computers

MDPI

Article

A Context-Aware Doorway Alignment and Depth Estimation Algorithm for Assistive Wheelchairs

Shanelle Tennekoon [1,*], Nushara Wedasingha [2], Anuradhi Welhenge [1], Nimsiri Abhayasinghe [1] and Iain Murray [1,*]

[1] School of Electrical Engineering, Computing & Mathematical Sciences, Curtin University, Bentley, WA 6102, Australia; anuradhi.welhenge@curtin.edu.au (A.W.); k.abhayasinghe@curtin.edu.au (N.A.)
[2] Department of Electrical and Electronic Engineering, Center of Excellence in Informatics (CIET), Electronics & Transmission, Faculty of Engineering, Sri Lanka Institute of Information Technology, Malabe 10115, Sri Lanka; nushara.w@sliit.lk
* Correspondence: h.tennekoon@postgrad.curtin.edu.au (S.T.); i.murray@curtin.edu.au (I.M.)

Abstract

Navigating through doorways remains a daily challenge for wheelchair users, often leading to frustration, collisions, or dependence on assistance. These challenges highlight a pressing need for intelligent doorway detection algorithm for assistive wheelchairs that go beyond traditional object detection. This study presents the algorithmic development of a lightweight, vision-based doorway detection and alignment module with contextual awareness. It integrates channel and spatial attention, semantic feature fusion, unsupervised depth estimation, and doorway alignment that offers real-time navigational guidance to the wheelchairs control system. The model achieved a mean average precision of 95.8% and a F1 score of 93%, while maintaining low computational demands suitable for future deployment on embedded systems. By eliminating the need for depth sensors and enabling contextual awareness, this study offers a robust solution to improve indoor mobility and deliver actionable feedback to support safe and independent doorway traversal for wheelchair users.

Keywords: assistive navigation; context-aware; doorway detection; indoor navigation; vision-based navigation; wheelchair guidance; YOLOv8 segmentation

check for updates

Academic Editors: Amit Kumar Mishra and Deepak Puthal

Received: 18 June 2025
Revised: 11 July 2025
Accepted: 14 July 2025
Published: 17 July 2025

Citation: Tennekoon, S.; Wedasingha, N.; Welhenge, A.; Abhayasinghe, N.; Murray, I. A Context-Aware Doorway Alignment and Depth Estimation Algorithm for Assistive Wheelchairs. *Computers* **2025**, *14*, 284. https://doi.org/10.3390/computers14070284

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Individual autonomy is widely recognized as a fundamental factor that fosters intrinsic motivation [1] and supports mental health [2]. Since mobility and enabling environments are essential to exercising individual autonomy, the ability to move independently plays a critical role in supporting self-reliance and enhancing overall well-being [3–5]. However, individuals with conditions such as spasticity [6], tremor [7], paresis [8], or visual impairments [9] often face severe challenges in maintaining this autonomy. These conditions can hinder the precise control and navigation of wheelchairs, especially in complex environments with narrow or misaligned spaces like doorways [10]. As a result, affected individuals are frequently forced to depend on others for assistance, which can reduce their independence, increase both cognitive and physical strain, and lead to feelings of social stigma. Given the rising global prevalence of visual and mobility impairments [11–14], there is a growing need for intelligent, automated navigating algorithms that minimize the reliance on caregivers and support individuals in maintaining autonomy and well-being when using wheelchairs.

In response to this need, researchers have focused on developing automated, sensor-based doorway detection systems for powered wheelchairs. Early navigation systems primarily relied on sensors such as ultrasonic modules [15,16], LiDAR [17,18], and infrared sensors [19] to detect doors and doorways. These approaches typically utilized range and depth data to identify flat vertical surfaces and open spaces, often incorporating geometric modeling to delineate doorway boundaries. For example, Grewal et al. [20] used a 2D LiDAR sensor integrated with simultaneous localization and mapping [21] to map the surrounding environment, detect walls, and infer doorway locations based on spatial discontinuities. This information was then combined with path planning algorithms to enable autonomous wheelchair navigation through the identified doorways. While these sensor-based methods have shown effectiveness in controlled settings, their performance in real-world environments remains limited. Challenges include reduced detection accuracy in dynamic or cluttered spaces, limited sensing range, and vulnerability to environmental noise and interference [22–24]. Additionally, the reliance on multiple sensors increases system complexity, hardware cost, and power consumption, thereby limiting the practicality, scalability, and widespread adoption of such systems [25].

While sensor-based approaches laid the groundwork for autonomous doorway detection, the need for richer spatial understanding prompted researchers to explore vision-based alternatives. Various methods have been proposed using stereo and depth imaging to enhance environmental perception. For instance, Derry and Argall [26] utilized the Microsoft Kinect to extract 3D point clouds and applied geometric constraints to identify open doorways. Their approach eliminated the need for pre-existing maps but was limited to fully open doors and required high-quality depth data. Similarly, Rusu et al. [27] extracted planar surfaces representing doors and walls from 3D point clouds. Anguelov et al. [28] employed stereo vision by integrating 2D laser scanners with panoramic cameras and used probabilistic models to segment doors from walls. Although this method incorporated visual information, it remained dependent on prior maps and was computationally demanding. Despite these advances, vision-based doorway detection methods continue to face significant challenges. Their performance is often hindered by variations in lighting conditions, occlusions, and a lack of robust spatial reasoning capabilities, limiting their reliability in dynamic and unstructured environments.

To overcome the limitations of traditional sensor-based mechanisms, researchers have increasingly focused on vision-based systems, leveraging recent advances in deep learning to enhance the accuracy and autonomy of doorway detection. These approaches commonly utilize monocular RGB inputs to recognize and interpret structural features in indoor environments. Lecronsnier et al. [29] introduced a deep learning framework to support doorway traversal for smart wheelchairs, integrating the You Only Look Once Version 3 (YOLOv3) algorithm [30] with Intel RealSense sensors to incorporate depth perception. The system also employed the simple online real-time tracking algorithm [31] for robust 3D object tracking, enabling the accurate detection of doors and handles and enhancing the semi-autonomous functionality of the wheelchair. Expanding on this direction, Zhang et al. [32] proposed DenseNet Spatial Pyramid Pooling (DSPP-YOLO), an enhanced YOLOv3-based architecture designed for the detection of doors and windows in unfamiliar indoor settings. This model incorporated DenseNet blocks [33] and spatial pyramid pooling [34] to improve multi-scale feature extraction. As a result, DSPP-YOLO achieved a 3.3% improvement in door detection accuracy and an 8.8% increase in window detection. Similarly, Mochurad and Hladun [35] developed a real-time neural network for door handle detection using RGB-D cameras. Their model utilized a MobileNetV2 backbone [36] with a custom decoder, enabling processing speeds of up to 16 frames per second. With the growing demand for lightweight and efficient real-time detection models, recent work

has turned to YOLOv8 [37] to enhance scene understanding in wheelchair navigation, including doorway recognition [38–40]. However, many of these models focus primarily on object detection without effectively capturing the contextual or structural cues necessary for robust spatial understanding [23,41,42]. In particular, these architectures often lack dedicated mechanisms to highlight the distinct visual and geometric features of structural elements, such as doors. As a result, they are prone to misclassifying visually similar patterns such as wall segments or cabinets as doorways, leading to navigation errors and potential safety risks for users.

Despite the growing interest in both sensor- and vision-based navigation systems, accurate doorway detection and alignment guidance for autonomous wheelchairs remain relatively underexplored. This study specifically addresses the doorway detection and traversal guidance component of a broader unified navigation system currently being developed for assistive wheelchairs. We propose a vision-based architecture that detects doorways and provides directional feedback to the controller using only monocular RGB input. The proposed model improves the YOLOv8-seg backbone by integrating spatial and channel-wise attention mechanisms via the Convolutional Block Attention Module (CBAM) [43] and the Content-Guided Convolutional Attention Fusion (CGCAFusion) module [44], improving multi-scale feature aggregation. To compensate for the lack of physical depth sensors, we introduce a lightweight, RGB-based depth estimation head capable of inferring spatial layouts directly from 2D images. This module is optimized for real-time inference in low-power embedded systems, facilitating practical deployment on assistive mobility platforms in the future. The core contributions of this work include the following:

- Enhanced spatial and channel-wise focus on structural features of doorways by integrating an attention module within the YOLOv8 backbone.
- Improved accuracy in detecting semantically similar regions by utilizing multi-scale contextual refinement.
- Development of a lightweight dual head structure for simultaneous door detection and depth estimation.
- Construction of a module to detect and correct door misalignment and provide real-time guidance to the wheelchair control system.

2. Materials and Methods

In this section, we present the proposed model, YOLOv8-seg-CA (Context-Aware), which specifically addresses the doorway detection and alignment guidance component within a larger assistive wheelchair navigation system currently under development (Figure 1). The primary objective of this module is to detect doorways and generate directional cues that assist the wheelchair's control system using only video input.

As shown in Figure 1, the YOLO architecture is structured into three main stages as follows: the Backbone, Neck, and Head. The Backbone is responsible for initial feature extraction from the input RGB frames. The extracted features are then passed through the Neck, which performs multi-scale feature aggregation using a series of convolutional and upsampling layers. Finally, a dual Head produces the segmentation output and depth estimation by decoding the fused features. The overall architecture of the proposed model is composed of the following four key sub-modules:

(1) **CBAM:** This sub-module enhances feature representation within the YOLOv8 backbone by enabling the model to more effectively focus on salient regions associated with doorways, thereby improving detection accuracy and robustness (Section 2.2).
(2) **CGCAFusion:** This sub-module is responsible for dynamically modeling contextual relationships based on the input feature content, thereby enhancing struc-

tural segmentation performance, particularly in cluttered and visually complex environments (Section 2.3).

(3) **Depth Estimation Module (DEM):** The objective of this sub-module is to predict relative depth information directly from RGB inputs, while incorporating spatial context to support accurate alignment estimation (Section 2.4).

(4) **Alignment Estimation Module (AEM):** This sub-module computes the offset of the detected doorway relative to the image center, enabling the generation of precise directional guidance for the controller (Section 2.5).

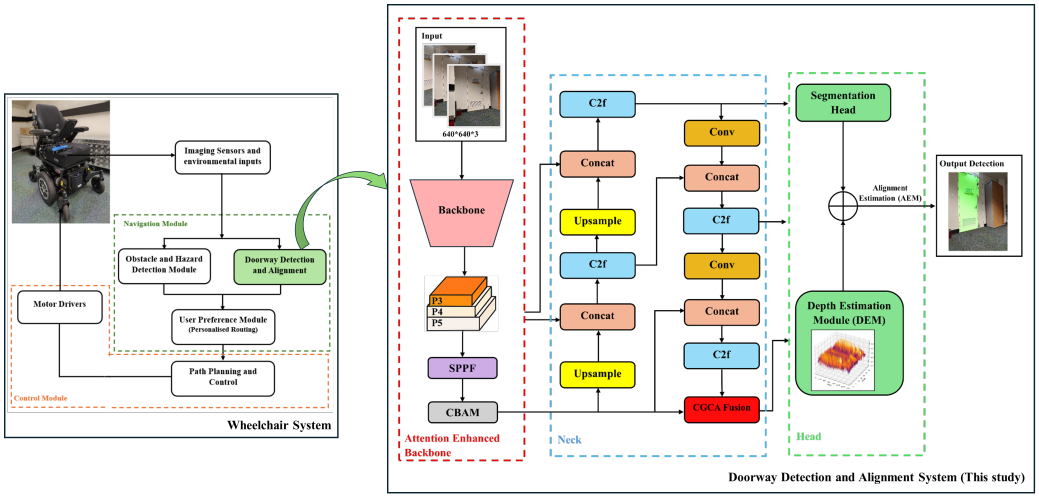


Figure 1. Architecture of the proposed doorway detection and alignment guidance algorithm.

2.1. Dataset Description

The datasets used in this study are designed to support semantic segmentation and object detection of door structures in diverse environments and conditions. A detailed description of the datasets is listed in Table 1.

Table 1. Dataset description.

Attribute	Description
Dataset source	DeepDoorsv2 [45] + Custom RGB images
Total images	3100
Resolution	480 × 640 pixels
Door states	Closed, Semi-open, Open
Door types	Single, Double, Sliding, Glass
Conditions	Diverse lighting, occlusions, angles, and layouts
Segmentation pixels	Door/Frame: (192, 224, 192); Background: (0, 0, 0)
Data split	7:1.5:1.5 (train:val:test)

The choice of these datasets was motivated by the need for a standardized and widely accepted benchmark to ensure a fair comparison with existing algorithms.

2.2. CBAM

Standard detectors often struggle to highlight structural cues unique to doorways, leading to the misclassification of visually similar features [23,41,42]. To address this limitation, we incorporate a lightweight attention mechanism that enhances the ability of the network to focus on spatially and semantically relevant information, that is, the CBAM, a widely adopted technique introduced by Woo et al. in 2018 [43]. CBAM refines feature representations by sequentially applying channel and spatial attention, allowing the network to selectively emphasize informative regions. In our model, CBAM is integrated

immediately after the SPPF (Spatial Pyramid Pooling–Fast) layer at the end of the YOLOv8 backbone to strengthen attention to doorway-specific features (Figure 2). The detailed subcomponents of CBAM are shown in Figure 3, where the attention operations refine feature quality before fusion.

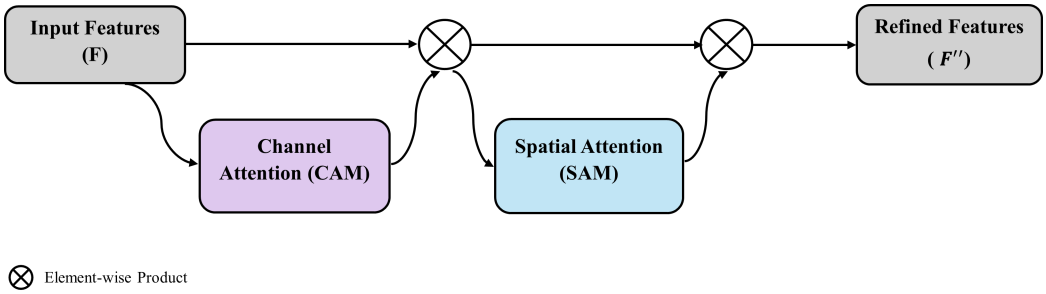


Figure 2. Overall architecture of the CBAM model.

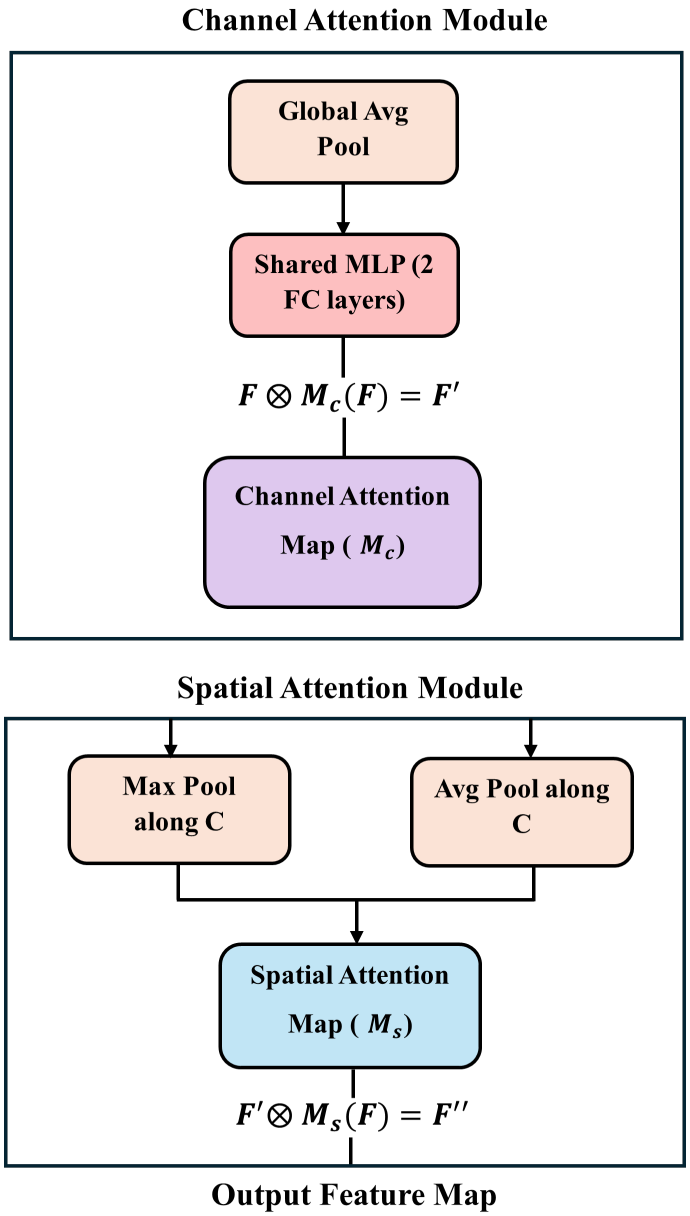


Figure 3. Channel attention module and spatial attention module of CBAM.

Let $F \in \mathbb{R}^{C\times H\times W}$ be the input feature map output from SPPF, where C represents the channel dimension and $H \times W$ indicates spatial resolution. CBAM refines this feature map through two sequential sub-modules.

Firstly, the Channel Attention Module (CAM) infers channel-wise importance using both global average pooling (GAP) and global max pooling (GMP), followed by a shared multi-layer perceptron (MLP) [46] (Equations (1) and (2)).

$$M_c(F) = \sigma[\mathrm{MLP}(\mathrm{GAP}(F)) + \mathrm{MLP}(\mathrm{GMP}(F))] \tag{1}$$

$$F' = M_c(F) \otimes F \tag{2}$$

where $M_c(F) \in \mathbb{R}^{C\times 1\times 1}$, σ denotes the sigmoid function, and $\otimes$ denotes element-wise multiplication.

Next, the spatial attention module generates a spatial attention map based on the average and max projections across the channel dimension (Equations (3) and (4)).

$$M_s(F) = \sigma\Big(f^{7\times 7}([\mathrm{AvgPool}(F); \mathrm{MaxPool}(F)])\Big) \tag{3}$$

$$F'' = M_s(F) \otimes F' \tag{4}$$

where $M_s(F) \in \mathbb{R}^{1\times H\times W}$ and $f^{7\times 7}$ represents a convolutional kernel.

The final refined feature map F'' is then passed for further multi-scale fusion. By inserting CBAM at this stage, we allow the network to selectively emphasize features that are structurally and semantically relevant to door detection, improving segmentation accuracy while maintaining computational efficiency.

2.3. CGCAFusion

To address the limitations of conventional fusion methods [47–50] in detecting structurally small and complex objects such as doorframes, we adopt the CGCAFusion module [44]. This is a content-guided convolutional attention mechanism designed to enhance both local and global feature modeling with low computational overhead. This mechanism is particularly beneficial for real-time doorway detection in resource-constrained environments.

This module adopts a two-branch fusion strategy designed to enhance multi-scale feature integration (see Figure 4). The first branch, Content-Guided Attention (CGA), applies a coarse-to-fine spatial attention mechanism that refines each feature channel by learning saliency patterns within channels. The second branch, the Convolutional Attention Fusion Module (CAFM) [51], incorporates a simplified transformer-inspired structure that captures global dependencies through self-attention while retaining local context using depthwise convolutions. The model achieves multimodal feature fusion by dynamically adjusting attention weights according to the input content, enabling fine-grained focus on doorway edges and suppressing irrelevant features.

As illustrated in Figure 4, given two input tensors $F'' \in \mathbb{R}^{C\times H\times W}$ (output from CBAM), their corresponding elements are first aggregated (Equation (5)).

$$F_1 = F'' \oplus F'' \tag{5}$$

where C represents the channel dimension, $H \times W$ indicates spatial resolution, and $\oplus$ denotes element-wise addition.

The combined feature F_1 is then passed through the CAFM module (M_{CAFM}) to capture global and local relationships (Equation (6)).

$$F_2 = F_1 \oplus M_{CAFM}(F_1) \tag{6}$$

The output F_2 is refined by the CGA module (M_{CGA}), which applies space-, channel-, and pixel-level attention to further emphasize critical regions (Equation (7)).

$$F_3 = \sigma[M_{CGA}(F_1, F_2)] \tag{7}$$

The fusion process concludes with the fusion module (M_{Fusion}), which performs weighted summation and dimensionality alignment via 1 × 1 convolution, yielding the final fused output (Equation (8)).

$$F_{out} = Conv[M_{Fusion}(F_1 \oplus F_2 \oplus F_3)] \tag{8}$$

This output effectively balances global semantic context with local detail, ensuring the enhanced segmentation of door structures even under challenging lighting or cluttered backgrounds.

Simultaneously, the pixel attention module dynamically modulates feature importance (low-level and high-level features), guiding the model to concentrate on the most relevant regions within the image. This refined fusion strategy (see Figure 5) substantially enhances the model's accuracy and efficiency in real-time detection and segmentation applications (Equation (9)).

$$F_{fusion} = Conv[M_{CAFM}(F_{low} + F_{high}) \cdot w + M_{CAFM}(F_{low} + F_{high}) \cdot (1 - w) + F_{low} + F_{high}] \tag{9}$$

where F_{low} denotes low-level features, F_{high} denotes high-level features, and w denotes feature weights.

This fusion module improves the ability of the model to capture both fine-grained and global contextual features, making it well suited for detecting doorways in complex environments with varying structural layouts.

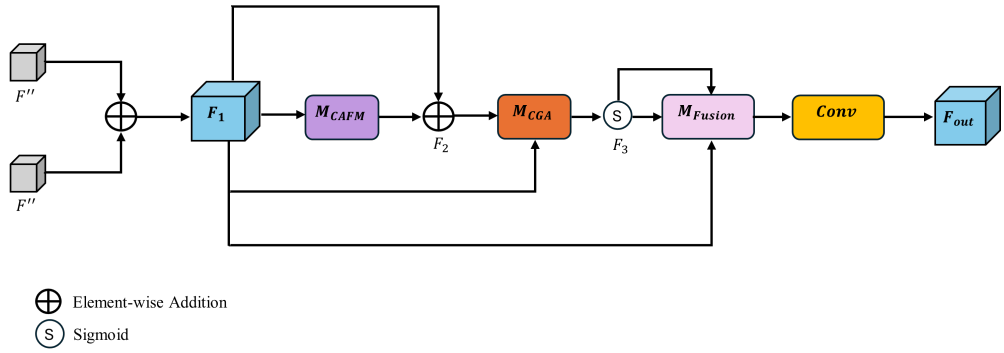


Figure 4. Architecture of the CGCAFusion [44].

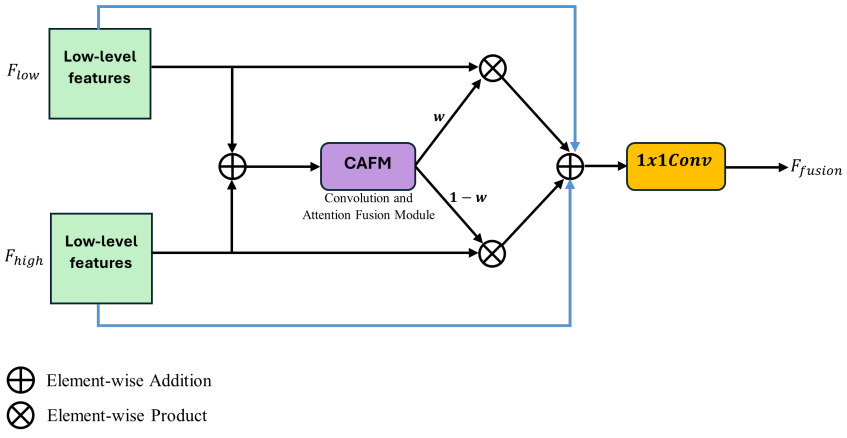


Figure 5. Architecture of the fusion module [44].

2.4. DEM

One of the primary limitations in many existing doorway detection systems is their reliance on external depth sensors such as LiDAR to estimate object proximity. Even though these hardware-based solutions are effective, they increase system cost, complexity, and power consumption, making them less suitable for lightweight mobile platforms such as assistive wheelchairs.

To overcome the limitations of external depth sensors to estimate doorway proximity, we incorporate a lightweight DEM that learns to predict relative depth directly from monocular RGB input by unsupervised learning. The module leverages the spatial and semantic features produced by CBAM and CGCAFusion to determine depth. The DEM takes as input the fused feature tensor $F_{fusion} \in \mathbb{R}^{C \times H \times W}$ generated by the CGCAFusion module, which integrates multi-scale context. This tensor is processed by a shallow convolutional decoder composed of two stacked convolutional layers, batch normalization, and sigmoid activation (Equation (10)).

$$D = \sigma[\text{Conv}_{3\times3}[\text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(F_{\text{fusion}})))]] \tag{10}$$

The output $D \in \mathbb{R}^{1 \times H \times W}$ is a normalized depth map that captures relative distances within the scene (Figure 6) to capture the geometric layout of the environment.

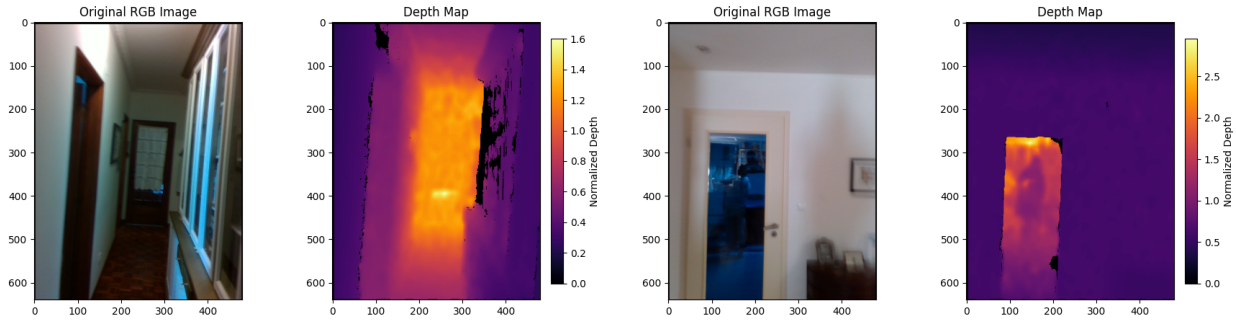


Figure 6. Depth maps generated by the DEM module. (**a**) Original RGB image (**Left**) and predicted normalized depth map (**Right**) of doorway. (**b**) Original RGB image (**Left**) and predicted normalized depth map (**Right**) of an open door.

2.5. AEM

One of the key challenges in autonomous wheelchair navigation is not only detecting the presence of doors, but also assessing their relative position and orientation to ensure safe navigation. Traditional vision-based methods tend to stop at detection, lacking the capability to provide directional guidance, such as adjusting to the left or right for optimal alignment. This limitation can result in collisions or failed attempts to navigate through narrow doorways. To address this challenge, we propose an AEM that operates on the output of the segmentation mask to determine the horizontal offset between the doorway center and the center of the camera frame. This offset is used to issue directional guidance to the wheelchair controller to either "Move Left", "Move Right", or "Aligned", based on a predefined tolerance margin.

Let x_f denote the horizontal center of the frame and x_d the horizontal center of the detected doorway (bounding box center). The offset is defined as follows:

$$\Delta x = x_f - x_d$$

A threshold θ is used to decide alignment as follows:

$$\text{Guidance} = \begin{cases} \text{Aligned}, & |\Delta x| < \theta \\ \text{Move Left}, & \Delta x > \theta \\ \text{Move Right}, & \Delta x < -\theta \end{cases}$$

The output of the AEM provides interpretable visual guidance cues to the control system. By integrating AEM into our pipeline, we bridge the gap between static doorway detection and actionable mobility decisions, enabling intelligent guidance that is crucial for real-world assistive navigation.

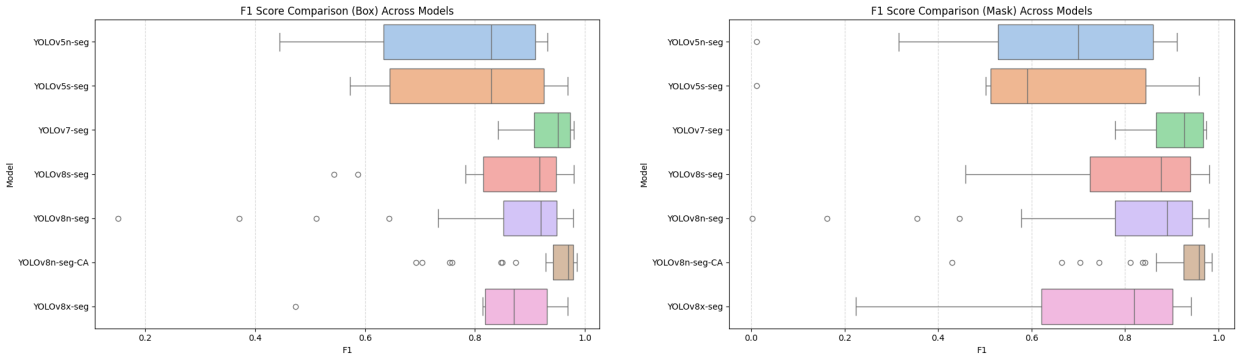
3. Results and Discussion

In this section, we discuss the results obtained by assessing the constructed model in the following four main areas: structural attention enhancement through the CBAM (Section 3.1), improved precision and semantic fusion using CGCAFusion (Section 3.2), performance of unsupervised DEM (Section 3.3), and finally, the evaluation of the model's ability to provide accurate alignment estimation for intelligent guidance (Section 3.4). All qualitative results were obtained using real-world environments and working conditions to assess the generalization of the model, while quantitative evaluations were performed on the datasets.

3.1. Structural Attention Enhancement Through CBAM

In this section, we discuss the tests performed to evaluate the performance of the CBAM in improving structural attention. We compared its performance with the following six state-of-the-art deep learning segmentation approaches: YOLOv5n-seg [52], YOLOv5s-seg [52], YOLOv7-seg [53], YOLOv8n-seg [54], YOLOv8s-seg [54], and YOLOv8x-seg [54].

To evaluate the impact of CBAM on the focus of structural characteristics, we focused on two tests to evaluate its performance. Firstly, we compared the F1 scores of the outputs of the bounding box and the segmentation masks (Figure 7). Secondly, we present qualitative results (Figure 8) obtained in real working conditions and environments to illustrate how the CBAM enhances spatial focus by suppressing false positives compared to the baseline model.



(**a**) (**b**)

Figure 7. The proposed model (YOLOv8n-seg-CA) demonstrates consistently high F1 scores with reduced variability compared to baseline methods: (**a**) F1 score comparison (bounding box) across different segmentation models. (**b**) F1 score comparison (mask) across different segmentation models.

The box plots of the F1 score comparisons (Figure 7) clearly demonstrate that, in both the bounding box and segmentation masks, the proposed model (YOLOv8n-seg-CA) consistently outperforms other deep learning models in both accuracy and stability.

In particular, YOLOv8n-seg-CA achieves a higher median F1 score and exhibits tighter interquartile ranges, suggesting greater reliability under varied image conditions. Compared to YOLOv8n-seg, which demonstrates a wider dispersion and more outliers, the attention-enhanced version yields more concentrated and stable predictions.

Moreover, the presence of fewer outliers in YOLOv8n-seg-CA indicates robustness to extreme cases, such as partial occlusion or low contrast boundaries. This improvement is critical in doorway detection for wheelchair navigation, where false positives, such as misclassifying walls or furniture as doors, can compromise navigational safety.

The results of the qualitative analysis (Figure 8) in real-world environments illustrate how the CBAM enhances spatial focus by suppressing false positives. In particular, Figure 8a–c demonstrate three sample cases of different types of doors where baseline models misclassified flat wall surfaces and furniture as doors, which was correctly resolved in our model. This highlights its improved ability to capture the contextual and geometric signals of doorway structures.

Overall, these findings validate the effectiveness of CBAM in enabling the network to focus on meaningful structural cues such as doorway frames, edges, and contours, thus improving both detection precision and generalization.

3.2. Improved Precision and Semantic Fusion Using CGCAFusion

In this section, to validate the effectiveness of the proposed CGCAFusion module, we evaluate its performance in terms of both detection accuracy and computational efficiency. We compared its performance with the following seven state-of-the-art segmentation approaches: Mask R-CNN [55], YOLOv5n-seg [52], YOLOv5s-seg [52], YOLOv7-seg [53], YOLOv8n-seg [54], YOLOv8s-seg [54], and YOLOv8x-seg [54]. The hyperparameter settings of the training process are outlined in Table 2.

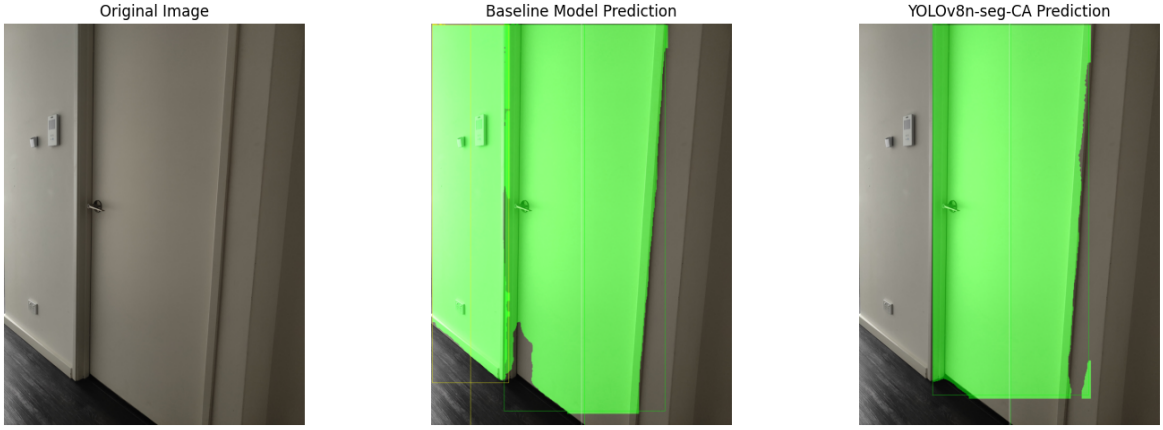
The performance of the model was evaluated using main indicators such as the mean average precision mAP (Bounding Box), parameters, FPS (Frames Per Second), model size, and inference time, as shown in Table 3.

- *Mean Average Precision (mAP50)* quantifies detection accuracy by averaging precision of all detections. A higher mAP50 indicates more accurate localization and classification of objects.
- *Params (M)* refers to the total count of learnable weights (in Millions) within the model, reflecting its computational complexity. Models with fewer parameters are more suitable for resource-constrained environments.
- *Frames Per Second (FPS)* measures the real-time processing capability of the model. The model exhibits higher FPS, signifying that it can process video frames more rapidly, which is critical for responsive assistive navigation.
- *Model Size (MB)* represents the storage footprint of the trained model. Smaller models are more efficient for deployment on embedded hardware.
- *Inference Time (ms)* denotes the time required to process a single frame. Lower inference times are essential for real-time operation in time-sensitive applications.

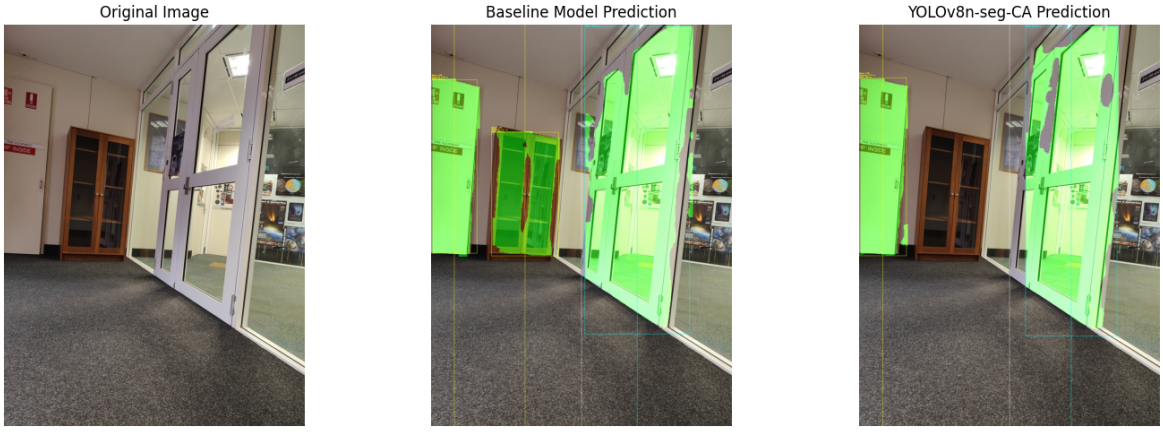
While YOLOv7-seg achieves a competitive mAP50 of 95.3%, its higher parameter count and larger model size make it less suitable for deployment on edge devices. In contrast, the proposed model is built upon the YOLOv8n-seg baseline, specifically optimized for lightweight deployment. Through the integration of CBAM, CGCAFusion, and a DEM, the proposed model achieves a superior mAP50 of 95.8% while maintaining only 2.96 M parameters and a compact 3.6 MB footprint. Furthermore, it delivers an inference time of just 0.42 ms. These results demonstrate that the proposed model offers an optimal balance of accuracy, efficiency, and speed, making it well suited for real-time deployment in embedded assistive systems, a key direction for future implementation.



(**a**) Case 1: Original RGB image input, segmentation prediction from the baseline model (misclassifications), and prediction of proposed model.



(**b**) Case 2: Original RGB image input, segmentation prediction from the baseline model (misclassifications), and prediction of proposed model.



(**c**) Case 3: Original RGB image input, segmentation prediction from the baseline model (misclassifications), and prediction of proposed model.

Figure 8. Qualitative comparison of doorway segmentation between the original image, the baseline model, and the YOLOv8n-seg-CA model in real working conditions. The enhanced model demonstrates improved boundary accuracy and reduces false positives: (**a**) Case 1 comparison. (**b**) Case 2 comparison. (**c**) Case 3 comparison.

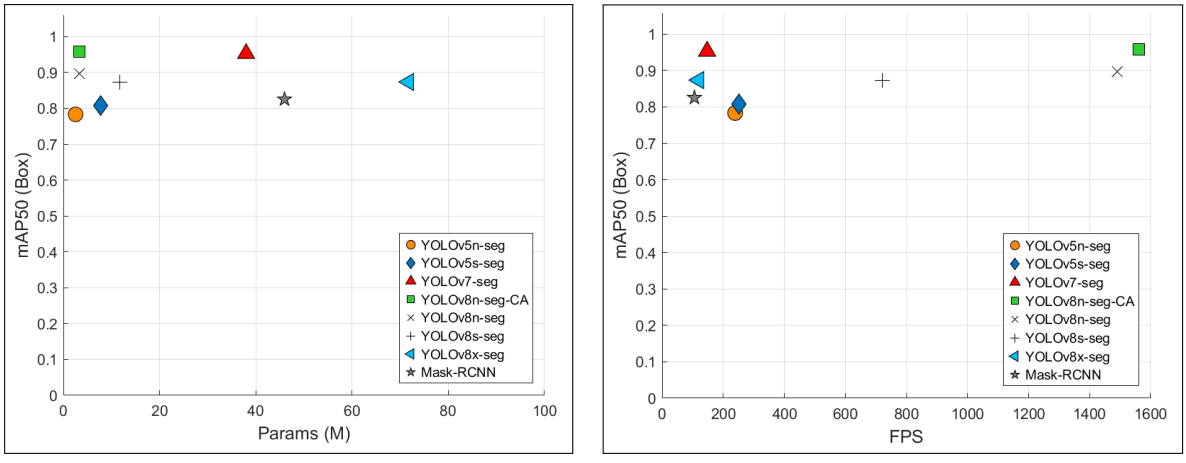
Table 2. Hyperparameter settings.

Parameters	Value
Epochs	150
Ir0	0.002
Irf	0.002
Momentum	0.9
Batchsize	16
Cache	False
Input image size	640 × 640
Optimizer	AdamW

Table 3. Quantitative comparison of segmentation models using metrics, mAP50 Box, params, FPS, model size, and inference time.

Model	mAP50 (Bounding Box)	Params (M)	FPS	Model Size (MB)	Inference Time (ms)
Mask R-CNN	0.825	45.96	105.7	346.52	9.33
YOLOv5n-seg	0.783	2.53	239	5.14	4.62
YOLOv5s-seg	0.808	7.74	252.1	15.6	4.26
YOLOv7-seg	0.953	37.98	147.23	78.1	6.9
YOLOv8n-seg	0.896	3.26	1490	6.45	0.64
YOLOv8s-seg	0.872	11.79	720	22.73	1.39
YOLOv8x-seg	0.873	71.75	120	137.26	8.62
Proposed Model	**0.958**	**2.96**	**1560**	**3.6**	**0.42**

To further illustrate the trade-offs between accuracy and model complexity, Figure 9a presents a marker plot of the mean average precision (mAP50) vs. parameter count, and Figure 9b demonstrates mAP50 vs. FPS across models. These plots demonstrate that the CGCAFusion module enables more accurate segmentation at a minimal cost to model size or speed. Even though larger models such as YOLOv8x-seg exhibit marginally higher accuracy, they do so at a significant cost in inference speed and memory usage. In contrast, our model remains well suited for deployment in embedded devices without compromising detection quality.



(**a**) (**b**)

Figure 9. Cross comparison of computational metrics of different segmentation models: (**a**) Accuracy vs. parameters. (**b**) Accuracy vs. frame rate.

Overall, it is evident that by fusing contextual information from low- and high-resolution feature maps with a global attention mechanism, CGCAFusion effectively en-

hances structural awareness while preserving real-time performance, which is a critical requirement for doorway detection in assistive applications such as wheelchair navigation.

3.3. Performance of Unsupervised DEM

In this section, we evaluate the effectiveness of the proposed DEM, which generates relative depth maps from RGB images without additional hardware or supervision. Unlike conventional sensor-based methods, our approach operates in a fully unsupervised manner using only monocular visual inputs, making it lightweight and cost-effective for deployment on assistive platforms.

Figure 10 presents the distribution of absolute depth estimation errors across the test set with a mean absolute error (MAE) of 0.69 m. The error histogram exhibits a clear right-skewed profile with a sharp peak in the 0–0.5 m range, indicating that most predicted depth values closely align with the ground truth values. While the distribution shows a long tail that extends to higher error values, these represent only a small fraction of the total predictions.

To assess the practical utility of the proposed DEM, we conducted a comparative analysis against the following two widely used monocular depth estimation models: Mixed Datasets for Monocular Depth Estimation (MiDaS) [56] and Dense Prediction Transformer (DPT) [57], as shown in Table 4.

Table 4. Comparison of DEM with MiDaS and DPT in terms of computational performance and depth estimation accuracy.

Model	Size (MB)	Inference Time (ms)	Memory Usage (MB per FPS)	MAE
MiDaS	100	0.017	1.7	3.19
DPT	350	0.046	16	0.08
Proposed DEM	<3.6	0.026	0.1	0.69

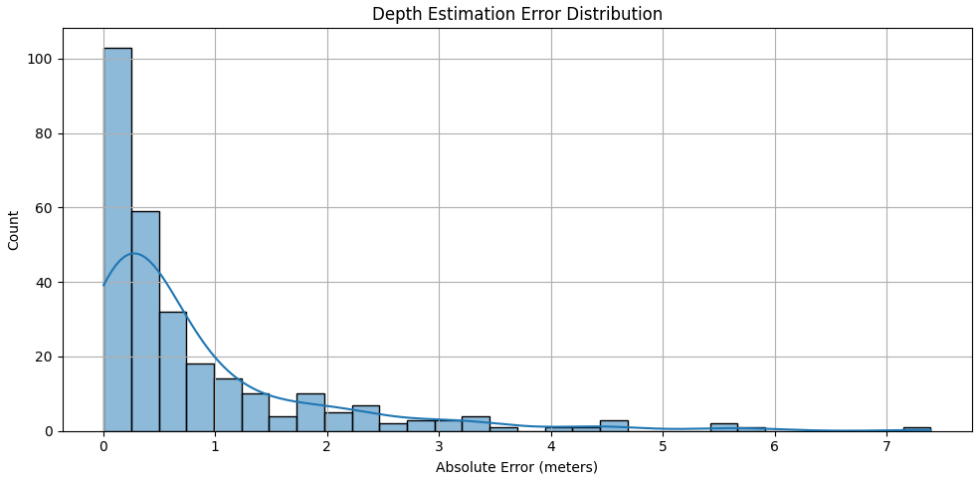


Figure 10. Distribution of absolute depth estimation errors produced by the proposed DEM. The histogram shows that most predictions fall within a low error range (0–0.5 m), with a right-skewed tail indicating occasional higher error outliers

Although DPT achieves the lowest MAE of 0.08, this comes at a high computational cost, with a model size of 350 MB and memory usage of 0.6 MB per FPS. While MiDaS performs better than the proposed DEM in terms of speed, it suffers from a high error (MAE) of accurate depth estimation.

In contrast, the proposed DEM demonstrates a favorable balance between accuracy and efficiency. It achieves an MAE of 0.69, while maintaining a minimal size of 3.6 MB,

an inference time of 0.026 s, and the lowest memory footprint. Furthermore, it operates with just 0.1 MB per frame, making it highly suitable for future edge deployment.

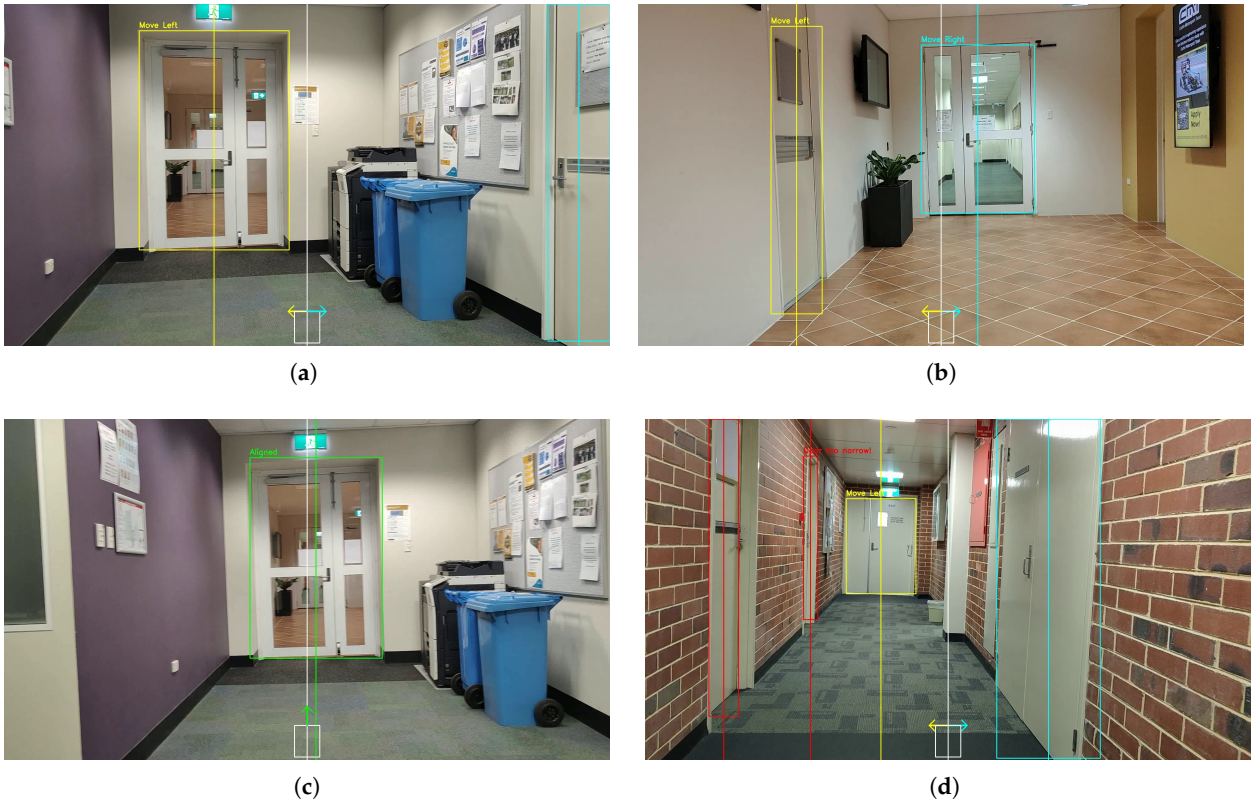
These results confirm that the proposed DEM delivers a compelling trade-off between accuracy and efficiency. While it is not as precise as DPT, it is significantly more compact and computationally efficient. This makes it ideal for integration into embedded systems and real-time assistive navigation platforms, where hardware resources are limited.

Overall, the DEM demonstrates sufficient precision for guiding alignment decisions and estimating doorway proximity. It forms a foundational module in the ongoing development of a unified assistive wheelchair navigation system and will be further validated in real-world edge deployments.

3.4. Accurate Alignment Estimation for Intelligent Guidance

In this section, we evaluate the performance of the proposed AEM in real-world environments, which is designed to provide interpretable navigation cues to the control system. This is achieved by estimating the positional alignment of detected doorways relative to the center of the frame. This module enables the system to make actionable decisions such as "Move Left", "Move Right", or "Aligned", which are essential to guide an assistive wheelchair safely and efficiently through constrained spaces.

To validate the effectiveness of AEM, we present four real-world visualizations that demonstrate the directional decision making of the system in varied indoor environments. As shown in Figure 11, the module provides discrete alignment decisions, "Move Left", "Move Right", "Aligned", or "Door Too Narrow", based on the relative location and geometry of detected doorways.



(**a**) (**b**)

(**c**) (**d**)

Figure 11. Visual outputs of the proposed AEM across different indoor navigation scenarios. The module provides directional guidance (**a**) "Move Left", (**b**) "Move Left/Right" with multiple doors, (**c**) alignment confirmation ("Aligned"), and (**d**) passability checks ("Door too narrow!").

Figure 11a shows an off-center door with respect to the center of the frame. The AEM identifies this misalignment and suggests a "Move Left" adjustment to center the wheelchair. It is evident that the decision is consistent with the actual geometric layout and demonstrates the spatial awareness of the model.

Figure 11b illustrates a dual-door scenario where two distinct doorways are detected simultaneously. The AEM correctly evaluates their position with respect to the center of the frame, issuing a "Move Right" instruction for the centrally located door and "Move Left" for the side door. This confirms the module's ability to disambiguate between multiple doors and provide appropriate direction guidance for each.

Figure 11c demonstrates an ideal alignment scenario where the detected door is well centered in the frame and the AEM outputs an "Aligned" message, indicating that no directional adjustment is necessary.

In contrast, Figure 11d represents detections of multiple doors in a complex hallway. A narrow door on the left is marked with the warning "Door Too Narrow" due to its insufficient width to allow safe passage. The larger exit door is detected correctly and the module recommends "Move Left" to better center the user.

Overall, the evaluation confirms the robustness of the AEM in issuing interpretable and context-sensitive guidance across varied geometries and lighting conditions. By incorporating spatial thresholds for both alignment and navigability, the module transforms detection into actionable feedback, which is critical for autonomous wheelchair assistance.

In summary, the combination of quantitative metrics and qualitative, real-world analyses confirms the robustness and generalization of our proposed model in diverse indoor environments. The integration of attention mechanisms, semantic fusion, depth estimation, and alignment evaluation enables accurate and interpretable doorway detection, while providing real-time directional guidance. These findings have strong implications for assistive navigation, particularly in enhancing autonomy and safety for wheelchair users by enabling context-aware environmental understanding and decision making.

4. Conclusions

In this research, we present the YOLOv8n-seg-CA model as the doorway detection and alignment module of a larger unified vision-based navigation system currently being developed for assistive wheelchairs. This model is designed to operate solely on RGB input and integrates the following four key modules: CBAM for refined spatial and channel-wise feature attention, the CGCAFusion module for multi-scale semantic feature fusion, a lightweight depth estimation module for unsupervised monocular depth prediction, and a doorway alignment estimation module that provides real-time directional feedback to the control system. Together, these modules improve the ability of the model to interpret complex indoor environments and deliver guidance in real-time, without the need for additional hardware such as LiDAR or depth sensors.

To assess the effectiveness of the algorithm, we conducted extensive evaluations that compared our model with existing baselines. The results showed that YOLOv8n-seg-CA achieves higher segmentation accuracy, lower inference complexity, and enhanced spatial awareness. Qualitative evaluations in real-world indoor settings further demonstrated the robustness of the algorithm in detecting doorways and assessing safe passage. The DEM effectively predicted depth in the absence of ground truth, while the AEM translated spatial cues into actionable instructions for the wheelchair controller, making the model particularly suited for unfamiliar or dynamically changing environments.

This vision-based approach offers a practical and scalable solution for doorway detection and alignment for assistive technologies. By eliminating the need for heavy or expensive sensor arrays, the proposed doorway detection algorithm supports lightweight

deployment on mobile robotic platforms and smart wheelchairs. Its minimal hardware footprint promotes broader accessibility, especially in low-resource healthcare and rehabilitation settings.

As this study focuses on the algorithmic development of a doorway detection and alignment guidance module, future work will involve its integration into a complete assistive navigation system. This includes incorporating dynamic obstacle detection, user preference modeling, path planning, and tailored control strategies for wheelchair users. The unified system will be implemented and validated on embedded hardware platforms such as NVIDIA Jetson, ensuring compliance with real-time performance and power constraints. Further investigations will include pilot trials in real-world environments (e.g., hospitals, homes, and aged care facilities) in collaboration with disability service providers to assess usability, safety, and user trust. Comprehensive testing against medical device safety standards will also be conducted as a part of on going research. These steps will ensure that the developed system meets the practical requirements for safe, robust, and personalized navigation support, ultimately advancing mobility and autonomy for individuals with physical impairments.

Author Contributions: Conceptualization, S.T., N.W., A.W., N.A. and I.M.; methodology, S.T.; software, S.T.; validation, S.T.; formal analysis, S.T.; resources, S.T.; writing—original draft preparation, S.T.; writing—review and editing, N.W., A.W., N.A. and I.M.; supervision, N.W., A.W., N.A. and I.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this study is publicly available at https://github.com/gasparramoa/DeepDoors2 accessed on 20 January 2025.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

YOLO	You Only Look Once
DSPP	DenseNet Spatial Pyramid Pooling
SPPF	Spatial Pyramid Pooling-Fast
CBAM	Convolutional Block Attention Module
CGCAFusion	Content-Guided Convolutional Attention Fusion Module
DEM	Depth Estimation Module
AEM	Alignment Estimation Module
CAM	Channel Attention Module
GAP	Global Average Pooling
GMP	Global Max Pooling
MLP	Multi-Layer Perceptron
SAM	Spatial Attention Module
CGA	Content-Guided Attention
CAFM	Convolutional Attention Fusion Module
mAP	Mean Average Precision
MAE	Mean Absolute Error
MiDaS	Mixed Datasets for Monocular Depth Estimation
DPT	Dense Prediction Transformer

References

1. Dickinson, L. Autonomy and motivation a literature review. *System* **1995**, *23*, 165–174. [CrossRef]
2. Atkinson, J. Autonomy and mental health. In *Ethical Issues in Mental Health*; Springer: Berlin/Heidelberg, Germany, 1991; pp. 103–126.

3. Mayo, N.E.; Mate, K.K.V. Quantifying Mobility in Quality of Life. In *Quantifying Quality of Life: Incorporating Daily Life into Medicine*; Wac, K., Wulfovich, S., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 119–136. [CrossRef]
4. Meijering, L. Towards meaningful mobility: A research agenda for movement within and between places in later life. *Ageing Soc.* **2021**, *41*, 711–723. [CrossRef]
5. World Health Organization (WHO); United Nations Children's Fund (UNICEF). Global Report on Assistive Technology. Licence: CC BY-NC-SA 3.0 IGO. 2022. Available online: https://www.who.int/publications/i/item/9789240074521 (accessed on 10 July 2025).
6. Marsden, J.F. Spasticity. *Rheumatol. Rehabil.* **2016**, *2*, 197–206.
7. Elias, W.J.; Shah, B.B. Tremor. *JAMA* **2014**, *311*, 948–954. [CrossRef] [PubMed]
8. MacDonald, A.E. General paresis. *Am. J. Psychiatry* **1877**, *33*, 451–482. [CrossRef]
9. Cooper, R.A.; Cooper, R.; Boninger, M.L. Trends and issues in wheelchair technologies. *Assist. Technol.* **2008**, *20*, 61–72. [CrossRef] [PubMed]
10. Tsao, C.C.; Mirbagheri, M.R. Upper limb impairments associated with spasticity in neurological disorders. *J. NeuroEng. Rehabil.* **2007**, *4*, 45. [CrossRef] [PubMed]
11. Rizzo, J.R.; Beheshti, M.; Hudson, T.E.; Mongkolwat, P.; Riewpaiboon, W.; Seiple, W.; Ogedegbe, O.G.; Vedanthan, R. The global crisis of visual impairment: An emerging global health priority requiring urgent action. *Disabil. Rehabil. Assist. Technol.* **2023**, *18*, 240–245. [CrossRef] [PubMed]
12. Pascolini, D.; Mariotti, S.P. Global estimates of visual impairment: 2010. *Br. J. Ophthalmol.* **2012**, *96*, 614–618. [CrossRef] [PubMed]
13. Chang, K.y.J.; Rogers, K.; Lung, T.; Shih, S.; Huang-Lung, J.; Keay, L. Population-based projection of vision-related disability in australia 2020–2060: Prevalence, causes, associated factors and demand for orientation and mobility services. *Ophthalmic Epidemiol.* **2021**, *28*, 516–525. [CrossRef] [PubMed]
14. Stevens, G.A.; White, R.A.; Flaxman, S.R.; Price, H.; Jonas, J.B.; Keeffe, J.; Leasher, J.; Naidoo, K.; Pesudovs, K.; Resnikoff, S.; et al. Global prevalence of vision impairment and blindness: Magnitude and temporal trends, 1990–2010. *Ophthalmology* **2013**, *120*, 2377–2384. [CrossRef] [PubMed]
15. Kim, E.Y. Wheelchair navigation system for disabled and elderly people. *Sensors* **2016**, *16*, 1806. [CrossRef] [PubMed]
16. Sanders, D.; Tewkesbury, G.; Stott, I.J.; Robinson, D. Simple expert systems to improve an ultrasonic sensor-system for a tele-operated mobile-robot. *Sens. Rev.* **2011**, *31*, 246–260. [CrossRef]
17. Zheng, T.; Duan, Z.; Wang, J.; Lu, G.; Li, S.; Yu, Z. Research on distance transform and neural network lidar information sampling classification-based semantic segmentation of 2d indoor room maps. *Sensors* **2021**, *21*, 1365. [CrossRef] [PubMed]
18. Gallo, V.; Shallari, I.; Carratù, M.; Laino, V.; Liguori, C. Design and Characterization of a Powered Wheelchair Autonomous Guidance System. *Sensors* **2024**, *24*, 1581. [CrossRef] [PubMed]
19. Perra, C.; Kumar, A.; Losito, M.; Pirino, P.; Moradpour, M.; Gatto, G. Monitoring Indoor People Presence in Buildings Using Low-Cost Infrared Sensor Array in Doorways. *Sensors* **2021**, *21*, 4062. [CrossRef] [PubMed]
20. Grewal, H.; Matthews, A.; Tea, R.; George, K. LIDAR-based autonomous wheelchair. In Proceedings of the 2017 IEEE Sensors Applications Symposium (SAS), Glassboro, NJ, USA, 13–15 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
21. Durrant-Whyte, H.; Bailey, T. Simultaneous localization and mapping: Part I. *IEEE Robot. Autom. Mag.* **2006**, *13*, 99–110. [CrossRef]
22. Sahoo, S.; Choudhury, B. Voice-activated wheelchair: An affordable solution for individuals with physical disabilities. *Manag. Sci. Lett.* **2023**, *13*, 175–192. [CrossRef]
23. Sahoo, S.K.; Choudhury, B.B. Autonomous navigation and obstacle avoidance in smart robotic wheelchairs. *J. Decis. Anal. Intell. Comput.* **2024**, *4*, 47–66. [CrossRef]
24. Ess, A.; Schindler, K.; Leibe, B.; Van Gool, L. Object detection and tracking for autonomous navigation in dynamic environments. *Int. J. Robot. Res.* **2010**, *29*, 1707–1725. [CrossRef]
25. Qiu, Z.; Lu, Y.; Qiu, Z. Review of ultrasonic ranging methods and their current challenges. *Micromachines* **2022**, *13*, 520. [CrossRef] [PubMed]
26. Derry, M.; Argall, B. Automated doorway detection for assistive shared-control wheelchairs. In Proceedings of the 2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, 6–10 May 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 1254–1259.
27. Rusu, R.B.; Marton, Z.C.; Blodow, N.; Holzbach, A.; Beetz, M. Model-based and learned semantic object labeling in 3D point cloud maps of kitchen environments. In Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 10–15 October 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 3601–3608.
28. Anguelov, D.; Koller, D.; Parker, E.; Thrun, S. Detecting and modeling doors with mobile robots. In Proceedings of the IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04, New Orleans, LA, USA, 26 April–1 May 2004; IEEE: Piscataway, NJ, USA, 2004; Volume 4, pp. 3777–3784.

29. Lecrosnier, L.; Khemmar, R.; Ragot, N.; Decoux, B.; Rossi, R.; Kefi, N.; Ertaud, J.Y. Deep learning-based object detection, localisation and tracking for smart wheelchair healthcare mobility. *Int. J. Environ. Res. Public Health* **2021**, *18*, 91. [CrossRef] [PubMed]
30. Ju, M.; Luo, H.; Wang, Z.; Hui, B.; Chang, Z. The application of improved YOLO V3 in multi-scale target detection. *Appl. Sci.* **2019**, *9*, 3775. [CrossRef]
31. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 3464–3468.
32. Zhang, T.; Li, J.; Jiang, Y.; Zeng, M.; Pang, M. Position detection of doors and windows based on dspp-yolo. *Appl. Sci.* **2022**, *12*, 10770. [CrossRef]
33. Iandola, F.; Moskewicz, M.; Karayev, S.; Girshick, R.; Darrell, T.; Keutzer, K. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv* **2014**, arXiv:1404.1869.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
35. Mochurad, L.; Hladun, Y. Neural network-based algorithm for door handle recognition using RGBD cameras. *Sci. Rep.* **2024**, *14*, 15759. [CrossRef] [PubMed]
36. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
37. Hussain, M. YOLOv5, YOLOv8 and YOLOv10: The Go-To Detectors for Real-time Vision. *arXiv* **2024**, arXiv:2407.02988.
38. Wei, L.; Tong, Y. Enhanced-YOLOv8: A new small target detection model. *Digit. Signal Process.* **2024**, *153*, 104611. [CrossRef]
39. Sharma, P.; Tyagi, R.; Dubey, P. Bridging the Perception Gap A YOLO V8 Powered Object Detection System for Enhanced Mobility of Visually Impaired Individuals. In Proceedings of the 2024 First International Conference on Technological Innovations and Advance Computing (TIACOMP), Bali, Indonesia, 29–30 June 2024; pp. 107–117. [CrossRef]
40. Choi, E.; Dinh, T.A.; Choi, M. Enhancing Driving Safety of Personal Mobility Vehicles Using On-Board Technologies. *Appl. Sci.* **2025**, *15*, 1534. [CrossRef]
41. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
42. Tennekoon, S.; Wedasingha, N.; Welhenge, A.; Abhayasinghe, N.; Murray Am, I. Advancing Object Detection: A Narrative Review of Evolving Techniques and Their Navigation Applications. *IEEE Access* **2025**, *13*, 50534–50555. [CrossRef]
43. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference On Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
44. Zhang, Z.; Zou, Y.; Tan, Y.; Zhou, C. YOLOv8-seg-CP: A lightweight instance segmentation algorithm for chip pad based on improved YOLOv8-seg model. *Sci. Rep.* **2024**, *14*, 27716. [CrossRef] [PubMed]
45. Ramôa, J.; Lopes, V.; Alexandre, L.; Mogo, S. Real-time 2D–3D door detection and state classification on a low-power device. *SN Appl. Sci.* **2021**, *3*, 590. [CrossRef] [PubMed]
46. Kruse, R.; Mostaghim, S.; Borgelt, C.; Braune, C.; Steinbrecher, M. Multi-layer perceptrons. In *Computational Intelligence: A Methodological Introduction*; Springer: Cham, Switzerland, 2022; pp. 53–124.
47. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
48. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
49. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W. Biformer: Vision transformer with bi-level routing attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10323–10333.
50. Xu, W.; Wan, Y. ELA: Efficient local attention for deep convolutional neural networks. *arXiv* **2024**, arXiv:2403.01123. [CrossRef]
51. Hu, S.; Gao, F.; Zhou, X.; Dong, J.; Du, Q. Hybrid convolutional and attention network for hyperspectral image denoising. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5504005. [CrossRef]
52. Jocher, G. *YOLOv5 by Ultralytics*; Zenodo: Geneva, Switzerland, 2020. [CrossRef]
53. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
54. Jocher, G.; Qiu, J.; Chaurasia, A. Ultralytics YOLO. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 28 December 2024).
55. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.

56. Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1623–1637. [CrossRef] [PubMed]
57. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12179–12188.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.