in real-world (Hong et al., 2018; Chaplot et al., 2020). In a similar manner, we propose to use a simple semantic view of the scene as an input for bridging domain gap. However our approach is not designed to be deployed on fully autonomous agents, and focuses on extraction of low-dimensional navigation cues that could be interpreted by humans with haptic or auditory feedback. This approach enables the model to learn better obstacle avoidance as shown in Section 4, and considerably reduces the domain gap with real-world data. Compared with photorealism approaches, the use of semantic views also reduces the input dimensionality providing faster training and avoiding texture biases.

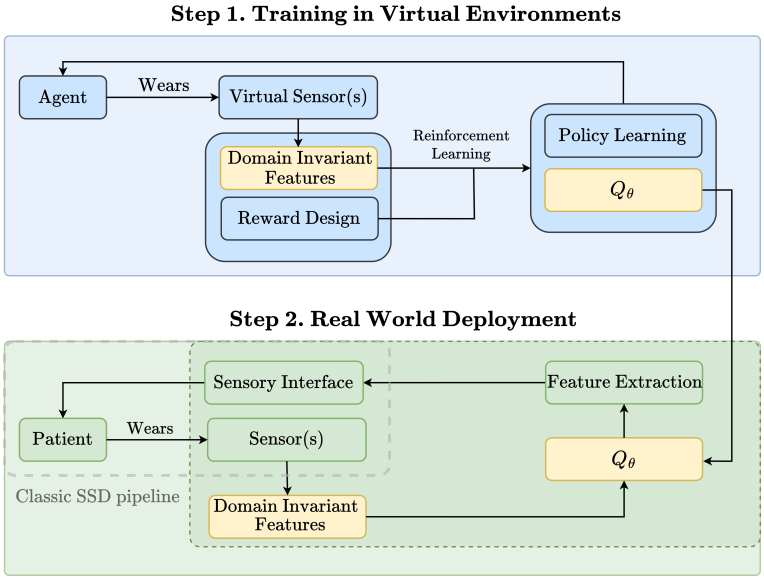

Figure 2: Method overview : we leverage training in virtual environments with reinforcement learning, by equipping a digital twin with sensor(s) to master navigation tasks from domain-invariant features. Post-training, utilizing the acquired knowledge encoded in Q_θ model, we extract navigation features from real-world data. Such features can finally be conveyed through a sensory substitution device (SSD). Virtual environment elements are denoted in blue, real-world elements in green, while cross-domain elements are highlighted in yellow.

EMBODIED AI PLATFORMS FOR ROBOTICS

AI-embodied platforms have primarily been developed for robotics navigation tasks, utilizing either synthetic assets or the scanning of real-world scenes. Scanning real-world scenes has led to the development of tools including Gibson (Xia et al., 2018; 2019), Habitat (Szot et al., 2021; Ramakrishnan et al., 2021), and Openroom (Li et al., 2021), providing hundreds of virtual scenes for training. Unity-based environments have also been developed by the Allen Institute for AI (AI2) (Ehsani et al., 2021; Deitke et al., 2020; 2022).

As proposed in ProcTHOR (Deitke et al., 2022), NavIndoor leverages procedural generation as a key element. In the context of mobility assistance, SSDs are expected to function in a wider range of environments

than those typically encountered in robotics, which necessitates superior generalization abilities. Opting for procedurally generated environments addresses this need by providing a far greater variety of data compared to photorealistic simulations, thereby increasing reinforcement learning models' robustness to unseen scenes, as demonstrated across various 2D (Cobbe et al., 2018; Johansen et al., 2019; Cobbe et al., 2019) and 3D (Jaderberg et al., 2018; Juliani et al., 2019) environments.

NavIndoor provides automatic generation of both semantic segmentation maps and depth maps as domain-agnostic features within procedurally generated environments, without the need for additional annotations or being constrained by a finite number of environments, contrary to existing platforms (Yadav et al., 2022) (Szot et al., 2021). Also, its design is specifically oriented for blind digital twin navigation including automatic generation of labeled collision instances. NavIndoor also includes parametrization for both agents (movement physics, action space, sensors) and environments (size, obstacle filling) and allows very high-speed rendering by leveraging various optimization features.

3 Materials and Method

We propose a scalable approach that leverages virtual environments, reinforcement learning, and transferable features to extract navigation-related features for SSDs, as illustrated in Figure 2. The navigation task is initially learned within a virtual environment by a digital twin equipped with sensor(s) and implemented through a reward function.

We used semantic segmentation masks as input for learning blind navigation within virtual environments, because they have lower dimensionality compared with depth maps, but still capture enough information for allowing navigation as shown in robotics (Hong et al., 2018; Chaplot et al., 2020).

Collection of synthetic navigation data was performed within virtual environments in the form of (s_t, a_t, r_t), where s_t, a_t, and r_t represent a semantic view, memory of actions taken by a blind digital twin, and reward values at time-step t, respectively. The navigation task is learned through policy learning from the semantic segmentation maps. Leveraging reinforcement learning, we estimate a parametric model $Q_\theta(s_t, a_t)$ using Q-learning. The Q-value function represents the expected sum of future rewards an agent would achieve in state s_t, choosing action a_t, and navigating optimally thereafter. Following the training, we evalyated the performance of Q_θ with real-world states.

To show the effectiveness of our approach, we propose to learn obstacle avoidance from a single head-mounted visual sensor, aligning with current start-of-the-art (see Figure 1) but without extensive real-world data collection and annotation. Semantic segmentation maps were processed to regroup obstacles in a single class and processed as $1 \times 128 \times 128$ tensors to reduce input dimensionality. We also leveraged procedural generation, allowing for randomization of virtual environments, semantic-segmentation specific data augmentation and used a compact convolutional neural network for Q-value estimation (1.6M parameters). These features aimed at improve model robustness and allow extraction of relevant feedback in real-time in a sensor-lightweight and energy-efficient pipeline.

For synthetic data generation, we developed a virtual environment, NavIndoor, encompassing the necessary features, and in which training was performed (Section 4). The following section presents the experimental setup employed for learning obstacle avoidance, including design of reward function, virtual environment, data randomization mechanisms, model architecture, hyperparameters optimisation as well as the model's results within virtual setting.

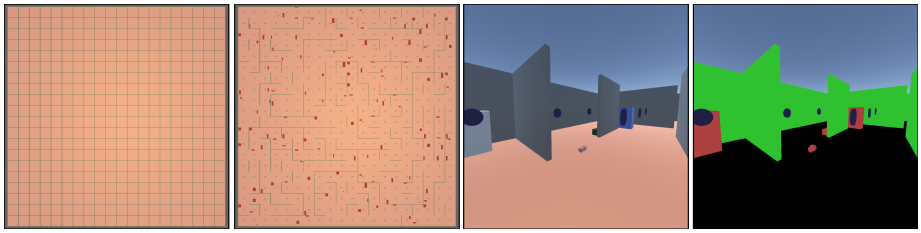

Figure 3: From left to right. The maze generation procedure initiates with a grid of closed cells and a randomly positioned agent. Subsequently, depth-first search algorithm is applied to its graph structure and cells are randomly filled with obstacles and collectibles. The agent wears a forehead RGB camera and a semantic segmentation camera.

4 Training in Virtual Environment

3D Environment and blind digital twin

NavIndoor is built upon Unity and MLAgents library (Juliani et al., 2018). It was developed as a platform for generation of synthetic data from navigation of a blind digital twin. NavIndoor is designed to create sequential, partially observable, static, procedurally generated environments filled with walls, obstacles, and collectibles. The environment generation includes maze generation and the sampling of a random starting point where the agent begins its exploration, in order to prevent memorization biases (Zhang et al., 2018). The generation procedure for mazes is based on the Depth First Search algorithm (Tarjan, 1972) and depicted in Figure 3. Obstacles include both cuboids of various shapes and open-source low-polygons assets. The environment was designed in Unity to allow for easy parametrization integration through Python for generation of mazes, agents and sensors with different properties.

A blind digital twin was designed to navigate these mazes. It has a discrete action space $\mathcal{A} = \{\text{forward, backward, rotate left, rotate right}\}$. The agent is equipped with a frontal monocular camera hat has a field of view (FOV) of $115° \times 100°$. The camera sensor is configured to return 128×128 semantic segmentation maps along with RGB views of the current scene.

Reward Design and Observations

We propose using collectibles located at the center of the maze's cells to design the reward function. The agent receives a positive reward when collecting a coin, which encourages exploration of the maze. It receives a negative reward when colliding with a wall or an obstacle. At each timestep, the environment returns a semantic map of the current view seg_t as well as an RGB view f_t. The semantic map has five labels: *floor, obstacles, walls, coins, other*. At timestep t, the state $s_t = (seg_{t-2}, seg_{t-1}, seg_t, (a_i)_{i\in[t-m,t-1]})$ is extracted, where seg_i is the semantic segmentation map at timestep i, a_i is a one hot encoding of action taken at timestep i, and m is an action memory length parameter. Encoding semantic maps as 2D arrays by associating a single value to walls (-0.5), floor (0.5), and obstacles (-1) provided better stability during training compared to multi-channel semantic encoding. Additionally, our experiments showed that providing time context through previously performed actions allowed the agent to escape situations where it would get stuck during learning, leading to better obstacle avoidance strategies.

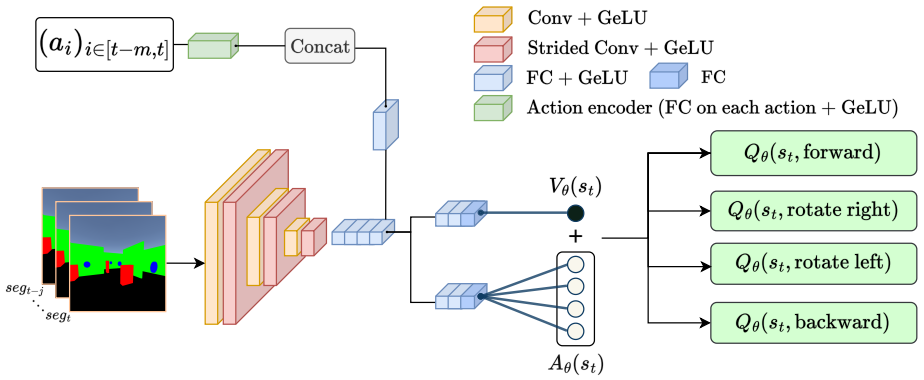

Figure 4: The Q-network architecture consists of six convolutional layers, followed by fully connected layers to estimate $Q_\theta(s)$. Given the state s_t, the deep Q-network estimates both the value and advantage functions, which can be interpreted as indicators of safety and guidance on potential actions, respectively..

TRAINING

Obstacle avoidance was learned using Double Dueling Deep Q Network (D3QN) (Wang et al., 2015), with each episode taking place in a new procedurally generated maze. The model architecture processing s_t is presented in Figure 4. It is composed of image and actions encoders followed with fully connected layers. The D3QN architecture offers an estimation V_θ of expected sum of discounted rewards, which relates directly to the safety level for obstacle avoidance. An extensive grid search was conducted over training hyper-parameters and then agent and environment parameters (speed, obstacle proportion, progressive increase in difficulty) to determine the best training setting. Details on training and grid search are given in Appendix B. The best trained model (VC) setup was then used to trained another model (VCD) using data augmentation. Data augmentation included erosion and/or dilatation morphological operators around obstacle shapes using a fixed 3x3 square kernel with probability $p = 0.2$ for each image and operator. Morphological augmentation was followed by random uniform changes in each pixel label with probability $p = 0.05$. These changes aim to simulate potential errors during semantic segmentation of real-world images and to make the model robust to various obstacle shapes. Training setups included a model was trained from standard RGB views of the camera and a model trained with unvisible collectibles (NVC).

Two human individuals were also trained to collect the maximum amount of coins while avoiding collisions in to assess human performance and compare it to our models. Human performance was evaluated under settings with both visible and invisible collectibles, as well as RGB inputs, with 10 episodes for each individual. Each episode lasted for 24 seconds (corresponding to 400 decision timesteps) and humans navigated in the NavIndoor using the keyboard's directional arrows.

RESULTS IN NAVINDOOR

Table 1: Mean reward for final models *vs.* humans

	Best Model	Humans	Ratio
Visible coins	**74.34** (VC)	**102.36**	**0.73**
	73.18 (VCD)	/	0.71
Invisible coins	47.44 (NVC)	83.82	0.56
RGB views	63.98	101.34	0.63

Results after training are presented in Table 1. The best-trained model (VC) reached 73% of human-level performance. It had a mean reward of 74.34 and was obtained using visible collectibles and progressive increase in difficulty during training. The model trained with RGB views achieved a mean reward of 57.73, which was the lowest among models trained using semantic segmentation maps, showcasing the relevance of semantic segmentation for obstacle avoidance during visual-based navigation.

5 RESULTS

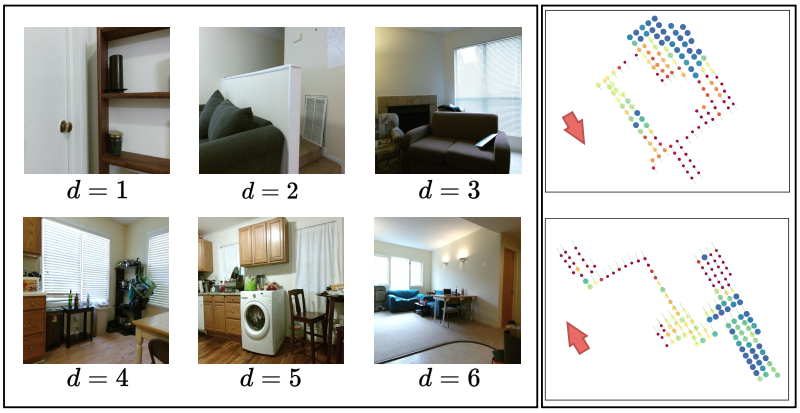

Figure 5: (a) AVD samples showing various distances to the navigation boundary. (b) A top-down view of two AVD rooms. Points mark the locations where photographs were captured. $Q_\theta(s, \text{forward})$ was computed for each state s, indicating the same direction (orange arrows). Blue (big) points indicates high values of $Q_\theta(s, \text{forward})$ and red (small) points low values.

We identified the Active Vision Dataset (AVD) as a good dataset for evaluating our model, because it offers indoor views coupled with spatial metadata, including coordinates and viewpoint angles for each navigable location. We computed for each image its distance to navigation boundary d, representing the number of possible forward steps from a given view. This value can be seen as a path clearance level. Samples from the AVD are illustrated in Figure 5 (a).

We conducted image segmentation as a pre-processing step using SegFormer-b2 on each AVD image. Next, we extracted features from VC and VCD models trained in NavIndoor using the initial state configuration (no previous action in the action memory buffer, and stacking the 3 same images). The process ran at 179 FPS in our setting (RTX 4090) without further optimization.

The model's output for VC correlated well with the path clearance level. In particular, the safety level $V_\theta(s)$ and the rotation advantage relative to going forward, $R_\theta(s)$, defined as the difference between

$Q_\theta(s, \text{forward})$ and the maximum between $Q_\theta(s, \text{rotate left})$ and $Q_\theta(s, \text{rotate right})$, correlated well with the path clearance level d as depicted in Figure 6. We observed a mean increase of V_θ with respect to d, indicating the model's ability to assign better values to states with clearer pathways. Similarly, we noticed a mean decrease of $R_\theta(s)$ with respect to d, suggesting its potential as a navigation insight to indicate when rotfation is a favorable option. Additionally, we generated schemes of two AVD scenes and colored photograph locations for a specific direction based on $Q_\theta(s, \text{forward})$ in Figure 5 (b), showcasing high correlation with the forward pathway clearance.

The same processing method was applied on a video captured by a sighted individual wearing a camera on their forehead, and results showcased a high correlation between the forward distance to walls or obstacles with $Q_\theta(s)$ and $V_\theta(s)$. As depicted in Figure 8, $V_\theta(s)$ decreases when the individual approached obstacles, showcasing features relevance for feedback integration in real sensory substitution devices.

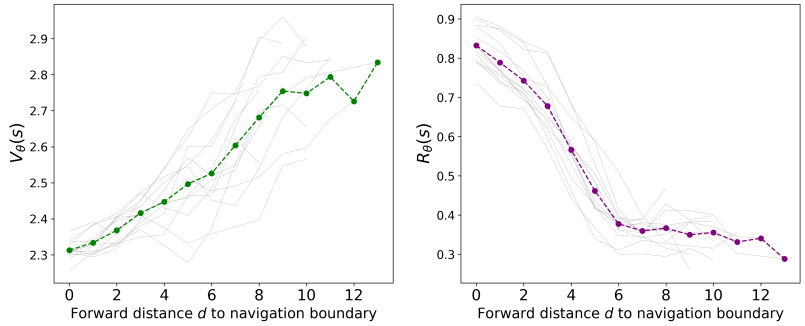

Figure 6: (Left) Mean values of V_θ on images within each scene (gray) and the overall average (green) are plotted with respect to d. (Right) Mean values of R_θ on images within each scene (gray) and the overall average (purple) are shown with respect to d (left). Both V_θ and R_θ are single-dimensional features that significantly correlate with the distance to the navigation boundary in the real-world.

LINEAR PROBING

Quantitative evaluation was conducted through linear probing, involving binary classification of images based on their forward navigation boundary. Labels y_i^d were determined by applying a threshold to the maximum reachable forward distance before encountering a wall or an obstacle. Specifically, y_i^d was set to 1 if $d_i > d$, indicating a boundary within the next d forward steps. This approach simulates the need to alert users when they approach obstacles with a binary feedback with alert distances d varying from 0 to 6. Linear classifiers were trained on 9 indoor scenes and tested on 5 unseen scenes from different buildings (details in Appendix C).

Performance comparisons were made with other state-of-the-art models, including self-supervised backbones (DINOv2 distilled (Oquab et al., 2024), ConvNext V2 (Woo et al., 2023)) and supervised models (SegFormer-b2 (Xie et al., 2021), EfficientNet-b7 (Tan & Le, 2019)). For transformers models, the latent space of every image patch was used because it gave better performances compared with using only the *cls* token.

The results for AUCs of each classifier are depicted in Figure 7. Table 2 presents the mean evaluation metrics accross all trained classifiers. VCD classifiers consistently outperformed the other models for $d > 2$. Indeed, the use of morphological operators coupled with random changes in the unified semantic segmentation images thus provided with significantly better generalization with real-world semantic segmentation maps, which are naturally prone to errors. Although evaluation of VCD relied on pre-processing using SegFormer-

Table 2: Linear probing binary classifiers for forward navigation boundary detection. Mean metrics on test set.

Model	Features	F1	AUC
ConvNext V2 (Woo et al., 2023)	15680	0.63	0.82
SegFormer-b2 (Xie et al., 2021)	131072	0.72	<u>0.87</u>
EfficientNet-b7 (Tan & Le, 2019)	231040	0.71	0.80
DINOv2-B (Oquab et al., 2024)	197376	<u>0.75</u>	0.86
(Ours) VCD	**768**	**0.77**	**0.88**
(Ours) VC	**768**	0.74	0.86
(Ours) NVC	**768**	0.69	0.85

b2, the results show significantly higher AUCs using VCD for $d > 3$ and higher F1 score, demonstrating our model performance in understanding structure of real-world indoor places.

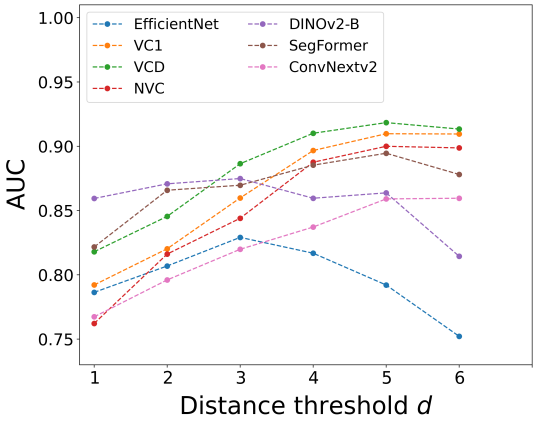

Figure 7: AUCs for different distance thresholds (binary classification) on test set for state-of-the-art models and VC, NVC, VCD.

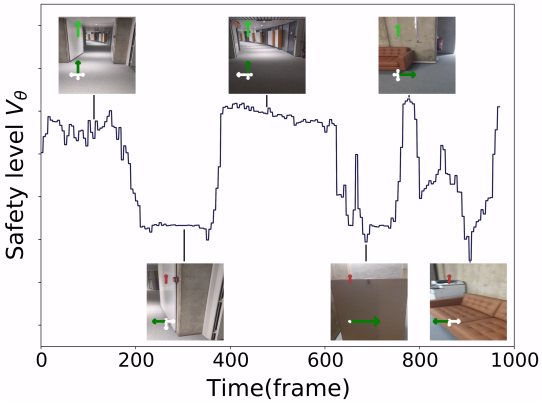

Figure 8: Value function estimate V_θ across the real-world video sample. Image samples from the video are displayed with their associated V_θ (top arrow) and A_θ (bottom arrows) outputs.

6 Conclusion

In this study, we introduced a new framework aimed at improving mobility for visually impaired people through the use of synthetic data. We proposed a virtual environment specifically designed for generating human-like navigation data, which can be used for training and evaluating deep learning models for SSDs. Compared to previous approaches, our method offers scalability and real-time extraction of low-dimensional features for safe navigation from a single visual sensor. Indeed, the proposed method ensures a high compatibility with the lightweightness, cognitive and hardware constraints associated with SSDs. These advances represent a significant step in the development of robust AI-based assistive technologies and pave the way for further research aimed at improving mobility for visually impaired individuals.

REFERENCES

Mouna Afif, Riadh Ayachi, Yahia Said, Edwige Pissaloux, and Mohamed Atri. An Evaluation of RetinaNet on Indoor Object Detection for Blind and Visually Impaired Persons Assistance Navigation. *Neural Processing Letters*, 51(3):2265–2279, June 2020. ISSN 1370-4621, 1573-773X. doi: 10.1007/s11063-020-10197-9. URL http://link.springer.com/10.1007/s11063-020-10197-9.

Rohit Agarwal, Nikhil Ladha, Mohit Agarwal, Kuntal Kr. Majee, Abhijit Das, Subham Kumar, Subham Kr. Rai, Anand Kr. Singh, Somen Nayak, Shopan Dey, Ratul Dey, and Himadri Nath Saha. Low cost ultrasonic smart glasses for blind. In *2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 210–213, 2017. doi: 10.1109/IEMCON.2017.8117194.

Patricia Arno, Christian Capelle, Marie-Chantal Wanet-Defalque, Mitzi Catalan-Ahumada, and Claude Veraart. Auditory Coding of Visual Patterns for the Blind. *Perception*, 28(8):1013–1029, August 1999. ISSN 0301-0066, 1468-4233. doi: 10.1068/p281013. URL http://journals.sagepub.com/doi/10.1068/p281013.

Bismark Kweku Asiedu Asante and Hiroki Imamura. Towards Robust Obstacle Avoidance for the Visually Impaired Person Using Stereo Cameras. *Technologies*, 11(6):168, November 2023. ISSN 2227-7080. doi: 10.3390/technologies11060168. URL https://www.mdpi.com/2227-7080/11/6/168.

Malika Auvray, Sylvain Hanneton, and J Kevin O'Regan. Learning to perceive with a visuo — auditory substitution system: Localisation and object recognition with 'the voice'. *Perception*, 36(3):416–430, 2007. doi: 10.1068/p5631. URL https://doi.org/10.1068/p5631. PMID: 17455756.

P. Bach-y Rita, K. A. Kaczmarek, M. E. Tyler, and J. Garcia-Lara. Form perception with a 49-point electrotactile stimulus array on the tongue: a technical note. *Journal of Rehabilitation Research and Development*, 35(4):427–430, October 1998. ISSN 0748-7711.

Paul Bach-y Rita. *Brain mechanisms in sensory substitution*. Acad. Press, New York, 1972. ISBN 978-0-12-071040-9.

Jinqiang Bai, Dijun Liu, Guobin Su, and Zhongliang Fu. A Cloud and Vision-based Navigation System Used for Blind People. In *Proceedings of the 2017 International Conference on Artificial Intelligence, Automation and Control Technologies*, pp. 1–6, Wuhan China, April 2017. ACM. ISBN 978-1-4503-5231-4. doi: 10.1145/3080845.3080867. URL https://dl.acm.org/doi/10.1145/3080845.3080867.

Pranab Gajanan Bhat, Deepak Kumar Rout, Badri Narayan Subudhi, and T. Veerakumar. Vision sensory substitution to aid the blind in reading and object recognition. In *2017 Fourth International Conference on Image Information Processing (ICIIP)*, pp. 1–6, Shimla, December 2017. IEEE. ISBN 978-1-5090-6733-6 978-1-5090-6734-3. doi: 10.1109/ICIIP.2017.8313754. URL http://ieeexplore.ieee.org/document/8313754/.

Shripad Bhatlawande, Swati Shilaskar, Aditi Kumari, Mahi Ambekar, Mohit Agrawal, Amit Raj, and Siddhi Amilkanthwar. AI Based Handheld Electronic Travel Aid for Visually Impaired People. In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pp. 1–5, Mumbai, India, April 2022. IEEE. ISBN 978-1-66542-168-3. doi: 10.1109/I2CT54291.2022.9823962. URL https://ieeexplore.ieee.org/document/9823962/.

Audun Brunes and Trond Heir. Serious life events in people with visual impairment versus the general population. *International Journal of Environmental Research and Public Health*, 18(21):11536, November 2021. ISSN 1660-4601. doi: 10.3390/ijerph182111536. URL http://dx.doi.org/10.3390/ijerph182111536.

Jaroslaw Bulat and Andrzej Glowacz. Vision-based navigation assistance for visually impaired individuals using general purpose mobile devices. In *2016 International Conference on Signals and Electronic Systems (ICSES)*, pp. 189–194, Krakow, Poland, September 2016. IEEE. ISBN 978-1-5090-2667-8. doi: 10.1109/ICSES.2016.7593849. URL http://ieeexplore.ieee.org/document/7593849/.

Sahar Busaeed, Iyad Katib, Aiiad Albeshri, Juan M. Corchado, Tan Yigitcanlar, and Rashid Mehmood. LidSonic V2.0: A LiDAR and Deep-Learning-Based Green Assistive Edge Device to Enhance Mobility for the Visually Impaired. *Sensors*, 22(19):7435, September 2022. ISSN 1424-8220. doi: 10.3390/s22197435. URL https://www.mdpi.com/1424-8220/22/19/7435.

Simona Caraiman, Anca Morar, Mateusz Owczarek, Adrian Burlacu, Dariusz Rzeszotarski, Nicolae Botezatu, Paul Herghelegiu, Florica Moldoveanu, Pawel Strumillo, and Alin Moldoveanu. Computer Vision for the Visually Impaired: the Sound of Vision System. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1480–1489, Venice, Italy, October 2017. IEEE. ISBN 978-1-5386-1034-3. doi: 10.1109/ICCVW.2017.175. URL http://ieeexplore.ieee.org/document/8265385/.

Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object Goal Navigation using Goal-Oriented Semantic Exploration, 2020. URL https://arxiv.org/abs/2007.00643. Version Number: 2.

Amit Chaudhary and Dr. PrabhatVerma. Road Surface Quality Detection Using Light Weight Neural Network for Visually Impaired Pedestrian. *Evergreen*, 10(2):706–714, June 2023. ISSN 2189-0420, 2432-5953. doi: 10.5109/6792818. URL https://hdl.handle.net/2324/6792818.

Dian Chen and Philipp Krähenbühl. Learning from all vehicles, 2022.

Karl Cobbe, Oleg Klimov, Christopher Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. *CoRR*, abs/1812.02341, 2018. URL http://arxiv.org/abs/1812.02341.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. *CoRR*, abs/1912.01588, 2019. URL http://arxiv.org/abs/1912.01588.

Carter Collins. Tactile Television¿Mechanical and Electrical Image Projection. *IEEE Transactions on Man Machine Systems*, 11(1):65–71, March 1970. ISSN 0536-1540. doi: 10.1109/TMMS.1970.299964. URL http://ieeexplore.ieee.org/document/4081932/.

James Coughlan and Roberto Manduchi. FUNCTIONAL ASSESSMENT OF A CAMERA PHONE-BASED WAYFINDING SYSTEM OPERATED BY BLIND AND VISUALLY IMPAIRED USERS. *International Journal on Artificial Intelligence Tools*, 18(03):379–397, June 2009. ISSN 0218-2130, 1793-6349. URL https://www.worldscientific.com/doi/abs/10.1142/S0218213009000196.

D. Dakopoulos, S. K. Boddhu, and N. Bourbakis. A 2D Vibration Array as an Assistive Device for Visually Impaired. In *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, pp. 930–937, Boston, MA, USA, October 2007. IEEE. ISBN 978-1-4244-1509-0. doi: 10.1109/BIBE.2007.4375670. URL http://ieeexplore.ieee.org/document/4375670/.

YURI DANILOV and MITCHELL TYLER. Brainport: An alternative input to the brain. *Journal of Integrative Neuroscience*, 04(04):537–550, 2005. doi: 10.1142/S0219635205000914. URL https://doi.org/10.1142/S0219635205000914.

Raghavendra Singh Dasila, Meet Trivedi, Shubham Soni, M. Senthil, and M. Narendran. Real time environment perception for visually impaired. In *2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, pp. 168–172, Chennai, April 2017. IEEE. ISBN 978-1-5090-4437-5. doi: 10.1109/TIAR.2017.8273709. URL http://ieeexplore.ieee.org/document/8273709/.

Ton De Jong. Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science*, 38(2):105–134, March 2010. ISSN 0020-4277, 1573-1952. doi: 10.1007/s11251-009-9110-0. URL http://link.springer.com/10.1007/s11251-009-9110-0.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. Robothor: An open simulation-to-real embodied AI platform. *CoRR*, abs/2004.06799, 2020. URL https://arxiv.org/abs/2004.06799.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation, 2022.

Aya Dernayka, Michel-Ange Amorim, Roger Leroux, Lucas Bogaert, and René Farcy. Tom pouce iii, an electronic white cane for blind people: Ability to detect obstacles and mobility performances. *Sensors*, 21(20), 2021. ISSN 1424-8220. doi: 10.3390/s21206854. URL https://www.mdpi.com/1424-8220/21/20/6854.

Kiana Ehsani, Winson Han, Alvaro Herrasti, Eli VanderBilt, Luca Weihs, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Manipulathor: A framework for visual object manipulation. *CoRR*, abs/2104.11213, 2021. URL https://arxiv.org/abs/2104.11213.

Giulia V. Elli, Stefania Benetti, and Olivier Collignon. Is There a Future for Sensory Substitution Outside Academic Laboratories? *Multisensory Research*, 27(5-6):271–291, 2014. ISSN 2213-4794, 2213-4808. doi: 10.1163/22134808-00002460. URL https://brill.com/view/journals/msr/27/5-6/article-p271_2.xml.

Wael Elloumi, Kamel Guissous, Aladine Chetouani, Raphael Canals, Remy Leconge, Bruno Emile, and Sylvie Treuillet. Indoor navigation assistance with a Smartphone camera based on vanishing points. In *International Conference on Indoor Positioning and Indoor Navigation*, pp. 1–9, Montbeliard, France, October 2013. IEEE. ISBN 978-1-4799-4043-1. doi: 10.1109/IPIN.2013.6817911. URL http://ieeexplore.ieee.org/document/6817911/.

Jothi Ganesan, Ahmad Taher Azar, Shrooq Alsenan, Nashwa Ahmad Kamal, Basit Qureshi, and Aboul Ella Hassanien. Deep Learning Reader for Visually Impaired. *Electronics*, 11(20):3335, October 2022. ISSN 2079-9292. doi: 10.3390/electronics11203335. URL https://www.mdpi.com/2079-9292/11/20/3335.

Gladys Garcia and Ani Nahapetian. Wearable computing for image-based indoor navigation of the visually impaired. In *Proceedings of the conference on Wireless Health*, pp. 1–6, Bethesda Maryland, October 2015. ACM. ISBN 978-1-4503-3851-6. doi: 10.1145/2811780.2811959. URL https://dl.acm.org/doi/10.1145/2811780.2811959.

Noriaki Hirose, Fei Xia, Roberto Martín-Martín, Amir Sadeghian, and Silvio Savarese. Deep visual mpc-policy learning for navigation. *CoRR*, abs/1903.02749, 2019. URL http://arxiv.org/abs/1903.02749.

Zhang-Wei Hong, Chen Yu-Ming, Shih-Yang Su, Tzu-Yun Shann, Yi-Hsiang Chang, Hsuan-Kung Yang, Brian Hsi-Lin Ho, Chih-Chieh Tu, Yueh-Chuan Chang, Tsu-Ching Hsiao, Hsin-Wei Hsiao, Sih-Pin Lai, and Chun-Yi Lee. Virtual-to-Real: Learning to Control in Visual Semantic Segmentation, 2018. URL https://arxiv.org/abs/1802.00285. Version Number: 4.

Di Hu, Dong Wang, Xuelong Li, Feiping Nie, and Qi Wang. Listen to the Image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7964–7973, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00816. URL https://ieeexplore.ieee.org/document/8953471/.

Volodymyr Ivanchenko, James Coughlan, and Huiying Shen. Crosswatch: A Camera Phone System for Orienting Visually Impaired Pedestrians at Traffic Intersections. In Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer (eds.), *Computers Helping People with Special Needs*, volume 5105, pp. 1122–1128. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-70539-0 978-3-540-70540-6. URL http://link.springer.com/10.1007/978-3-540-70540-6_168. ISSN: 0302-9743, 1611-3349 Series Title: Lecture Notes in Computer Science.

Max Jaderberg, Wojciech M. Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio García Castañeda, Charles Beattie, Neil C. Rabinowitz, Ari S. Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Z. Leibo, David Silver, Demis Hassabis, Koray Kavukcuoglu, and Thore Graepel. Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. *CoRR*, abs/1807.01281, 2018. URL http://arxiv.org/abs/1807.01281.

Crescent Jicol, Tayfun Esenkaya, Michael Proulx, Simon Lange-Smith, Meike Scheller, Eamonn O'Neill, and Karin Petrini. Efficiency of sensory substitution devices alone and in combination with self-motion for spatial navigation in sighted and visually impaired. *Frontiers in Psychology: Perception Science*, 11, July 2020. doi: 10.3389/fpsyg.2020.01443.

Mads Johansen, Martin Pichlmair, and Sebastian Risi. Video game description language environment for unity machine learning agents. In *2019 IEEE Conference on Games (CoG)*, pp. 1–8, 2019. doi: 10.1109/CIG.2019.8848072.

Lise A. Johnson and Charles M. Higgins. A Navigation Aid for the Blind Using Tactile-Visual Sensory Substitution. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6289–6292, New York, NY, August 2006. IEEE. ISBN 978-1-4244-0032-4. doi: 10.1109/IEMBS.2006.259473. URL http://ieeexplore.ieee.org/document/4463247/.

Arthur Juliani, Vincent-Pierre Berges, Esh Vckay, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange. Unity: A general platform for intelligent agents. *CoRR*, abs/1809.02627, 2018. URL http://arxiv.org/abs/1809.02627.

Arthur Juliani, Ahmed Khalifa, Vincent-Pierre Berges, Jonathan Harper, Hunter Henry, Adam Crespi, Julian Togelius, and Danny Lange. Obstacle tower: A generalization challenge in vision, control, and planning. *CoRR*, abs/1902.01378, 2019. URL http://arxiv.org/abs/1902.01378.

Niels Justesen, Ruben Rodriguez Torrado, Philip Bontrager, Ahmed Khalifa, Julian Togelius, and Sebastian Risi. Procedural level generation improves generality of deep reinforcement learning. *CoRR*, abs/1806.10729, 2018. URL http://arxiv.org/abs/1806.10729.

Hiroyuki Kajimoto, Naoki Kawakami, T Maeda, and S Tachi. Electro-tactile display with tactile primary color approach. *Graduate School of Information and Technology, The University of Tokyo*, 10 2004.

Hiroyuki Kajimoto, Yonezo Kanno, and Susumu Tachi. Forehead retina system. In *ACM SIGGRAPH 2006 Emerging technologies on - SIGGRAPH '06*, pp. 11, Boston, Massachusetts, 2006. ACM Press. ISBN 978-1-59593-364-5. doi: 10.1145/1179133.1179145. URL http://portal.acm.org/citation.cfm?doid=1179133.1179145.

Hiroyuki Kajimoto, Masaki Suzuki, and Yonezo Kanno. HamsaTouch: tactile vision substitution with smartphone and electro-tactile display. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, pp. 1273–1278, Toronto Ontario Canada, April 2014. ACM. ISBN 978-1-4503-2474-8. doi: 10.1145/2559206.2581164. URL https://dl.acm.org/doi/10.1145/2559206.2581164.

Katie Kang, Suneel Belkhale, Gregory Kahn, Pieter Abbeel, and Sergey Levine. Generalization through simulation: Integrating simulated and real data into deep reinforcement learning for vision-based autonomous flight. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 6008–6014, 2019. doi: 10.1109/ICRA.2019.8793735.

R. Kavitha, S. Shree Sandhya, Praveena Betes, P. Rajalakshmi, and E. Sarubala. Deep learning-based image captioning for visually impaired people. *E3S Web of Conferences*, 399:04005, 2023. ISSN 2267-1242. doi: 10.1051/e3sconf/202339904005. URL https://www.e3s-conferences.org/10.1051/e3sconf/202339904005.

Bengaluru Kavya and Lskshmikantha G C. Virtual eye for blind using iot. *International Journal of Engineering and Technical Research*, 8:116, 01 2020.

Hamideh Kerdegari, Yeongmi Kim, and Tony J. Prescott. Head-Mounted Sensory Augmentation Device: Comparing Haptic and Audio Modality. In Nathan F. Lepora, Anna Mura, Michael Mangan, Paul F.M.J. Verschure, Marc Desmulliez, and Tony J. Prescott (eds.), *Biomimetic and Biohybrid Systems*, volume 9793, pp. 107–118. Springer International Publishing, Cham, 2016. ISBN 978-3-319-42416-3 978-3-319-42417-0. URL http://link.springer.com/10.1007/978-3-319-42417-0_11. Series Title: Lecture Notes in Computer Science.

Atif Khan, Febin Moideen, Juan Lopez, Wai L. Khoo, and Zhigang Zhu. KinDectect: Kinect Detecting Objects. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Klaus Miesenberger, Arthur Karshmer, Petr Penaz, and Wolfgang Zagler (eds.), *Computers Helping People with Special Needs*, volume 7383, pp. 588–595. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-31533-6 978-3-642-31534-3. URL http://link.springer.com/10.1007/978-3-642-31534-3_86. Series Title: Lecture Notes in Computer Science.

Mooseop Kim, YunKyung Park, KyeongDeok Moon, and Chi Yoon Jeong. Analysis and Validation of Cross-Modal Generative Adversarial Network for Sensory Substitution. *International Journal of Environmental Research and Public Health*, 18(12):6216, June 2021. ISSN 1660-4601. doi: 10.3390/ijerph18126216. URL https://www.mdpi.com/1660-4601/18/12/6216.

Mooseop Kim, Yunkyung Park, Kyeongdeok Moon, and Chi Yoon Jeong. Deep Learning-Based Optimization of Visual–Auditory Sensory Substitution. *IEEE Access*, 11:14169–14180, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3243641. URL https://ieeexplore.ieee.org/document/10041118/.

Eunjeong Ko and Eun Kim. A Vision-Based Wayfinding System for Visually Impaired People Using Situation Awareness and Activity-Based Instructions. *Sensors*, 17(8):1882, August 2017. ISSN 1424-8220. doi: 10.3390/s17081882. URL https://www.mdpi.com/1424-8220/17/8/1882.

Thibault Kruse, Amit Kumar Pandey, Rachid Alami, and Alexandra Kirsch. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12):1726–1743, 2013. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2013.05.007. URL https://www.sciencedirect.com/science/article/pii/S0921889013001048.

R. Kuc. Binaural sonar electronic travel aid provides vibrotactile cues for landmark, reflector motion and surface texture classification. *IEEE Transactions on Biomedical Engineering*, 49(10):1173–1180, October 2002. ISSN 0018-9294. doi: 10.1109/TBME.2002.803561. URL http://ieeexplore.ieee.org/document/1035967/.

Jonás Kulhánek, Erik Derner, Tim de Bruin, and Robert Babuska. Vision-based navigation using deep reinforcement learning. *CoRR*, abs/1908.03627, 2019. URL http://arxiv.org/abs/1908.03627.

A. Rohith Kumar, K. Sanjay, and M. Praveen. EchoGuide: Empowering the Visually Impaired with IoT-Enabled Smart Stick and Audio Navigation. In *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pp. 1770–1774, Pudukkottai, India, December 2023. IEEE. ISBN 9798350340235. doi: 10.1109/ICACRS58579.2023.10405085. URL https://ieeexplore.ieee.org/document/10405085/.

Cheng-Lung Lee. An evaluation of tactile symbols in public environment for the visually impaired. *Applied Ergonomics*, 75:193–200, February 2019. ISSN 00036870. doi: 10.1016/j.apergo.2018.10.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0003687018304939.

Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, Sai Bi, Hong-Xing Yu, Zexiang Xu, Kalyan Sunkavalli, Milos Hasan, Ravi Ramamoorthi, and Manmohan Chandraker. Openrooms: An open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7190–7199, June 2021.

Lorena Lobo, Patric C. Nordbeck, Vicente Raja, Anthony Chemero, Michael A. Riley, David M. Jacobs, and David Travieso. Route selection and obstacle avoidance with a short-range haptic sensory substitution device. *International Journal of Human-Computer Studies*, 132:25–33, December 2019. ISSN 10715819. doi: 10.1016/j.ijhcs.2019.03.004. URL https://linkinghub.elsevier.com/retrieve/pii/S1071581918301046.

Jagadish K. Mahendran, Daniel T. Barry, Anita K. Nivedha, and Suchendra M. Bhandarkar. Computer vision-based assistance system for the visually impaired using mobile edge artificial intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2418–2427, June 2021.

Shachar Maidenbaum, Shlomi Hanassy, Sami Abboud, Galit Buchs, Daniel-Robert Chebat, Shelly Levy-Tzedek, and Amir Amedi. The "EyeCane", a new electronic travel aid for the blind: Technology, behavior & swift learning. *Restorative Neurology and Neuroscience*, 32(6):813–824, 2014. ISSN 09226028. doi: 10.3233/RNN-130351. URL https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/RNN-130351.

Adriano Mancini, Emanuele Frontoni, and Primo Zingaretti. Mechatronic System to Help Visually Impaired Users During Walking and Running. *IEEE Transactions on Intelligent Transportation Systems*, 19(2):649–660, February 2018. ISSN 1524-9050, 1558-0016. doi: 10.1109/TITS.2017.2780621. URL http://ieeexplore.ieee.org/document/8253603/.

P.B.L. Meijer. An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering*, 39(2):112–121, 1992a. doi: 10.1109/10.121642.

P.B.L. Meijer. An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering*, 39(2):112–121, February 1992b. ISSN 00189294. doi: 10.1109/10.121642. URL http://ieeexplore.ieee.org/document/121642/.

Mukhriddin Mukhiddinov and Jinsoo Cho. Smart Glass System Using Deep Learning for the Blind and Visually Impaired. *Electronics*, 10(22):2756, November 2021. ISSN 2079-9292. doi: 10.3390/electronics10222756. URL https://www.mdpi.com/2079-9292/10/22/2756.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

Andrew Port, Chelhwon Kim, and Mitesh Patel. Deep sensory substitution: Noninvasively enabling biological neural networks to receive input from artificial neural networks, 2021. URL https://arxiv.org/abs/2005.13291.

Mohammad Marufur Rahman, Md. Milon Islam, Shishir Ahmmed, and Saeed Anwar Khan. Obstacle and Fall Detection to Guide the Visually Impaired People with Real Time Monitoring. *SN Computer Science*, 1(4):219, July 2020. ISSN 2662-995X, 2661-8907. doi: 10.1007/s42979-020-00231-x. URL https://link.springer.com/10.1007/s42979-020-00231-x.

Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M. Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3D): 1000 large-scale 3d environments for embodied AI. *CoRR*, abs/2109.08238, 2021. URL https://arxiv.org/abs/2109.08238.

Alejandro Rituerto, Giovanni Fusco, and James M. Coughlan. Towards a Sign-Based Indoor Navigation System for People with Visual Impairments. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 287–288, Reno Nevada USA, October 2016. ACM. ISBN 978-1-4503-4124-0. doi: 10.1145/2982142.2982202. URL https://dl.acm.org/doi/10.1145/2982142.2982202.

Alberto Rodríguez, J. Javier Yebes, Pablo Alcantarilla, Luis Bergasa, Javier Almazán, and Andrés Cela. Assisting the Visually Impaired: Obstacle Detection and Warning System by Acoustic Feedback. *Sensors*, 12(12):17476–17496, December 2012. ISSN 1424-8220. doi: 10.3390/s121217476. URL http://www.mdpi.com/1424-8220/12/12/17476.

Luis F. Rodríguez-Ramos. Virtual Acoustic Space: Space perception for the blind. *Proceedings of the International Astronomical Union*, 5(S260):556–563, January 2009. ISSN 1743-9213, 1743-9221. doi: 10.1017/S1743921311002845. URL https://www.cambridge.org/core/product/identifier/S1743921311002845/type/journal_article.

Mriyank Roy and Purav Shah. Internet of Things (IoT) Enabled Smart Navigation Aid for Visually Impaired. In Leonard Barolli, Farookh Hussain, and Tomoya Enokido (eds.), *Advanced Information Networking and Applications*, volume 451, pp. 232–244. Springer International Publishing, Cham, 2022. ISBN 978-3-030-99618-5 978-3-030-99619-2. URL https://link.springer.com/10.1007/978-3-030-99619-2_23. Series Title: Lecture Notes in Networks and Systems.

Florian Scalvini, Camille Bordeau, Maxime Ambard, Cyrille Migniot, and Julien Dubois. Outdoor Navigation Assistive System Based on Robust and Real-Time Visual–Auditory Substitution Approach. *Sensors*, 24(1):166, December 2023. ISSN 1424-8220. doi: 10.3390/s24010166. URL https://www.mdpi.com/1424-8220/24/1/166.

Tobias Schwarze, Martin Lauer, Manuel Schwaab, Michailas Romanovas, Sandra Bohm, and Thomas Jurgensohn. An Intuitive Mobility Aid for Visually Impaired People Based on Stereo Vision. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 409–417, Santiago, Chile, December 2015. IEEE. ISBN 978-1-4673-9711-7. doi: 10.1109/ICCVW.2015.61. URL http://ieeexplore.ieee.org/document/7406410/.

Hervé Segond, Déborah Weiss, and Eliana Sampaio. Human Spatial Navigation via a Visuo-Tactile Sensory Substitution System. *Perception*, 34(10):1231–1249, October 2005. ISSN 0301-0066, 1468-4233. doi: 10.1068/p3409. URL http://journals.sagepub.com/doi/10.1068/p3409.

Dhruv Shah and Sergey Levine. Viking: Vision-based kilometer-scale navigation with geographic hints. In *Robotics: Science and Systems XVIII*, RSS2022, pp. –. Robotics: Science and Systems Foundation, June 2022. doi: 10.15607/rss.2022.xviii.019. URL http://dx.doi.org/10.15607/RSS.2022.XVIII.019.

Tarun Sharma, J H M Apoorva, Ramananathan Lakshmanan, Prakruti Gogia, and Manoj Kondapaka. NAVI: Navigation aid for the visually impaired. In *2016 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 971–976, Greater Noida, India, April 2016. IEEE. ISBN 978-1-5090-1666-2. doi: 10.1109/CCAA.2016.7813856. URL http://ieeexplore.ieee.org/document/7813856/.

S. Shoval, J. Borenstein, and Y. Koren. The NavBelt-a computerized travel aid for the blind based on mobile robotics technology. *IEEE Transactions on Biomedical Engineering*, 45(11):1376–1386, November 1998. ISSN 00189294. doi: 10.1109/10.725334. URL http://ieeexplore.ieee.org/document/725334/.

Rajeev Ranjan Singh and Priya Maurya. Visual impairment and falls among older adults and elderly: evidence from longitudinal study of ageing in india. *BMC Public Health*, 22(1), December 2022. ISSN 1471-2458. doi: 10.1186/s12889-022-14697-2. URL http://dx.doi.org/10.1186/s12889-022-14697-2.

Muhammad Sulaman, S.U.Bazai, Muhammad AKram, and Muhammad Akram Khan. The Deep Learning based Smart Navigational Stick for Blind People. *UMT Artificial Intelligence Review*, 2(No 2), March 2023. ISSN 2791-1276, 2791-1268. doi: 10.32350/umtair.22.05. URL https://journals.umt.edu.pk/index.php/UMT-AIR/article/view/3488.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimír Vondruš, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 251–266. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/021bbc7ee20b71134d53e20206bd6feb-Paper.pdf.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. URL http://arxiv.org/abs/1905.11946.

Ruxandra Tapu, Bogdan Mocanu, Andrei Bursuc, and Titus Zaharia. A Smartphone-Based Obstacle Detection and Classification System for Assisting Visually Impaired People. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 444–451, Sydney, Australia, December 2013. IEEE. ISBN 978-1-4799-3022-7. doi: 10.1109/ICCVW.2013.65. URL http://ieeexplore.ieee.org/document/6755931/.

Ruxandra Tapu, Bogdan Mocanu, and Titus Zaharia. DEEP-SEE: Joint Object Detection, Tracking and Recognition with Application to Visually Impaired Navigational Assistance. *Sensors*, 17(11):2473, October 2017. ISSN 1424-8220. doi: 10.3390/s17112473. URL http://www.mdpi.com/1424-8220/17/11/2473.

Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972. doi: 10.1137/0201010. URL https://doi.org/10.1137/0201010.

Ender Tekin and James M. Coughlan. An algorithm enabling blind users to find and read barcodes. In *2009 Workshop on Applications of Computer Vision (WACV)*, pp. 1–8, Snowbird, UT, USA, December 2009. IEEE. ISBN 978-1-4244-5497-6. doi: 10.1109/WACV.2009.5403098. URL http://ieeexplore.ieee.org/document/5403098/.

Anne Theurel, Arnaud Witt, Philippe Claudet, Yvette Hatwell, and Edouard Gentaz. Tactile picture recognition by early blind children: The effect of illustration technique. *Journal of Experimental Psychology: Applied*, 19(3):233–240, September 2013. ISSN 1939-2192, 1076-898X. doi: 10.1037/a0034255. URL https://doi.apa.org/doi/10.1037/a0034255.

Marin Toromanoff, Émilie Wirbel, and Fabien Moutarde. End-to-end model-free reinforcement learning for urban driving using implicit affordances. *CoRR*, abs/1911.10868, 2019. URL http://arxiv.org/abs/1911.10868.

I. Ulrich and J. Borenstein. The GuideCane-applying mobile robot technologies to assist the visually impaired. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 31(2):131–136, March 2001. ISSN 10834427. doi: 10.1109/3468.911370. URL http://ieeexplore.ieee.org/document/911370/.

Pablo Vera, Daniel Zenteno, and Joaquín Salas. A smartphone-based virtual white cane. *Pattern Analysis and Applications*, 17(3):623–632, August 2014. ISSN 1433-7541, 1433-755X. doi: 10.1007/s10044-013-0328-8. URL http://link.springer.com/10.1007/s10044-013-0328-8.

Ziyu Wang, Nando de Freitas, and Marc Lanctot. Dueling network architectures for deep reinforcement learning. *CoRR*, abs/1511.06581, 2015. URL http://arxiv.org/abs/1511.06581.

Tess Winlock, Eric Christiansen, and Serge Belongie. Toward real-time grocery detection for the visually impaired. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 49–56, San Francisco, CA, USA, June 2010. IEEE. ISBN 978-1-4244-7029-7. doi: 10.1109/CVPRW.2010.5543576. URL http://ieeexplore.ieee.org/document/5543576/.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders, 2023.

Joanne M. Wood, Philippe Lacherez, Alex A. Black, Michael H. Cole, Mei Ying Boon, and Graham K. Kerr. Risk of Falls, Injurious Falls, and Other Injuries Resulting from Visual Impairment among Older Adults with Age-Related Macular Degeneration. *Investigative Ophthalmology Visual Science*, 52(8):5088–5092, 07 2011. ISSN 1552-5783. doi: 10.1167/iovs.10-6644. URL https://doi.org/10.1167/iovs.10-6644.

Thomas Wright and Jamie Ward. The evolution of a visual-to-auditory sensory substitution device using interactive genetic algorithms. *Quarterly Journal of Experimental Psychology*, 66(8):1620–1638, August 2013. ISSN 1747-0218, 1747-0226. doi: 10.1080/17470218.2012.754911. URL http://journals.sagepub.com/doi/10.1080/17470218.2012.754911.

Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. *CoRR*, abs/1808.10654, 2018. URL http://arxiv.org/abs/1808.10654.

Fei Xia, William B. Shen, Chengshu Li, Priya Kasimbeg, Micael Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. Interactive gibson: A benchmark for interactive navigation in cluttered environments. *CoRR*, abs/1910.14442, 2019. URL http://arxiv.org/abs/1910.14442.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José M. Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203, 2021. URL https://arxiv.org/abs/2105.15203.

Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation, 2022.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. *arXiv preprint arXiv:2210.05633*, 2022. URL https://arxiv.org/abs/2210.05633.

Chiyuan Zhang, Oriol Vinyals, Rémi Munos, and Samy Bengio. A study on overfitting in deep reinforcement learning. *CoRR*, abs/1804.06893, 2018. URL http://arxiv.org/abs/1804.06893.

Yuhang Zhao, Elizabeth Kupferstein, Hathaitorn Rojnirun, Leah Findlater, and Shiri Azenkot. The Effectiveness of Visual and Audio Wayfinding Guidance on Smartglasses for People with Low Vision. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, Honolulu HI USA, April 2020. ACM. ISBN 978-1-4503-6708-0. doi: 10.1145/3313831.3376516. URL https://dl.acm.org/doi/10.1145/3313831.3376516.

Zejia Zheng and Juyang Weng. Mobile Device Based Outdoor Navigation with On-Line Learning Neural Network: A Comparison with Convolutional Neural Network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 11–18, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-5090-1437-8. doi: 10.1109/CVPRW.2016.9. URL http://ieeexplore.ieee.org/document/7789499/.

Jinjing Zhu, Haotian Bai, and Lin Wang. Patch-mix transformer for unsupervised domain adaptation: A game perspective, 2023.

A LITERATURE REVIEW

To assess the relevance of AI methods for sensory substitution devices, we conducted a literature review on sensory substitution devices and electronic travel aids published since 1970. The review was performed in PubMed using binary strings that combined one sensory substitution-related term (e.g., sensory substitution, electronic travel aid) with a signal-processing-related term (e.g., deep learning, computer vision, machine learning, image processing, signal processing, artificial intelligence, neural networks). The results were supplemented with research from Elicit, various search engines, and arXiv. We included papers proposing new systems for sensory substitution, resulting in a total of 61 studies. Each paper was analyzed and labeled according to the information detailed in Table 3, including the type of image or signal processing used, as well as additional information such as the device's purpose, type of sensors, and sensor location. The full review results are presented in Tables 4,5.

Reviewed Property	**Description and acronyms**
Year	Publish year.
Sensor (S)	Sensors used for environment information acquisition. *RGB Camera (CAM), RGB-Depth Camera(s) (D-CAM), Ultrasonic (US), LIDAR, Infrared (IR), Externally added Geographic Information (GI), GPS, Laser Beam (LB).*
Purpose	Purpose of the SSD. *Generic Use for any application (GU), Obstacle Avoidance (OA), Navigation (NAV), Object Recognition (OR), Localization (LOC).*
Processing Tools (TOOLS)	Methodological Tools used for information extraction. *Neural Networks (NN), Computer Vision (CV), Machine Learning traditional approach (ML), Signal Processing (SP).*
Sensor Location (SLOC)	Location of the acquisition sensors.

Table 3: List of reviewed aspects for prosthetic vision studies and their acronyms

Ref	Year	S	Purpose	Tools	SLOC
Collins (1970)	1970	CAM	GU	CV	back
Meijer (1992b)	1992	CAM	GU	CV	NA
Bach-y Rita et al. (1998)	1998	CAM	GU	CV	NA
Shoval et al. (1998)	1998	US	OA	SP	belt
Rodríguez-Ramos (2009)	1999	D-CAM	GU	CV	glasses
Arno et al. (1999)	1999	CAM	GU	CV	Head
Ulrich & Borenstein (2001)	2001	US	OA	SP	A-Cane
Kuc (2002)	2002	US	OA,OR	SP	NA
Segond et al. (2005)	2005	CAM	OA	CV	abdomen
DANILOV & TYLER (2005)	2006	CAM	GU	NA	head
Kajimoto et al. (2006)	2006	CAM	GU	CV	head
Johnson & Higgins (2006)	2006	D-CAM	OA	CV	belt
Dakopoulos et al. (2007)	2007	D-CAM	OA	CV	glasses
Tekin & Coughlan (2009)	2009	CAM	OR	ML	NA
Coughlan & Manduchi (2009)	2009	CAM,GI	WF	CV	hand
Winlock et al. (2010)	2010	CAM	OR	ML	NA
Khan et al. (2012)	2012	D-CAM	OR,OA	CV	belt
Rodríguez et al. (2012)	2012	D-CAM	OA	CV	chest
Vera et al. (2014)	2013	CAM,LB	OA	SP	hand
Tapu et al. (2013)	2014	CAM	OA	ML	chest
Elloumi et al. (2013)	2014	CAM,GI	WF	ML	chest
Maidenbaum et al. (2014)	2014	IR	OA	SP	hand
Kajimoto et al. (2014)	2014	CAM	GU	CV	hand
Garcia & Nahapetian (2015)	2015	CAM	OA	CV	glasses
Zheng & Weng (2016)	2016	CAM,GPS	OA	NN	hand
Bulat & Glowacz (2016)	2016	D-CAM	OA	CV	NA
Ivanchenko et al. (2008)	2016	CAM	OR	ML	hand
Schwarze et al. (2015)	2016	D-CAM,IS	OA	CV	helmet
Kerdegari et al. (2016)	2016	US	OA	NN	helmet
Ko & Kim (2017)	2017	CAM,IS,GI	WF	ML	hand
Rituerto et al. (2016)	2017	CAM,IS,GI	LOC	CV	chest
Tapu et al. (2017)	2017	CAM	OA	NN	belt
Bai et al. (2017)	2017	D-CAM, CAM,GI	NAV	NN	helmet
Sharma et al. (2016)	2017	CAM	OR,OA	NN	chest
Agarwal et al. (2017)	2017	US	OA	SP	glasses
Dasila et al. (2017)	2018	D-CAM	OR,OA	CV	NA

Table 4: Review Results part.1

Ref	Year	S	Purpose	Tools	SLOC
Caraiman et al. (2017)	2018	IR,D-CAM	OA,OR, TR	ML	abdomen
Mancini et al. (2018)	2018	CAM	WF	CV	Gloves
Bhat et al. (2017)	2018	CAM	OR,TR	ML	NA
Lobo et al. (2019)	2019	IR	OA	SP	hand
Hu et al. (2019)	2019	CAM	GU	NN	NA
Rahman et al. (2020)	2020	US,IR, IS,GI	OA	SP	shin
Port et al. (2021)	2020	CAM	OR	NN	NA
Afif et al. (2020)	2020	CAM	OR	NN	None
Zhao et al. (2020)	2020	CAM,GPS	WF,OA	NA	glasses
Kavya & G C (2020)	2020	CAM,US, GPS	OA,OR, TR	NN	A-Cane
Kim et al. (2021)	2021	CAM	GU	NN	glasses
Dernayka et al. (2021)	2021	LIDAR,IR	OA	SP	A-Cane
Mukhiddinov & Cho (2021)	2021	CAM	OR,TR	NN	NA
Wright & Ward (2013)	2021	CAM	GU	ML	NA
Busaeed et al. (2022)	2022	LIDAR,LB, US,GPS	OA	NN	glasses
Ganesan et al. (2022)	2022	CAM	OR	NN	NA
Roy & Shah (2022)	2022	GPS,US	OA,LOC	SP	A-Cane
Bhatlawande et al. (2022)	2022	CAM,US	OR,OA	NN	Hand
Scalvini et al. (2023)	2023	D-CAM, GPS,IS	NAV	NN	helmet
Asiedu Asante & Imamura (2023)	2023	D-CAM	OA	NN	abdomen
Kim et al. (2023)	2023	CAM	GU	NN	NA
Sulaman et al. (2023)	2023	CAM,US, GPS,IF	OA,OR, LOC	NN	A-Cane
Chaudhary & Dr. PrabhatVerma (2023)	2023	CAM	OR	NN	None
Kavitha et al. (2023)	2023	CAM	OR	NN	None
Kumar et al. (2023)	2023	CAM,US, GPS,IS	OA,OR, LOC	CV	A-Cane

Table 5: Review Results part.2

B GRID SEARCH IN NAVINDOOR

Training was conducted alongside the agent's exploration of the environment, utilizing an ϵ-greedy policy.

Each episode took place in a new maze and lasted for 400 decision timesteps, for both training and evaluation. At a fixed frequency, data was sampled from a replay buffer and used for Q_θ parameters update. Semantic maps used for training were encoded as 2D arrays with values: 1 for coins, 0.5 for floor, -0.5 for walls, -1 for obstacles, and 0 elsewhere. Training lasted for 250 episodes. The model was implemented with PyTorch and trained using NVIDIA RTX 4090.

Obstacle avoidance was learned using Double Dueling Deep Q Network (D3QN) Wang et al. (2015) which the associated loss function L is given by:

$$L(\theta) = \mathbb{E}\left[\left(y - Q_\theta(s,a)\right)^2\right], \tag{1}$$

where the target value y for the Double Q-Learning approach is defined as:

$$y = r + \gamma Q_{\theta^-}\left(s', \underset{a'}{\operatorname{argmax}}\, Q_\theta(s',a')\right), \tag{2}$$

and in the Dueling Network architecture, the Q function is decomposed into:

$$Q_\theta(s,a) = V_\theta(s) + \left(A_\theta(s,a) - \frac{1}{|\mathcal{A}|}\sum_{a'} A_\theta(s,a')\right), \tag{3}$$

where $V_\theta(s)$ represents the value function, $A(s,a;\theta)$ the advantage function, $|\mathcal{A}|$ the number of possible actions, θ and θ^- the parameters for current and target networks, respectively.

Considering the complexity of reinforcement learning in 3D environments, we performed a grid search to understand the model's sensitivity to training parameters. The parameters considered for the grid search, along with their associated **best** values, are as follows: model size (1.1, **1.6M**, 2.4 parameters), learning rate (0.0001,0.0005,**0.001**), batch size (64,**128**,256), discount factor γ (0.95,**0.97**,0.99), action memory length m (10,**20**), ϵ_{min} (0.15,**0.3**), training proportion of linear decrease ratio for ϵ (0.25,**0.33**,0.5), training iteration frequency (**5**,10), coin reward value (1,**5**,10), update type of Q_{θ^-} (***hard*** **update every 50 training steps**, *hard* update every 10 training steps, *soft* update). This grid search, as well as all evaluation results use a default setting specified below. After the grid search, we conducted training sessions with the best-found hyper-parameters. A total of 16 training sessions were performed, considering 4 binary settings, and the best models for visible/invisible collectibles were designated as VC/NVC, respectively. Then, using the VC settings, 2 additional training sessions were conducted to train the RGB model with data augmentation for VCD.

The first binary setting examined fixed moving speed (default, fixed to 1) *vs* random moving speed (sampled between 0.75 and 1.25) to assess the model's ability to generalize to agents with different speeds. The second binary setting investigated the need for visual cues by training with visible (default) *vs* invisible coins. Visual cues assist the model in navigation but also introduce a bias for generalization to real-world semantic segmentation maps, which may not include such visual cues. The third binary setting implemented decreasing rewards only at collision time (default) *vs*decreasing rewards when staying near an obstacle. This study aims to determine if this approach could prevent situations where the agent becomes stuck by colliding with an obstacle.

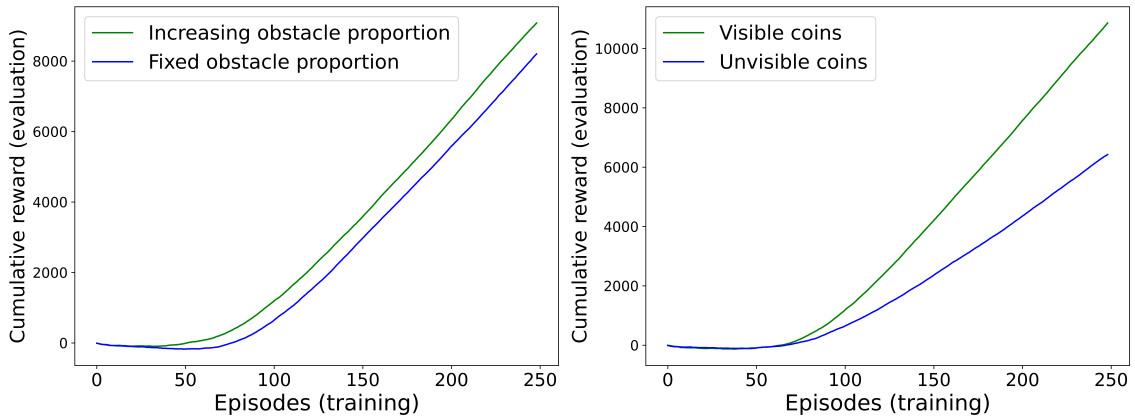

Figure 9: Mean cumulative rewards are shown for fixed *vs* variable difficulty during training (left) and for visible *vs* invisible coin settings (right). The mean cumulative reward was computed over the 8 models in each group.

The last binary setting aimed to study potential improvements in performance by progressively increasing task difficulty through obstacle proportion, in accordance with previous studies Kulhánek et al. (2019); Justesen et al. (2018). We compared a fixed proportion of obstacles throughout training (default) with a linear increase in obstacle proportion during training. Following training, each of the 16+2 models was evaluated on 500 unseen mazes under the default setting.

Two subgroups showed significant differences compared to their binary counterparts. First, the coin visibility group demonstrated better performance (mean reward of 72.76), while the model still exhibited learning capabilities even without visual cues (mean reward of 45.22). Although the gradual increase in difficulty did not significantly enhance the mean reward, it accelerated the convergence of the models. These results can be observed in the cumulative rewards of these subgroups throughout training, as shown in Figure 9. Cumulative rewards and mean rewards after training showed no significant differences for the other two binary settings.

C Train/Test Split in Active Vision Dataset

Linear probing was performed using the currently available data from the Active Vision Dataset Houses. Scenes from Houses 1, 2, 3, 4, and 7 were used for training, while Houses 10, 11, 13, 15, and 16 were used for testing.