Article

Mobile Phone Indoor Scene Recognition Location Method Based on Semantic Constraint of Building Map

Liu Jianhua[1, 2,*], Feng Guoqiang[1], Luo Jingyan[3], Wen Danqi[1], Chen Zheng[1], Wang Nan[1], Zeng Baoshan[1], Wang Xiaoyi[1], Li Xinyue[1], Gu Botong[1]

[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 100044, China;
[2] Key Laboratory for Urban Geomatics of National Administration of Surveying, Mapping and Geoinformation, Beijing 100044, China;
[3] Fujian College of Water Conservancy and Electric Power, Fujian 366000, China;
* Correspondence: liujianhua@bucea.edu.cn; Website: www.dxkjs.com.

Abstract: At present, indoor localization is one of the core technologies of location-based services (LBS), and there exist numerous scenario-oriented application solutions. Visual features, as the main semantic information to help people understand the environment and thus occupy the dominant part, many techniques about indoor scene recognition are widely adopted. However, the engineering application problem of cell phone indoor scene recognition and localization has not been well solved due to insufficient semantic constraint information of building map and the immaturity of building map location anchors (MLA) matching positioning technology. To address the above problems, this paper proposes a cell phone indoor scene recognition and localization method with building map semantic constraints. Firstly, we build a library of geocoded entities for building map location anchors (MLA), which can provide users with "immersive" real-world building maps on the one hand and semantic anchor point constraints for cell phone positioning on the other. Secondly, using the improved YOLOv5s deep learning model carried on the mobile terminal, we recognize the universal map location anchors (MLA) elements in building scenes by cell phone camera video in real-time. Lastly, the spatial location of the scene elements obtained from the cell phone video recognition is matched with the building MLA to achieve real-time positioning and navigation. The experimental results show that the model recognition accuracy of this method is above 97.2%, and the maximum localization error is within the range of 0.775 m, and minimized to 0.5 m after applying the BIMPN road network walking node constraint, which can effectively achieve high positioning accuracy in the building scenes with rich MLA element information. In addition, the building map location anchors (MLA) has universal characteristics, and the positioning algorithm based on scene element recognition is compatible with the extension of indoor map data types, so this method has good prospects for engineering applications.

Keywords: cell phone indoor positioning; scene recognition; building map; map location anchor; YOLOv5; geocoding matching

1. Introduction

Buildings are the main space for human activities, such as office buildings, libraries, shopping centers, hospitals, train stations, airports, etc. According to research, humans spend about 87% of their time in indoor spaces [1,2]. However, the widely used Global Navigation Satellite System (GNSS) cannot be used indoors or in urban environments, where GNSS signals are blocked by buildings, trees, or other obstructions. Compared with outdoor positioning, indoor positioning is more challenging. Because indoor spaces are more complex than outdoor environments in terms of layout, topology, and spatial constraints [3], indoor positioning requires higher accuracy [4]. In recent years, many indoor

doi:10.20944/preprints202201.0431.v1

positioning systems have been proposed by researchers, which use different techniques, such as infrared [5], Wi-Fi [6], Bluetooth [7], optical [8], and inertial sensors [9]. However, each of these techniques has its application scenarios when considering accuracy, cost, coverage, complexity, and applicability. On the one hand, a certain number of signal access points need to be deployed in advance. On the other hand, the complex indoor space blocks the effective transmission of some signals, which makes pervasive indoor localization services more challenging. The current scene recognition visual localization technology incorporating multi-source sensors provides a new way to solve these problems, and it is quickly becoming one of the important directions in the research field of mobile phone indoor localization.

The purpose of the scene recognition research is to use recognition algorithms to effectively process the semantic information contained in the image data, in order to extract the image features and determine the valid information of the category to which the scene image belongs. In the face of complex scene recognition problems, traditional scene recognition methods [10] gradually show limitations. Deep neural networks are able to learn the deep characteristics of images from a large number of sample images and show significant advantages in the field of image recognition, which better achieves low cost, high accuracy, and more stable navigation services [11]. Deep neural networks are network structures containing multiple hidden layers and multiple perceptrons, which can describe the properties and features of objects at a more abstract and deeper level [12]. They are widely used with the advantages of strong feature extraction ability, high recognition accuracy, and good real-time performance. Deep learning based object detection methods can be divided into three categories. First, candidate region-based object detection methods, such as Hybrid Task Cascade[13], CenterMask [14], PolyTransform[15], etc; second, regression-based object detection methods, such as YOLO [16,17], SSD [18], FPN [19], etc; third, search-based object detection methods, such as AttentionNet [20] and reinforcement learning-based object detection algorithms [21]. Many scholars have incorporated deep learning into the technical solutions for indoor positioning and navigation: A fingerprint localization algorithm based on Deep Belief Networks (DBN) with noise reduction is used to achieve target localization in specific indoor environments [22]; using deep learning methods to automatically encode and extract deep features from Wi-Fi fingerprint data, and create a deep feature location fingerprint database with one-to-many relationships for indoor localization [23]; adding the scene recognition classification process to a visual localization system [24], etc. At present, the image quality, pixel resolution, sensor, and aperture performance of the video frames obtained by the cell phone camera have been significantly improved. And with the rapid development of artificial intelligence, the smartphone camera sensor adds intelligent anti-vibration, super night scene, backlighting, and other auxiliary functions to make the video image more clearer. The use of more efficient and suitable for scene recognition and geocoding of the semantic constraints of the building map information to assist, and then achieve cell phone camera scene recognition and localization is a field with great potential.

Building maps is an effective type of information representation of interior spatial elements, in which semantic information refers to that information that enables cell phones to better understand user movement rules, perceive user scenes, plan navigation routes, and is covered in multi-level and rich dimensionality in high precision maps of buildings [25-27]. Semantic information in building maps can better represent the user's scene [28]. Semantically rich indoor maps are an indispensable part of geolocation-based indoor services [29]. Markers and contextual information in indoor maps can better understand user movement rules, perceive user scenarios, correct indoor positioning errors, and plan indoor navigation paths [30]. The research on the genealogical semantic features of building map models still has much room for development, among which how to effectively organize the semantic information of building entity elements, and construct and improve building map models for cell phone indoor positioning and navigation with universal applications, which is a key problem that needs to be solved urgently [31,32].

In summary, for the problems of insufficient semantic constraint information of building map and matching localization of anchor points of building map, this paper proposes a cell phone indoor scene recognition and localization method with the semantic constraint of building map. Through the attribute association relationship of each element of building indoor space, the building indoor map location anchor map is constructed [32,33]. The map location anchors (MLA) is distributed in the indoor environment of the building along with the step nodes on road network [33]. And in the process of user movement, deep learning is used to obtain more accurate underlying features of the scene map, to achieve semantic recognition of building scenes, to match the recognition results with MLA. At the same time, logical reasoning using scene element matching results realizes the deep integration of information from various parts of perception, semantics, localization, and element management [34,35]. To further obtain more accurate location coordinate information of instantiated scene elements, and achieve semantically constrained indoor scene recognition and localization of building maps for cell phones.

The organization of this paper is as follows. Section 2 describes the construction method of indoor scene recognition and localization method for cell phones with semantic constraints of building maps in detail. Section 3 presents the experiment and results. Section 4 discusses the usability and advantage. Finally, Section 5 concludes this study.

2. Methods

In this section, we will illstrate the implementation of cell phone indoor scene recognition and localization, under the semantic constraints of building maps [33]. Firstly, the effective information is extracted from the BIM model, and the building map and map location anchor points are constructed respectively. The building map model part consists of a solid model and a network model, which are mainly used for the visualization of building information and the abstract expression of topological relationships. The map localization anchor point part consists of two parts: the geometric information location anchors MLA (S) that senses each sensor signal in the fused multi-source sensors, and the geometric location anchors MLA (C) that is regarded as having recognizable elements in scene recognition. Next, the identification method of cell phone indoor scene location anchor point class elements in building maps are proposed. Lastly, the semantic constraint information of the building map is geocoded to match with the recognition results of mobile phone indoor scene elements, to implement real-time positioning and navigation at the cell phone.

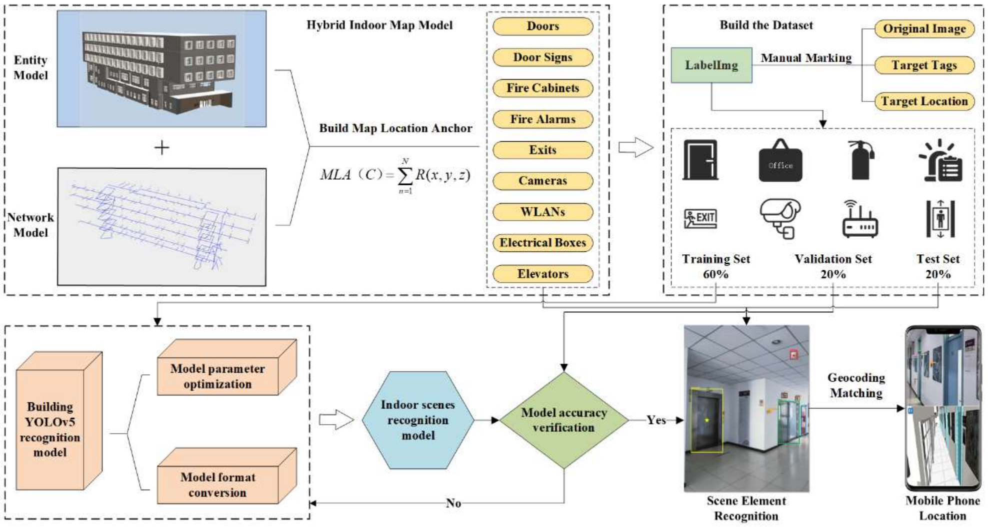


Figure 1. Technical flowchart of indoor scene recognition and localization method for cell phones with semantic constraints of building maps

doi:10.20944/preprints202201.0431.v1

2.1. Semantic Constraint Information Construction of Building Maps

2.1.1. Construction method of building map model

Building map model is the prerequisite for building map semantic constraint recognition scene construction. Firstly, based on the existing 3D building solid model, a 3D building component solid model with the spatial concept is proposed, combined with the texture information collected by UAV tilt photography technology and cell phone close-up photography technology, to improve the advantages of 3D building component solid in the real-world simulation of building map, and at the same time, referring to the spatial representation in the geometric boundary model, select building components that belong to a specific space and have a boundary relationship to represent a certain space, so as to meet the needs of spatial representation of building maps. Secondly, based on the abstract structure of the "edge-node" relationship, the edge-node elements are further divided and organized for the inconsistency of spatial topology relationship in different building scenes. The multi-level organization of main road-secondary road-connection relationship is adopted for abstract representation, and the spatial relationship description of the grid model is referred, to abstract mixed open spaces in buildings, to further satisfy the universality of building maps in spatial topology representation. Lastly, according to the difference of spatial expression between the network model and the solid model, the spatial relationship and semantic association of elements in the network model and the solid model are referred to. Their spatial linkage relationships are formed by a combination of direct and indirect links, to construct an interactive building map model (BIMPN) based on the network model and the solid model [33], to create the digital twin of the building map at the level of refined space and instantiated objects. The construction method of building map model BIMPN is shown in Figure 2.

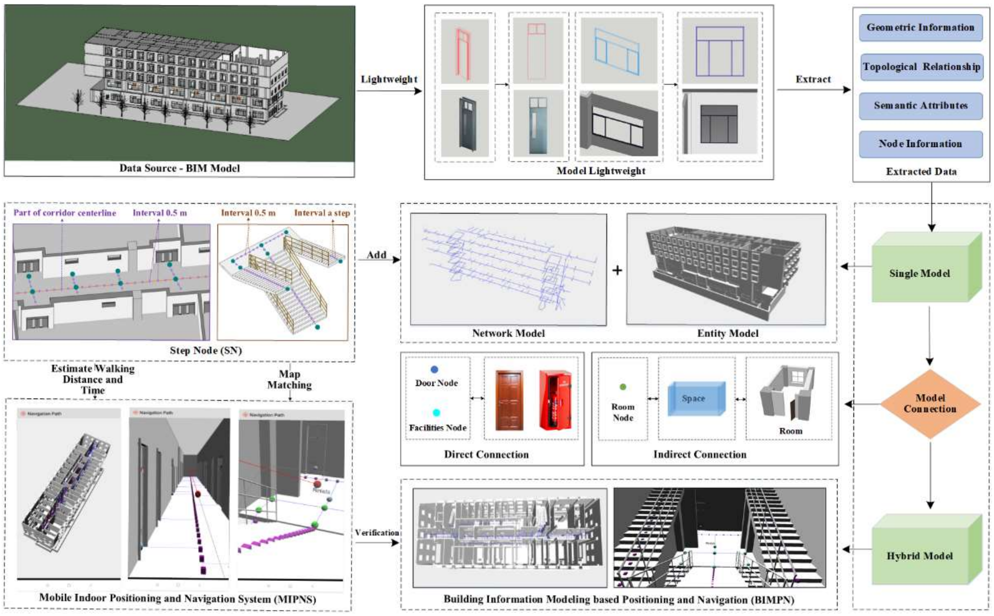


Figure 2. The construction method of building map model BIMPN [33]

2.1.2. Method for constructing map positioning anchor points in building map model

On the one hand, indoor localization can enhance location estimation by building maps and indoor features. On the other hand, it can also leverage the potential value of indoor landmarks, to provide semantic localization capabilities with spatial constraints. This paper constructs MLA for semantic and geometric information representation in each scene of the building map, including geometric information location anchors MLA (S) where the cell phone cooperates with multi-source sensors to sense each sensor‘s signal, and geometric location anchors MLA (C) which are considered as having identifiable elements in scene recognition. First, we selectively construct the semantic information of pervasive accessory facilities (doors, door signs, fire cabinets, fire alarms, safety exits, cameras, WLAN, electrical boxes, elevators, etc.) within the building map, and then obtain the starting position of fused multi-source sensors (such as Bluetooth, etc.) for cooperative positioning through the interface, and associate and record it with the geometric position of scene recognition elements MLA (C) in the map location anchor points for subsequent user positioning and navigation movement process using deep learning algorithms to identify and match the elements of the scene video frame images taken by cell phone cameras. Mathematically we define the map location anchors MLA as shown in (1 and 2):

$$\text{MLA} = \{ S(x,y,z), C(x,y,z) \mid x,y,z \in R \} \tag{1}$$

$$C(x,y,z) = \{ P(x,y,z)\,, \textstyle\sum_{i}^{n} R_i(x,y,z) \mid x,y,z \in R \} \tag{2}$$

In Equation 1, the localization anchor point MLA consists of two parts, S and C. $S(x,y,z)$ represents the geocoded information part of the coordinate position (set) corresponding to the built-in sensor signal feature pattern of the cell phone that can be used for matching localization in the building map, and $C(x,y,z)$ represents the geocoded information part of the coordinate position corresponding to the identifiable pervasive elements in the scene that can be used for matching localization in the building map. In Equation 2, $P(x,y,z)$denotes the position coordinates of the pervasive element location anchor points in the building map that can be used as cell phone video image recognition, $\sum_{i}^{n} R_i(x,y,z)$ denotes the sequence of coordinates of the elements acquired by recognition used in the scene for the matching localization calculation, where $R_i(x,y,z)$ denotes the position coordinates of the ith element acquired by recognition in the scene, n is the number of elements acquired by recognition, and R denotes the real number field. The acquisition of coordinates $P(x,y,z)$ of a current location needs to be solved using the aid of one or more identification elements $R_i(x,y,z)$. The construction of a map location anchor point map is to provide a service interface to the building map engine for the implementation of indoor location navigation and location services for cell phones under semantic constraints.

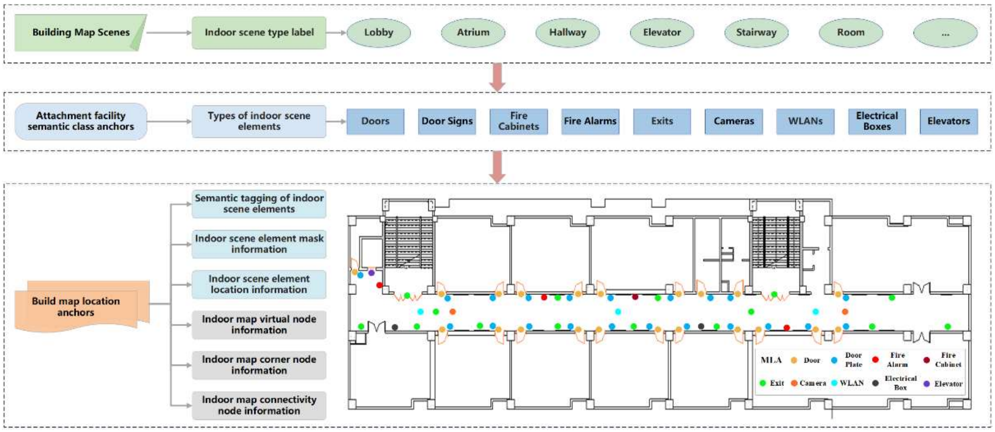

Building Map Scenes
Indoor scene type label
Lobby
Atrium
Hallway
Elevator
Stairway
Room
...
Attachment facility semantic class anchors
Types of indoor scene elements
Doors
Door Signs
Fire Cabinets
Fire Alarms
Exits
Cameras
WLANs
Electrical Boxes
Elevators
Build map location anchors
Semantic tagging of indoor scene elements
Indoor scene element mask information
Indoor scene element location information
Indoor map virtual node information
Indoor map corner node information
Indoor map connectivity node information
MLA
Door
Door Plate
Fire Alarm
Fire Cabinet
Exit
Camera
WLAN
Electrical Box
Elevator

Figure 3. Outline of map positioning anchor point construction method in the building map model

2.2. *Method for identifying indoor scene elements in building map model*

Researchers have now widely deployed applied deep learning recognition models on mobile devices [36,37], and YOLO (You Only Look Once) is one of the SOTA deep convolutional neural models in the field of target detection. YOLOv1 [16] was proposed in 2016, and the latest version is currently YOLOv5 (Glenn Jocher, 2020) [38]. This paper uses the deep learning open source framework Pytorch to model, train, test, validate and deploy the YOLO v5 algorithm on cell phones to achieve the recognition of localization

doi:10.20944/preprints202201.0431.v1

anchor elements in indoor scenes. The YOLOv5 network architecture contains four network models, YOLOv5s [38], YOLOv5m [38], YOLOv5l [38], and YOLOv5x [38]. The main difference between them is the different number of feature extraction modules and convolution kernels at specific locations for each network model, and the sequential increase in the size and number of parameters for each network model. There are nine targets to be recognized for the experiments in this paper, and there are high requirements for the real-time and lightweight nature of this recognition model. Therefore, this paper comprehensively considers the accuracy, efficiency, and size of the recognition model, and ultimately improves the recognition network of building map location anchor elements in indoor scenes based on the YOLOv5s [38] architecture.

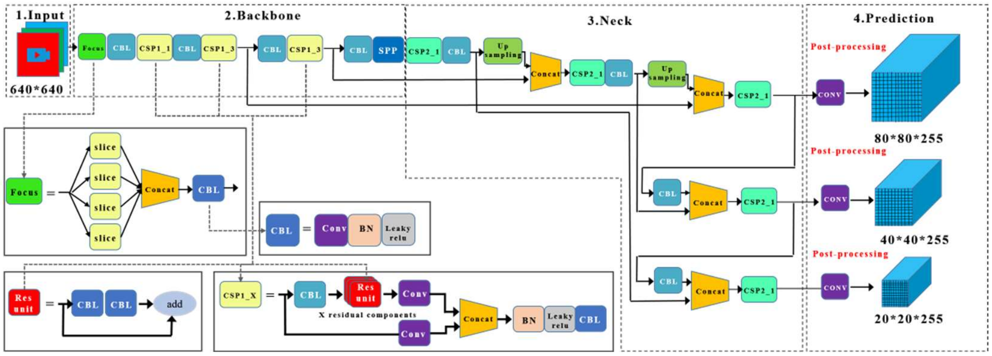


Figure 4. Adaptation of YOLOv5 network structure for scene localization anchor element recognition

As shown in Figure 4, the YOLO v5s [38] architecture mainly consists of four parts: input side, backbone network, neck network, and prediction network. Mosaic data enhancement, adaptive anchor frame calculation, and adaptive image scaling are used on the input side to optimize the input image and reduce the computation to improve the target detection speed. The backbone network is a convolutional neural network that aggregates and forms image features at different image granularity, aiming to reduce the computation of the model and accelerate the training speed. Firstly, using the slice operation, the input three-channel image (3 × 640 × 640) is segmented into four slices, each of size 3 × 320 × 320. Secondly, the four sections are connected in depth using the Concat operation, and the output feature map is of size 12 × 320 × 320. Thirdly, a convolutional layer consisting of 32 convolutional kernels is used to generate a 32 × 320 × 320 output feature map. Lastly, the result is output to the next layer through the BN layer (batch normalization) and the Hardswish activation function. The neck network is a series of feature aggregation layers that mix and combine image features. It is mainly used to generate FPN (Feature Pyramid Network), and then transmit the output feature maps to the detection network (Prediction Network). Since the feature extractor of this network adopts a new PAN structure with enhanced bottom-up paths, improved transmission of low-level features, and enhanced detection of targets at different scales. As a result, the same target object of different sizes and scales can be accurately identified. The prediction network is mainly used for the final prediction of the model, which applies the anchor frame to the feature map output from the previous layer and outputs a vector with the class probability of the target object, the target score, and the location of the bounding box around the target. The prediction network of YOLOv5s [38] architecture consists of three prediction layers and its input is a feature map of dimensions 80 × 80, 40 × 40, and 20 × 20 for detecting image objects of different sizes. In our modified version, multiple arrays are used to store the candidate frame parameters in the post-processing process. We need to remove the original multi label and add the best class only part, and then generate the predicted bounding boxes and target classes in the original image and label them to fit the recognition task of this paper and achieve the detection of building map location anchor point element targets in indoor scene images.

2.3. Method for Mobile phone indoor scene recognition and localization under semantic constraints of building map

As shown in Figure 5, the cell phone indoor scene recognition and localization method under the semantic constraints of building maps mainly includes steps of model quantification, element identification, map matching, and visualization of localization results.

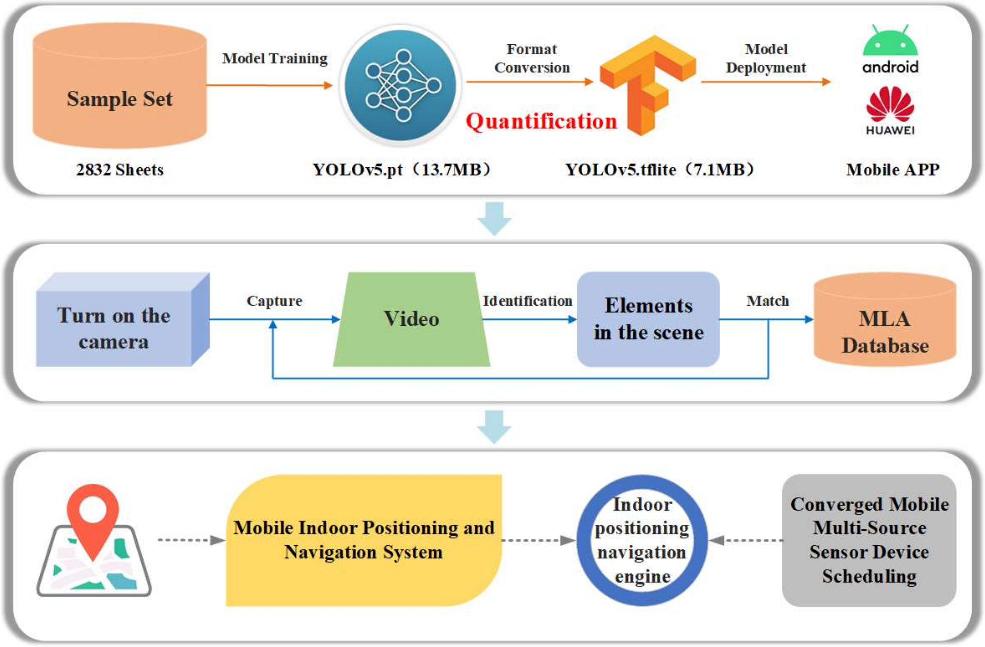


Figure 5. Mobile phone indoor scene recognition and localization method under the semantic constraints of building map

2.3.1. Quantification

In the mobile indoor positioning and navigation system MINPS 2.0 [39], the YOLOv5.pt model is first converted into a tflite model, and the Flatbuffer serialized model file format is used to make it more suitable for mobile piggybacking. At the same time, in order to reduce the computational pressure on the mobile terminal, the model is compressed by quantization, and the weight parameters stored in the model file are converted from Float32 to FP16. The quantization formula is shown below.

$$X_{quantized} = X_{float} \div X_{scale} + X_{zeropint} \tag{3}$$

$$X_{scale} = \frac{X_{float}^{max} - X_{float}^{min}}{X_{quantized}^{max} - X_{quantized}^{min}} \tag{4}$$

$$X_{zeropint} = X_{quantized}^{max} - X_{float}^{max} \div X_{scale} \tag{5}$$

$$X_{float} = X_{scale} \times (X_{quantized} - X_{zeropoint}) \tag{6}$$

Equation (3) is the quantization of the floating-point value to the fixed-point value, and Equation (6) is the inverse quantization of the fixed point value to the floating-point value, where X_{floa} denotes the true floating-point value, $X_{quantized}$ denotes the quantized fixed-point value, X_{scale} denotes the compression ratio of the quantization interval, X_{float}^{max} denotes the maximum floating-point value, X_{float}^{min} denotes the minimum floating-point value, $X_{quantized}^{max}$ denotes the maximum fixed-point value, $X_{quantized}^{min}$ denotes the minimum fixed-point value, and $X_{zeropint}$ denotes the quantized fixed-point value corresponding to the zero floating-point value.

2.3.2. Identification

The quantized model file is deployed to the mobile APP MINPS2.0 [39], and the user takes a video of the scene through the built-in optical camera of the smartphone, as shown in Figure 6. Each frame of the video is used as the input image for scene element recognition, and then performs the element extraction and the real-time fast solution of the 3D coordinates of the elements of the map location anchor points in the scene in the smartphone-side APP. As shown in Figure 7.

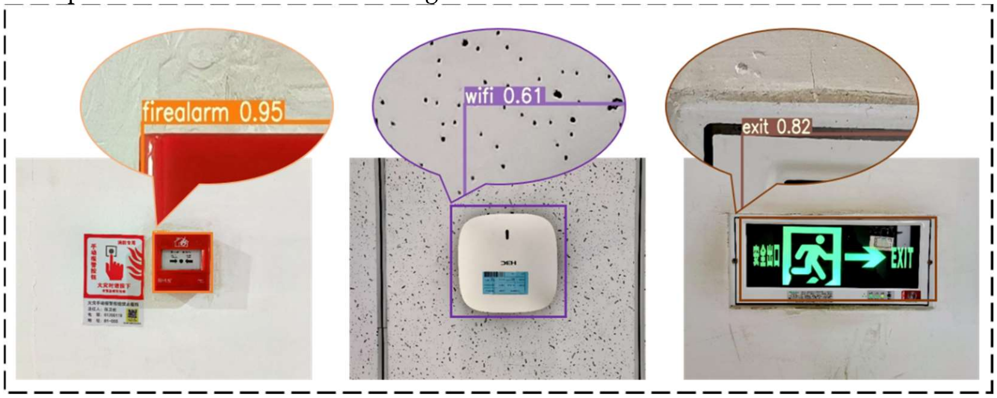


Figure 6. The recognition effect of each element in different scenes of the building

2.3.3. Matching

The mobile app calculates the exact pose of the positioning image locally by using the building map location anchors MLA stored locally in SQLite in advance. Then performs the initial positioning result on the cell phone to match the map with the walking nodes in the building map road network. The initial positioning result point and the building map positioning anchor point are in the same user coordinate system. The distance between the map positioning anchor point and the initial positioning point is calculated based on stereometric and linear algebra, and the unique solution is determined by the P3P algorithm [40]. Next, the nearest walking node to the positioning result point is determined by the calculation as the positioning matching result in the road network. Ultimately, the position information of the localization image is displayed in the user's smartphone, thus realizing the self-localization of the instantaneous positional position of the smartphone camera. The algorithm flow is shown in Figure 7.

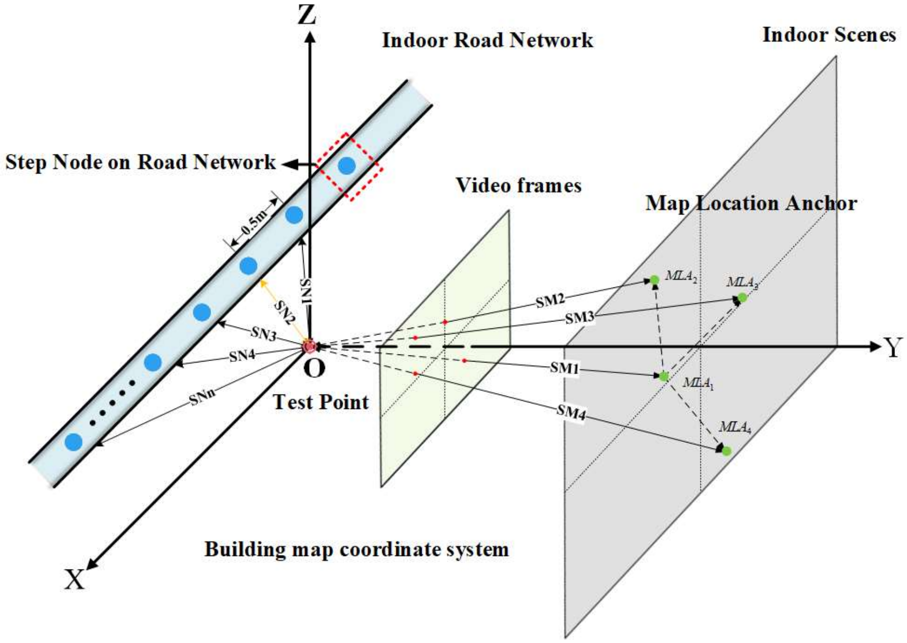


Figure 7. Building map positioning anchor point geocoding matching positioning method

The process of the matching location algorithm is described as follows:

Algorithm 1. Matching location pseudocode for scene element recognition and building map location anchor point geocoding

Input: Initial Bluetooth location points $p0(x0, y0, z0)$; map location anchor points obtained from scene element recognition; road network walking node data set

Output: Positioning point matching position $pt(xt, yt, zt)$; distance error D_e

Steps:

1) Calculate the distance from the map positioning anchor points (the user must determine that there must be and not less than 4 in the scene) obtained from the indoor scene element recognition to the initial Bluetooth positioning point $p0(x0, y0, z0)$. Generally, multiple sets of distance solutions SM_i are obtained until all positioning anchor nodes have been processed and then stop.
2) After obtaining the map positioning anchor point distance, the $P3P$ algorithm is used to determine a unique set of solutions, to obtain the distance SM_t of the user camera, and to obtain the coordinates $pc(xc, yc, zc)$ corresponding to the current walking node SN_c corresponding to SM_t.
3) Create a buffer centered on the current walking node $pc(xc, yc, zc)$ coordinates, the radius of this buffer is the maximum error range E_{max} plus the step size S_l. Use equation (2-7) to obtain the walking node SN_n in the buffer.

$$SN_n = Buffer\ (\ p0\ , radius\ ) \tag{2-7}$$

In the equation, $radius$ is the buffer radius, $radius = E_{max} + S_l$.

4) Calculate the location $pt(xt, yt, zt)$ of the matching locus: the distance from each SM_n to the coordinates, pc of the walking node in the current road network will be calculated, and then the minimum value D_{min} will be obtained from it, and the walking node $pt(xt, yt, zt)$ corresponding to D_{min} will be obtained.
5) Save D_{min} as distance error D_e, count the distance error D_e obtained from each calculation and obtain the

maximum error range E_{max} after adaptive correction, and output D_e and $pt(xt, yt, zt)$.

3. Experiment

3.1. Data

3.1.1. Building map data

The building spatial geometric model is the expression of 3D data to the real world, and also the basis for indoor location-oriented service applications. In this paper, the building F (longitude 116.29606E, latitude 39.751892N) of the School of Geomatics and Urban Spatial Informatics of Beijing University of Civil Engineering and Architecture is used as the experimental area, with an area of about 2800 square meters. The object of the experimental study is a composite solid building, consisting of six floors, five above ground and one underground. The outdoor structure consists mainly of side and top elevations. The interior space distribution includes rich geometric structures such as lobbies, atriums, corridors, elevators, stairs, and rooms, as well as universal signs such as doors, door signs, fire cabinets, fire alarms, safety exits, cameras, WLAN, electrical boxes and elevators. Figure 8(a) shows the construction process and results of the network model, including the construction of the single-level horizontal network and the connection between the horizontal network and the vertical transportation mode. Figure 8(b) shows the construction process and results of the physical model, including the construction of basic elements, marking elements, and thematic elements in the physical model. The construction result of the data of the building map hybrid model BIMPN [33] example is shown in Figure 8(c), while the local situation of the visualization of the building map elements is shown in Figure 8(c). The visualization of building maps and other related data in this paper can be accessed via the Internet [41].

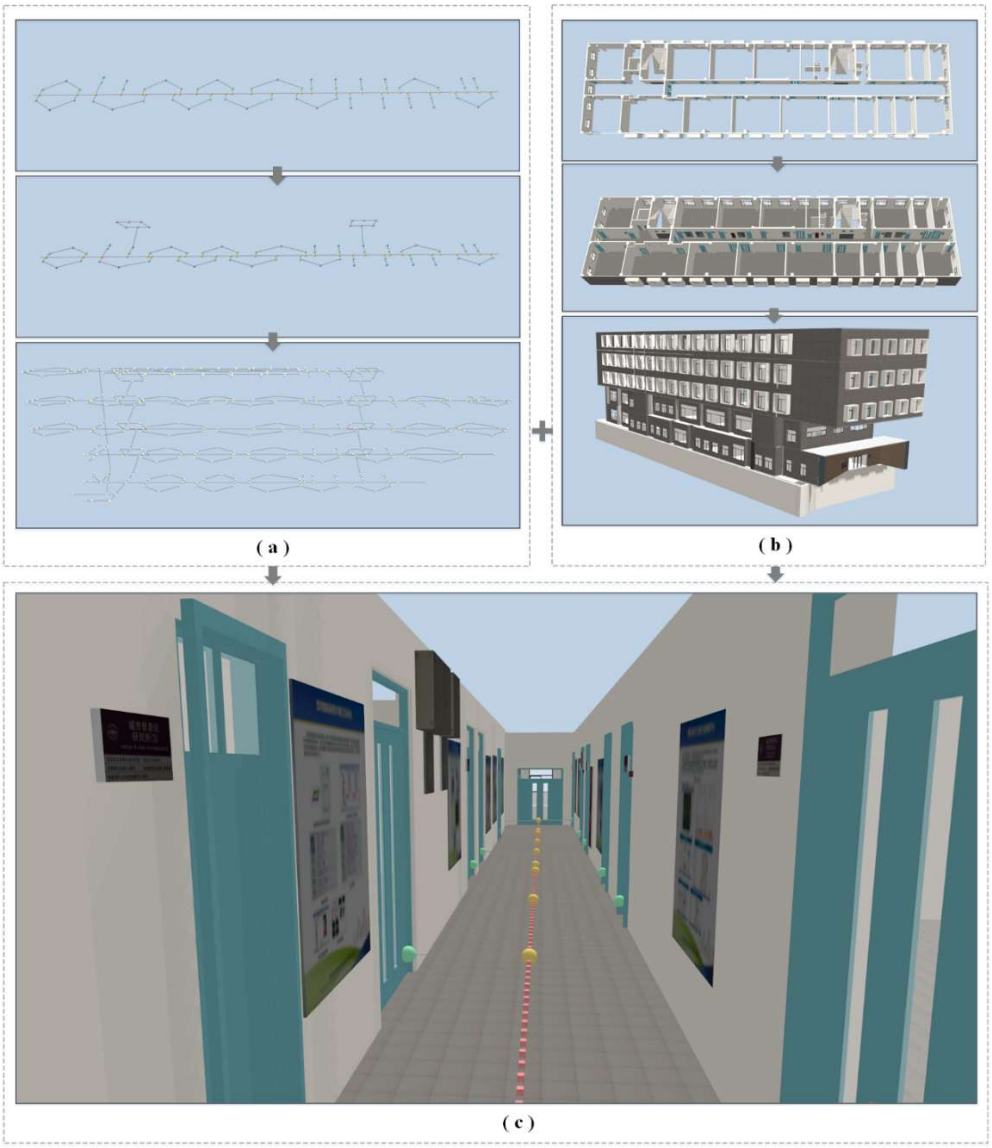
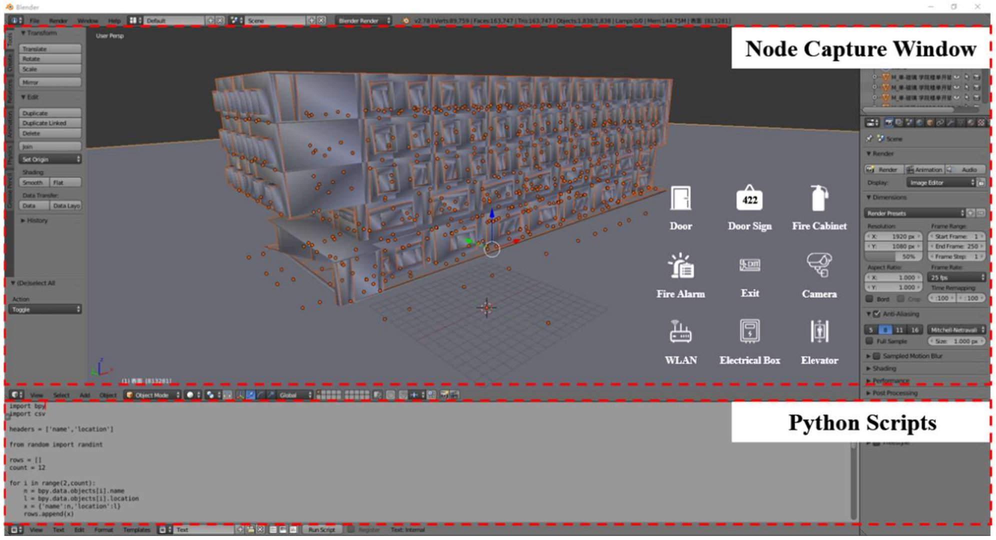


Figure 8. Building map of Building F of the School of Geomatics and Urban Spatial Informatics of Beijing University of Civil Engineering and Architecture [33]

3.1.2. Anchor Point Data

The extraction of map positioning anchor point (semantic) coordinate information required in the building map is processed by Blender software [42], which is an open-source cross-platform 3D production software toolkit that supports a series of operations such as modeling, animation, materials, rendering, node capture, etc., as shown in Figure 3-1-2. At present, Blender does not support the IFC format, so it needs to export the BIM model built by Revit to the universal 3D format FBX and import it into Blender V2.78 for capturing the map positioning anchor points. The map positioning anchor points to be captured in this paper mainly include the geometric centers of universal building components such as doors, door signs, fire cabinets, fire alarms, security exits, cameras, WLANs, electrical boxes, elevators, etc. The capture is processed automatically by the Python script developed in this paper.



```
import bpy
import csv

headers = ['name','location']

from random import randint

rows = []
count = 12

for i in range(2,count):
    n = bpy.data.objects[i].name
    l = bpy.data.objects[i].location
    x = {'name':n,'location':l}
    rows.append(x)
```

Figure 9. Extraction of coordinate information of map positioning anchor points (semantic) in the building map

3.1.3. Element sample dataset

For the current study, a dataset of building indoor scene elements that can be used for YOLOV5 model training is needed. We consider geometric location points with certain identifiable elements in scene recognition as map localization anchor points. However, the public datasets are not suitable for the objectives of this study. Therefore, it is necessary to customize and build pervasive accessory facility datasets within the building map to continue this research. We have selected nine types of building indoor scene pervasive elements as the identification targets for this project. The building map localization anchor elements are doors, door signs, fire cabinets, fire alarms, security exits, cameras, WLAN, electrical boxes, and elevators. Pictures with universal elements in the building scenes need to be taken and collected, with different angles and distances according to the user's pose requirements during recognition, in order to build the element information required by the recognition algorithm. The dataset has a total of 2832 images and 7610 element target samples, as shown in Table 1.

Table 1. Sample subset of building interior scene features.

Categories	Number of elemental samples	Percentage
Doors	2460	32.33%
Door Signs	1200	15.77%
Fire Cabinets	450	5.91%
Fire Alarms	540	7.10%
Exits	1380	18.14%
Cameras	320	4.20%
WLANs	440	5.78%
Electrical Boxes	660	8.67%
Elevators	160	2.10%
Total	7610	100%

3.2. Experimental results

3.2.1. Building map and map location anchor point construction results

In this paper, the semantic information of map localization anchor points required in the building map is extracted based on the FBX format data of the target building (F building of the School of Geomatics and Urban Spatial Informatics, Beijing University of Civil

Engineering and Architecture), which is obtained by using Python scripting tool in Blender [42] software environment, stored as CSV format data and then imported into PostgreSQL database, and the data statistics are shown in Table 2, with a total of 586 element semantic constraints for building map positioning anchor points.

Table 2. Statistics of Map Location Anchor s for semantic constraints of building maps.

F-building	Doors	Door Signs	Fire Cabinets	Fire Alarms	Exits	Cameras	WLANs	Electrical Boxes	Elevators	Total
B1	39	11	6	4	14	5	1	8	1	89
F1	36	11	4	4	14	5	4	3	1	82
F2	37	11	3	5	16	4	6	3	1	86
F3	46	1	4	4	14	4	7	4	1	85
F4	50	19	4	5	15	4	7	5	1	110
F5	47	33	4	7	18	6	7	11	1	134
Total	255	86	25	29	91	28	32	34	6	586

3.2.2. Recognition results of indoor scene elements of cell phones

In the experiments of cell phone video frame recognition of building indoor scene elements, the improved YOLOv5s model is used for the recognition of building indoor elements. All experiments in this study are implemented in the framework of Torch 1.7.0, driven by CUDA, running on a single NVIDIA GeForce RTX 3070 GPU, with the specific hyperparameter information shown in Table 3.

Table 3. Hyperparameters information.

Hyperparameters	Values
GPU_COUNT	1
CFG	Yolov5s.yaml
Data	Scence.yaml
Weights	Yolov5s.pt
Unm-Classes	9
Epochs	1000
Batch Size	32
Img Size	640*640
Evolve	true
Cache images	true
Single cls	false

The experiments use 2256 images (video frames) from a sample set of 2832 images for training and 288 images for validation and testing in the field. Experiment 1 uses the YOLOv5s model for training and takes 63 hours 3 minutes and 20 seconds to complete 1000 epochs. Experiment 2 uses the improved YOLOv5s model for training and takes 64 hours and 10 minutes to complete 1000 epochs. The quantitative comparison of the models in terms of precision, recall, mAP@0.5, and mAP@0.5:0.95 is shown in Figure 10. The blue corresponds to the YOLOv5s model, and the orange corresponds to the improved YOLOv5s model.

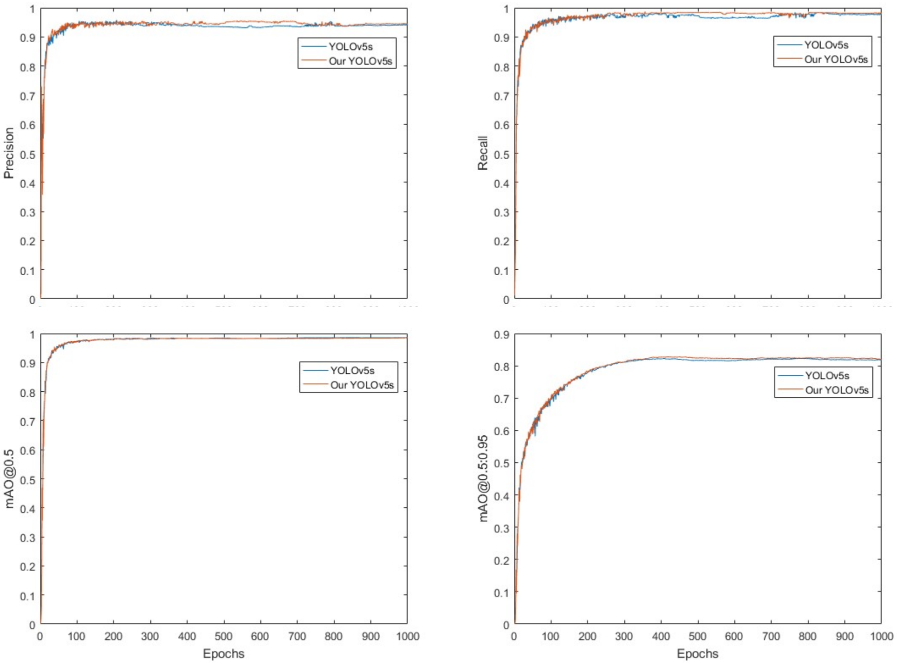


Figure 10. Comparison of performance metrics between YOLOv5s and modified YOLOv5s during training

As shown in Figure 10, in terms of the speed of recognition performed by the scene video, the total duration of a video 758 frames is about 25s, the recognition time of the YOLOv5s model is 23.793s (31.858 frames/s), and the recognition time of the improved YOLOv5s model is 22.818s (33.219 frames/s), the results show that in terms of recognition speed the improved model can achieve the effect of real-time availability. In terms of accuracy, the improved YOLOv5s model is slightly better than the original model overall, and it is obvious that the improved model is better than the original model in 500 to 720 epochs. The result shows that the improved model recognition effect is more suitable for the application of such scenes mainly due to the influence of the type (single door, double door, glass door, fire door, etc.) and complexity of the door. The learning performance of the model gradually improves with iterations, and the convergence speed is very fast, and the curve has stabilized by 1000 epochs. The experiments in this paper use the training results of 1000 epochs to demonstrate, and the actual production and engineering applications can be adjusted and optimized based on the actual situation.

The loss function describes the performance of a given predictor in classifying the input data points in a dataset. The smaller the loss, the better the classifier is at modeling the representation of the relationship between the input data and the output target. Figure 11 plots the effect of two different types of losses, which represent losses related to the predicted bounding box and losses associated with a given cell containing objects during training. ValBox and valObjectness plots represent their validation scores, with training losses measured in the middle of each stage and validation losses measured after each stage. The results show that the improved YOLOv5s model loss function is smoother and converges faster than the original model loss function, which is more suitable for the application in the scenario of this paper.

doi:10.20944/preprints202201.0431.v1

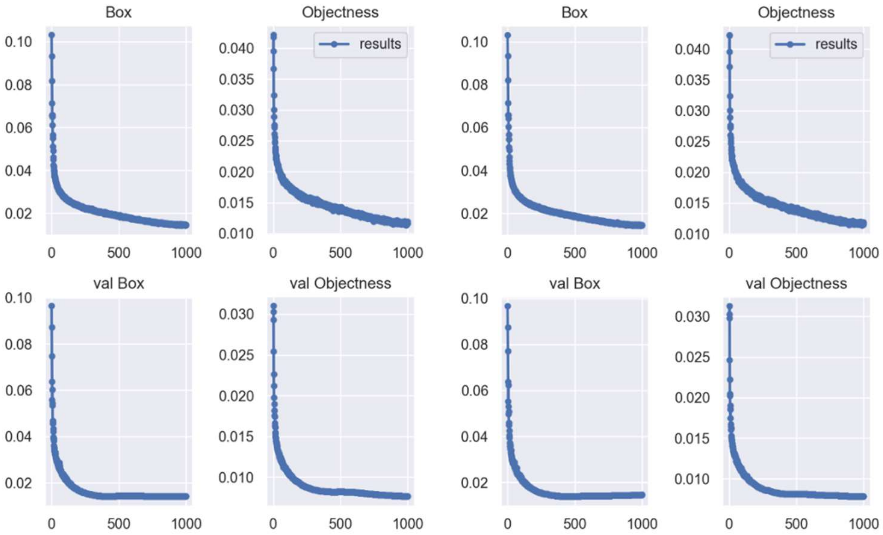


Figure 11. Comparison of performance metrics between YOLOv5s and improved YOLOv5s during training

Figure 12 gives an example of some of the results of element recognition for indoor scenes of buildings under different lighting and angle conditions. The proposed model in this paper is not only applicable to detecting the elements of interest captured in each frame of the scene video when the line of sight is in frontal view but also to localize the anchor elements captured under the condition that the line of sight is shifted by a certain angle during walking. In addition, using the proposed model, the localization anchor elements can also be well-identified under the conditions of sunny daytime, dusk, and indoor lighting at night.

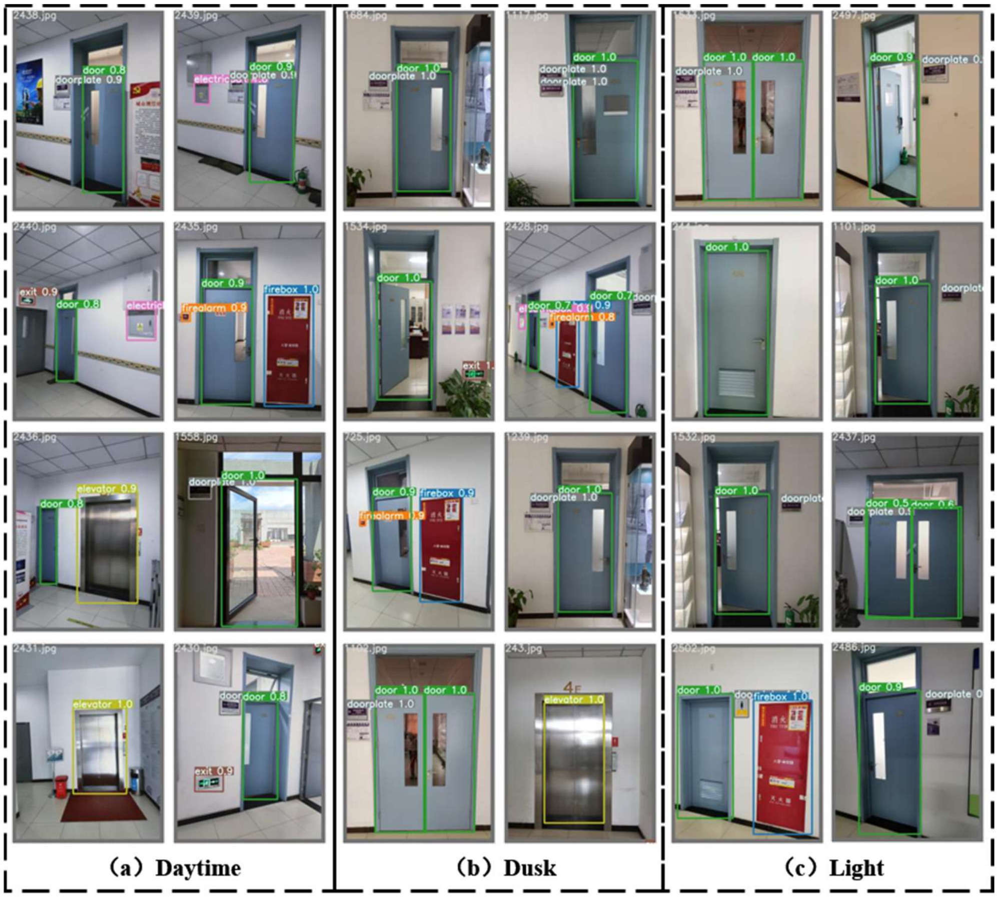


Figure 12. Recognition results of building interior elements using the improved YOLOv5s network in different time series

3.2.3. Localization results of indoor scene recognition for cell phones

The goal of the experiment is to verify that this method has good localization results under the constraints of map localization anchor point information and in buildings with rich spatial structure semantic information. The system focuses on indoor scene localization under the condition of the known motion starting point. The user starting position is obtained by Bluetooth and fused multi-source sensor localization, which is input to this method as a known condition. Figure 13 shows the visualization effect of real-time positioning starting from a certain starting position in the building area. Yellow is the trajectory of the user walking along the corridor path, and blue is the trajectory depicted after the video scene element identification localization anchor point and the building map road network node (corridor centerline) for map matching. When the input video data can be solved in real-time to output accurate positioning coordinates, it will be matched with the road network walking nodes to obtain the fusion results of positioning points and road network and draw the segment trajectory map. The experimental results show that the richer the semantic constraint information in the building map scene and the richer the element information obtained from the element recognition in the field video, the more information that can be matched between the identifiable elements of the building space scene and the anchor points of the building map positioning, and the higher the accuracy of the completed positioning in the scene walking will be.

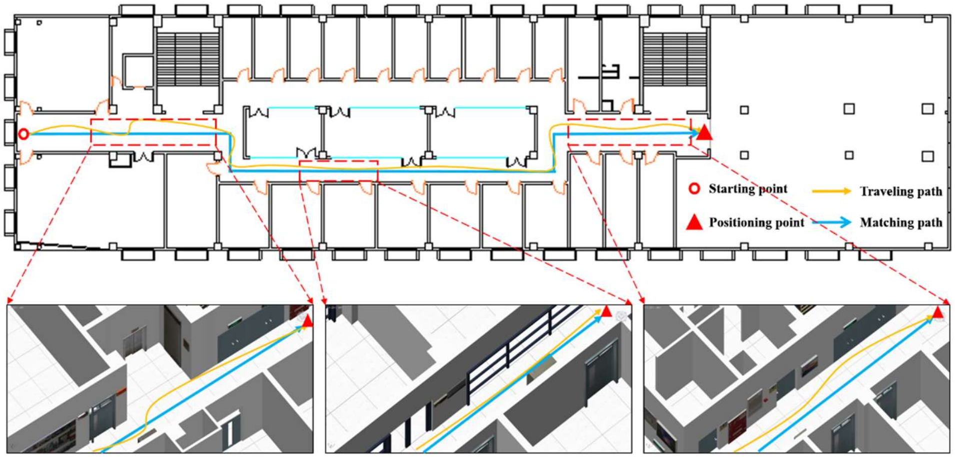


Figure 13. Localization results of indoor scene recognition for cell phones with semantic constraints of building maps

In order to analyze the effectiveness of this method quantitatively, a total of 103 coordinate points were collected during the experimental matching positioning process, and the deviations from the x and y directions of the matching coordinates of the road network are shown in Figure 14. The deviation points are mainly concentrated in the x negative half-axis. Since the user will face the camera toward the semantic information-rich wall in the corridor scene during the recognition process through the cell phone camera, and thus will be closer to the opposite semantic information-less wall, resulting in the x direction deviation is mostly negative. Because the corner direction is the direction where the y-axis is located and the user will temporarily miss the semantic information constraint points in the building during the cornering process, the y-direction deviation is larger than the x-direction deviation. The experiments do not measure the deviations in the z-direction. The z value of the final positioning point coordinates is the z value of the matching walking node SN (Step Node), and the 10 pairs of point coordinates with the largest deviations in the x and y directions of the path coordinates are selected from them for typicality analysis.

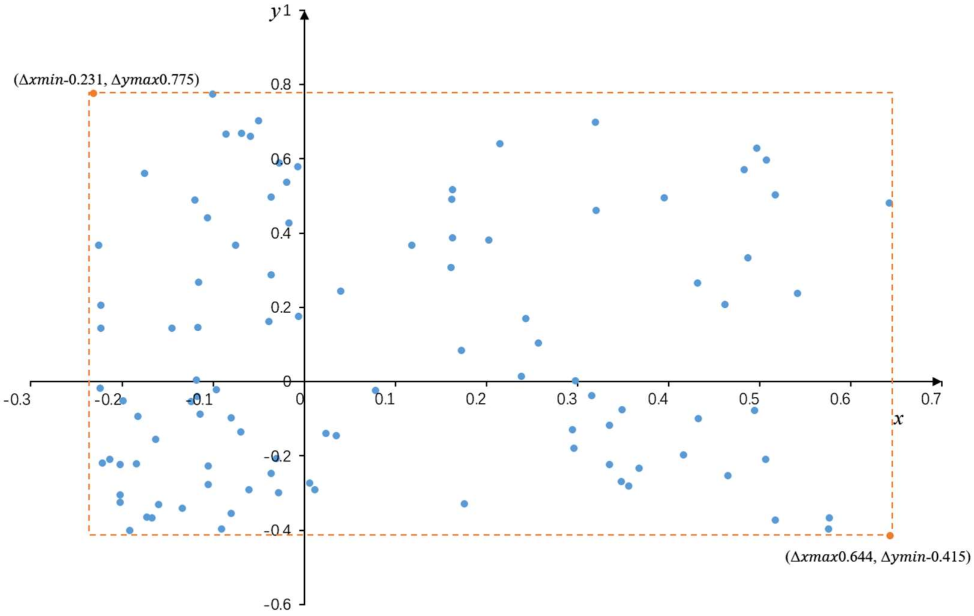


Figure 14. Coordinate deviation statistics of pedestrian walking trajectory and map matching trajectory

As shown in Tables 4, the quantified analysis of the x and y coordinate deviations of the coordinate point pairs shows that the maximum interval of deviation variation is $\Delta x \in [-0.231,0.644\,]$, $\Delta y \in [-0.415,0.775]$. The analysis shows that the large deviation is a result of less information of identifiable elements within the field of view of pedestrians at the corner. Since the span of accuracy unit scale (m VS cm) between the arbitrary oscillation of pedestrians during walking (meter level) and the deviation of recognition algorithm (centimeter-level) is large, the error range of this method is controlled in the maximum range which is acceptable in practice. Therefore, the visualization of the guidance information in the form of matching scene recognition positioning anchor points with road network nodes does not cause any disturbance to the user's positioning and navigation process. The method has good feasibility and engineering application value.

Table 4. Statistics of coordinate deviation between pedestrian walking trajectory and map-matched trajectory (partial).

Track point number	Coordinates of pedestrian walking track points	Map Matching Track Point Coordinates	Deviation values （$\Delta x, \Delta y$）
Starting Point	(-12.145,21.343)	(-12.0,21.2,17.550)	(-0.145,0.143)
1	(-3.231,21.975)	(-3.0,21.2,17.550)	(-0.231,0.775)
2	(29.401,21.474)	(29.5,20.7,17.550)	(-0.099,0.774)
3	(7.950,18.903)	(8.0,18.2,17.550)	(-0.050,0.703)
4	(39.931,21.868)	(40.0,21.2,17.550)	(-0.069,0.668)
5	(31.915,21.867)	(32.0,21.2,17.550)	(-0.085,0.667)
6	(38.941,21.860)	(39.0,21.2,17.550)	(-0.059,0.660)
7	(-0.858,21.681)	(-1.5,21.2,17.550)	(0.642,0.481)
8	(4.077,19.333)	(3.5,19.7,17.550)	(0.577,-0.367)
9	(21.017,17.828)	(20.5,18.2,17.550)	(0.517,-0.372)
10	(0.144,20.785)	(-0.5,21.2,17.550)	(0.644,-0.415)
End Point	(41.661,21.507)	(41.5,21.2,17.550)	(0.161,0.307)

4. Discussion

In this paper, a building map semantic constrained cell phone indoor scene recognition and localization method is proposed. The scene element recognition method is based on the improved YOLOv5 model, where the element information in the building scene is recognized in real-time through the cell phone camera, and then the map location anchor points with geographic coordinates are matched. This paper constructs MLA with universal scene elements in building interior, so the scene element recognition model does not need to manually research a lot of element information of other building interior scenes, and it does not need to maintain or update the scene element recognition information for a long time, therefore this method is less dependent and more universal in multi-application scenes. The comparison experiments show that the improved YOLOv5s network model outperforms the YOLOv5s model in identifying nine different types of pervasive element anchors in building scenes, and the recall rate in the test set is consistently above 97.2%, indicating that the method is suitable for indoor scenes of buildings with rich scene element information.

This paper constructs map location anchors based on the geometric and semantic information of building spatial elements to provide spatial semantic constraints for scene element recognition results. The elements of building map location anchor points are all types of elements with universal characteristics, and these elements generally have a long life cycle after the building is put into use, and thus have the advantages of stability and long-term availability. At the same time, the location anchor point of the building map proposed in this paper contains not only the geometric location anchors MLA(C), which is considered as a recognizable element in scene recognition, but also the geometric information location anchors MLA(S), which is used by the cell phone to sense the signal of each sensor in cooperation with multi-source sensors and can be applied to assist the cooperative localization method of other built-in sensors of the cell phone, so as to achieve the effect of cooperative localization application for complex multi-scene. The experimental results show that map location anchor points can provide very effective reference coordinate location information in the process of cell phone video recognition localization. In addition, the data sample collection scheme is oriented to the geometric and semantic constraint process of building map model, so the method in this paper can easily implement a crowdsourcing-based approach to aggregate building location anchor data, and efficiently integrate indoor scene data from different buildings to form a shared building map sample library.

The experiments match the element information obtained by recognition with the map location anchors MLA in the SQLite database to locate the position of the constrained user in the road network. The maximum interval of the deviation change of the scene element recognition matching localization method is $\Delta x \in [-0.231, 0.644]$, $\Delta y \in [-0.415, 0.775]$, which is within the acceptable range of the arbitrary oscillation (meter level) error during the pedestrian walking process, and the real-time matching process of this method can eliminate the error in the early pedestrian movement without cumulative error generation, which significantly enhances the robustness of the method calculation process. In addition, the building indoor scene recognition model on the cell phone not only provides input video data but also can quickly retrieve the building map data source locally on the mobile side, which is a significant advantage of offline recognition and map matching quickly on the mobile side. This method not only allows real-time browsing of realistic holographic maps of buildings with real feelings on the cell phone but also facilitates the further enhancement of related applications utilizing AR-enhanced semantic element information in building maps, etc.

5. Conclusions and Future Work

In this paper, we propose an indoor scene recognition and localization method for cell phones with semantic constraints of building maps. This paper provides semantic constraint information for indoor positioning by constructing a geocoded entity library of building map location anchor points (MLA), and then identifies the semantic constraint element information in the scene based on the improved YOLOv5s model, and matches the identified element information with the database map location anchor points MLA, and lastly constrains the location of the user in the road network corresponding to the location information from the scene element feature points, thus Realize real-time positioning and navigation. The experimental results show that the improved YOLOv5s model network model can identify 9 different types of pervasive element anchors in building scenes by comparison, and the recall rate is consistently above 97.2% in the test set, and the method can be extended and applied to other building map models, and the maximum localization error is within the range of 0.775 m, and up to about 0.5 m after applying the BIMPN road network walking node constraint, which can effectively achieve high positioning accuracy in the building scenes with rich MLA element information.

The solution proposed in this paper is not a solution that particularly requires indoor environmental data. The video for scene element recognition is obtained through cell phone camera shooting, and the key to cell phone scene element recognition is an efficient lightweight network model. In the future, it is necessary to consider a more efficient and robust generalized training element anchor model, and apply it to more complex and large-scale environments. The main goal is to interact building maps with augmented reality and to visually represent the semantic information in building maps, thus providing more accurate and richer services to users for real-time location navigation.

Funding: This research is funded by the National Natural Science Foundation of China (No. 41301489), Beijing Natural Science Foundation (No. 4192018, No. 4142013), Outstanding Youth Teacher Program of Beijing Municipal Education Commission (No. YETP1647, No. 21147518608), Outstanding Youth Researcher Program of Beijing University of Civil Engineering and Architecture (No. 21082716012), the Fundamental Research Funds for Beijing Universities (No. X20099, No. X18282, No. X20077), and Beijing University of Civil Engineering and Architecture Education Scientific Research Projects (No. Y2111).

References

1. Zhang, D.; Xia, F.; Yang, Z.; et al. Localization Technologies for Indoor Human Tracking [J]. *IEEE Communications Surveys & Tutorials*. **2010**, *11*, 1-6.
2. Chen, R.Z.; & Chen, L. Smartphone-Based Indoor Positioning Technologies [C]. *Urban Informatics*. **2021**.
3. MMD, H.; Orozco-Barbosa, L.; García-Varea, I. A smartphone-based multimodal indoor tracking system [J]. *Information Fusion*. **2021**, *76(6)*, 36-45.
4. Delfa, G.; Catania, V.; Monteleone, S.; et al. Computer Vision Based Indoor Navigation: A Visual Markers Evaluation [J]. *Advances in Intelligent Systems & Computing*. **2015**, *376*, 165-173.
5. Martin-Gorostiza, E.; Garcia-Garrido, M.A.; Pizarro, D.; et al. Infrared and Camera Fusion Sensor for Indoor Positioning[C]. *2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. **2019**.
6. Qin, F.; Zuo, T.; Wang, X. CCpos: WiFi Fingerprint Indoor Positioning System Based on CDAE-CNN [J]. *Sensors*. **2021**, *21(4)*, 1114-1131.
7. Lie, M.; Kusuma, G.P. A fingerprint-based coarse-to-fine algorithm for indoor positioning system using Bluetooth Low Energy [J]. *Neural Computing and Applications*. **2021**, *33(7)*, 2735-2751.
8. Dawood, M.A.; Saleh, S.S; El-Badawy, E.; et al. A comparative analysis of localization algorithms for visible light communication [J]. *Optical and Quantum Electronics*. **2021**, *53(2)*, 108-133.
9. Niu, X.; Li, Y.; Kuang, J.; et al. Data Fusion of Dual Foot-Mounted IMU for Pedestrian Navigation [J]. *IEEE Sensors Journal*. **2019**, *99*, 1109-1119.
10. Sowmya, V.; Govind, D.; Soman, K.P. Significance of processing chrominance information for scene classification: a review [J]. *Artificial Intelligence Review*. **2020**, 53(2):811-842.
11. Ding, X.; Luo, Y.; Yu, Q.; et al. Indoor object recognition using pre-trained convolutional neural network [C]. *2017 23rd International Conference on Automation and Computing (ICAC). IEEE*. **2017**.
12. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks [J]. *Science*. **2006**, *313(5786)*, 504-507.
13. Kai, C.; Pang, J.; Wang, J.; et al. Hybrid Task Cascade for Instance Segmentation [C]. *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE*. **2019**.
14. Lee, Y.; Park, J. CenterMask: Real-Time Anchor-Free Instance Segmentation [C]. *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE*. **2020**.
15. Liang, j.; Homayounfar, N.; Ma, W.C.; et al. PolyTransform: Deep Polygon Transformer for Instance Segmentation [C]. *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE*.**2020**.
16. Redmon, J.; Divvala, S.; Girshick, R.; et al. You Only Look Once: Unified, Real-Time Object Detection [C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE*. **2016**.
17. Liu, Y.; Wang, L.; Liu, M. YOLOStereo3D: A Step Back to 2D for Efficient Stereo 3D Detection [J]. *2021 IEEE International Conference on Robotics and Automation (ICRA)*. **2021**, *95(6)*, 1423-1432.
18. Sun, P.; Zhao, Y.; Zhu, S. An approach to improve SSD through mask prediction of multi-scale feature maps [J]. *Pattern Analysis and Applications*. **2021**, *24(3)*, 1357-1366.
19. Yang, L.; Zhang, X.; Wang, L.; et al. Lite-FPN for Keypoint-based Monocular 3D Object Detection [J]. *Computer Vision and Pattern Recognition*. **2021**, *21(5)*, 268-278.
20. Zhang, K.; Lv, G.; Wu, L.; et al. LadRa-Net: Locally-Aware Dynamic Re-read Attention Net for Sentence Semantic Matching [J]. *IEEE Transactions on Neural Networks and Learning Systems*, pp. (99)1-14
21. Elgendy, I.A.; Zhang, W.Z.; He, H.; et al. Joint computation offloading and task caching for multi-user and multi-task MEC systems: reinforcement learning-based algorithms[J]. *Wireless Networks*, **2021**, *27(3)*, 2023-2038.
22. Liu, P.; Zhang, Z.; Wu, L.; et al. Fingerprint-Based Indoor Localization Algorithm with Extended Deep Belief Networks[C].*2020 Information Communication Technologies Conference (ICTC)*. **2020**.

23. Chen, Y.; Du, T.; Jiang, C.; et al. Indoor location method of interference source based on deep learning of spectrum fingerprint features in Smart Cyber-Physical systems[J]. *EURASIP Journal on Wireless Communications and Networking*. **2019**, *2019*, 47-59.
24. Cheng, R.; Wang, K.; Bai, J. et al. Unifying Visual Localization and Scene Recognition for People With Visual Impairment [J]. *IEEE Access*. **2020**, *8*, 64284-64296.
25. Lin, H.; Peng, L.; Chen, S.; et al. Indexing for Moving Objects in Multi-Floor Indoor Spaces That Supports Complex Semantic Queries[J]. *ISPRS International Journal of Geo-Information*. **2016**, *5(10)*, 1-30.
26. Yan, Z.; Zheng, X.; Xiong, H.; et al. Robust Indoor Mobile Localization with a Semantic Augmented Route Network Graph [J]. *ISPRS International Journal of Geo-Information*. **2017**, *6(7)*, 221-245.
27. Mortari, F.; Clementini, E.; Zlatanova, S.; et al. An indoor navigation model and its network extraction [J]. *Applied Geomatics*. **2019**, *7(2)*, 273-288.
28. Guo, R.; Chen, Y.; Zhao, Z.; et al. A Theoretical Framework for the Study of Pan-Maps [J]. *Journal of Geomatics*, **2021**, *46(1)*, 9-15.
29. Elhamshary, M.; Youssef, M. SemSense: Automatic construction of semantic indoor floorplans[C]. International Conference on Indoor Positioning & Indoor Navigation. *IEEE*. **2015**, *9978*, 1-11.
30. Gu, F.; Hu, X.; Ramezani, M.; et al. Indoor Localization Improved by Spatial Context—A Survey [J]. *ACM Computing Surveys (CSUR)*. **2019**, *52(3)*, 1-35.
31. Hu, X.; Fan, H.; Noskov, A.; et al. Feasibility of Using Grammars to Infer Room Semantics [J]. *Remote Sens*. **2019**, *11(13)*, 1535-1561.
32. Gu, F.; Valaee, S.; Khoshelham, K.; et al. Landmark Graph-Based Indoor Localization [J]. *IEEE Internet of Things Journal*. **2020**, *7(9)*, 8343-8355.
33. Liu, J.; Luo, J.; Hou, J.; et al. A BIM Based Hybrid 3D Indoor Map Model for Indoor Positioning and Navigation [J]. *International Journal of Geo-Information*. **2020**, *9(12)*, 747-768.
34. Chen, S.; Liu, J.; Liang, X.; et al. A Novel Calibration Method between a Camera and a 3D LiDAR with Infrared Images [J]. *IEEE International Conference on Robotics and Automation*. **2020**, *10(11)*, 4963-4969.
35. Li, M.; Chen, R.; Liao, X.; et al. A Precise Indoor Visual Positioning Approach Using a Built Image Feature Database and Single User Image from Smartphone Cameras [J]. *Remote Sens*. **2020**, *12(5)*, 869-893.
36. Qin, W.; Song, T.; Liu, J.; WANG, H.W.; LIANG, Z.; Remote Sensing Military Target Detection Algorithm Based on Lightweight YOLOv3 [J]. *CEA*. **2021**, *57(21)*, 263-269.
37. Wang, D.; He, D.; Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning [J]. *Biosystems Engineering*. **2021**, *210(6*), 271-281.
38. Ultralytics. Available online:https://github.com/ultralytics/yolov5.
39. MINPS2.0, http://www.dxkjs.com:8080/BuildingMap/apkversionupdate/MIPNSv2.0.apk, **2022**, www.dxkjs.com.
40. Bai, L.; Yang, Y.; Feng, C.; et al. A Novel Received Signal Strength Assisted Perspective-three-Point Algorithm for Indoor Visible Light Positioning [J]. *Optics Express*. **2020**, *28(19)*, 1162-1175.
41. MLA Building F, http:// www.dxkjs.com:8080/BuildingMap/MLA/F.html, **2022**, www.dxkjs.com.
42. Blander. Available online:https://www.blendercn.org.