Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.063985

ARTICLE

Zero-Shot Based Spatial AI Algorithm for Up-to-Date 3D Vision Map Generations in Highly Complex Indoor Environments

Sehun Lee, Taehoon Kim and Junho Ahn*

AI Computer Engineering Department, Kyonggi University, Suwon, 16227, Republic of Korea

*Corresponding Author: Junho Ahn. Email: jha@kyonggi.ac.kr

Received: 31 January 2025; Accepted: 21 May 2025; Published: 03 July 2025

ABSTRACT: This paper proposes a zero-shot based spatial recognition AI algorithm by fusing and developing multi-dimensional vision identification technology adapted to the situation in large indoor and underground spaces. With the expansion of large shopping malls and underground urban spaces (UUS), there is an increasing need for new technologies that can quickly identify complex indoor structures and changes such as relocation, remodeling, and construction for the safety and management of citizens through the provision of the up-to-date indoor 3D site maps. The proposed algorithm utilizes data collected by an unmanned robot to create a 3D site map of the up-to-date indoor site and recognizes complex indoor spaces based on zero-shot learning. This research specifically addresses two major challenges: the difficulty of detecting walls and floors due to complex patterns and the difficulty of spatial perception due to unknown obstacles. The proposed algorithm addresses the limitations of the existing foundation model, detects floors and obstacles without expensive sensors, and improves the accuracy of spatial recognition by combining floor detection, vanishing point detection, and fusion obstacle detection algorithms. The experimental results show that the algorithm effectively detects the floor and obstacles in various indoor environments, with F1 scores of 0.96 and 0.93 in the floor detection and obstacle detection experiments, respectively.

KEYWORDS: Spatial AI; vision; foundation model; zero-shot learning; image segmentation

1 Introduction

The global market for large shopping malls is expected to grow from $6.23 trillion in 2024 to $9.86 trillion in 2032, and the demand for UUS is also expected to grow rapidly, with a compound annual growth rate (CAGR) of 11.8% from 2024 to 2032. According to a survey of AI-based customer traffic analysis platforms, 45% of shopping malls are expected to be developed for mixed use by 2030. Precise 3D map generation technology is essential for efficient management and sustainable operation of these large indoor and underground spaces. In particular, discrepancies between the actual indoor structure and the design drawings frequently occur due to relocation, remodeling, and construction, and regular spatial recognition and updates are required.

Some related studies suggest that a solution to the issue of discrepancies between the design drawings at the time of construction and the actual indoor structure is research on real-time data collection using unmanned robots [1,2]. Some studies attempt to solve the problem of structural inconsistency by generating an indoor 3D map based on the collected video data [3–5]. In addition, existing indoor spatial recognition technologies mainly use expensive sensors such as light detection and ranging (LiDAR) and red, green, blue-depth (RGB-D) cameras to collect data and use finetuned models to detect indoor structures [6–8]. However,

these methods have several limitations, including high costs, limited performance in specific environments and obstacle types, and difficulty in detecting complex patterns and sparse obstacles.

The emergence of the foundation model [9] has presented a general-purpose approach that can be applied to various environments, but technical limitations such as noise problems caused by complex textures and patterns, difficulty in detecting sparse obstacles, and performance degradation due to environmental changes remain unresolved. In particular, obstacle detection is an important element in spatial recognition, but it is a difficult task to learn and detect all obstacles in various indoor spaces. Against this backdrop, it is necessary to develop a zero-shot based spatial recognition algorithm that can be applied to complex and changeable indoor environments and detect unlearned obstacles. This study aims to overcome these limitations by developing an algorithm that reduces reliance on expensive sensors and enables the generation of precise 3D maps in various indoor environments.

In order to develop the proposed zero-shot indoor spatial recognition algorithm, two primary challenges have been established. The first challenge involves mitigating the noise introduced by complex textures and patterns on walls and floors, a factor that significantly impedes accurate detection. The second challenge arises from the detection of "sparse obstacles," which, unlike fully obstructive wall-like barriers, possess void spaces that allow partial visibility—examples include steel frames, chairs, and desks. Through preliminary analyses, it was determined that these sparse or partially incomplete obstacles are especially difficult to detect using traditional obstacle detection approaches, particularly when situated on floor surfaces with complex patterns. Therefore, the overarching objective of this study is to detect both conventional obstacles and these unknown, sparse obstacles with high precision, thereby ensuring broad versatility and practicality even in structurally complex indoor environments.

This study follows a three-stage research process to develop a general-purpose, zero-shot indoor spatial-recognition algorithm capable of generating precise 3D maps in complex indoor environments. First, we review existing research on foundation models and spatial artificial intelligence to identify current limitations and technological gaps; these insights provide the theoretical basis for algorithm development. Second, we propose an integrated algorithm composed of three modules. The floor-detection module fuses a segmentation algorithm with a depth-estimation algorithm to suppress noise from complex patterns and textures, thereby achieving robust floor detection. The vanishing-point-detection module leverages image-processing techniques and depth data, employing a dynamic weighted pyramid for accurate vanishing-point estimation. The obstacle-detection module combines gap-based, dispersion-based, and depth-based methods to detect both sparse and atypical obstacles with high precision. Third, we experimentally verify the effectiveness of the proposed algorithm. To this end, we collected data with a smartphone camera mounted on an entry-level robot and compared the performance of the proposed algorithm with that of a baseline method in a complex indoor environment.

The study presents a new suggestion for indoor spatial recognition technology that can resolve the limitations of existing zero-shot based foundation model technology and expand its commercialization potential. This provides a practical and universal solution that can be used in a variety of environments that require regular spatial recognition and precise 3D map generation. Moreover, previous learning-based approaches and algorithms, which aim to detect a wide variety of objects, often suffer from substantially reduced accuracy or frequent misclassifications when employed in environments that lack sufficient training data or differ significantly from trained datasets. To address these shortcomings, this paper introduces an algorithm built on a zero-shot based spatial AI framework capable of providing foundational object information in advance. First, a highly general model prioritizes the detection of spaces and objects; then, learning-based methods refine these initial detection outcomes with additional predictions, reducing false

positives and improving overall accuracy. This approach is designed to overcome the limitations of earlier learning-based solutions and broaden their applicability in real-world scenarios.

2 Related Works

2.1 Foundation Model

Foundation model is an AI model trained with a large dataset and is used in various studies in the field of computer vision, including classification, detection, segmentation, and spatial recognition. The foundation model, which was trained using zero-shot learning, shows performance similar or higher than that of fully supervised learning AI models. Currently, as the field of computer vision based on the foundation model is being actively researched, there is research that defines new evaluation methods and challenges [10]. Additionally, there is research on learning large datasets and zero-shot object segmentation in images and videos based on user interaction [11,12]. There is a specialized study on foundation model-based object tracking to solve the object occlusion and object id switch problems that occur when objects are divided [13]. This study predicts the next motion of the object and performs precise object tracking through the refinement process of the segmented mask. In addition to the ViT-based model, there is also research using the CNN-based Foundation Model with transformers [14]. This study proposes a new CNN-based foundation model that improves performance as the training dataset increases, similar to the ViT model. Foundation model has been studied not only in the field of vision but also in the convergence of computer vision and language, and the number of studies showing various uses is increasing [15–17]. This research enables foundation model, which was previously limited to the field of computer vision, to perform various tasks such as visual question answering (VQA) [18], image captioning, behavior recognition, and video search through various prompts. Moreover, there is research that proposes a new benchmark for evaluating the constructive ability of the foundation model in the field of vision-language [19]. Furthermore, there is a study that conducted learning to align vision-language expressions through dual-encoder-based contrastive learning [20]. The utilization of the computer vision and vision-language-based foundation model is also active in various industrial fields. There is a study that developed spatial reasoning on the surface of a planet by using the foundation model in space robotics [21]. There is also a study on improving the performance of the foundation model in object classification, detection, and segmentation in the field of remote sensing (RS) [22]. The study demonstrates superior performance through a new attention technique that reduces computational cost and memory usage.

2.2 Interactive Image Segmentation

Recently, in the field of referring image segmentation (RIS), there has been an active effort to enhance object segmentation capabilities by combining natural language sentences with visual information, leveraging complex context and reasoning processes. Table 1 summarizes key methods in this domain, including their datasets and performance metrics. Early studies simply fused features from CNNs and RNNs in the later stages (e.g., LSCM, EFN), or used transformer-based multimodal decoders to integrate linguistic and visual information at a delayed stage (e.g., VLT). However, these approaches were limited in enabling sufficient interaction between multi-scale visual features and language representations. To overcome these limitations, recent methods such as LAVT, PolyFormer, and GLaMM have proposed early fusion strategies that integrate language information within the transformer architecture or employed autoregressive approaches that predict segmentation masks sequentially, achieving improvements in representation learning and segmentation accuracy. Moreover, models like PixelLM, LISA, PSALM, and MCN have introduced large language models (LLMs) to leverage rich reasoning capabilities from text, significantly enhancing performance on benchmarks such as RefCOCO [23] and RefCOCO+. These efforts have expanded to unified models capable

of addressing panoptic and interactive segmentation tasks simultaneously. Further, studies such as u-LLaVA, UniLSeg, and UniRef++ have explored methods to adapt across various granularities and multi-task environments, while EVF-SAM integrates a text encoder with the Segment Anything Model (SAM), greatly extending the multimodal perception range to include interactive segmentation. Meanwhile, approaches like SimpleClick have minimized interaction for rapid fine-grained segmentation, and methods like MCN, LSCM, and VLT have actively exploited syntactic and tree-structured information in text to achieve more precise alignment between linguistic and visual modalities. Overall, these trends indicate a clear movement toward solving all segmentation tasks related to RIS within a unified framework by leveraging large-scale multimodal models, early fusion transformer techniques, and diverse interactive modules. Nevertheless, this study identifies the following three limitations in existing approaches and algorithms:

Table 1: Literature survey

Ref.	Year	Method	Dataset	Best Acc.
Hui et al. [24]	2020	LSCM	RefCOCOg	48.0
Luo et al. [25]	2020	MCN	RefCOCOg	49.4
Ding et al. [26]	2021	VLT	RefCOCOg	56.6
Feng et al. [27]	2021	EFN	RefCOCOg	51.9
Yang et al. [28]	2022	LAVT	RefCOCOg	62.1
Liu et al. [29]	2023	SimpleClick	SBD	4.15 NoC@90
Xu et al. [30]	2023	u-LLaVA-7B	RefCOCOg	77.97
Liu et al. [31]	2024	UniLSeg-100	RefCOCOg	80.54
Lai et al. [32]	2025	LISA-7B	RefCOCOg	70.6
Liu et al. [33]	2023	PolyFormer-L	RefCOCOg	71.17
Zhang et al. [34]	2024	PSALM	RefCOCOg	74.4
Wu et al. [35]	2023	UniRef++L	RefCOCOg	72.84
Zhang et al. [36]	2024	EVF-SAM	RefCOCOg	77.4
Rasheed et al. [37]	2024	GLaMM	RefCOCOg	74.9
Ren et al. [38]	2024	PixelLM	RefCOCOg	70.5

First, prior studies show degraded accuracy when reasoning and detecting floor regions with high complexity. In real-world indoor environments, floors often exhibit physical irregularities and complex patterns of mixed colors and textures, making it difficult for single models or algorithms to process them effectively. Consequently, existing methods frequently suffer from false positives, mistakenly recognizing non-obstacle floor areas as obstacles, thus revealing a critical limitation. Second, prior works also exhibit reduced detection accuracy for "sparse obstacles," as defined in the Introduction of this paper. Sparse obstacles differ from typical fully-enclosed obstacles in that they possess partially open structures, requiring fine-grained segmentation. However, it remains challenging to accurately detect such complex obstacles solely with learning-based models, highlighting limitations in achieving both high precision and high accuracy. Third, previous methods generally achieve strong performance by assuming specific prompts or click-based coordinates for reasoning and detecting floor and obstacle regions. However, in our study, when provided with unspecified, broad-range prompts or ambiguous boundary clicks, existing methods fail to robustly detect obstacles that interfere with driving or navigation. This reveals the inherent limitations in the generalizability of approaches reliant on specific prompts or click information.

2.3 Spatial Artificial Intelligence

A lot of research is being conducted in the field of indoor space recognition, including 3D mapping, indoor autonomous robot, and indoor monitoring. First, for research on indoor space recognition and mapping, a study has proposed a method that combines floor and ceiling detection for corridor segmentation and evaluates guideline consistency through boundary intersection-based vanishing point analysis [39]. There is a study that developed an initial segmentation based on Fuzzy C-Means (FCM) and a precise segmentation technique using CNN to automatically identify design elements in a floor plan image [40]. A framework that includes extracting building structure planes and creating 3D linear models using point cloud data has been proposed, and research has verified that it can be used for indoor robots and emergency response systems [41]. A study on 3D modeling and object detection through spatial recognition is the study that developed a style transfer technique for mesh reconstruction of indoor scenes [42]. Additionally, there is research on constructing indoor structures through spatial recognition using RGB-D-based image segmentation techniques [43,44]. Also, a voxel-based reconstruction method has been proposed that automatically generates semantically rich indoor 3D models from unstructured triangular meshes, and research has shown that it can be applied to complex indoor structures [45]. An example of indoor navigation and route estimation research is the research conducted by LEXIS, a real-time indoor SLAM system that uses LLMs [46]. V-Eye, a navigation system for the visually impaired that detects moving obstacles and provides accurate location and direction information, has been developed, and there is research that has proven its spatial awareness and walking safety [47]. There is a study that proposed a hybrid mapping method for constructing a dense 3D representation of a large indoor space in a general CPU environment [48]. One study has demonstrated that a digital twin framework combining BIM, IoT, and autonomous robots can be developed for facility management to automate indoor disaster response and rescue operations [49]. There is a study that proposes a method for generating a digital map of a disaster site that uses low-cost robots and AI algorithms to track collapse situations and detect obstacles [50]. There is a study that proposes a method of detecting and estimating the location of rescue targets and obstacles by applying an AI algorithm that combines a camera and 3D LiDAR [51].

Existing spatial recognition studies have limitations in that they only detect pre-trained objects or use expensive sensors. In addition, there is a limitation that it does not show high accuracy for detecting complex obstacles such as sparse obstacles. In this study, the zero-shot based Foundation Model can detect objects that have not been learned in advance, which is effective for recognizing and understanding various indoor spaces. It also has the advantage of being able to recognize space without using high-cost sensors such as LiDAR or RGB-D cameras. However, the existing foundation model showed a problem in which the noise caused by the complex textures and patterns on the floor and walls interfered with spatial perception, and the detection performance was reduced due to the large and small empty spaces in the obstacle area acting as noise when detecting sparse obstacles. To address this problem, the study proposes a spatial recognition algorithm that combines image processing techniques, depth estimation, a multi-scale pyramid, and a dynamic weighted pyramid structure. The proposed algorithm aims to increase the universality of various indoor spaces and improve the accuracy of spatial recognition.

3 Proposed Algorithms

In this study, we developed a zero-shot based indoor spatial recognition algorithm that enables the generation of precise 3D maps even in complex indoor environments. The proposed algorithm is designed by fusing three main modules for floor detection, vanishing point detection, and obstacle detection, which ensures high precision and versatility in various environments. The overall algorithm design and flow are shown in Figs. 1 and 2.

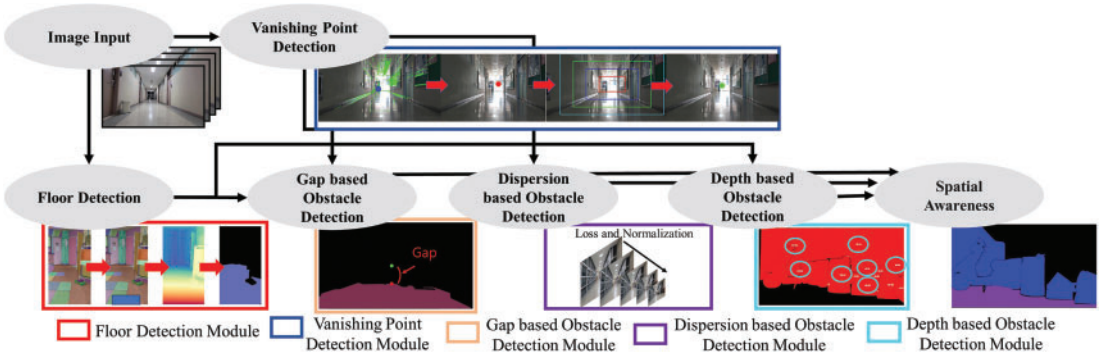

Figure 1: Proposed algorithm for the developed indoor spatial recognition

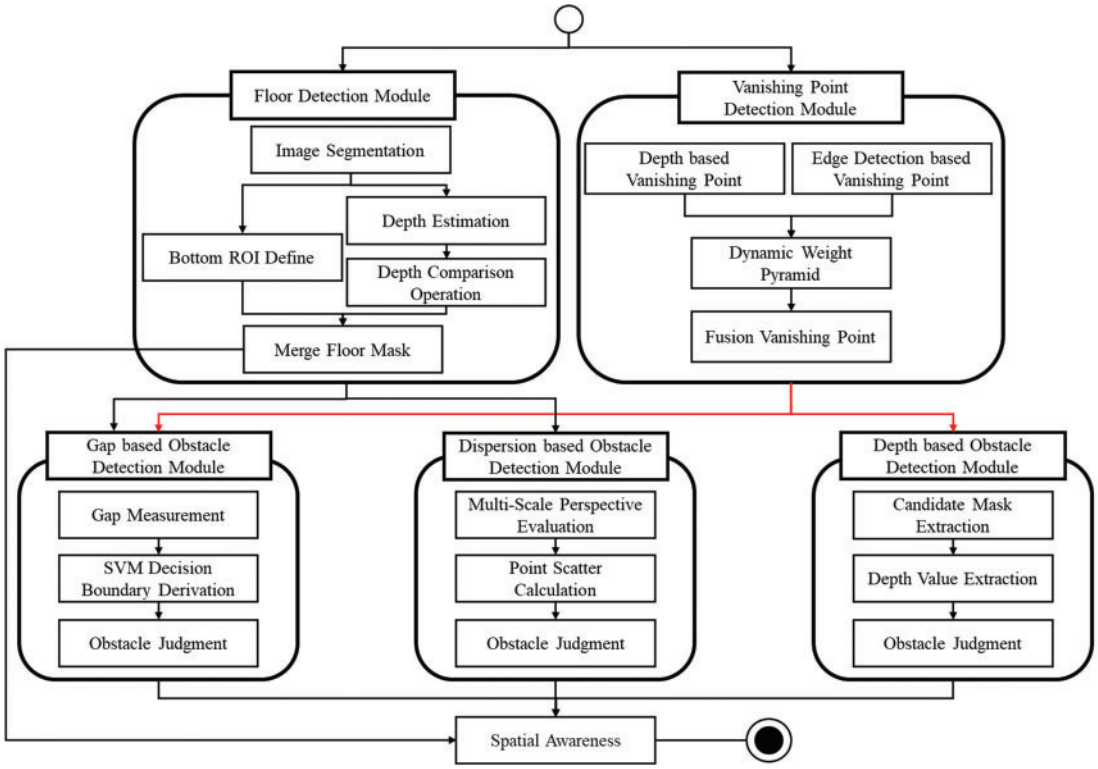

Figure 2: Flowchart of the proposed algorithm

Fig. 1 shows the overall architecture of the proposed indoor spatial recognition algorithm. First, images are obtained from a monocular RGB camera, followed by precise detection of the vanishing point and floor mask from the images. Next, the regions identified as the floor mask are excluded, and candidate areas for potential obstacles are extracted. These candidate areas are then processed by three detection modules, which selectively identify obstacles that interfere with navigation. The first module employs the previously detected vanishing point information to identify obstacles. The second module utilizes the floor mask data to analyze height differences or boundary information with respect to the floor, thereby assessing whether an object constitutes an obstacle. Lastly, the third module integrates depth information derived from a depth estimation algorithm to more accurately detect obstacles physically present in the robot's route. The outputs

of these three modules are ultimately fused to minimize false positive results and further enhance obstacle detection in indoor spaces.

3.1 Framework Overview

The proposed algorithm is designed to achieve high-precision and universal spatial recognition in indoor environments, and its main steps consist of floor detection, vanishing point detection, and obstacle detection. Fig. 2 visualizes the main steps of these algorithms in a flowchart, clearly showing the relationship between each step and the data flow.

First, the RGB video data collected is applied to the floor detection module to extract the floor area that can be passed through. To maximize the floor detection efficiency, the bottom region of interest (ROI) is defined, and in this process, the ROI focuses on the key areas of the analysis, such as the floor and nearby obstacles, to reduce unnecessary calculations and improve processing speed. Then, the image segmentation is performed using the segmentation technique, and the depth data of the segmented mask is calculated to extract the depth information of the space. Based on in-depth information, it removes the noise caused by complex patterns and extracts a merged single floor mask, providing important data for assessing whether it is passable. Second, the vanishing point detection module applies a dynamic weighted pyramid to dynamically assign different weights to each vanishing point in each image, thereby extracting improved fused vanishing points. These fused vanishing points contribute to increasing the precision of spatial directional analysis and obstacle detection. In the final stage, three detection modules based on gap, dispersion, and depth are integrated to make a comprehensive determination of whether there is an obstacle. The gap-based detection module analyzes the distance between the bottom mask and the vanishing point through a linear classification model [52] to distinguish between the cases where there is no obstacle and the cases where there is an obstacle, while the dispersion-based detection module analyzes the pattern in which points are dispersed due to obstacles and uses dispersion values to determine whether there are obstacles. The linear classification model employed in this study is a soft margin-based linear SVM, which allows for a certain degree of misclassification in order to find the optimal decision boundary even when the data is not perfectly separable. The depth-based detection module identifies obstacles that can be hit using depth data and supplements the results of the gap and dispersion-based modules. By fusing the results of the three detection modules mentioned above, the final result of whether the passage is available and the spatial recognition results can be derived, thereby providing reliable spatial information even in complex indoor environments.

3.2 Data Collection and Recognition

Recognizing complex indoor spaces is a key element in generating accurate site maps, which can effectively deal with situations where access is difficult. Fig. 3 shows the system structure for collecting spatial data in various indoor spaces and generating an indoor site map based on that data. This system uses mobile robots to collect spatial information in real time, and the collected data is transmitted to spatial recognition algorithms via the cloud. The transmitted data is analyzed by distinguishing between passable corridors and obstacles and is used to quickly identify and analyze indoor spaces, thereby providing precise spatial recognition data for the effective and regular generation of 3D indoor maps.

3.3 Floor Detection Module

Segmented floor masks play an important role in checking whether the hallway is passable. The algorithm proposed in this study developed the floor detection process shown in Fig. 4 to create a single merged floor mask. To this end, the floor mask was extracted using foundation model-based segmentation,

and in this process, the process of merging the segmented floors due to the pattern and pattern noise of the object was carried out. The merger process is as follows. First, set the ROI at the bottom of the image as a square area. It calculates the depth map [53] of the image and measures the similarity of the depth value between adjacent masks based on the defined ROI. If the similarity is high, the corresponding masks are merged to create a single unified floor mask. This mask clearly distinguishes between corridors that are open to traffic and corridors that are not. Fig. 4 shows the final merged mask results, clearly showing the difference between the mask patterns for when traffic is not possible and when it is possible. As a result, discontinuous and abnormal mask patterns were observed in corridors where passage was impossible due to obstacles, while continuous and uninterrupted mask patterns were observed in corridors where passage was possible. Based on this, it was possible to determine whether it was possible to pass through based on the floor surface, and it can be used as a basis for analyzing the presence of obstacles.

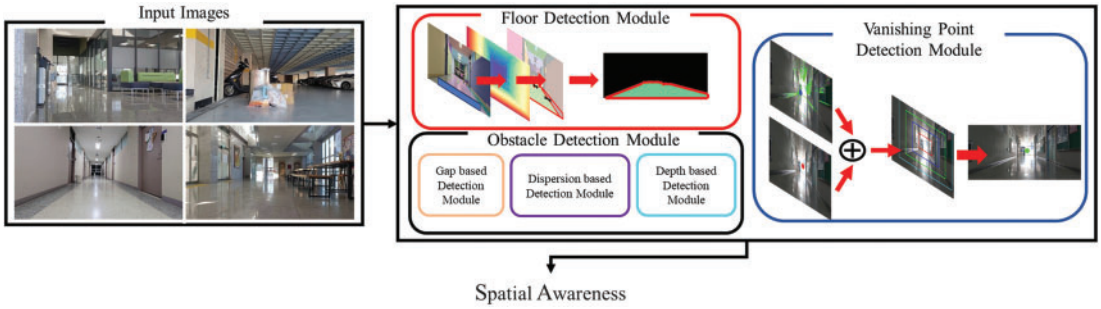

Figure 3: Robot-based data collection and spatial recognition system architecture

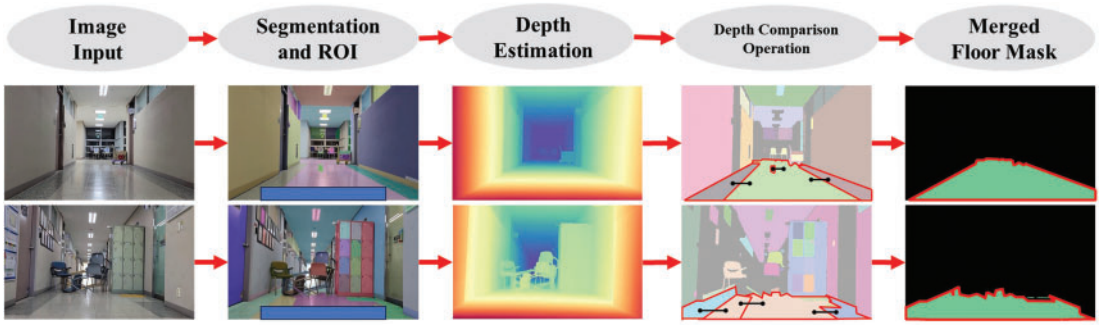

Figure 4: Process of detecting situations in general corridors and corridors blocked by obstacles through floor masks

3.4 Vanishing Point Detection Module

Previous object detection algorithms can detect commonly predefined obstacles such as chairs, desks, and boxes, but they have the limitation of not detecting complex obstacles that have not been pre-trained, such as those that appear in indoor disaster situations or complex indoor environments. The mask that detects without a label through the foundation model also cannot verify whether it is a real physical obstacle or a 2D object like the picture. We recognized the limitations of this existing obstacle detection and developed a fusion vanishing point detection module to improve it.

In this study, the fused vanishing point is estimated using a dynamic weighted pyramid. In the case of edge detection [54] based on vanishing point detection, it is inevitable that the appropriate parameter modification according to each indoor environment is difficult. Depth-based vanishing point detection has

difficulty in consistently predicting vanishing points in various environments. As a result of the experiment to find the exact vanishing point, it was confirmed that the vanishing point detection based on edge detection is accurate in the case of a general pattern without obstacles, and that the vanishing point detection based on depth is accurate in the case of a pattern with obstacles in front. This algorithm estimates the fusion vanishing point in an image through an algorithm that estimates the optimal vanishing point through two vanishing point estimations. Two separate initial vanishing point estimations are made: one method uses perspective to find a single point where the continuously extending edges come together based on the data that detects the edges of the input original image, and the other method estimates the depth value of each pixel of the input original image and estimates the vanishing point as the coordinate with the most depth. The edge-based vanishing point coordinates are defined as $(\boldsymbol{V_x}, \boldsymbol{V_y})$, and the depth-based vanishing point coordinates $(\boldsymbol{C_x}, \boldsymbol{C_y})$. Create an extended area based on depth-based vanishing points with high individual performance and create a dynamic weighted pyramid that sets the weighting ratio. First, for the red area in the center shown in Fig. 5, the weight ratio was designed as edge detection:depth = 5:5, the blue area as edge detection: depth = 4:6, the green area as edge detection:depth = 3:7, and the light blue area as edge detection:depth = 2:8. The overall sequence of the algorithm is shown in Fig. 6. Algorithm 1 is the pseudocode for the process of dynamically assigning weights between two points of depth and edge detection-based vanishing points as designed in advance according to the region where the edge detection-based vanishing point exists. The optimal fusion vanishing point is estimated by the weights of the two dynamically assigned points for the dynamic weighted pyramid region where the $(\boldsymbol{V_x}, \boldsymbol{V_y})$ coordinates exist. The fusion vanishing point is used to design the obstacle detection fusion algorithm that will be introduced later.

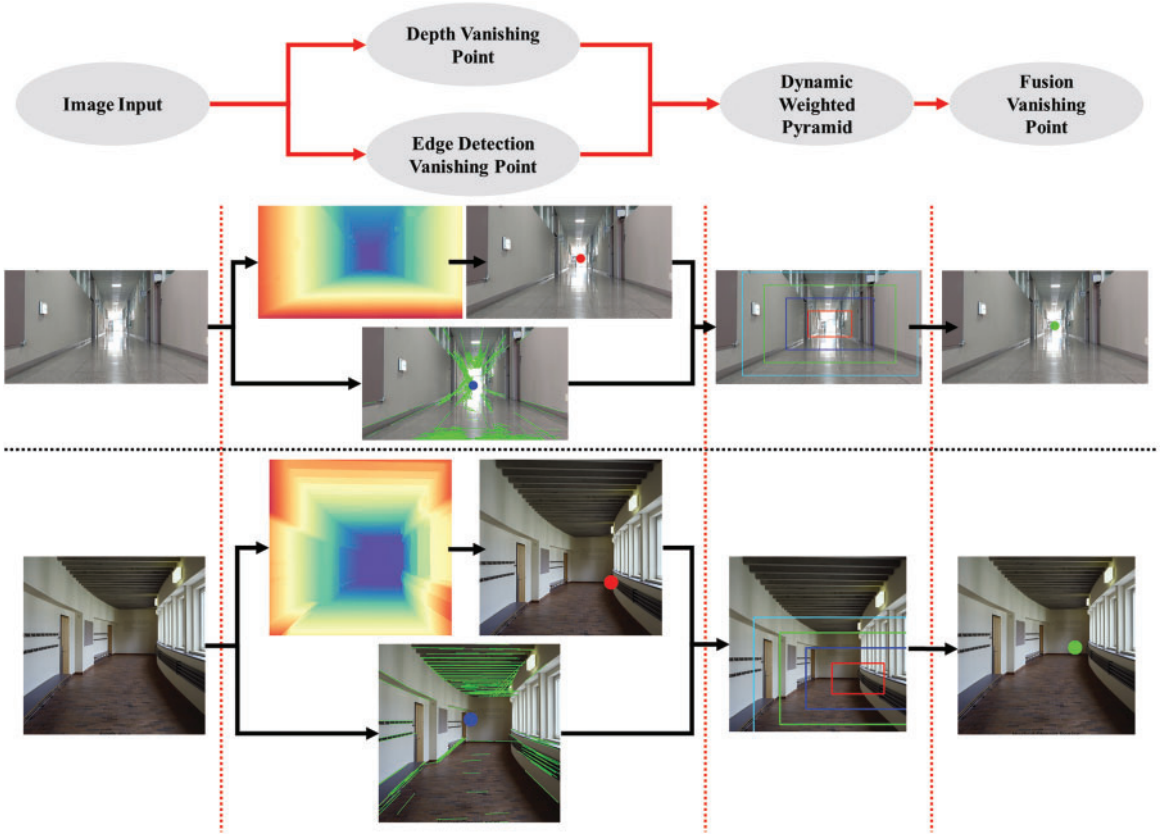

Figure 5: Depth and edge detection fusion vanishing point detection process

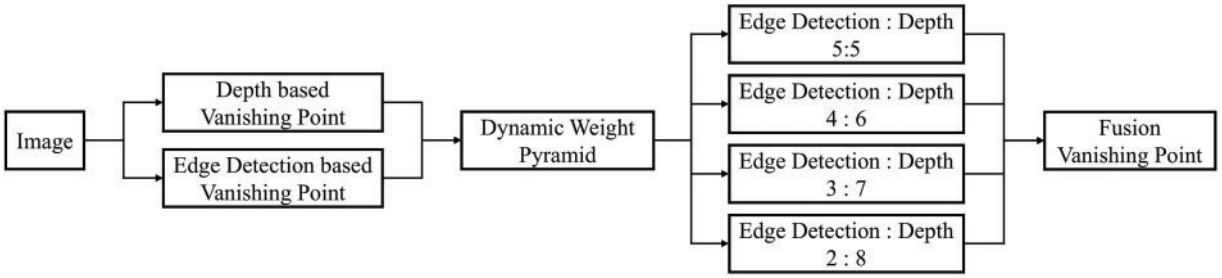

Figure 6: Depth and edge detection fusion vanishing detection process details

Algorithm 1: Dynamic weight assignment

for i, ((top_left_x, top_left_y), (bottom_right_ x, top_right_ y)) ← enumerate(regions_coordinates) **do**
  if top_left_$x \le vx \le$ bottom_right_ $x \wedge$ top_left_$y \le vy \le$ top_right_ y **then**
    region_found ← **true**
    if $i = 0$ **then**
      weight_v, weight_c ← 5, 5
    else if $i = 1$ **then**
      weight_v, weight_c ← 4, 6
    else if $i = 2$ **then**
      weight_v, weight_c ← 3, 7
    else if $i = 3$ **then**
      weight_v, weight_c ← 2, 8
    end if
    break
  end if
end for

3.5 Obstacle Detection Module

Obstacle recognition is important in spatial recognition because it is a key factor in determining whether a path is passable. However, in indoor environments, it is necessary to learn and detect various types of obstacles, especially to distinguish between sparse obstacles or 2D images such as pictures. To cope with these complex situations, a universal and efficient obstacle detection method is required. To solve this problem, we developed a fusion algorithm that can reduce the cost of learning and increase the detection rate of obstacles. The overall flow of the fusion algorithm is shown in Fig. 7. The modules that were developed in a fusion to detect obstacles include the gap-based obstacle detection module, the dispersion-based obstacle detection module, and the depth-based obstacle detection module.

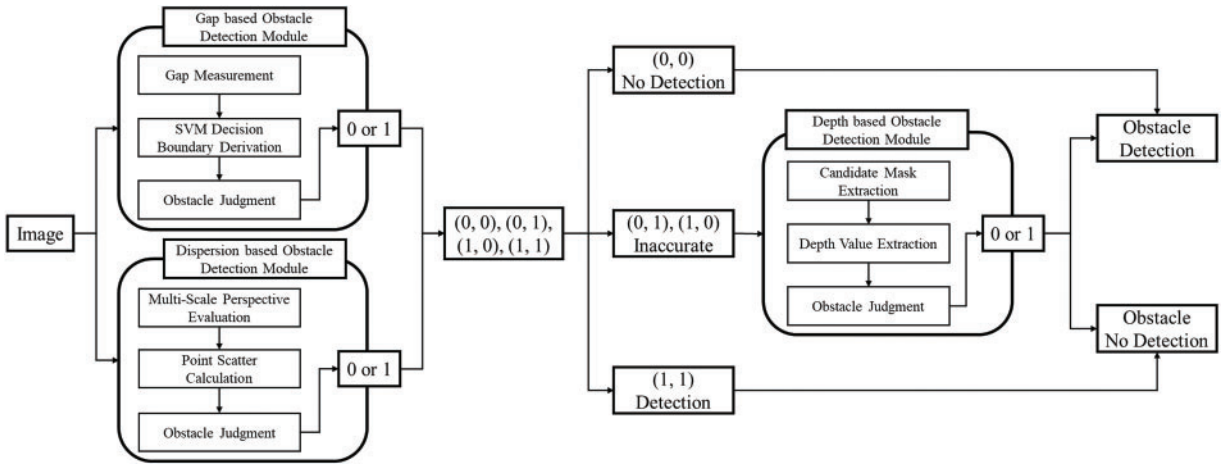

Figure 7: Flowchart of the fusion obstacle detection algorithm developed by fusing three modules

3.5.1 Gap-Based Obstacle Detection Module

The gap-based obstacle detection module detects obstacles using the floor mask data extracted from the 3.3 floor detection module and the fused vanishing data estimated by the 3.4 vanishing point detection module. The pattern of the bottom mask varies depending on the presence or absence of obstacles, which can be seen in Fig. 8a. A corridor with no obstacles that can be passed through is formed by the floor mask to the end point of the corridor. If there are obstacles, the floor mask is formed to be short due to the obstacles. The gap between the floor mask and the vanishing point depending on the presence or absence of obstacles is shown in Fig. 8b. When there are no obstacles, the gap between the bottom mask and the vanishing point appears narrow, and when there are obstacles, the gap between the bottom mask and the vanishing point appears wide. A linear classification model was applied to identify the presence of obstacles by using this gap difference. Fig. 8c is a graph that visualizes the results of the linear classification model. The x-axis represents the ratio of the gap distance between the floor mask and the vanishing point according to each image ratio, and the y-axis represents the presence or absence of obstacles. It is labeled as 1 if there is an obstacle and 0 if there is not. The blue dots represent images without obstacles, the red dots represent images with obstacles, and the green lines represent the linear classification line that distinguishes the presence or absence of gaps and obstacles. The decision boundary equation of the linear classification model is defined as Eq. (1). The graph shows that there tends to be a large variation in the spacing of images without obstacles. The gap-based obstacle detection module alone was difficult to achieve a universal and stable detection success rate due to the deviation of the gap. To compensate for this, an additional detection module was fused to overcome the limitations of the gap-based module and improve the overall obstacle detection performance.

$$\text{Decision Boundary Equation: } -4.431 + 23.518x = 0 \tag{1}$$

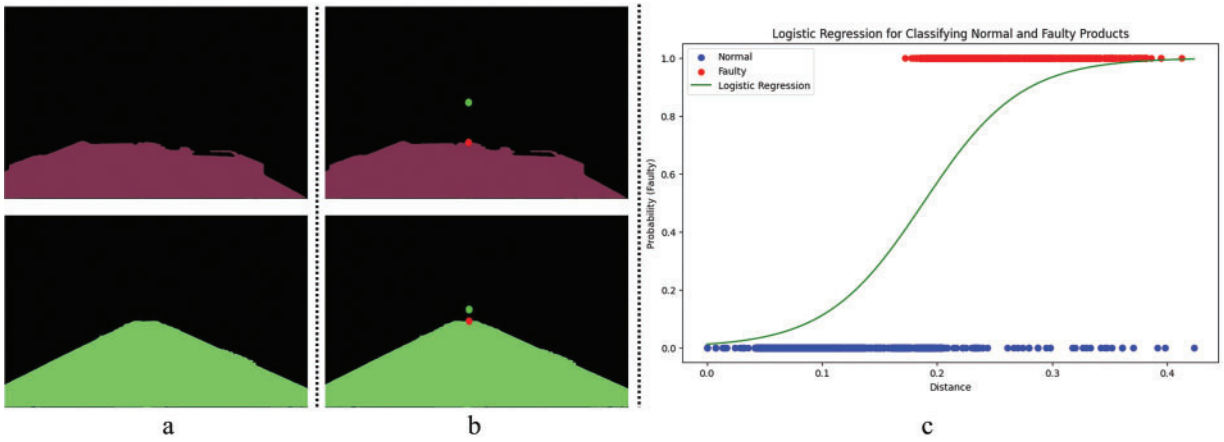

Figure 8: Floor mask with and without obstacles, vanishing point and linear classification

3.5.2 Vanishing Point Variance Based Obstacle Detection Module

The vanishing point detection module has developed and applied an algorithm that estimates the optimal vanishing point for a space based on edge detection-based vanishing points and depth-based vanishing point data. A multi-scale pyramid data augmentation technique was used to measure and adjust the dispersion of the algorithm's vanishing point. The multi-scale pyramid technique, as shown in Fig. 9, enlarges the number of images by first reducing the image to 1.0x, 0.8x, 0.6x, 0.4x, and 0.2x ratios, causing data loss in obstacle information within the image and performing image normalization, and then enlarging it back to the original image size to maintain its original dimensions. This data augmentation was applied to the algorithm as shown in (2) by observing that the dispersion value of the vanishing point varies depending on the data loss when receiving images of various environments and the presence of obstacles that block the floor path that can be driven. This has made it possible to measure the dispersion value of the estimated vanishing point by reflecting the factors that have a significant impact on the dispersion value of the distributed values. The five augmented images generated using the algorithm for estimating vanishing points are used to calculate the dispersion value for the euclidean distance based on the five estimated coordinates of the vanishing points (x1, y1), (x2, y2), ... (x5, y5) using (3). Based on all the calculated dispersion values, a linear classification model was used to classify the dispersion values according to the presence or absence of obstacles in the frontal driving path, as shown in Fig. 9. The x-axis of the graph is the log10 value of the variance of the augmented image, and the y-axis is defined as 0 for images without obstacles and 1 for images with obstacles. The blue dots represent images without obstacles, and the red dots represent images with obstacles. The green lines represent the classification lines of the models. Based on this, a significant classification of the dispersion value when an obstacle is present and when it is not achieved, and the decision boundary equation of the linear classification model used for the classification is as shown in (4).

$$\text{Obstacle Present: Obstacle Absent} = 0.001510 : 0.000340 \tag{2}$$

$$\text{Euclidean Based Var Equation: } Var(d) = \frac{1}{\left(\frac{N}{2}\right)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} - \overline{d} \right)^2 \tag{3}$$

$$\text{Linear SVM Equation: } y = 0.6624^{*} x1 + -1.0000 = 0 \tag{4}$$

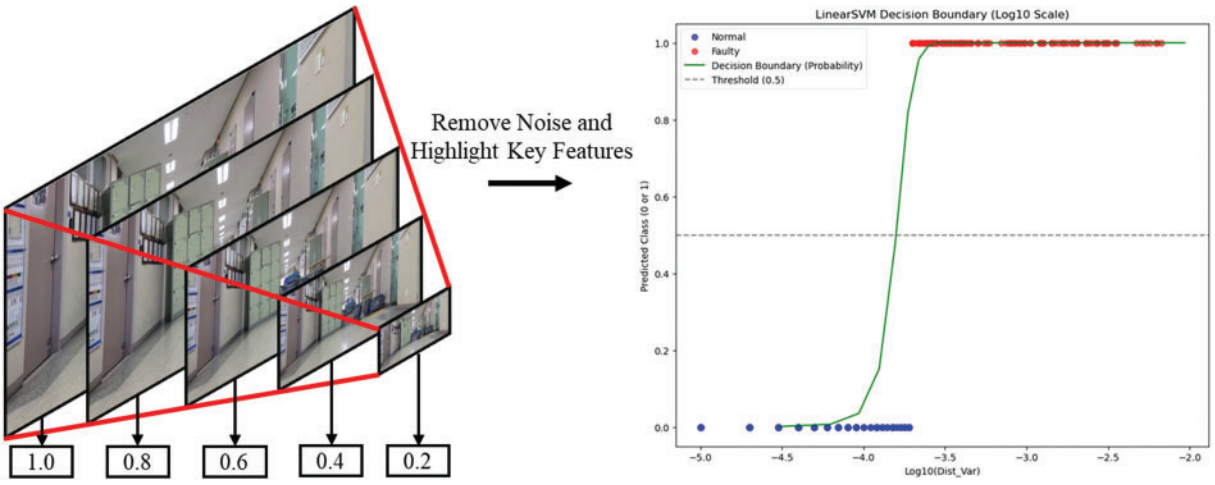

Figure 9: Identification of the presence or absence of obstacles based on the linear SVM model and the multi-scale pyramid data augmentation technique

The gap-based obstacle detection module and the dispersion-based obstacle detection module are used to predict the presence or absence of obstacles, and the prediction result is defined as "obstacle present" as 1 and "obstacle absent" as 0. The results of these two modules are combined, and if both modules indicate the presence of an obstacle (1, 1), it is classified as "final obstacle present," and if both modules indicate that there is no obstacle (0, 0), it is classified as "final obstacle absent." However, there are discrepancies such as (1, 0) or (0, 1) where one module predicts that there is an obstacle and the other module does not. The inconsistent results make it difficult to make accurate judgments on obstacle detection, so we are developing a depth-based obstacle detection module as a final solution.

3.5.3 Depth Based Obstacle Detection Module

The results (1, 0) and (0, 1), which are combined from the gap-based detection module and the distributed-based detection module, are the detection results of rare obstacles such as desks or chairs that are difficult to recognize or of spaces that are not completely blocked but have gaps. To this end, a depth-based obstacle detection module using depth value was introduced. In the initial stage, the mask data of the floor detection module was used to set the rest of the masks except the floor mask as the candidate masks for obstacles. After that, depth data was used to filter out obstacles such as ceilings and distant objects that were unlikely to collide.

Depth data is used to extract the depth value of the obstacle candidate mask from each image, and the median is set as the threshold value to determine the presence of an obstacle. Masks with a depth value greater than the median are identified as potential obstacles in front, meaning that they are not background elements such as paintings or distant scenery. The results of applying this process to an image with actual obstacles are shown in Fig. 10. In addition, by fusing this depth-based obstacle detection module with the gap and dispersion-based obstacle detection module, we are developing a versatile algorithm that can effectively detect obstacles even in complex environments. This algorithm supplements the inaccurate detection results (1, 0) and (0, 1) provided by the initial modules to provide more accurate obstacle detection results.

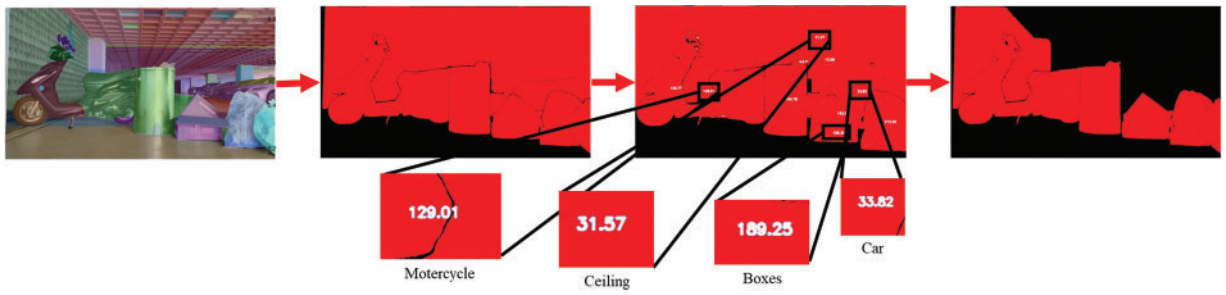

Figure 10: Depth based obstacle detection procedure

4 Experiments and Results

In Section 4, the data collection and experimental environment were systematically designed to evaluate the performance of each developed algorithm, and on this basis, performance verification experiments were conducted. The experiment used a driving robot to collect data in various indoor environments, and the performance of each algorithm was analyzed based on the collected data to evaluate and compare it with the existing algorithms.

4.1 Experiments Setup and Data Collection

The data required for the experiment was collected by utilizing a controllable driving robot [55] and a Samsung Galaxy S21+ [56]. A smartphone was attached to the camera holder of the driving robot to capture the data in front of it, and the driving direction and speed of the robot were kept constant (0.4 m/s) to standardize the data collection conditions. Data was collected in two ways, divided into environments with and without obstacles.

- Environment without obstacle: Data was collected in various spaces, including straight corridors, squares, underground parking lots, T-shaped corridors, and L-shaped corridors.
- Environment with obstacles: Data was collected under conditions that included various obstacles such as boxes, chairs, desks, cars, and trash cans.

Additionally, the MIT indoor scenes dataset [57] was used to evaluate the universality of the experiment. This dataset consists of 67 indoor categories based on indoor images collected from Google and Altavista, and in this study, images from the lobby and corridor categories were selected and utilized. The experiment was proceeded in a high-performance hardware environment such as the i9-13900K and i9-10900K CPUs and RTX 4090 and RTX 3090 GPUs, and the software used the PyTorch 2.3.1 and PyTorch 1.7.1 environments. Fig. 11 shows the robot configuration used for data collection.

4.1.1 Floor Detection Experiments and Results

To evaluate whether the performance of the proposed floor detection algorithm is robust to complex textures or pattern noises, this study compared the suggesting algorithm with existing segmentation algorithms (SimpleClick [46], SAM1, and SAM2). The collected data was utilized to generate floor masks through the floor detection module, which was used to clearly distinguish the passable floor area from obstacles and walls. Fig. 12 is a visual comparison of the results of the bottom detection of the existing segmentation method and the proposed algorithm. Existing models had the following limitations:

- The detection results are subdivided by the patterns or textures of the floor, and there is a lack of continuity.

- An error that detects objects other than the floor (walls, obstacles, etc.).
- A separate prompt must be provided to detect only the floor.

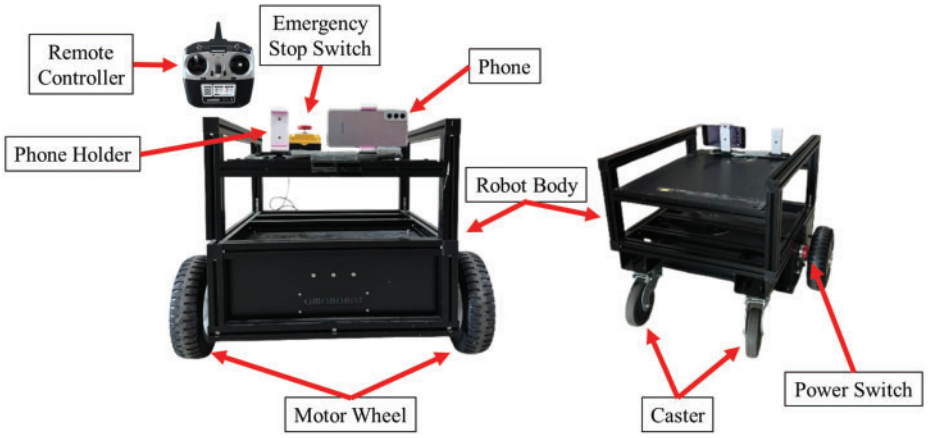

Figure 11: Robot configuration used in the experiment

Figure 12: Floor mask results for each algorithm image

The proposed algorithm addresses these limitations. By integrating the complex textures and pattern noises of the floor to generate a continuous and accurate floor mask, the area that can be driven without

confusion with obstacles or walls was clearly detected. In this process, it was shown that floor detection is possible without the help of a separate human prompt, such as specifying a specific pixel and entering a text prompt, and that the existing algorithm's errors have been reduced, and the precision has been greatly improved by integrating the segmented masks due to pattern noise. The proposed algorithm provided stable performance in various indoor environments such as corridors, lobbies, and indoor parking lots, and robust detection was possible regardless of shadows, sunlight, and the presence of obstacles.

In the floor detection module, complex patterns and textures segmented into multiple regions are merged into a single floor mask through ROI and depth comparison operations. Table 2 presents a quantitative evaluation comparing cases with and without depth comparison operations. Fragmentation refers to the number of inferred masks within the ground truth mask area; a value closer to 1 indicates that the floor regions were successfully merged into a single mask. max ratio denotes the highest overlap ratio between the inferred masks and the ground truth mask, where a higher value indicates better performance. Mask dispersion measures the variance in area among the segmented masks; a higher dispersion suggests that the floor mask is fragmented into many small or large regions, making accurate floor inference more challenging. Therefore, a lower mask dispersion is preferable.

Table 2: Floor detection module ablation experiment

Method	Fragmentation	Max ratio	Mask dispersion
SAM1 (without depth comparison)	2.96	0.9337	158.86
SAM2 (without depth comparison)	1.08	0.9717	16.00
Ours (with depth comparison)	**1.04**	**0.9954**	**2.36**

To compare the quantitative performance, the performance of the proposed algorithm and the existing models (SimpleClick, SAM1, SAM2) was evaluated based on the precision, recall, accuracy, and F1 Score indicators. Table 3 shows the results of comparing the performance of each algorithm.

Table 3: Performance comparison of the module for indoor floor detection and the existing segmentation algorithm

Algorithms	Precision	Recall	Accuracy	F1 score	Inference time (s)
SimpleClick [28] (CocoLvis_ViT_Base) [58–60]	0.8952	0.7933	0.7259	0.8412	0.37
SAM1 [11] (ViT-H) [60]	0.7656	0.7246	0.5930	0.7445	4.17
SAM2 [12] (SAM 2.1_Hiera_Large) [61]	0.9680	0.7819	0.7622	0.8651	0.22
Our Algorithm	**0.9831**	**0.9476**	**0.9324**	**0.9650**	**1.65**

The proposed algorithm showed the highest performance in all metrics, and in particular, it showed superior performance compared to existing models in precision (98.31%) and F1 Score (96.50%). This is attributable to its capability of effectively handling complex floor patterns and pattern noises and maintains high accuracy by clearly distinguishing obstacles and walls.

4.1.2 Vanishing Point Experiments and Results

The proposed vanishing point detection algorithm is to estimate accurate vanishing point data that will be incorporated in obstacle detection algorithms in the future and have been evaluated by comparing its

performance with existing vanishing point detection algorithms. Detection of vanishing points is based on two main methods: edge detection and depth estimation. Edge detection detects linear boundaries in the input image, and depth estimation calculates depth information by estimating distance data for near and far distances. To combine the results of these two methods, a dynamic weighted pyramid was employed to derive the final vanishing point.

To compare the results, the performance of the proposed vanishing point detection algorithm was evaluated in two scenarios using data with and without obstacles. Each data set was utilized to identify the strengths and limitations of the algorithm under various environmental conditions. Data without obstacles was used to evaluate the basic performance of the algorithm under general conditions, while data with obstacles was used to verify the robustness and adaptability of the algorithm in complex environments. Through these two scenarios, the study compared and analyzed how much more accurate and stable the proposed algorithm is than the existing methods based on edge detection and depth estimation.

The dynamic weighted pyramid, which was developed to resolve these limitations, was able to maintain a certain level of accuracy regardless of the presence of obstacles. The proposed algorithm combines the strengths of edge detection and depth estimation to complement for the shortcomings of both methods. This approach has greatly improved the reliability of vanishing point detection, with stable performance in various environments. Fig. 13 is the result of a visual comparison of the coordinates of the detected missing points for each data according to the multi-scale pyramid. The blue dots represent the coordinates of the vanishing point based on edge detection, the yellow dots represent the coordinates of the vanishing point based on depth estimation, and the red dots represent the final coordinates of the vanishing point derived from the dynamic weighted pyramid. As a result, the proposed algorithm fused the strengths of both methods to provide a more accurate and stable vanishing point.

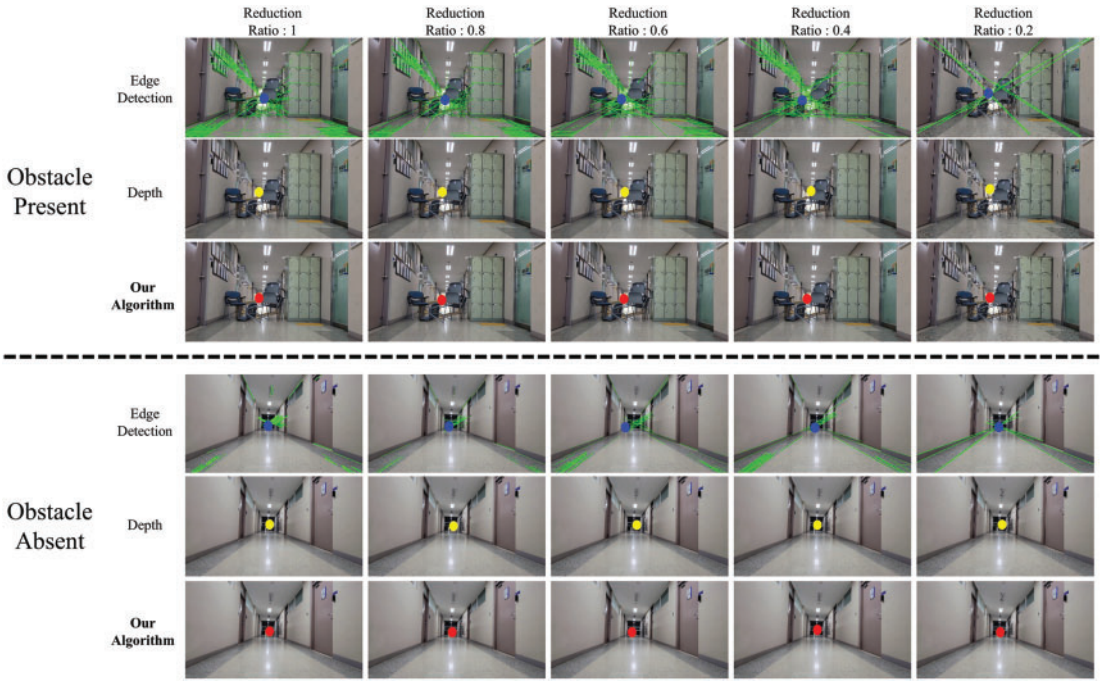

Figure 13: Image of the detection results of the vanishing for each algorithm

To compare the quantitative performance, the results of the proposed algorithm and the existing edge detection methods (based on edge detection and based on depth estimation) were evaluated based on the indicators of precision, recall, accuracy, and F1 Score. These metrics allow the overall performance of the algorithm to be evaluated by measuring the detection accuracy of each algorithm, the percentage of missing vanishing points, and overall accuracy. Performance evaluation was conducted in two scenarios: with and without obstacles. The strengths and stability of the proposed algorithm were verified by comparing and analyzing the results under each condition. Table 4 shows the quantitative performance results of each algorithm according to the obstacle presence scenario. This demonstrates that the proposed dynamic weighted pyramid provides overall higher performance compared to both the existing methods and their simple fusion, and can produce consistent results even in various environments.

Table 4: Performance comparison of vanishing point detection algorithms based on the presence or absence of obstacles

	Algorithm	Precision	Recall	Accuracy	F1 Score
With obstacles	Edge detection based vanishing point	0.6184	0.7344	0.5054	0.6714
	Depth based vanishing point	0.9707	0.8749	0.8524	0.9203
	Edge+Depth (without dynamic weighted pyramid)	0.5079	0.2152	0.1781	0.3023
	Our vanishing point	**0.9314**	**0.8902**	**0.8354**	**0.9103**
Without obstacles	Edge detection based vanishing point	0.9515	0.8932	0.8543	0.9214
	Depth based vanishing point	0.9869	0.8035	0.7951	0.8858
	Edge+Depth (without dynamic weighted pyramid)	0.9214	0.9043	0.9043	0.9128
	Our vanishing point	**0.9944**	**0.8859**	**0.8815**	**0.937**

4.1.3 Obstacle Detection Experiments and Results

To evaluate the performance of the proposed fusion obstacle detection algorithm, we compared its performance with the existing foundation model-based obstacle detection technique. The experiment was conducted with not only common obstacles but also rare obstacles that are difficult to detect, and the performance of the individual detection modules based on gap, distribution, and depth and the proposed fusion obstacle detection algorithms were compared.

The proposed fusion algorithm is designed to complement the limitations of individual detection modules (gap-based, distributed-based, and depth-based) and combine their strengths to provide more precise and reliable obstacle detection results. Each individual module performs the task of detecting obstacles independently, and the fusion algorithm integrates their results to produce the final obstacle detection result.

- The gap-based detection module classifies the presence of obstacles by calculating the ratio of the euclidean distance between the floor mask and the vanishing point. This data is stored as label 1 if there is an obstacle and label 0 if there is no obstacle, and the prediction performance is evaluated using the linear SVM classification model.
- The distributed detection module uses the dispersion data of the final missing point. After augmenting the data through a reduced pyramid structure, the presence of obstacles is classified by calculating the

variance. If the variance value is large, an obstacle is present; if it is small, no obstacle is present. The results are stored as label 1 and label 0 in the same way and evaluated using a linear SVM.

- The depth-based detection module additionally analyzes the results of the gap-based and distributed detection modules to detect obstacles. By using depth data to precisely analyze the information of obstacles, the detection results of individual modules are supplemented and their performance is improved.

The operation of the fusion algorithm is as follows:

1. If both the gap-based detection module and the distributed-based detection module detect an obstacle, it is classified as (1, 1), which is considered to be a clear detection of an obstacle.
2. If both modules fail to detect an obstacle, it is classified as (0, 0), which is considered to be free of obstacles.
3. If only one of the two modules detects an obstacle, it is classified as (1, 0) or (0, 1). In this case, a depth-based detection module is additionally applied to determine the presence of obstacles.
4. The final result is calculated by integrating the results of the three modules. Evaluation method the performance of the fusion algorithm was evaluated using the precision, recall, accuracy, and F1 Score indicators. Using these indicators, we analyzed how much better the fusion algorithm performs compared to individual modules, especially in complex environments, and whether it can detect stably. Table 5 shows the performance comparison results of each module and the fusion algorithm in detail. In particular, the fusion obstacle detection module recorded high performance of 0.9 or higher in the precision, recall, accuracy, and F1 Score metrics when obstacles were present and when they were not. This indicates that the fusion obstacle detection performance is superior compared to both the results of individual modules and the results of fusing only two out of the three modules.

Table 5: Comparison of obstacle detection performance of each module and fusion algorithm

Algorithms	Presence or absence of obstacles	Precision	Recall	Accuracy	F1 Score
Gap based detection module	Obstacle present	0.86	0.60	0.76	0.71
	Obstacle absent	0.70	0.91	0.776	0.79
Dispersion based detection module	Obstacle present	0.67	0.91	0.73	0.77
	Obstacle absent	0.86	0.56	0.73	0.68
Depth based detection module	Obstacle present	0.90	0.82	0.82	0.86
	Obstacle absent	0.84	0.91	0.90	0.87
Gap + Dispersion	Obstacle present	0.63	0.87	0.87	0.73
	Obstacle absent	0.78	0.48	0.48	0.59
Dispersion + Depth	Obstacle present	0.94	0.83	0.83	0.88
	Obstacle absent	0.85	0.94	0.94	0.89
Gap + Depth	Obstacle present	0.82	0.75	0.75	0.78
	Obstacle absent	0.77	0.83	0.83	0.80
Our obstacle detection algorithms	**Obstacle present**	**0.91**	**0.95**	**0.94**	**0.93**
	Obstacle absent	**0.95**	**0.90**	**0.90**	**0.92**

4.1.4 Obstacle Detection Performance Comparison

The last experiment compared the obstacle detection performance between the foundation model based on text prompts and our fusion obstacle detection algorithm. For comparison in this experiment, SAM1, SAM2, UniLSeg, and PolyFormer were selected, as they are widely recognized and readily accessible foundation models based on text prompts that are applicable to various domains. The text prompt input to these models was compared using "Obstacle" and "Object" to meet the objective of increasing the versatility of obstacle detection. As a result, it was confirmed that the prompt "Object" detected more of the intended obstacles in the input image, so it was adopted as the prompt with better detection performance and used as the final input. On the other hand, our fusion obstacle detection algorithm did not use text prompt, but detected obstacles based solely on image data and data from the existing detection module. This approach, unlike the other models, fuses the analysis results of each module to determine the presence of obstacles, making it possible to detect obstacles without additional information such as human prompts. The results of the module-by-module comparison are shown in Fig. 14, and the performance comparison results are shown in Table 6.

Figure 14: Obstacle detection results for each algorithm Image

Table 6: Comparison of obstacle detection performance between conventional general-purpose algorithms and fusion algorithms

Algorithms	Precision	Recall	Accuracy	F1 Score	Inference time (s)
SAM 1 [11] (ViT-H) [60]	0.39	0.13	0.11	0.20	6.22
SAM 2 [12] (SAM 2.1_Hiera_Large) [61]	0.99	0.19	0.19	0.31	2.20
UniLSeg [31] (ViT-B, Swin Transformer) [60,62]	0.41	0.87	0.38	0.56	1.19
PolyFormer [33] (Bert, PolyFormer-L) [63]	0.36	0.91	0.35	0.51	19.90
Our obstacle detection algorithms	**0.94**	**0.92**	**0.87**	**0.93**	2.37

As shown in Fig. 14, the results of SAM1, SAM2, UniLSeg and PolyFormer show inaccurate detection results, such as detecting only some obstacles or segmenting and detecting some areas of a single obstacle, for all obstacles that should be detected in the intended input image. Since such detection errors can act as fatal flaws in the recognition and understanding of indoor spaces, there are limitations to adopting such models for such situations.

On the other hand, the fusion obstacle detection algorithm we developed can clearly detect obstacles to be detected in the input image without prompting, and can also be used to prioritize and mask close obstacles among multiple candidate obstacles.

In the quantitative performance comparison, our algorithm performed well in precision (94%), recall (92%), accuracy (87%), and F1 Score (93%), significantly outperforming SAM1 (precision: 39%, recall: 13%, accuracy: 11%, F1 Score: 20%) and SAM2 (precision: 99%, recall: 19%, accuracy: 19%, F1 Score: 31%). In particular, the proposed algorithm demonstrated excellent performance in terms of the ratio of actual obstacles detected (precision) and the ratio of obstacles detected without omission (recall), proving that it provides stable and reliable results without relying on prompt. This has proven its usefulness and versatility in real-world applications. Furthermore, the algorithm of this study proved that it met the purpose of the Foundation model, which is highly versatile and can operate stably in various environments and conditions, better than existing models.

5 Conclusion

This study proposed a zero-shot based spatial recognition algorithm that supports the creation of precise 3D maps in complex indoor environments by fusing various vision-based algorithms. The proposed algorithm consists of three modules: floor detection, lost-point-based spatial analysis, and obstacle detection, each of which is designed to effectively respond to the complexity and variation of indoor spaces.

The floor detection module combines segmentation and depth algorithms to enable accurate detection even on floors with complex patterns and textures, and the vanishing point-based spatial analysis uses image processing techniques to accurately identify the indoor structure. The obstacle detection module has improved the accuracy of spatial recognition by fusing gap-based, distributed-based, and depth-based detection techniques to distinguish between sparse obstacles or unusual structures and analyze whether they are passable.

The experimental results show that the proposed algorithm delivers higher detection accuracy and efficiency than existing methods. First, we introduced a novel indoor floor-detection algorithm and validated its effectiveness by comparing it with a conventional segmentation approach. The proposed method achieved

an F1 score of 0.965, outperforming the baseline. We also developed an obstacle-detection algorithm that fuses gap, dispersion, and depth cues. Compared with the individual detectors, the fused model achieved stable performance, posting an F1 score of 0.93 in obstacle-present scenes and 0.92 in obstacle-free scenes. Furthermore, against foundation-model-based baselines, the proposed algorithm again proved superior, recording an F1 score of 0.93, which exceeds those of SAM-1 and SAM-2. These findings are expected to contribute significantly to the advancement of precise floor-detection and obstacle-recognition algorithms for indoor environments.

This algorithm can quickly reflect structural changes in indoor spaces and efficiently collect key data for creating 3D site maps. This allows the driving robot to precisely recognize the space ahead, derive the optimal route, and analyze changes in the indoor environment in real time. In particular, it provides practical contributions in that it enables precise facility management and maintenance by automatically detecting discrepancies between the actual internal structure and the design drawings after the building is completed.

The uniqueness of this study is that it has enabled precise floor and wall recognition even in increasingly complex indoor environments through detection techniques using floor masks and vanishing points. In addition, by combining gap-based, distributed-based, and depth-based algorithms, it has been possible to analyze whether it is possible to pass through stably even in various abnormal situations, including unusual obstacles. These achievements demonstrate the potential for use in various applications of indoor space management, including internal construction, accidents, and disaster situations. Additionally, by attaching a smartphone to a low-cost robot and running an algorithm, it is possible to create accurate 3D maps even in construction sites, old buildings, and disaster areas that are difficult for humans to access, making it highly practical.

However, this algorithm is specialized for environments such as indoor and underground spaces, where structural boundaries are clearly defined and lighting conditions are relatively stable. In contrast, outdoor environments often involve frequent weather- and lighting-related variations and have fewer obvious structural reference points, which makes it difficult to accurately detect the vanishing point and may significantly degrade the algorithm's performance. For instance, in expansive open-air areas lacking walls or pillars, the accuracy of vanishing point detection can decline, and frequent changes in lighting due to backlighting or strong shadows can destabilize object detection. Therefore, while this algorithm exhibits stable performance in structured indoor or underground settings, additional measures—such as optical calibration or sensor fusion—would be required for direct application to outdoor environments. Furthermore, although the Foundation Model-based detection technique excels at recognizing previously unknown objects, it has limited capacity to provide detailed classifications of those detected objects. As a result, in scenarios where specific object identification or classification is critical, an additional object recognition module must be incorporated. This supplementary approach is particularly necessary in outdoor environments, where the wide variety of possible objects often demands more sophisticated recognition capabilities.

Future research will apply spatial segmentation techniques to ensure that the proposed algorithm can operate effectively in a multi-room environment and improve the performance of detecting sparse obstacles. In addition, this algorithm can be expanded to disaster response and indoor safety management systems, and it can support the detection of structural changes and the automatic creation of emergency evacuation routes in the event of an earthquake or collapse. In addition, future research will integrate a pre-trained object detection model into the foundation model to add the ability to classify the type of obstacle in addition to its presence, thereby establishing a detailed classification and detection indoor space recognition system.

Acknowledgement: The authors would like to sincerely express their gratitude to each other for their collaboration on this research. In particular, the authors extend their deep appreciation to the advisor for their valuable insights, support, and guidance, which greatly contributed to the completeness of this study.

Funding Statement: This work was supported by Kyonggi University Research Grant 2024.

Author Contributions: Research Design: Sehun Lee, Taehoon Kim, Junho Ahn; Methodology: Sehun Lee, Taehoon Kim, Junho Ahn; Data Collection: Sehun Lee, Taehoon Kim; Data Analysis and Results Integration: Sehun Lee, Taehoon Kim; Drafting of the Manuscript: Sehun Lee, Taehoon Kim, Junho Ahn; Visualization: Sehun Lee, Taehoon Kim; Supervision: Junho Ahn. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used in the research can be made available to the corresponding author, Junho Ahn, upon request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Luleci F, AlGadi A, Necati Catbas F. Multimodal data collection using mobile robotics for rapid structural assessment. In: Bridge maintenance, safety, management, digitalization and sustainability. London: CRC Press; 2024. p. 742–9. doi:10.1201/9781003483755-86.
2. Wang R, Veloso M, Seshan S. Active sensing data collection with autonomous mobile robots. In: 2016 IEEE International Conference on Robotics and Automation (ICRA); 2016 May 16–21; Stockholm, Sweden: IEEE. 2016. p. 2583–8. doi:10.1109/ICRA.2016.7487415.
3. Wierzbicki D, Stogowski P. Application of stereo cameras with wide-angle lenses for the indoor mapping. Int Arch Photogramm Remote Sens Spatial Inf Sci. 2022;XLIII–B2–2022:477–84. doi:10.5194/isprs-archives-xliii-b2-2022-477-2022.
4. Wietrzykowski J. PlaneLoc2: indoor global localization using planar segments and passive stereo camera. IEEE Access. 2022;10:67219–29. doi:10.1109/access.2022.3185732.
5. Zimmerman N, Sodano M, Marks E, Behley J, Stachniss C. Constructing metric-semantic maps using floor plan priors for long-term indoor localization. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2023 Oct 1–5; Detroit, MI, USA: IEEE. 2023. p. 1366–72. doi:10.1109/IROS55552.2023.10341595.
6. Zhong X, Pan Y, Behley J, Stachniss C. SHINE-mapping: large-scale 3D mapping using sparse hierarchical implicit neural representations. In: 2023 IEEE International Conference on Robotics and Automation (ICRA); 2023 May 29–Jun 2; London, UK: IEEE. 2023. p. 8371–7. doi:10.1109/ICRA48891.2023.10160907.
7. Huang H, Li L, Cheng H, Yeung SK. Photo-SLAM: real-time simultaneous localization and photorealistic mapping for monocular, stereo, and RGB-D cameras. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA: IEEE. 2024. p. 21584–93. doi:10.1109/CVPR52733.2024.02039.
8. Liu Z, Li Z, Liu A, Sun Y, Jing S. Fusion of binocular vision, 2D lidar and IMU for outdoor localization and indoor planar mapping. Meas Sci Technol. 2023;34(2):025203. doi:10.1088/1361-6501/ac9ed0.
9. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv:2108.07258. 2021.
10. Awais M, Naseer M, Khan S, Anwer RM, Cholakkal H, Shah M, et al. Foundation models defining a new era in vision: a survey and outlook. IEEE Trans Pattern Anal Mach Intell. 2025;47(4):2245–64. doi:10.1109/TPAMI.2024.3506283.
11. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France: IEEE. 2023. p. 3992–4003. doi:10.1109/ICCV51070.2023.00371.

12. Ravi N, Gabeur V, Hu YT, Hu R, Ryali C, Ma T, et al. Sam 2: segment anything in images and videos. arXiv:2408.00714. 2024.
13. Yang CY, Huang HW, Chai W, Jiang Z, Hwang JN. SAMURAI: adapting segment anything model for zero-shot visual tracking with motion-aware memory. arXiv:2411.11922. 2024.
14. Wang W, Dai J, Chen Z, Huang Z, Li Z, Zhu X, et al. InternImage: exploring large-scale vision foundation models with deformable convolutions. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada: IEEE. 2023. p. 14408–19. doi:10.1109/CVPR52729.2023.01385.
15. Singh A, Hu R, Goswami V, Couairon G, Galuba W, Rohrbach M, et al. FLAVA: a foundational language and vision alignment model. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA: IEEE. 2022. p. 15617–29. doi:10.1109/CVPR52688.2022.01519.
16. Yuan L, Chen D, Chen YL, Codella N, Dai X, Gao J, et al. Florence: a new foundation model for computer vision. arXiv:2111.11432. 2021.
17. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. PMLR; 2021. Vol. 139, p. 8748–63.
18. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, et al. VQA: visual question answering. In: 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile: IEEE. 2015. p. 2425–33. doi:10.1109/ICCV.2015.279.
19. Ma Z, Hong J, Gul MO, Gandhi M, Gao I, Krishna R. @ CREPE: can vision-language foundation models reason compositionally? In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada: IEEE. 2023. p. 10910–21. doi:10.1109/CVPR52729.2023.01050.
20. Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, et al. Scaling up visual and vision-language representation learning with noisy text supervision. International conference on machine learning. In: Proceedings of the 38th International Conference on Machine Learning. PMLR; 2021. Vol. 139, p. 4904–16.
21. Foutter M, Bhoj P, Sinha R, Elhafsi A, Banerjee S, Agia C, et al. Adapting a foundation model for space-based tasks. arXiv:2408.05924. 2024.
22. Wang D, Zhang Q, Xu Y, Zhang J, Du B, Tao D, et al. Advancing plain vision transformer toward remote sensing foundation model. IEEE Trans Geosci Remote Sens. 2022;61:5607315. doi:10.1109/TGRS.2022.3222818.
23. Yu L, Poirson P, Yang S, Berg AC, Berg TL. Modeling context in referring expressions. In: Computer vision–ECCV 2016. Cham: Springer International Publishing; 2016. p. 69–85. doi:10.1007/978-3-319-46475-6_5.
24. Hui T, Liu S, Huang S, Li G, Yu S, Zhang F, et al. Linguistic structure guided context modeling for referring image segmentation. In: Computer vision–ECCV 2020. Cham: Springer International Publishing; 2020. p. 59–75. doi:10.1007/978-3-030-58607-2_4.
25. Luo G, Zhou Y, Sun X, Cao L, Wu C, Deng C, et al. Multi-task collaborative network for joint referring expression comprehension and segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA: IEEE. 2020. p. 10031–40. doi:10.1109/CVPR42600.2020.01005.
26. Ding H, Liu C, Wang S, Jiang X. VLT: vision-language transformer and query generation for referring segmentation. IEEE Trans Pattern Anal Mach Intell. 2023;45(6):7900–16. doi:10.1109/tpami.2022.3217852.
27. Feng G, Hu Z, Zhang L, Lu H. Encoder fusion network with co-attention embedding for referring image segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA: IEEE. 2021. p. 15501–10. doi:10.1109/cvpr46437.2021.01525.
28. Yang Z, Wang J, Ye X, Tang Y, Chen K, Zhao H, et al. Language-aware vision transformer for referring segmentation. IEEE Trans Pattern Anal Mach Intell. 2024. doi:10.1109/TPAMI.2024.3468640.
29. Liu Q, Xu Z, Bertasius G, Niethammer M. SimpleClick: interactive image segmentation with simple vision transformers. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France: IEEE. 2023. p. 22233–43. doi:10.1109/ICCV51070.2023.02037.
30. Xu J, Xu L, Yang Y, Li X, Wang F, Xie Y, et al. U-LLaVA: unifying multi-modal tasks via large language model. In: European Conference on Artificial Intelligence ECAI 2024. Amsterdam, Netherlands: IOS Press; 2024. doi:10.3233/faia240541.

31. Liu Y, Zhang C, Wang Y, Wang J, Yang Y, Tang Y. Universal segmentation at arbitrary granularity with language instruction. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA: IEEE. 2024. p. 3459–69. doi:10.1109/CVPR52733.2024.00332.
32. Lai X, Tian Z, Chen Y, Li Y, Yuan Y, Liu S, et al. LISA: reasoning segmentation via large language model. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA; 2024. p. 9579–89.
33. Liu J, Ding H, Cai Z, Zhang Y, Kumar Satzoda R, Mahadevan V, et al. PolyFormer: referring image segmentation as sequential polygon generation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada: IEEE. 2023. p. 18653–63. doi:10.1109/CVPR52729.2023.01789.
34. Zhang Z, Ma Y, Zhang E, Bai X. PSALM: pixelwise SegmentAtion with large multi-modal model. In: Computer vision–ECCV 2024. Cham: Springer Nature Switzerland; 2024. p. 74–91. doi:10.1007/978-3-031-72754-2_5.
35. Wu J, Jiang Y, Yan B, Lu H, Yuan Z, Luo P. Segment every reference object in spatial and temporal spaces. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France: IEEE. 2023. p. 2538–50. doi:10.1109/ICCV51070.2023.00240.
36. Zhang Y, Cheng T, Zhu L, Hu R, Liu L, Liu H, et al. EVF-SAM: early vision-language fusion for text-prompted segment anything model. arXiv:2406.20076. 2024.
37. Rasheed H, Maaz M, Shaji S, Shaker A, Khan S, Cholakkal H, et al. GLaMM: pixel grounding large multimodal model. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA: IEEE. 2024. p. 13009–18. doi:10.1109/CVPR52733.2024.01236.
38. Ren Z, Huang Z, Wei Y, Zhao Y, Fu D, Feng J, et al. PixelLM: pixel reasoning with large multimodal model. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA: IEEE. 2024. p. 26364–73. doi:10.1109/CVPR52733.2024.02491.
39. Lafuente-Arroyo S, Maldonado-Bascón S, Gómez-Moreno H, Alén-Cordero C. Segmentation in corridor environments: combining floor and ceiling detection. In: Pattern recognition and image analysis. Cham: Springer International Publishing; 2019. p. 485–96. doi:10.1007/978-3-030-31321-0_42.
40. Ravishankar K, Devaraj P, Yeliyur Hanumathaiah SK. Floor segmentation approach using FCM and CNN. Acadlore Trans AI Mach Learn. 2023;2(1):33–45. doi:10.56578/ataiml020104.
41. Wen C, Tan J, Li F, Wu C, Lin Y, Wang Z, et al. Cooperative indoor 3D mapping and modeling using LiDAR data. Inf Sci. 2021;574:192–209. doi:10.1016/j.ins.2021.06.006.
42. Höllein L, Johnson J, Nießner M. StyleMesh: style transfer for indoor 3D scene reconstructions. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA: IEEE. 2022. p. 6188–98. doi:10.1109/CVPR52688.2022.00610.
43. Wu Z, Zhou Z, Allibert G, Stolz C, Demonceaux C, Ma C. Transformer fusion for indoor RGB-D semantic segmentation. Comput Vis Image Underst. 2024;249:104174. doi:10.1016/j.cviu.2024.104174.
44. Hu X, Assaad RH. A BIM-enabled digital twin framework for real-time indoor environment monitoring and visualization by integrating autonomous robotics, LiDAR-based 3D mobile mapping, IoT sensing, and indoor positioning technologies. J Build Eng. 2024;86:108901. doi:10.1016/j.jobe.2024.108901.
45. Muhammad Yasir S, Muhammad Sadiq A, Ahn H. 3D instance segmentation using deep learning on RGB-D indoor data. Comput Mater Contin. 2022;72(3):5777–91. doi:10.32604/cmc.2022.025909.
46. Kassab C, Mattamala M, Zhang L, Fallon M. Language-EXtended indoor SLAM (LEXIS): a versatile system for real-time visual scene understanding. In: 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan: IEEE. 2024. p. 15988–94. doi:10.1109/ICRA57147.2024.10610341.
47. Duh PJ, Sung YC, Chiang LF, Chang YJ, Chen KW. V-eye: a vision-based navigation system for the visually impaired. IEEE Trans Multimed. 2020;23:1567–80. doi:10.1109/TMM.2020.3001500.
48. Gomez C, Fehr M, Millane A, Hernandez AC, Nieto J, Barber R, et al. Hybrid topological and 3D dense mapping through autonomous exploration for large indoor environments. In: 2020 IEEE International Conference on Robotics and Automation (ICRA); 2020 May 31–Aug 31; Paris, France: IEEE. 2020. p. 9673–9. doi:10.1109/icra40945.2020.9197226.

49. Hübner P, Weinmann M, Wursthorn S, Hinz S. Automatic voxel-based 3D indoor reconstruction and room partitioning from triangle meshes. ISPRS J Photogramm Remote Sens. 2021;181:254–78. doi:10.1016/j.isprsjprs.2021.07.002.
50. Kim D, Kim C, Ahn J. Vision-based recognition algorithm for up-to-date indoor digital map generations at damaged buildings. Comput Mater Contin. 2022;72(2):2765–81. doi:10.32604/cmc.2022.025116.
51. Kim D, Min J, Song Y, Kim C, Ahn J. Intelligent risk-identification algorithm with vision and 3D LiDAR patterns at damaged buildings. Intell Autom Soft Comput. 2023;36(2):2315–31. doi:10.32604/iasc.2023.034394.
52. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97. doi:10.1007/BF00994018.
53. Yang L, Kang B, Huang Z, Xu X, Feng J, Zhao H. Depth anything: unleashing the power of large-scale unlabeled data. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA: IEEE. 2024. p. 10371–81. doi:10.1109/CVPR52733.2024.00987.
54. Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell. 1986; PAMI-8(6):679–98. doi:10.1109/TPAMI.1986.4767851.
55. Omorobot. omorobot.com/r1v2 [Internet]. Ansan, Kyonggi, Republic of Korea [cited 2025 Jan 21]. Available from: https://www.omorobot.com/r1v2/.
56. Samsung. samsung.com/sec/mobile [Internet]. Suwon, Kyonggi, Republic of Korea [cited 2025 Jan 21]. Available from: https://www.samsung.com/sec/mobile/.
57. Massachusetts Institute of Technology. Indoor Scene Recognition [Internet]. Cambridge, MA, USA; 2009 [cited 2025 Jan 21]. Available from: https://web.mit.edu/torralba/www/indoor.html2009.
58. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Computer vision–ECCV 2014. Cham: Springer International Publishing; 2014. p. 740–55. doi:10.1007/978-3-319-10602-1_48.
59. Gupta A, Dollár P, Girshick R. LVIS: a dataset for large vocabulary instance segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA: IEEE. 2019. p. 5351–9. doi:10.1109/CVPR.2019.00550.
60. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
61. Meta. SA-V Dataset [Internet]. Menlo Park, CA, USA; 2024 [cited 2025 Jan 21]. Available from: https://ai.meta.com/datasets/segment-anything-video/.
62. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada: IEEE. 2021. p. 9992–10002. doi:10.1109/ICCV48922.2021.00986.
63. Denvlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). Minneapolis, MN, USA; 2019. p. 4171–86.