Review Article

Liu Jianhua*, Feng Guoqiang, Luo Jingyan, Wen Danqi, Chen Zheng, Wang Nan, Zeng Baoshan, Wang Xiaoyi, Li Xinyue, and Gu Botong

Mobile phone indoor scene features recognition localization method based on semantic constraint of building map location anchor

https://doi.org/10.1515/geo-2022-0427
received June 05, 2022; accepted October 05, 2022

Abstract: Visual features play a key role in indoor positioning and navigation services as the main semantic information to help people understand the environment. However, insufficient semantic constraint information and mismatching localization without building map have hindered the ubiquitous application services. To address the problem, we propose a smartphone indoor scene features recognition localization method with building map semantic constraints. First, based on Geographic Information System and Building Information Modeling techniques, a geocoded entity library of building Map Location Anchor (MLA) is constructed, which is able to provide users with "immersive" meta-building-map and semantic anchor constraints for mobile phone positioning when map matching. Second, using the MYOLOv5s deep learning model improved on indoor location scenario, the nine types of ubiquitous anchor features in building scenes are recognized in real time by acquiring video frames from the smartphone camera. Lastly, the spatial locations of the ubiquitous indoor facilities obtained using smartphone video recognition are matched with the MLA P3P algorithm to achieve real-time positioning and navigation. The experimental results show that the MLA recognition accuracy of the improved MYOLOv5s is 97.2%, and the maximum localization error is within the range of 0.775 m and confined to the interval of 0.5 m after applying the Building Information Modeling based Positioning and Navigation road network step node constraint, which can effectively achieve high positioning accuracy in the building indoor scenarios with adequate MLA and road network constraint.

Keywords: mobile phone indoor positioning, scene recognition, building map, map location anchor, geocoding matching

1 Introduction

Buildings, such as office buildings, libraries, shopping centres, hospitals, train stations, and airports, are the main space for human activities. Humans spend about 87% of their time in indoor spaces [1]. 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 [2]. 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 positioning systems have been proposed by researchers, who 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 limited application scenarios when considering accuracy, cost, coverage, complexity, and applicability. 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

* **Corresponding author: Liu Jianhua,** College of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing, 100044, China; Key Laboratory for Urban Geomatics of National Administration of Surveying, Mapping and Geoinformation, Beijing, 100044, China, e-mail: liujianhua@bucea.edu.cn, tel: +86-10-68322237, personal website: http://www.dxkjs.com.
Feng Guoqiang: College of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing, 100044, China; The First Geographic Information Mapping Institute, Ministry of Natural Resources, Xi'an 710054, China
Luo Jingyan: Department of Hydraulic Engineering, Fujian College of Water Conservancy and Electric Power, Fujian, 366000, China
Wen Danqi, Chen Zheng, Wang Nan, Zeng Baoshan, Wang Xiaoyi, Li Xinyue, Gu Botong: College of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing, 100044, China

services more challenging. The current visual localization technology incorporating multisource 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.

With the development of deep learning technology, many scholars currently have incorporated deep learning into the technical solutions for indoor positioning and navigation in the research fields of indoor positioning technology. Recognition algorithms are used to effectively process the semantic information contained in image data in order to extract image features and determine the effectiveness of obtaining valid information about the category to which the scene image belongs [10–13]. At present, the image quality, pixel resolution, sensor, and aperture performance of the video frames obtained by the smartphone camera have been significantly improved. And with the rapid development of artificial intelligence, smartphone camera sensors have gradually added intelligent ant vibration, super night scene, backlighting, and other auxiliary functions to make the video image clearer. A research direction with greater potential is the use of more efficient and suitable for ubiquitous indoor facilities recognition and geocoding of the semantic constraints of the building map information to assist and then achieve smartphone camera scene recognition and localization.

Building map is an effective type of information representation of interior spatial features, in which semantic information can better represent the user's scene [14]. Landmark anchors and contextual information in building maps can better understand user movement rules, perceive user scenarios, correct indoor positioning errors, and plan indoor navigation paths [15]. With the development of Building Information Modeling (BIM) building operation system, the information of indoor subsidiary features is getting richer and richer to meet the universal scene element recognition method, but 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 features, and construct and improve building map models for mobile phone indoor positioning and navigation with universal applications, is a key problem that needs to be solved urgently [16].

Aiming at City Information Modeling (CIM) and metaphase application scenarios, the indoor positioning landmark recognition algorithm still has the problem of an incomplete category system which hinder the steps. In this article, from the perspective of a building map hybrid model, we integrate Geographic Information System and BIM to propose a building map indoor location anchors classification system and combine State-of-the-Art (SOTA) indoor scene recognition algorithm for mobile phone indoor positioning.

The main contributions of this paper can be summarized as follows:

(1) Inspired by the geometric and semantic constraint information of building map, the building Map Location Anchor (MLA) for indoor scene element identification is constructed through the attribute association relationship of each geocoding element in the building indoor space. By constructing MLA, the pervasive scene element recognition method can be satisfied from the perspective of building map element classification.
(2) Based on the SOTA YOLOv5s model, our improved MYOLOv5s MLA recognition model for mobile phone indoor scene element recognition is proposed for different indoor scenarios. Since the pre-trained MYOLOv5s based on the ubiquitous indoor facilities sample library is relatively stable, it can be used as a powerful spatial features extractor in all kinds of geocoding building MLA recognition.
(3) We propose a mobile phone localization method with ubiquitous recognition of the representative nine types of features in indoor scenes for semantic constraints of building maps. In MLA-rich building scenarios, the geocoding anchor coordinate information in MLA is obtained through MYOLOv5s model identification and position solved by MLA P3P algorithm, then it is matched to the Building Information Modeling based Positioning and Navigation (BIMPN) road network Step Node (SN); in MLA-sparse building scenarios, the user position also can be spatially constrained in the road network SN by BIMPN.

The organization of this study is as follows. Section 2 provides a brief overview of related work. Section 3 details the specific implementation of indoor scene recognition for building map mobile phones and its application for indoor localization and navigation. Section 4 presents the experiment and results. Section 5 discusses the usability and advantage. Lastly, Section 6 concludes this study.

2 Related work

In order to provide ubiquitous location-based service from smartphone, the work in this article covers a number of areas such as map anchor-assisted localization

based on road networks, visual pose estimation, and scene recognition.

Map anchor-assisted indoor localization algorithms use the geometric and semantic information of the map as a constraint to correct the resultant errors in the solution of trajectory and location coordinates of pedestrians walking in buildings and to improve the accuracy of indoor positioning of pedestrians in buildings [17]. Bandyopadhyay et al. improved the accuracy of the system location navigation solution by mapping inertial trajectories with magnetometer or compass data to user-generated location estimates, combined with information such as road signs and floors of building maps; however, the method cannot determine the initial user heading information [18]. Zhou et al. proposed ALIMC, an indoor mapping system based on active landmarks, which can automatically construct indoor maps for buildings without any a priori knowledge [19]. Shang et al. proposed APFiLoc, a low-cost, smartphone-based framework for indoor localization. It detects organic landmarks in the environment in an unsupervised manner and uses enhanced particle filters to fuse them with measurements from smartphone sensors and map information for indoor localization [20]. Li et al. present a new holistic vision-based mobile assistive navigation system that develops an indoor map editor to parse the geometric information of building models and generate a semantic map consisting of a global 2D traversable grid map layer and a context-aware layer [21]. Gu et al. proposed LGLoc, a landmark map-based indoor positioning method for mobile phones, which constructs an initial landmark map consisting of landmarks such as stairs, lifts, corners, and turns in the offline phase, and the latest collected data from the user's smartphone sensors for location initialization, location estimation and location calibration in the online phase [22]. However, the performance of these systems is highly dependent on the integrity of the landmark anchor points, which will be badly affected by desertions of landmarks and usually mismatches in landmark anchor points can lead to large positioning errors.

Existing visual localization systems are usually divided into three main steps: image matching, pose solving, and coordinate solving, of which pose solving is the most critical step. A 2DTriPnP algorithm for querying camera poses using the Google Street View image database for localization was first proposed by Sadeghi et al., and the method uses Perspective n-Point (PnP) geometry to solve the feature matching problem in visual localization [23]. On this basis, numerous researchers have proposed improvements for solving the problem of P3P pose. Lepetit et al. have both successively proposed a noniterative method for efficiently solving PnP, the EPnP algorithm, and the method represents n coordinate points by a weighted sum of four coordinate points, which can reduce the complexity and achieve high accuracy [24]. Based on the EPnP algorithm, Kneip et al. proposed an UPnP algorithm that does not require the acquisition of smart sensor parameters in advance, which eliminates the need for extensive manual annotation in the early stage [25]. In the current research fields of computing the bit pose of video image sensors, numerous scholars have incorporated algebraic methods into visual localization coordinate-solving schemes [26], which can reduce the error between solving the PnP problem. The PnP algorithm typically requires at least four known coordinate position points for a unique positive solution, but usually, the view of the average user's mobile phone is limited. Therefore, the demand for rich, numerous, and more evenly distributed landmark anchor points in the application scenario is also a challenge to the coverage properties of the localization algorithm.

Scene recognition not only obtains semantic information about the building features but also helps to understand contextual information in other related vision tasks such as object detection or behavior recognition. Deep neural networks have shown significant advantages in feature extraction and scene recognition, making it possible for users to provide low-cost, high-precision location navigation applications [27,28]. Liu et al. proposed a method for constrained mobile localization of indoor scenes in buildings, by which users can upload the pictures of the scene taken by the mobile phone to the server for identification, and use the particle filter algorithm to fuse other sensor data of the mobile phone for positioning [29]. Obviously, with the increasing picture dataflow of the mobile localization tools available to the huge mobile Internet users, it is hard to afford the computation cost for the server to answer immediately. Shuang et al. proposed a method to fuse scene recognition results and PnP algorithm for indoor localization of buildings, and the method is a highly accurate and stable solution for indoor localization by solving the coordinate information of camera sensors with multiple reference points in the indoor space of a building [30]. Liu et al. combined the YOLOv3 model for natural scene recognition, which is effective in detecting small targets and reduces the training time and speed significantly [31]. The model style of YOLO takes the benefits of miniature and lightweight, compact size, which utterly meets the requirements to deploy it on smartphone to cut the battery energy consumption and server dataflow efficiently. In our MLA P3P algorithm, we calculate the current geocoding anchor position from MLA recognition by the modified model

MYOLOv5s, which will adapt the P3P algorithm to dynamic video frames.

In summary, address to the problems of insufficient semantic constraint information of building map and matching positioning of geocoding anchors for building maps, this study proposes a method for mobile phone indoor scene recognition and localization with the semantic constraint of building maps. The geocoding anchors of building indoor map are constructed through the association relationship of the attributes of each element of building indoor space [22]. MLA with road network SNs [32] are distributed in the indoor environment of buildings, and deep learning model MYOLOv5s is used to obtain more accurate scene features during user movement to achieve semantic recognition of building scenes, match the recognition results with MLA, and at the same time use the scene element matching results to logical reasoning to achieve deep integration of the information of each part of perception, semantics, localization, and element management [33,34]. Furthermore accurate location coordinate information of instantiated scene features is obtained to achieve semantically constrained indoor scene recognition and localization for smartphones with building maps.

3 Methods

In this section, we will illustrate the implementation of mobile phone indoor scene recognition and localization, under the semantic constraints of building maps [32]. First, the effective information is extracted from the BIM model from the perspective of CIM, and the building map and MLA 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 MLA consists of two parts: the geometric information location anchors MLA (S) that senses each sensor's signal in the fused multi-source sensors, and the sematic location anchors MLA (C) that is regarded as having recognizable features of geocoding elements in the scene recognition procedure. Next, we propose the identification method of MLA features in building maps. Lastly, the semantic constraint information of the MLA in the building map is geocoded to match the recognition results of indoor scene features, to implement real-time positioning and navigation at the smartphone terminal (Figure 1).

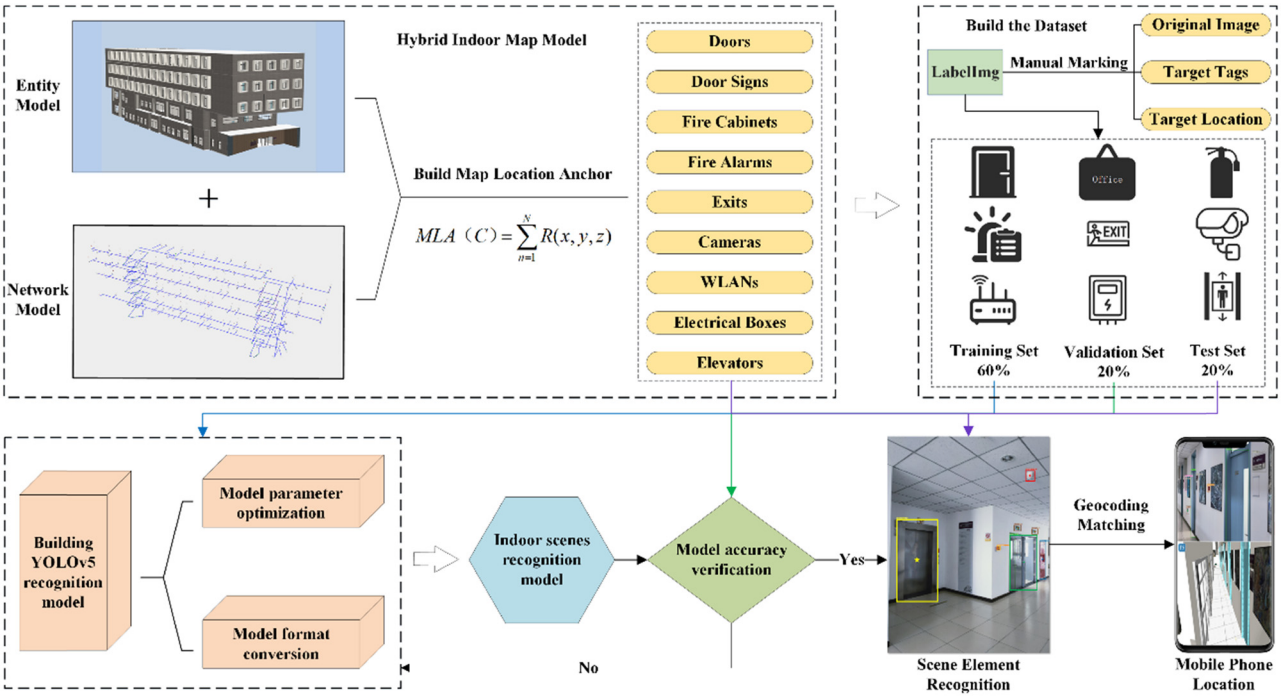

Figure 1: Technical procedure of indoor scene recognition localization method for mobile phone with semantic constraints of MLA building maps.

3.1 Semantic constraint information construction of building maps

3.1.1 Construction method of building map model

Building map model is the prerequisite for building map semantic constraint recognition scene construction. First, a 3D building component solid model is proposed, which is combined with the texture information collected by UAV tilt photography technology and smartphone. Considering the spatial representation in the geometric boundary model, we select building components that share boundary relationships to constitute a specific space. Second, based on the abstract structure of the "edge-node" relationship in network model, the building features are classified and organized in the spatial network topology relationship. Lastly, according to the difference in spatial expression between the network model and the solid model, the spatial relationship and semantic association of features in the network model and the solid model are merged to produce the hybrid building map model BIMPN [32]. The spatial linkage relationships of features among the solid model and network model are formed by a combination of direct and indirect links to create the digital twin of the building map at the level of refined space and instantiated objects. The construction procedure of building map model BIMPN is shown in Figure 2 [32].

3.1.2 Constructing MLA in building map

Indoor localization can enhance location estimation by building maps and indoor features. Furthermore, it can also leverage the potential value of indoor landmarks, to provide semantic localization capabilities with spatial constraints. This article constructs MLA for semantic and geometric information representation in each scene of the building map, including geometric information location anchors MLA (S) where the smartphone cooperates with multi-source sensors to sense each inner sensors' signal, and geometric location anchors MLA (C) which are considered as having identifiable features 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) for cooperative

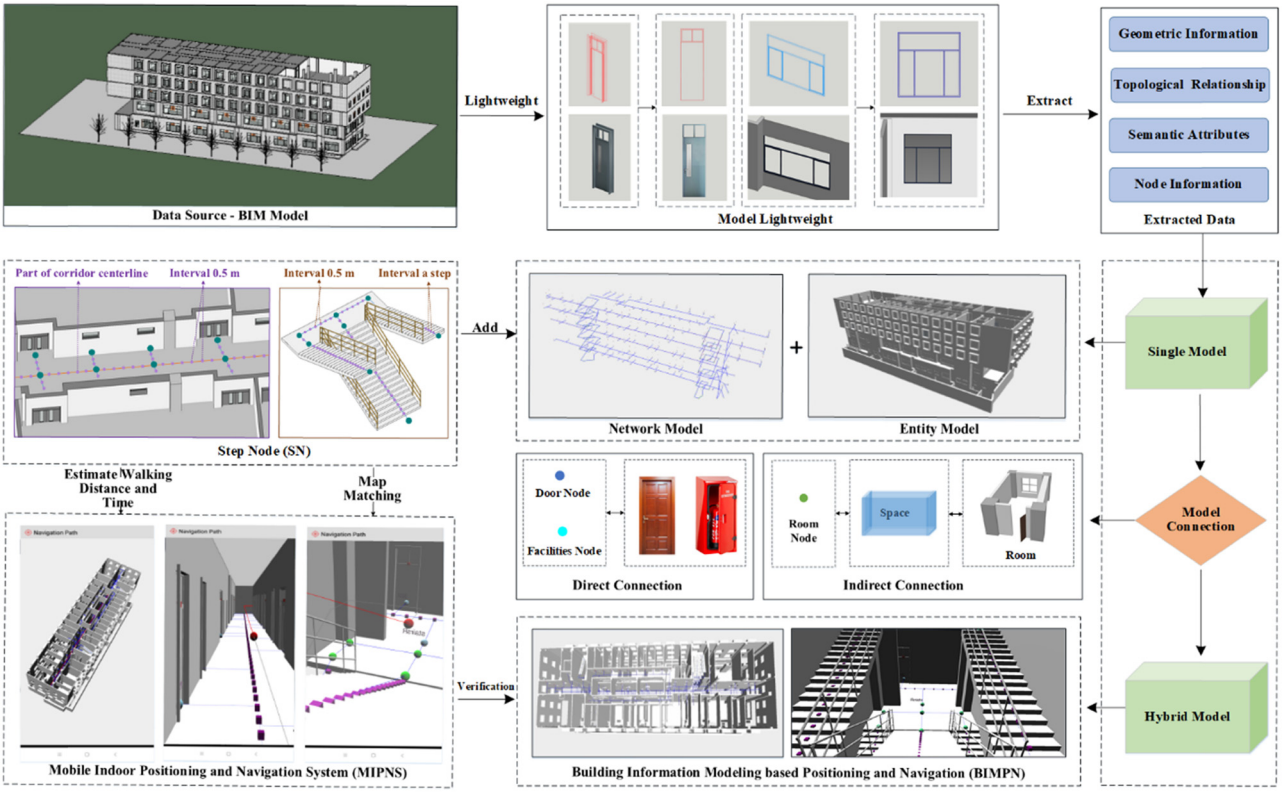

Figure 2: The construction procedure of building map model BIMPN [32].

positioning through the software interface API, and associate and record it with the geometric position of scene recognition features MLA (C) for subsequent user positioning and navigation movement process. Deep learning model MYOLOv5s is used to identify and match the features of MLA in the scene video frames taken by smartphone cameras. We define the MLA as shown in functions (equations (1) and (2)):

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

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

In equation (1), the MLA consists of two parts, S and C. $S(x, y, z)$ represent the geocoded information part of the coordinate position (set) corresponding to the built-in sensor signal feature pattern of the mobile phone that can be used for matching localization in the building map. $C(x, y, z)$ represents the geocoded information part of the coordinate position corresponding to the identifiable pervasive features 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 anchors in the building map that can be used for smartphone video image recognition. $\sum_i^n R_i(x, y, z)$ denotes the sequence of coordinates of the features 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 features 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 features $R_i(x, y, z)$. The construction of the MLA is to provide a service interface to the building map engine for the implementation of indoor location navigation and location-based services for mobilephone under semantic constraints (Figure 3).

3.2 Recognition of indoor scene MLA features in building map

Researchers have widely deployed and applied deep learning recognition models on mobile devices [35,36], and YOLO [37,38] is one of the SOTA deep convolutional neural models in the field of target detection. This article uses the deep learning opensource framework PyTorch to model, train, test, validate, and deploy the modified YOLOv5 algorithm on smartphone to achieve the recognition of location anchor features in indoor scenes. The SOTA 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 types of MLA features to be recognized for the experiments, and there need high requirements for the real-time and lightweight nature of this recognition model. This paper comprehensively considers the accuracy, efficiency, and size of the recognition model, and ultimately

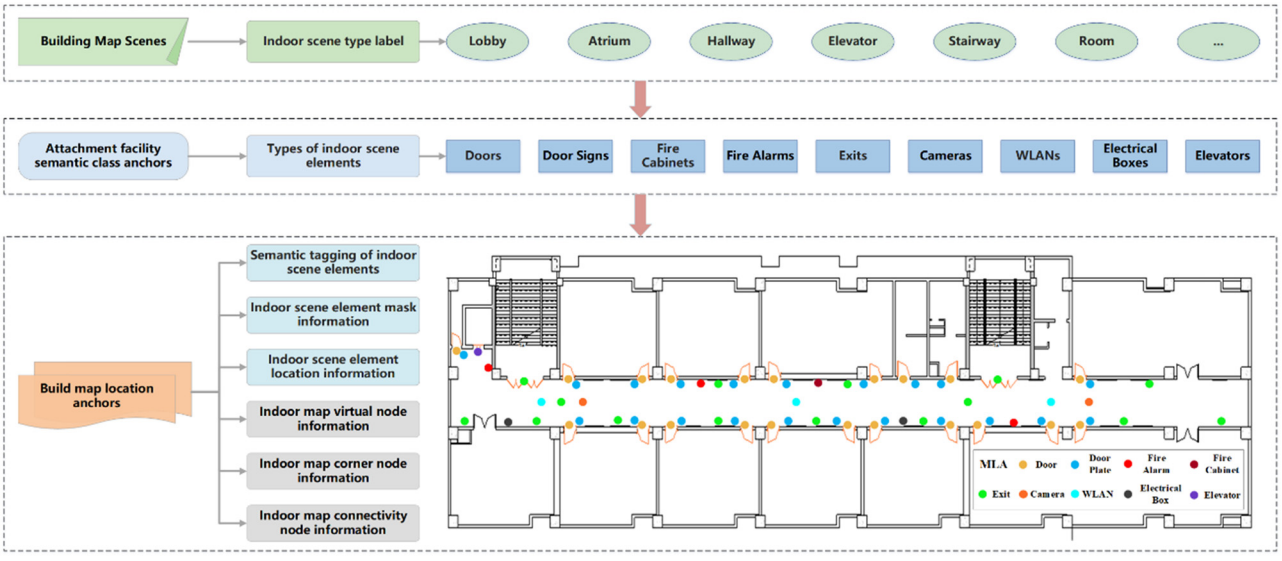

Figure 3: Outline of MLA construction procedure based on the building map model.

improves the recognition network of building MLA features in indoor scenes based on the modified YOLOv5s [38] architecture.

As shown in Figure 4, the MYOLO 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 cut the computation cost 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 accelerate the training speed. First, using the slice operation, the input three-channel image ($3 \times 640 \times 640$) is segmented into four slices, each of size ($3 \times 320 \times 320$). Second, the four sections are connected in depth using the Concat operation, and the output feature map is of size ($12 \times 320 \times 320$). Third, a convolutional layer consisting of 32 convolutional kernels is used to generate a ($32 \times 320 \times 320$) output feature map. Lastly, the result is output to the next layer through the BN (batch normalization) layer 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 Feature Pyramid Network (FPN) and then transmit the output feature maps to the detection network (Prediction Network). Since the feature extractor of this network adopts a new Pixel Aggregation Network (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 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.

Considering factors such as computational cost (related to battery energy consumption) and acceptable MLA recognition model size (related to limited storage space and internal CPU) for deployment on smartphones, multiple arrays are used to store the candidate frame parameters in the post-processing process. Meanwhile, we remove the original multi-label layers and only add the best class part, and then generate the predicted bounding boxes and target classes in the original image. Finally, we label them to fit the task of recognizing architectural MLA element targets in indoor scene images, and we define the improved model as MYOLOv5.

3.3 Mobile phone indoor localization under semantic constraints of building map MLA

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

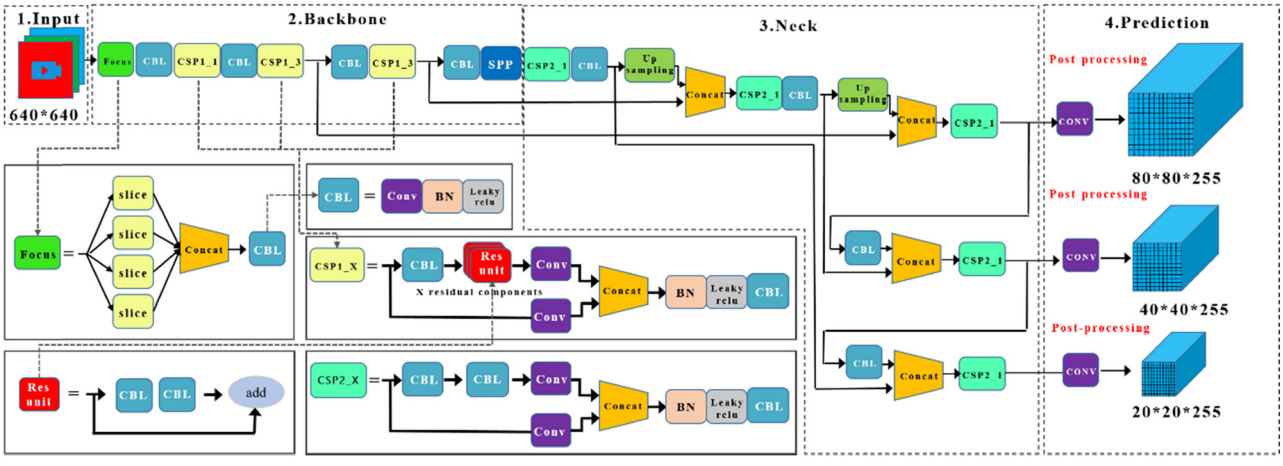

Figure 4: Adaptation of MYOLOv5 network structure for scene location anchor element recognition.

3.3.1 Quantification

In the proposed mobile indoor positioning and navigation system MIPNS 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 cost 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.

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

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

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

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

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_{float} denotes the true floating-point value, $X_{\text{quantized}}$ denotes the quantized fixed-point value, X_{scale} denotes the compression ratio of the quantization interval, $X_{\text{float}}^{\max}$ denotes the maximum floating-point value, $X_{\text{float}}^{\min}$ denotes the minimum floating-point value, $X_{\text{quantized}}^{\max}$ denotes the maximum fixed-point value, $X_{\text{quantized}}^{\min}$ denotes the minimum fixed-point value, and X_{zeropint} denotes the quantized fixed-point value corresponding to the zero floating-point value.

3.3.2 Identification

The quantified model file is deployed to the smartphone APP MIPNS2.0 [39], and the user shoots the scene video through the built-in optical camera of the smartphone, and each frame of the video is used as the input image for scene element recognition. As shown in Figure 6(a)–(i) for nine types of features: door, door sign, fire cabinet, fire alarm, security exit, camera, WLAN, electric box, and elevator. The text and values in the top left corner are the probability that the element belongs to that category. The recognition of the features is carried out on the APP MIPNS2.0 [39], and the 3D coordinates of the anchor points of the map positioning features in the scene are quickly solved in real time, as shown in Figure 7.

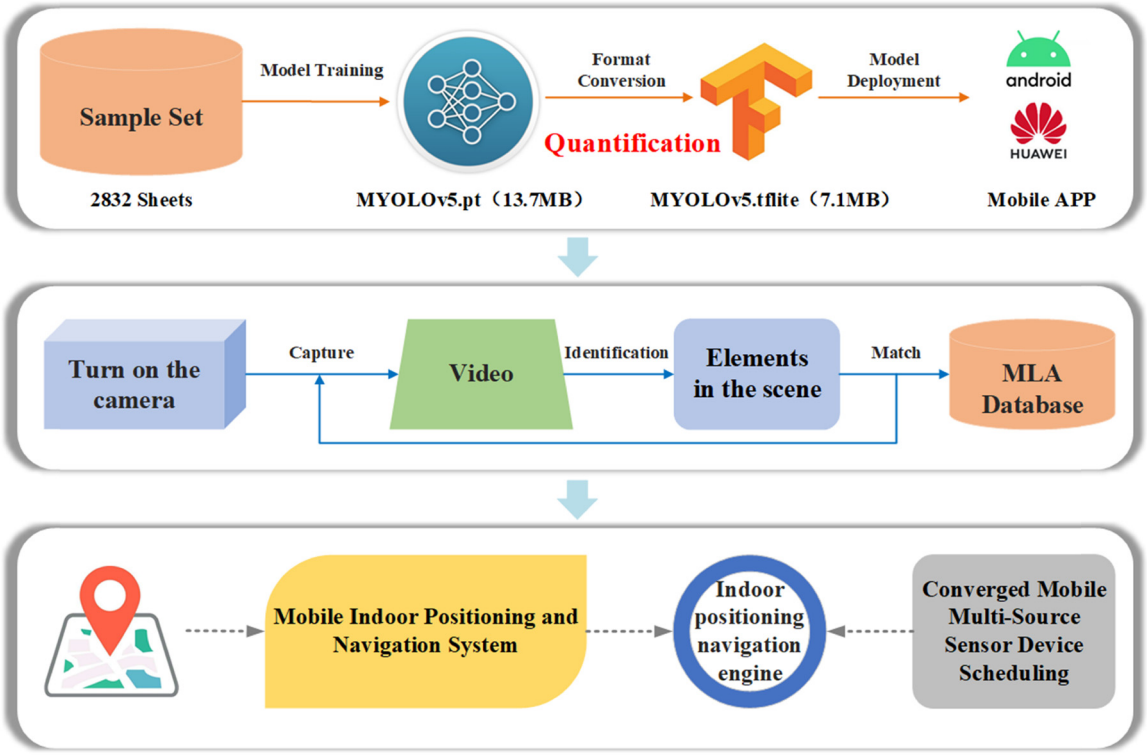

Figure 5: Mobile phone indoor scene recognition localization procedure under the semantic constraints of building map.

3.3.3 Matching

The mobile APP MIPNS2.0 [39] matches the recognized features by using the building MLA locally stored in SQLite which is downloaded from Server. Then, the mobile APP MIPNS2.0 performs the initial positioning result on the mobile phone to match the SN in the building map road network. The initial positioning result and the building MLA are in the same user coordinate system. The distance between the MLA and the initial positioning point is calculated using the P3P algorithm [40]. Next, the nearest SN [32] to the positioning result point is determined by the calculation as the positioning matching result in the road network. Ultimately, the position is displayed in the user's mobile phone, thus realizing the instantaneous localization of the smartphone camera. The algorithm flow is shown in Figure 7.

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

Algorithm 1. Matching localization pseudo code for scene element recognition and building MLA geocoding

Input: Initial Bluetooth location point $p0(x0, y0, z0)$; MLA obtained from scene element recognition; road network SN data set
Output: Positioning point matching position $pt(xt, yt, zt)$; distance error D_e
Steps:

Figure 6: The recognition results of MLA features in different scenes of the building.

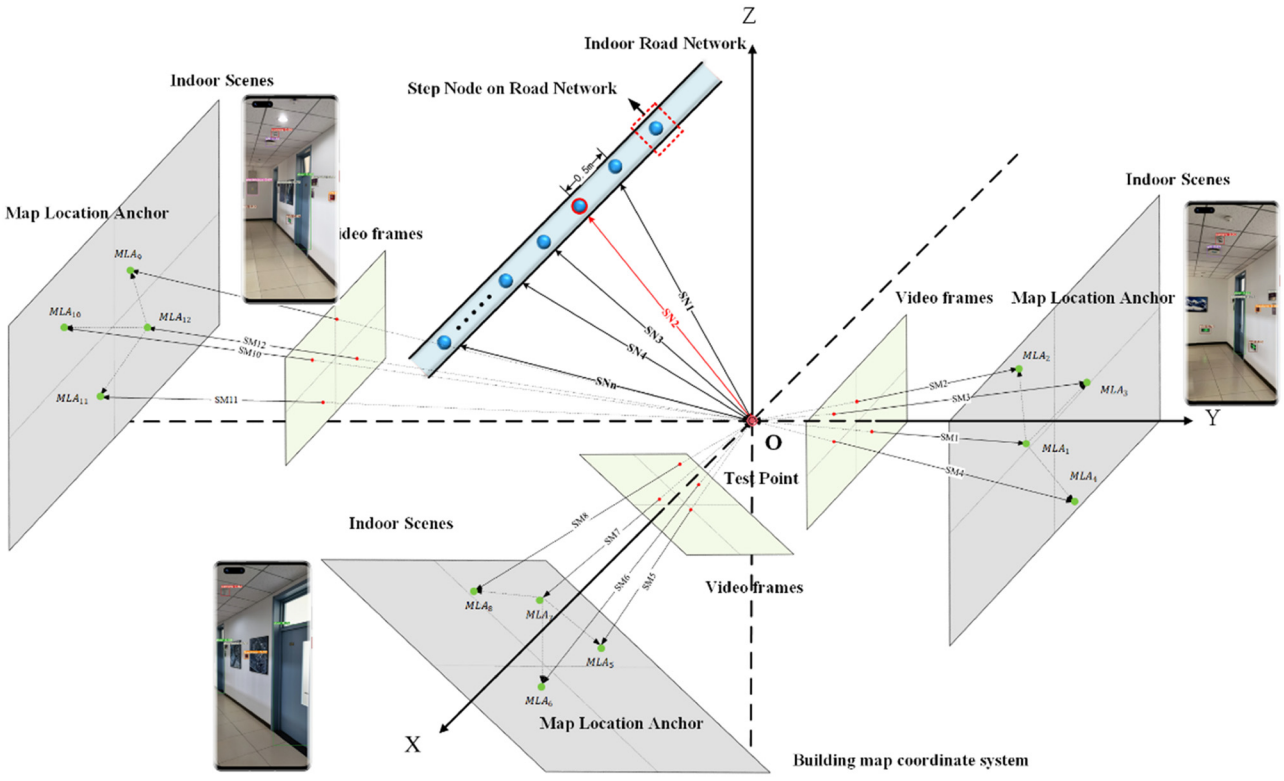

Figure 7: Matching positioning method based on geocoding Building MLA features.

1) Calculate the distances from the MLA (the user's cell phone shall interactively operate to 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 MLA distance set, the *P3P* algorithm is used to determine a unique set of solutions, to obtain the distance SM_i of the user camera, and to obtain the coordinates $pc(xc, yc, zc)$ corresponding to the current SN SN_c corresponding to SM_t
3) Create a buffer centred on the current SN $pc(xc, yc, zc)$ coordinates, the radius of this buffer is the maximum error range E_{max} plus the step size S_l. Use equations (2)–(7) to obtain the SN SN_n in the buffer

$$SN_n = \text{Buffer}(p0, \text{radius}) \quad (7)$$

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 SN in the current road network will be calculated, and then the minimum value D_{min} will be obtained from it, and the SN $pt(xt, yt, zt)$ 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 $pt(xt, yt, zt)$

4 Experiment

4.1 Data

4.1.1 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 study, the building F (longitude 116.29606E, latitude 39.751892 N) of the School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, is used as the experimental area, with an area of about 2,800 square metres. The object of the experimental study is a composite solid building, consisting of six floors, five above ground, and one underground. The outdoor

Figure 8: Building map of the School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture.

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 network model, including the construction of a single-level horizontal network and the connection between the horizontal network and vertical traffic patterns. The construction result of the data of the building map hybrid model BIMPN [32] example and the local situation of the visualization of the building map feature is shown in Figure 8(c). The visualization of building maps and other related data in this article can be accessed via the Internet [41].

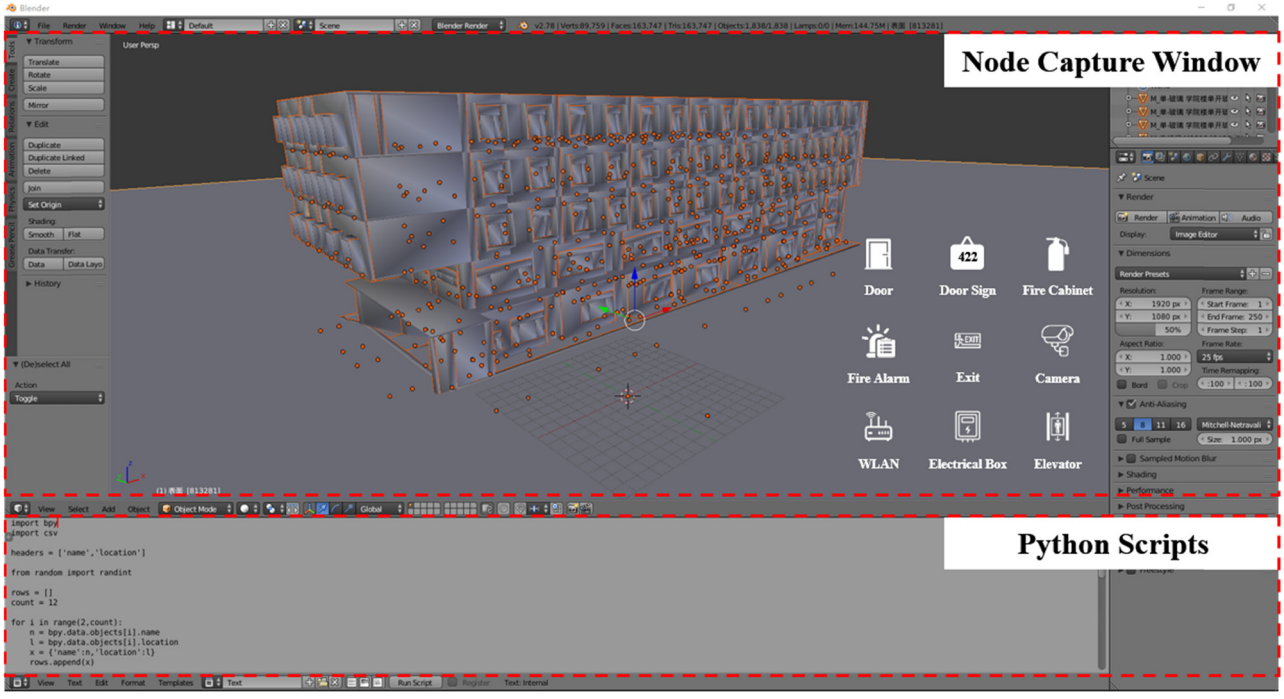

Figure 9: Extraction of coordinate information of MLA (semantic) in the building map.

Table 1: Sample subset of building interior scene features

Categories	Number of elemental samples	Percentage (%)
Doors	2,460	32.33
Door signs	1,200	15.77
Fire cabinets	450	5.91
Fire alarms	540	7.10
Exits	1,380	18.14
Cameras	320	4.20
WLANs	440	5.78
Electrical boxes	660	8.67
Elevators	160	2.10
Total	7,610	100

4.1.2 MLA data

The extraction of MLA (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, as shown in Figure 9. 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 MLA. The MLA to be captured in this article mainly include the geometric centres of universal building components such as doors, door signs, fire cabinets, fire alarms, security exits, cameras, WLANs, electrical boxes, and elevators. The capture is processed automatically by the Python script developed in this study.

4.1.3 Element sample data set

For the current study, a data set of building indoor scene features for MYOLOV5s model training is needed. We consider geometric location points with certain identifiable features in scene recognition as MLA [41]. However, the publicly available sample database found did not match the objectives of this study. Therefore, it is necessary to customize and build pervasive accessory facility data sets within the building map to continue this research. We have selected nine types of building indoor scene pervasive features as the identification targets for this project. The building MLA features are doors, door signs, fire cabinets, fire alarms, security exits, cameras, WLAN, electrical boxes, and elevators. Pictures with universal features in the building scenes need to be taken and collected, with different angles and distances according to the actual user's pose requirements during recognition, in order to build the element information required by the recognition algorithm. The data set has a total of 2,832 images and 7,610 element target samples, as shown in Table 1.

4.2 Experimental results

4.2.1 Building map and MLA construction results

In this study, the semantic information of MLA required in the building map is extracted based on the FBX format data of the target building (the 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 the data statistics are shown in Table 2, with a total of 586 element semantic constraints for building MLA [41].

To facilitate the subsequent management of the data, we imported the constructed building MLA data into a PostgreSQL database, which is an object-relational database management system (ORDBMS) with a wide range of features, and Figure 10 shows the result of the MLA (semantic) construction in the building map.

4.2.2 Recognition results of indoor scene features

The improved MYOLOv5s model is used for the video frame recognition of building indoor features. In this

Table 2: Statistics of MLA 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

	id [PK] integer	Revitid character varying(255)	scenex character varying	sceney character varying	scenez character varying	floor character varying	number character varying	order character varying	describe character varying
1	1	door<3>	4.921020508	-125.9853516	62.99206543	4	1	1	432
2	2	door<2>	4.928649902	0.006835938	62.99206543	4	2	2	433
3	3	exit<99>	-11.37463093	19.83551979	11.39435387	4	1	3	432
4	4	electrical box<88>	52.69203949	20.006464	13.98639584	4	1	4	430
5	5	door<6>	241.1417847	-125.984375	62.99206543	4	3	5	430
6	6	door sign<7>	-7.416632652	22.80401039	13.18045139	4	1	6	431
7	7	door sign<8>	-7.416632652	22.80401039	12.93742752	4	2	7	431
8	8	door<4>	241.1417847	-0.000244141	62.99206543	4	4	8	431
9	9	door<1>	-0.000152588	2.755371094	41.33850098	4	5	9	429
10	10	fire alarm<43>	-2.407995939	23.73365402	13.1806879	4	1	10	
11	11	door sign<47>	-5.240945339	19.83553505	13.48615932	4	3	11	430
12	12	door<10>	0.000259399	1.180908203	40.84643555	4	6	12	428
13	13	fire alarm<39>	-2.085511446	22.80401039	13.10357094	4	1	13	
14	14	door< 5>	-0.001933813	-2.756103516	41.33862305	4	7	14	
15	15	door< 22>	-0.001934052	-2.756103516	41.33862305	4	8	15	
16	16	fire alarm<42>	2.010787487	22.80401039	13.1806879	4	2	16	
17	17	exit<40>	-0.608837366	22.80401039	14.03495979	4	2	17	
18	18	exit<41>	0.726547778	22.80401039	14.03495979	4	3	18	
19	19	exit<98>	-2.176110506	19.83551979	11.39435387	4	4	19	426
20	20	door<37>	0.000259399	1.180908203	40.84643555	4	9	20	426

Figure 10: Results of MLA (semantic) construction in the building map.

study, Huawei Mate 40 Pro, Xiaomi Mix 2, and Huawei P9 are selected as mobile terminal devices, and their hardware parameters are shown in Table 3. All experiments 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 4.

Table 3: Mobile client hardware parameters

Phone Model	Parts	Specifications
Huawei Mate 40 Pro	CPU	Mali-G78 MP24
	Memory	8GB RAM + 128ROM
	Grid view	5 G network system
	Home screen resolution	FHD + 2,772 × 1,344 pixel
	Camera pixel	50million pixel super sensing camera
	Operating system	HarmonyOS 2.0
Xiaomi MIX 2	CPU	Qualcomm Snapdragon 835 (MSM8998)
	Memory	8GB RAM + 128ROM
	Grid view	4 G network system
	Home screen resolution	FHD + 2,160 × 1,080 pixel
	Camera pixel	12million pixel HD camera
	Operating system	MIUI 12.0.1
Huawei P 9	CPU	HiSilicon Kirin 955
	Memory	4GB RAM + 64ROM
	Grid view	4 G network system
	Home screen resolution	FHD + 1,920 × 1,080 pixel
	Camera pixel	12 million pixel HD camera
	Operating system	Android 8.0.0

The experiments use 2,256 images (video frames) from a sample set of 2,832 images for training and 288 images for validation and testing in the field. Experiment 1 uses the original YOLOv5s model for training and takes 63 h 3 min and 20 s to complete 1,000 epochs. Experiment 2 uses the improved MYOLOv5s model for training and takes 64 h and 10 min to complete 1,000 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 11. The blue curve corresponds to the YOLOv5s model, and the orange curve corresponds to the improved MYOLOv5s model.

In terms of the speed of recognition performed by the scene video, the total duration of a 758 frames video is around 25 s, the recognition time of the YOLOv5s model is 23.793 s (31.858 frames/s), and the recognition time of the improved MYOLOv5s model is 22.818 s (33.219 frames/s). The results show that in terms of recognition speed the improved model can achieve the effect of real-time availability. As shown in Figure 11(a) and (b), in terms of

Table 4: Hyperparameters information

Hyperparameters	Values
GPU_COUNT	1
Unm-Classes	9
Epochs	1,000
Batch Size	32
Img Size	640 * 640
Evolve	True
Cache images	True
Single cls	False

accuracy, the improved MYOLOv5s model is slightly better than the original model overall, and it is apparent that the improved model is better than the original model in 500–720 epochs. The analysis concluded 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. As shown in Figure 11(c) and (d), the learning performance of the model gradually improves with iterations, and the convergence speed is very fast, and the curve has stabilized by 1,000 epochs. The experiments in this paper use the training results of 1,000 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 data set. The smaller the loss, the better the classifier is at modeling the representation of the relationship between the input data and the output target. Figures 12 and 13 plot 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. The Box and Objectness plots represent the scores of the YOLOv5s model as shown in Figure 12(a) and (b), and the val-Box and val-Objectness plots

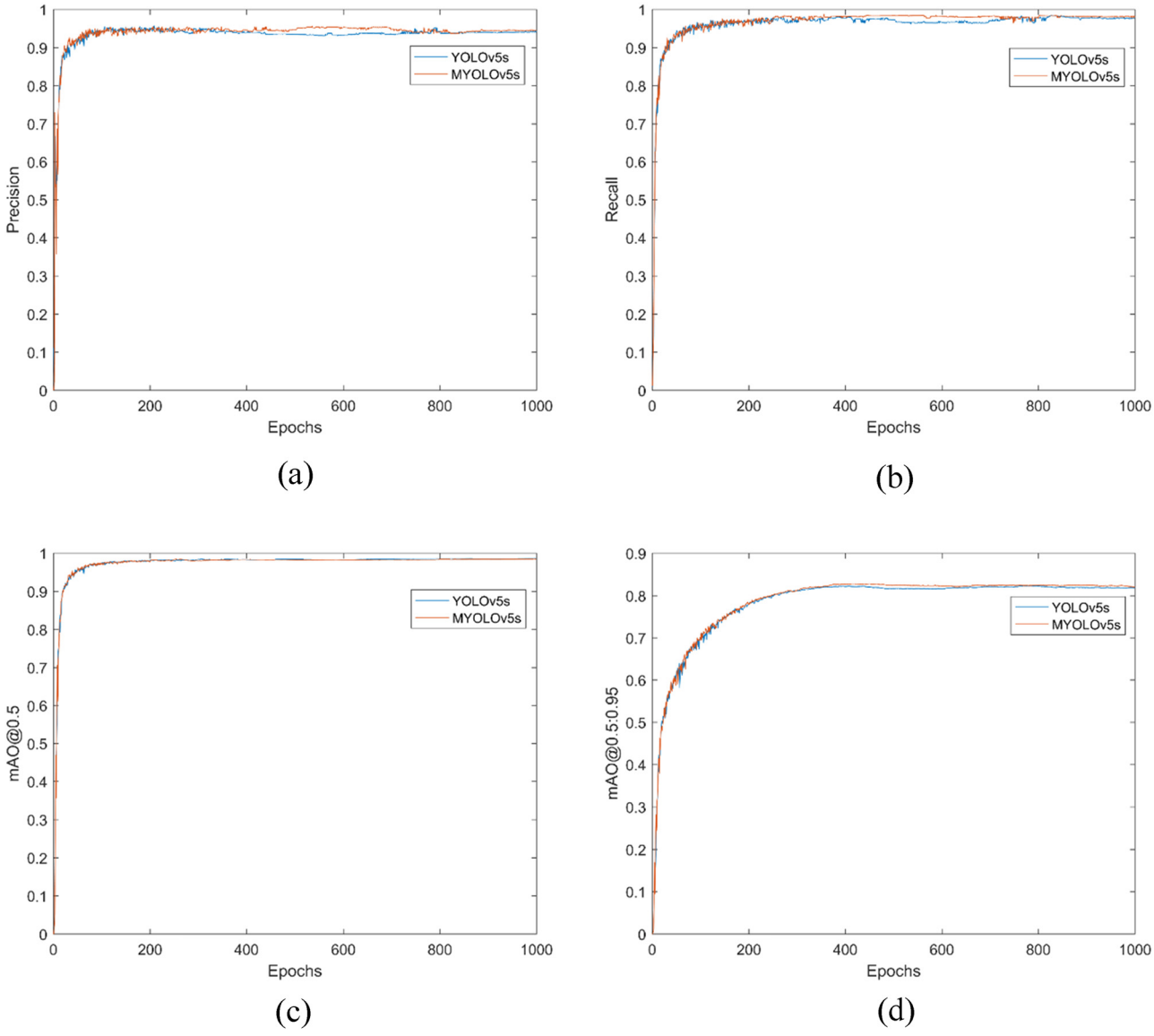

Figure 11: Comparison of performance metrics between YOLOv5s and modified MYOLOv5s during training: (a) model precision comparison, (b) model recall comparison, (c) model mAP@0.5 comparison, and (d) model mAP@0.5:0.95 comparison.

represent the validation scores of the YOLOv5s model as shown in Figure 12(c) and (d). The Box and Objectness plots represent the scores of the improved MYOLOv5s model as shown in Figure 13(a) and (b), and the val-Box and val-Objectness plots represent the validation scores of the improved MYOLOv5s model as shown in Figure 13(c) and (d). The training loss is measured during each stage, while the validation loss is measured after each stage. The results show that the improved MYOLOv5s model loss function is smoother and converges faster than the original model loss function; therefore, the MYOLOv5s model is more suitable for the application in the scenario of this article.

Figure 14 plots the accuracy and confidence of the improved MYOLOv5s model against the nine categories of MLA recognition models with generalizability in the building scenario defined in this study. Lifts are the most accurately identified category despite having the smallest number of samples due to their simple texture features. The seven categories of door signs, fire cabinets, fire alarms, security exits, cameras, WLANs, and electrical boxes also fall within the confidence interval of (0.1–0.2), with recognition accuracy tending to be stable. Only doors had a lower accuracy rate and a significant jump in the confidence interval of (0.4–0.7). This is because there are many different types of doors, such

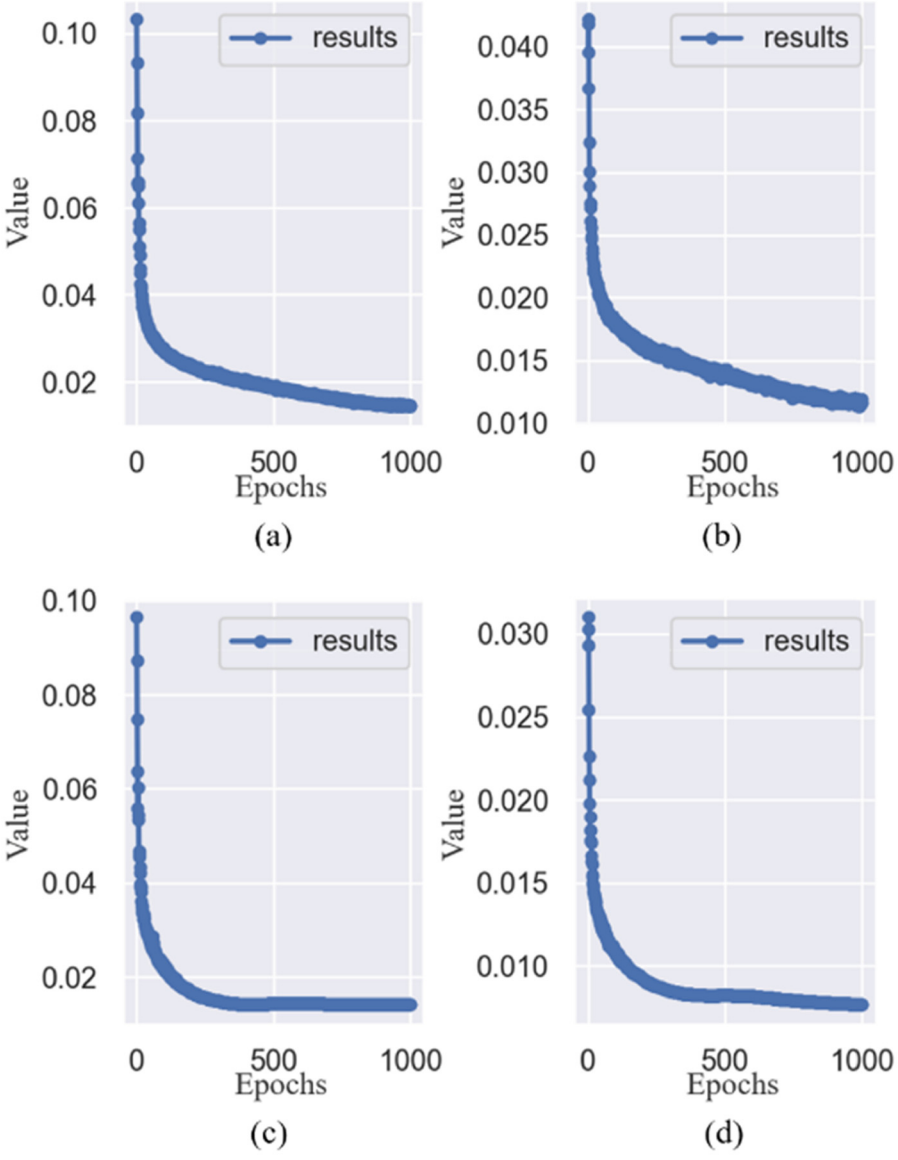

Figure 12: Loss effect of YOLOv5s during training. (a) Box, (b) objectness, (c) val Box, and (d) val objectness.

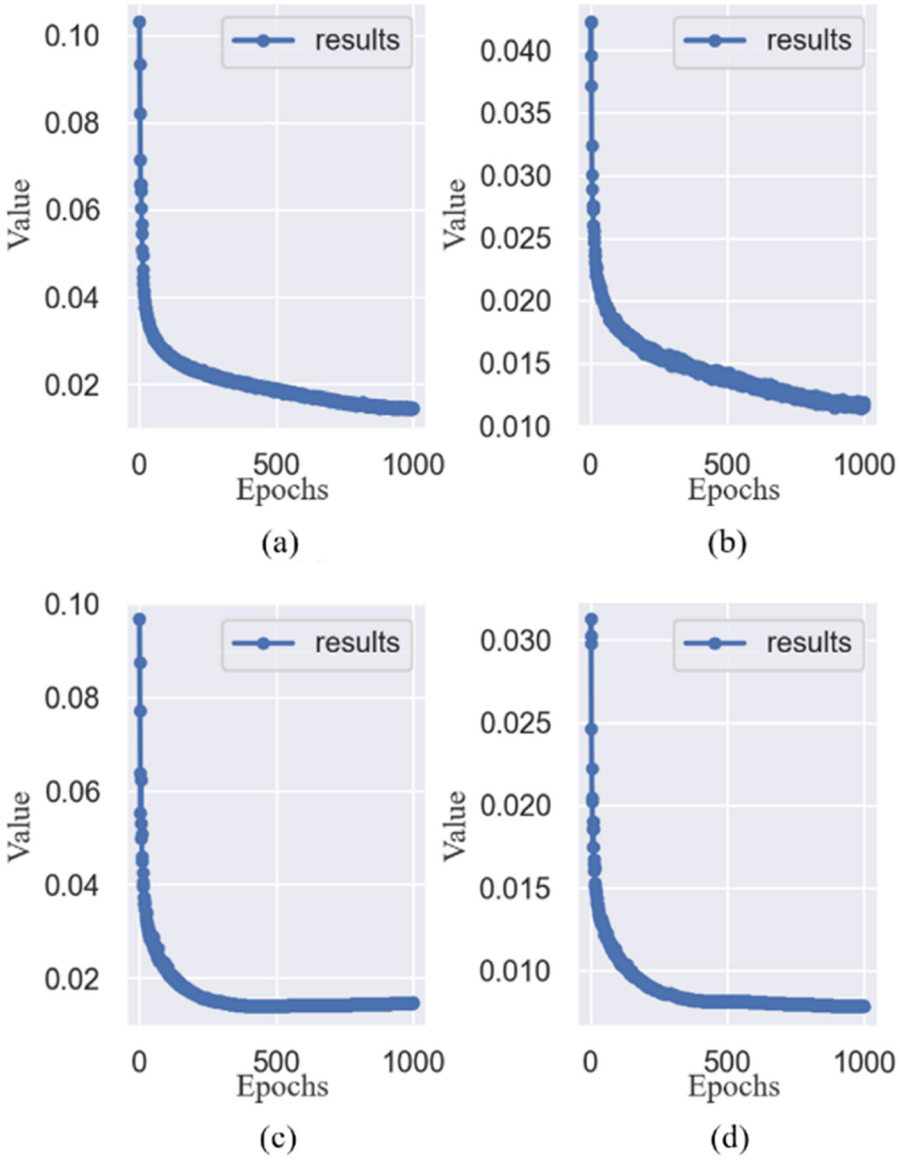

Figure 13: Loss effect of MYOLOv5s during training. (a) Box, (b) objectness, (c) val Box, and (d) val objectness.

as single, double, security and glass doors, and that the geometric area of the doors is somewhat influenced by the large size of the doors compared to the other features in the sample labelling. Overall, the accuracy of the model in identifying various types of features in building scenes meets the requirements of the scene identification and localization method.

Figure 15 shows some results of the recognition of the features of the indoor scene of the building under different time, light intensity, and angle experimental conditions. The proposed model is not only applicable to detecting the features 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 features captured under the condition that the line of sight is shifted by a certain angle during walking. In addition, the proposed MYOLOv5s model is able to identify the nine types of features of the proposed MLA under different conditions such as sunny day, dusk, night, and indoor lighting. Especially at night when the lighting conditions are not particularly adequate, it can be seen that these features are still identified very accurately. This is a basic application for users walking indoors in buildings with less favourable light conditions. Experiments show that the accuracy of the improved MYOLOv5s model achieves the requirements of scene features recognition and localization method for the recognition of indoor scene features of buildings under different time, lighting, and angles.

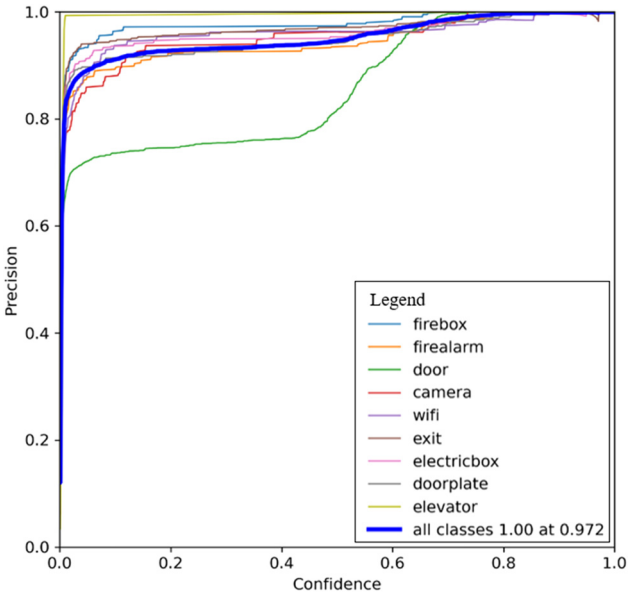

Figure 14: Comparison between accuracy and confidence relationships for each category of the improved MYOLOv5s model.

4.2.3 Localization results of indoor scene recognition

The goal of the experiment in this subsection is to verify that this method has good localization results under the constraints of MLA 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 starting position of the user is obtained by Bluetooth and fused multi-source sensor localization, and through the fusion of multi-source sensor positioning to obtain the user's floor location, we can use the barometer to obtain the air pressure of the user's location for threshold calculation, while the floor's Bluetooth ibeacon can provide us with the user's current floor, which is input to this method as a prerequisite condition. Figure 16 shows the visualization effect of real-time positioning starting from a certain starting position in the building area. Yellow line shows the trajectory of the user walking along the corridor path, and blue line shows the trajectory depicted after the video scene element identification location anchor and the building map road network node (corridor centre line) 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 SN to obtain the fusion results of positioning points, and then 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 frames, the more information that can be matched between the identifiable features of the building space scene and the building MLA, and the higher the accuracy of the completed positioning in the scene walking will be.

In order to analyse 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 17. The deviation points are mainly concentrated in the x-negative half-axis. Since the user will face the camera towards the semantic information-rich wall in the corridor scene during the recognition process through the smart phone camera, thus will be closer to the opposite semantic information-less wall, resulting in the x direction deviation being 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 SN, and the coordinates of the 10 pairs of points with the largest deviation in the x and y directions of the path coordinates are selected from 103 pairs of coordinate points for typicality analysis.

As shown in Table 5, 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 on identifiable features 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 (metre-level) and the deviation of the recognition algorithm (centimetre-level) is large, the deviation of this method is controlled in the maximum range which is acceptable in practice. In addition, for scene element identification and localization under sparse indoor scene element conditions, the system adopts geometric constraints of road network nodes and multi-source sensor MLA (S) cooperative localization, which can generally ensure the continuity of the localization and navigation process within a certain range. Therefore, the visualization of the guidance information in the form of matching scene recognition location anchors with road network nodes does not cause any disturbance to the user's positioning and navigation process. The method is feasible in engineering applications.

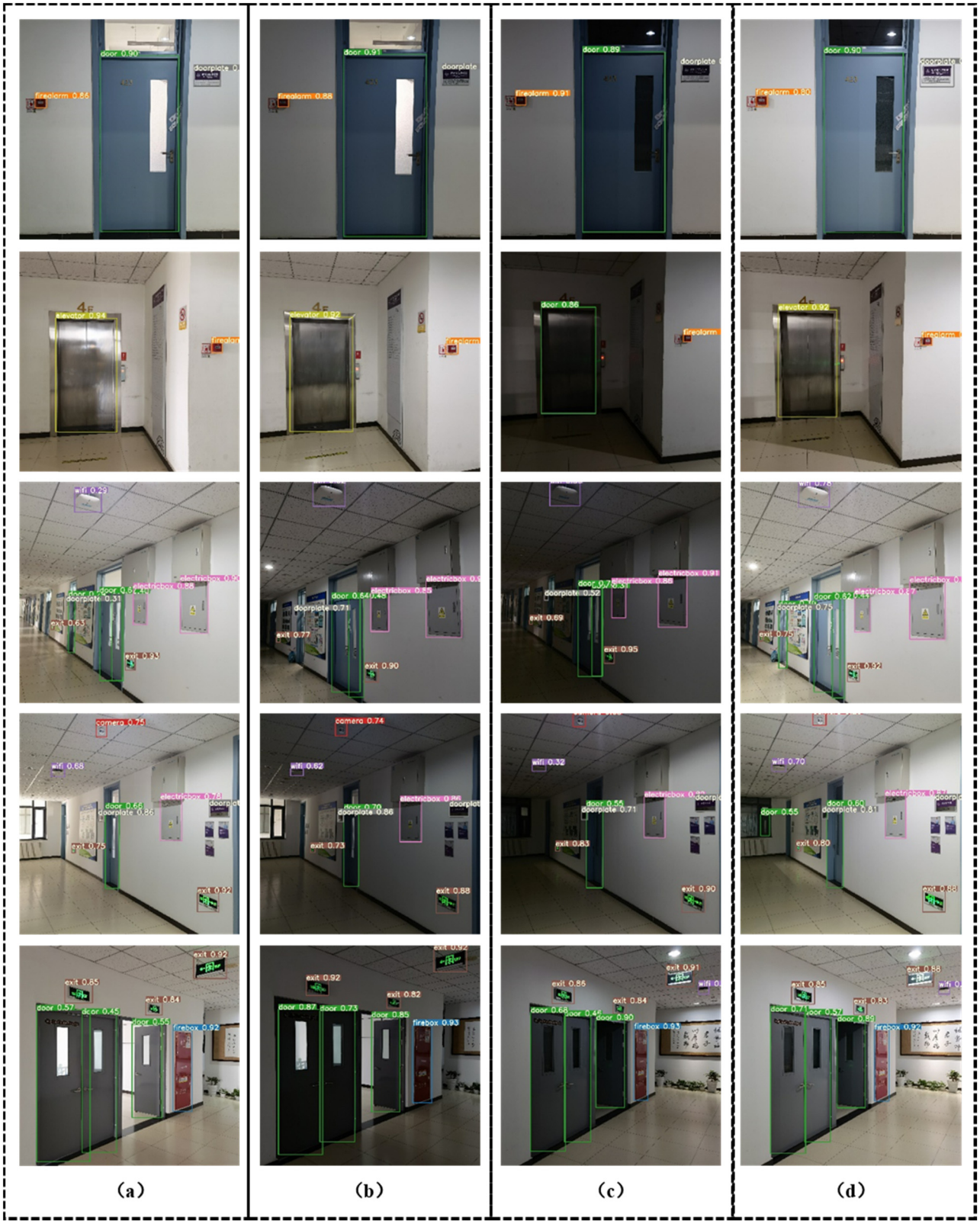

Figure 15: Results of the recognition of building interior features in different time series using the improved MYOLOv5s network. (a) Daytime, (b) dusk, (c) night, and (d) light.

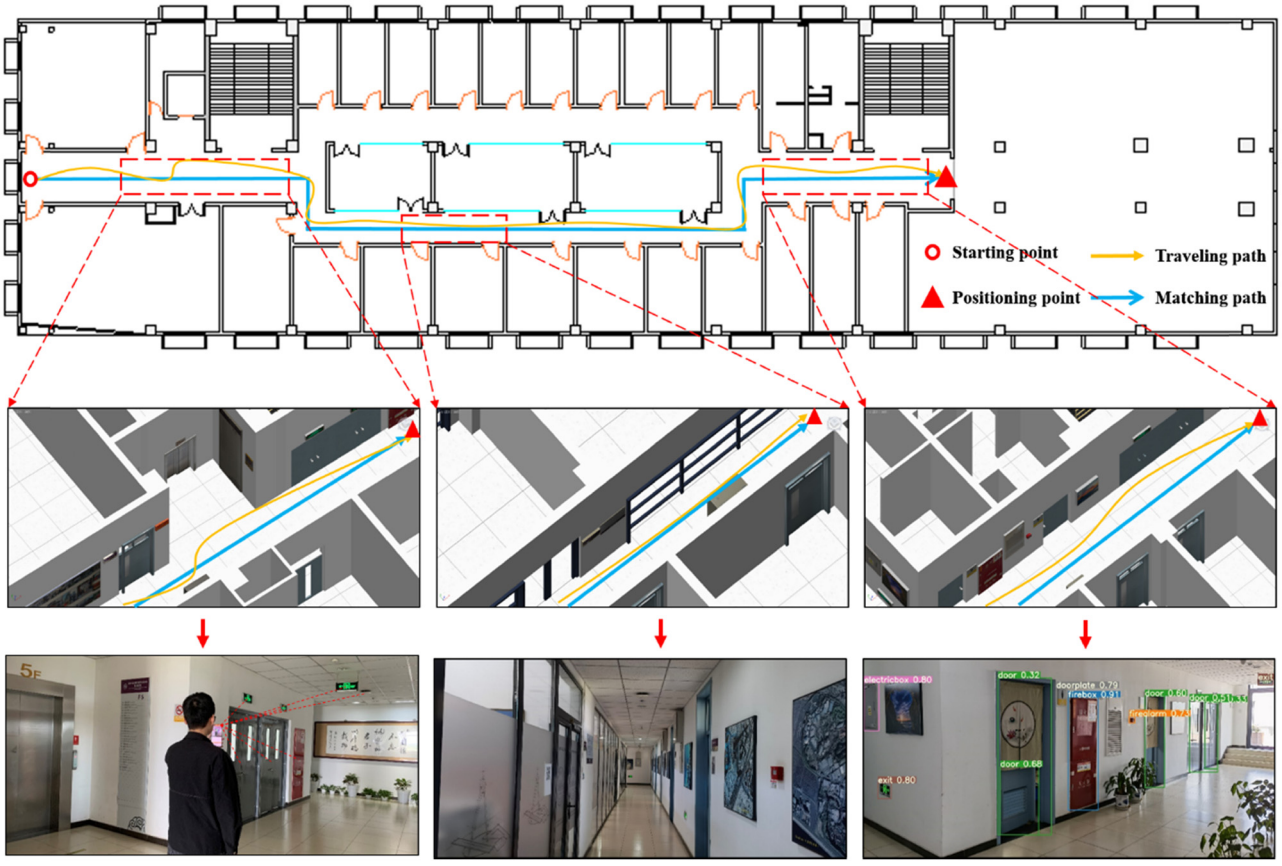

Figure 16: Localization results of indoor scene recognition for mobile phones with semantic constraints of building maps.

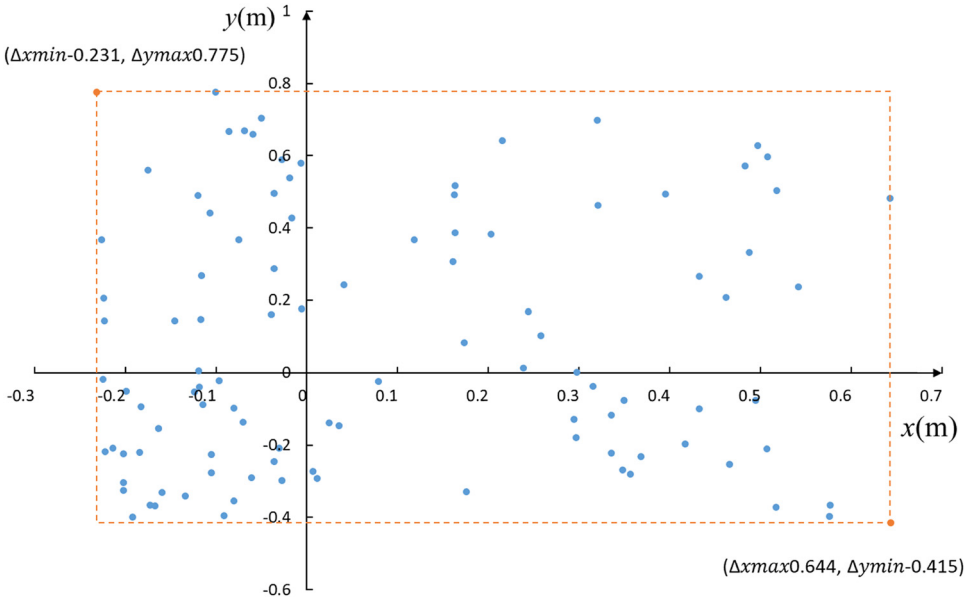

Figure 17: Coordinate deviation statistics of pedestrian walking trajectory and map matching trajectory.

Table 5: 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 (Δx, Δy)
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)

5 Discussion

In this study, we use the improved YOLOv5 model to identify the elements in the building scene in real time through the mobile phone camera, match them with the geographic coordinates of the map positioning anchor points for spatial resolution, and determine the unique solution through the P3P algorithm, finally matching to the nearest road network node to the positioning result. This article constructs MLA with universal scene features in building interiors, so the scene element recognition model does not need to manually deal with a lot of element information of other building interior scenes. And the model 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.

Our results are significantly better than the YOLOv5s model in terms of robustness due to the multi-scale and multi-granularity features of the identified elements. The recall rate in the test set is consistently above 97.2%, indicating that the method is suitable for architectural interior scenes with rich information on scene elements. The results of the positioning experiments show that the maximum localization error is within the range of 0.775 m, and it is confined to about 0.5 m after applying the BIMPN road network SN constraint, which is within the acceptable range of the arbitrary oscillation (metre 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.

Moreover, owing to the data sample collection scheme and the building map-based location matching process, our method has massive potential to extend to a crowdsourcing-based method. Integration of interior scene data from different buildings into a shared library of sample building maps. In engineering applications, the building indoor scene recognition model on the mobile 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 quickly map matching on the mobile side. This method not only allows real-time browsing of realistic holographic maps of buildings on the mobile phone but also facilitates the further enhancement of related applications utilizing AR-enhanced semantic element information in building maps, etc.

6 Conclusions

In this article, we propose an indoor scene features recognition and localization method for mobile phones with semantic constraints of building MLA. This article provides semantic constraint information for indoor positioning by constructing a geocoded entity library for building MLA, then identifies the semantic constraint element information in the scene based on the improved MYOLOv5s model, matches the identified element information with the database MLA, and finally, constrains the location of the user in the road network corresponding to the location information from the scene element feature points, thereby achieving real-time positioning and navigation.

The method proposed in this article is a solution that particularly requires indoor environmental data from CIM perspective. The video for building interior scene element recognition is obtained through smartphone camera

shooting, and the key to mobile 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 CIM and metaphase environments. The final goal is to merge building maps with augmented reality and to visually represent the semantic information in building maps, thereby providing more accurate and richer services to users for real-time location-based services.

Acknowledgments: The authors would like to thank the anonymous reviewers, the editors, and the scholars for their constructive comments and suggestions, which greatly improved the quality of the manuscript.

Funding information: This research is funded by the National Natural Science Foundation of China (No. 41301489), Beijing Natural Science Foundation (No. 4192018, No. 4142013), and 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).

Author contributions: Conceptualization, Liu Jianhua and Feng Guoqiang; Validation, Luo Jingyan and Wen Danqi; Data Curation, Chen Zheng, Wang Nan and Zeng Baoshan; Writing–Original Draft Preparation, Feng Guoqiang and Wang Nan; Writing–Review & Editing, Liu Jianhua; Visualization, Wang Xiaoyi, Li Xinyue ang Gu Botong. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: No potential conflict of interest was reported by the authors.

References

[1] Hu M, Giapis KP, Goicochea JV, Zhang X, Poulikakos D. Localization technologies for indoor human tracking. IEEE Commun Surv Tutor. 2010;11:1–6.
[2] Chen R, Chen L. Smartphone-based indoor positioning technologies. Urban Informatics; 2021.
[3] del Horno MM, Orozco-Barbosa L, García-Varea I. A smartphone-based multimodal indoor tracking system. Inf Fusion. 2021;76(6):36–45.
[4] La Delfa GC, Catania V, Monteleone S, De Paz JF, Bajo J. Computer vision based indoor navigation: A visual markers evaluation. Adv Intell Syst Comput. 2015;376:165–73.
[5] Martin-Gorostiza E, Garcia-Garrido MA, Pizarro D, Torres P, Miguel MO, Salido-Monzú D. Infrared and camera fusion sensor for indoor positioning. 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. Sensors. 2021;21(4):1114–31.
[7] Lie M, Kusuma GP. A fingerprint-based coarse-to-fine algorithm for indoor positioning system using Bluetooth Low Energy. Neural Comput Appl. 2021;33(7):2735–51.
[8] Dawood MA, Saleh SS, El-Badawy ESA, Aly MH. A comparative analysis of localization algorithms for visible light communication. Optical Quantum Electron. 2021;53(2):108–33.
[9] Niu X, Li Y, Kuang J, Zhang P. Data fusion of dual foot-mounted IMU for pedestrian navigation. IEEE Sens J. 2019;99:1109–19.
[10] Liu P, Zhang Z, Wu L, Dang J, Li Y, Jin X. Fingerprint-based indoor localization algorithm with extended deep belief networks. Information Communication Technologies Conference (ICTC); 2020. p. 91–7.
[11] Chen Y, Du T, Jiang C, Sun S. Indoor location method of interference source based on deep learning of spectrum fingerprint features in Smart Cyber-Physical systems. EURASIP J Wirel Commun Netw. 2019;2019:47–59.
[12] Cheng R, Wang K, Bai J, Xu Z. Unifying visual localization and scene recognition for people with visual impairment. IEEE Access. 2020;8:64284–96.
[13] Xiong Y, Liu H, Gupta S, Akin B, Bender G, Wang Y, et al. MobileDets: Searching for object detection architectures for mobile accelerators. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 3824–33.
[14] Guo R, Chen Y, Zhao Z, He B, Lv G, Li Z, et al. A theoretical framework for the study of pan-maps. J Geomat. 2021;46(1):9–15.
[15] Gu F, Hu X, Ramezani M, Acharya D, Khoshelham K, Valaee S, et al. Indoor localization improved by spatial context–A survey. ACM Comput Surv (CSUR). 2019;52(3):1–35.
[16] Hu X, Fan H, Noskov A, Zipf A, Wang Z, Shang J. Feasibility of using grammars to infer room semantics. Remote Sens. 2019;11(13):1535–61.
[17] Cossaboom M, Georgy J, Karamat T, Noureldin A. Augmented Kalman filter and map matching for 3D RISS/GPS integration for land vehicles. International Jouranl of Navigation and Observation; 2012. p. 2012.
[18] Bandyopadhyay A, Hakim D, Funk BE, Kohn EA, Teolis C, Weniger GB. System and method for locating, tracking, and/or monitoring the status of personnel and/or assets both indoors and outdoors: US, US8712686 B2[P]. 2016.
[19] Zhou B, Li Q, Mao Q, Tu W, Zhang X, Chen L. ALIMC: Activity landmark-based indoor mapping via crowdsourcing. IEEE Trans Intell Transp Syst. 2015;16(5):2774–85.
[20] Shang J, Gu F, Hu X, Kealy A. APFiLoc: An infrastructure-free indoor localization method fusing smartphone inertial sensors, landmarks and map information. Sensors. 2015;15(10):27251–72.
[21] Li B, Muñoz JP, Rong X, Chen Q, Xiao J, Tian Y, et al. Vision-based mobile indoor assistive navigation aid for blind people. IEEE Trans Mob Comput. 2019;18(3):702–14.

[22] Gu F, Valaee S, Khoshelham K, Shang J, Zhang R. Landmark graph-based indoor localization. IEEE Internet Things J. 2020;7(9):8343–55.
[23] Sadeghi H, Valaee S, Shirani S. 2DTriPnP: A robust two-dimensional method for fine visual localization using google street-view database. IEEE Trans Veh Technol. 2017;66(6):4678–90.
[24] Lepetit V, Moreno-Noguer F, Fua P. EPnP: An accurate O(n) solution to the PnP problem. Int J Comput Vis. 2009;81(2):155–66.
[25] Kneip L, Li H, Seo Y. UPnP: An optimal O(n) solution to the absolute pose problem with universal applicability. European Conference on Computer Vision. Cham: Springer; 2014.
[26] Ke T, Roumeliotis SI. An efficient algebraic solution to the perspective-three-point problem. IEEE; 2017.
[27] Ding X, Luo Y, Yu Q, Li Q, Cheng Y, Munnoch R, et al. Indoor object recognition using pre-trained convolutional neural network. 2017 23rd International Conference on Automation and Computing (ICAC). Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE; 2017.
[28] Guo W, Wu R, Chen Y, Zhu X. Deep learning scene recognition method based on localization enhancement. Sensors. 2018;18(10):3376–495.
[29] Liu M, Chen R, Li D, Chen Y, Guo G, Cao Z, et al. Scene recognition for indoor localization using a multi-sensor fusion approach. Sensors. 2017;17(12):2847.
[30] Shuang L, Xingli G, Ruihui Z, Ya'ning L. Scene recognition and PnP problem solution for indoor visual location. Radio Engineering; 2018.
[31] Liu J, Zhang X, Wang Y, Zhang HY. Pruning based deep network is used for text detection of natural scene images [C]//ICDSP 2020. 2020 4th International Conference on Digital Signal Processing; 2020.
[32] Liu J, Luo J, Hou J, Wen D, Feng G, Zhang X. A BIM based hybrid 3D indoor map model for indoor positioning and navigation. Int J Geo-Information. 2020;9(12):747–68.
[33] Chen S, Liu J, Liang X, Zhang S, Hyyppa J, Chen R. A novel calibration method between a camera and a 3D LiDAR with infrared images. IEEE Int Conf Robot Autom. 2020;10(11):4963–9.
[34] Li M, Chen R, Liao X, Guo B, Zhang W, Guo G. A precise indoor visual positioning approach using a built image feature database and single user image from smartphone cameras. Remote Sens. 2020;12(5):869–93.
[35] Qin W, Song T, Liu J, Wang H, Liang Z. Remote Sensing Military Target Detection Algorithm Based on Lightweight YOLOv3. CEA. 2021;57(21):263–9.
[36] Wang D, He D. Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning. Biosyst Eng. 2021;210(6):271–81.
[37] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016.
[38] Ultralytics. https://github.com/ultralytics/yolov5.
[39] MINPS2.0, https://mp.weixin.qq.com/s/gJYMj2vFfEOEoc2jsF45XQ, 2022, www.dxkjs.com.
[40] Bai L, Yang Y, Feng C, Guo C. A novel received signal strength assistedperspective-three-point algorithm for indoor visi-ble light positioning. Opt Express. 2020;28(19):1162–75.
[41] MLA Building F, http://www.dxkjs.com/indoorroad/indexF.html, 2022.
[42] Blander. https://www.blendercn.org.