CONSTRUCTION OF A DUAL-TASK MODEL FOR INDOOR SCENE RECOGNITION AND SEMANTIC SEGMENTATION BASED ON POINT CLOUDS

Jianwu Jiang[1,2], Zhizhong Kang[1*], Jingwen Li[2]

[1] School of Land Science and Technology, China University of Geosciences, Beijing 100083, China
[2] College of Geomatics and Geoinformation, Guilin University of Technology, Guilin 541004, China
[*] Corresponding author. E-mail address: zzkang@cugb.edu.cn (Zhizhong Kang).

KEY WORDS: Scene recognition , Semantic segmentation , Indoor cognition , Point cloud , Multi-tasking.

ABSTRACT:

Indoor scene recognition remains a challenging problem in the fields of artificial intelligence and computer vision due to the complexity, similarity, and spatial variability of indoor scenes. The existing research is mainly based on 2D data, which lacks 3D information about the scene and cannot accurately identify scenes with a high frequency of changes in lighting, shading, layout, etc. Moreover, the existing research usually focuses on the global features of the scene, which cannot represent indoor scenes with cluttered objects and complex spatial layouts. To solve the above problems, this paper proposes a dual-task model for indoor scene recognition and semantic segmentation based on point cloud data. The model expands the data loading method by giving the dataset loader the ability to return multi-dimensional labels and then realizes the dual-task model of scene recognition and semantic segmentation by fine-tuning PointNet++, setting task state control parameters, and adding a common feature layer. Finally, in order to solve the problem that the similarity of indoor scenes leads to the wrong scene recognition results, the rules of scenes and elements are constructed to correct the scene recognition results. The experimental results showed that with the assistance of scene-element rules, the overall accuracy of scene recognition with the proposed method in this paper is 82.4%, and the overall accuracy of semantic segmentation is 98.9%, which is better than the comparison model in this paper and provides a new method for cognition of indoor scenes based on 3D point clouds.

1. INTRODUCTION

Scene recognition is one of the most challenging tasks in image classification (Nascimento et al., 2015) and also a basic task of robot perception (Zhou et al., 2021). Scene recognition gives the robot the ability to describe the environment at a conceptual level (Shi et al., 2011), which provides robots with high-level indoor space semantic information and is of great significance for robots to understand the surrounding environment (Espinace et al., 2010). Effective scene understanding can help the robot make reasonable judgments and behaviors (Miao et al., 2021), better interact with people, and perform complex tasks (Shi et al., 2011). Although robot research has made great progress in recent years, indoor scenes have great spatial variation due to the diversity of internal elements and the layout of the scene. In addition, the occlusion and lighting factors prevalent in indoor scenes increase the difficulty of robot scene recognition.

Different from conventional image classification, the category of indoor scene is closely related to its internal elements, and optimal results cannot be obtained for scene recognition by only using global features (Xiong et al., 2020). Therefore, the design of an indoor scene recognition network should take into account global and object features (He et al., 2016) and build effective rules to describe the relationship between the scene and the indoor elements (Pal et al., 2019). At present, experts and scholars have proposed many indoor scene recognition models based on CNN (convolutional neural network), such as VGG16 (Simonyan and Zisserman, 2014) , ResNet50 (He et al., 2016), MobileNet (Howard et al., 2017), TRecgNet (Du et al., 2019), and FOSNet (Seong et al., 2020), etc. Most of the above models use RGB and RGB-D data as training data. Although good performance was achieved in the respective training environments, due to the complexity of the realistic scene, the robustness of scene recognition is not high. Especially in a dynamic scene, the data acquired by the robot usually has noise and occlusion. If the model cannot utilize the 3D information of the scene, it is difficult to design a more robust scene recognition model (Mosella-Montoro and Ruiz-Hidalgo., 2021, Romero-Gonzalez et al., 2016). Compared with 2D sensors, 3D sensors can better describe the geometric characteristics of indoor space, and when equipped with cameras, 3D sensors can also obtain color features, which is an advantage for indoor scene understanding. Therefore, it is very important to develop a scene recognition model based on 3D data.

This paper builds a two-task model for both scene recognition and semantic segmentation. The model also takes into account the global and object features of indoor scenes and improves the robustness of scene recognition. At the same time, in order to correct the results of the model output, this paper constructs the correlation rules between scenes and elements to further improve the accuracy of scene recognition. The main contributions of this paper are as follows:

(1) A dual-task model of indoor scene recognition and semantic segmentation is constructed that simultaneously meets the tasks of scene recognition and semantic segmentation through one training, and the state of the dual task can be controlled through hyperparameters to improve the efficiency of the model. The model takes into account the global and object-level features of indoor scenes, which improves the robustness of scene recognition.

(2) By analyzing the occurrence probability of elements within the scene, the correlation relationship between indoor scenes

and elements is constructed, which can effectively correct the results of scene recognition.

(3) The model storage strategy under the dual-task model is proposed, which proves that the single optimal storage strategy of the dual-task is more conducive to the improvement of the overall effect of the model, can be extended to the multi-task field, and provides the optimal model storage suggestions for the parallel training of the multi-task model.

The paper is organized as follows: Section 2 mainly introduces the related work of indoor scene recognition. In Section 3, the proposed method is described in terms of dataset loading, dual-task model construction, and scene recognition result correction. The proposed method is validated in Section 4.. In Section 5, the influences on the model are discussed from three aspects: multi-task, loss function, and indoor element category. Section 6 lists the conclusions of this paper.

2. RELATED WORKS

Indoor scene recognition is a hot topic in the field of computer vision. From the perspective of the data source, scene recognition methods are usually divided into three categories: RGB-based, RGB-D-based, and 3D-based. Experts and scholars have carried out some research on the three types of scene recognition methods.

(1) Indoor scene recognition based on RGB

RGB data can better reflect the color, texture, and other information of indoor scenes, and there are more open-source datasets, so there is more research on indoor scene recognition models based on RGB data. Due to the high variability of indoor environments, the performance of current scene recognition methods for indoor scenes has significantly decreased. Therefore, in the process of indoor scene recognition, multi-scale and deep-level features as well as advanced semantic information should be extracted to improve the performance of scene recognition. Zeeshan Ali et al. (2021) constructed an indoor scene recognition model based on the ResNet-18 model, adopted 18 convolution layers to extract the deep features of scenes, and applied transfer learning to the model to realize the recognition of indoor scenes. Pereira et al. (2020) constructed a two-branch-based Global and Semantic Feature Fusion Approach (GSF2App) based on RGB data to realize indoor scene recognition. Branch 1 is used to extract global features from RGB images, and Branch 2 is based on YOLO v3 to obtain semantic information in images and construct semantic feature vectors. Nascimento et al. (2015) aiming at the problem that traditional convolutional neural networks are prone to losing local detail features, proposed an extraction and representation method of scene global features and local features based on sparse coding methods to realize indoor scene recognition. This model also shows a certain robustness for scene recognition with noise and occlusion. Afif et al. (2020) developed a new model for indoor object recognition and scene classification by simplifying and fine-tuning the neural network hierarchy and achieved good results. However, in general, the more features obtained by the scene recognition model, the higher the computational cost. To solve this problem, Qiao et al. (2021) proposed a scene recognition model based on APM (Attention Pyramid Module), which simplifies the extraction of multi-scale features and can obtain local and global scene features at the same time. In addition, the scene recognition model needs to merge the acquired multi-scale features. Chen et al. (2020) used the full connection layer to extract the features output by the convolutional neural network model and fuse them. A high-level feature fusion method based on a multi-convolutional neural network (MultiCNN) was constructed to realize the scene recognition task. Wang et al. (2020) proposed a deep feature fusion through adaptive discriminative metric learning (DFF-ADML). This method solves the problem of the lack of complementarity and consistency information in traditional depth feature vector fusion. At present, a large number of research efforts focus on developing new auxiliary features and networks to improve indoor scene recognition accuracy. However, the relationship between scenes and their internal objects is also very important for indoor scene recognition. Espinace et al. (2010) proposed an indoor scene recognition method based on the generated probability hierarchy model. This method uses low-level visual features, associates the features with indoor elements, and extracts the context relationship between elements and scenes. Seong et al. (2020) proposed a framework of object and scene integration and constructed the fusion of object and scene net (FOSNet) model to realize scene recognition. Lopez-Cifuentes et al. (2020) proposed an end-to-end multi-modal that combines image and context information through an attention module. Xia et al. (2019) proposed a scene recognition approach based on a WS-AM-weakly supervised attention map. This method solves the problem that the existing CNN-based scene recognition methods do not fully consider the relationship between image regions and categories when selecting local areas, resulting in reduced recognition accuracy. This method uses gradient-weighted class-activation mapping (Grad-CAM) and weak supervision information to generate an AM (attention map) of scene images, which was helpful to distinguish scene recognition areas. Miao et al. (2021) proposed an object-to-scene (OTS) method that extracts object features and learns object relations to recognize indoor scenes. Zhou et al. (2021) took the improved object model (IOM) as the benchmark and expressed the IOM as a Bayesian object relation model (BORM). Combined with BORM and PlacesCNN, the Bayes object relation fusion model (CBORM) is constructed to realize scene recognition. The existing scene recognition model is mainly used in the real world. With the rise of the meta-universe, it is also very important to recognize the virtual and unknowable scenes. Njoku et al. (2022) built a model for automatic recognition of the virtual scene based on SimpleNet and AlexNet. The training data of the model is the real scene, and the model results can be transferred to the virtual scene. Chen et al. (2020) proposed a prototype-agnostic scene layout (PaSL) method to realize scene recognition, construct the spatial structure of each image when it does not conform to any prototype, and, based on PaSL, construct a LGN (Layout Graph Network), in which regions in PaSL are defined as nodes and two independent relationships between regions are encoded as edges.

(2) Indoor scene recognition based on RGB-D

Compared with RGB data, RGB-D data contains more depth information about the scene and can obtain richer scene features, which plays an important role in overcoming spatial variability. Zhu et al. (2016) proposed a new discriminant multi-mode fusion framework based on RGB-D to realize scene recognition. Xiong et al. (2021) proposed a framework based on RGB-D that could adaptively select important local features and capture the spatial variability of the scene. The framework was developed by constructing a local feature selection module where you can choose different local subject-level and object-level presentation features from RGB-D. Although a single scene contains a variety of elements and perspectives, some elements often reflect the core characteristics of the scene. It is difficult to obtain a robust model for indoor scene recognition tasks due to the variety of elements in a single scene. Song et al. (2020) studied the more discriminant image representation of object-object relation in scene recognition based on the object relation, the relation obtained by object detection technology, and proposed the cooccurrence frequency of object-object relation (COOR, co-occurring frequency of object-to-object relation) and the sequential representation of object-to-object relation (SOOR, sequential representation of object-to-object relation), which describes objects and their relative relations in different forms and encodes SOOR as features to provide classifiers to realize indoor scene recognition based on RGB-D. Compared with RGB data, RGB-D has depth information, and how to effectively use depth information is also very important for scene recognition. Xiong et al. (2020) proposed a network to extract RGB-D modal consistency features and modal-specific features at the same time. The model can determine the uniqueness, commonality, and correlation of different modal data so as to achieve better scene recognition. Romero-Gonzalez et al. (2016) proposed a method to generate global image descriptors based on RGB-D images. Spatial pyramid technology (Lazebnik et al., 2006) was used to integrate local three-dimensional features. Then, a 3D spatial pyramid (3DSP) descriptor was constructed and applied to indoor scene recognition to improve model accuracy. Du et al. (2019) proposed a unified framework to integrate trans-modal conversion and mode-specific recognition tasks, which is called TRecgNet (Translate-to-Recognize Network). The conversion task and recognition task share the same encoder network. With the help of the trans-module, the training of the recognition task can be explicitly regulated to improve its final generalization ability. Xiong et al. (2019) proposed a new RGB-D scene recognition framework that can clearly learn global modal specific features and local modal consistent features at the same time. The main principle is to learn modal consistent representation by considering local CNN features and designing a key feature selection (KFS) module that can adaptively select important local features from a high-semantic CNN feature map.

(3) Indoor scene recognition based on 3D data

There are few studies on indoor scene recognition based on 3D data. The main reason is that there is little 3D data available for scene-level labeling. However, the 3D information contained in the 3D data is of great significance for the scene recognition task to overcome the influence of illumination and noise. Therefore, it is also necessary to research the scene recognition model based on the 3D data. Mosella-Montoro and Ruiz-Hidalgo (2021) proposed a scene recognition model that combines 3D geometric features with 2D texture features obtained by 2D convolutional neural networks. Shi et al. (2011) propose a method to effectively classify indoor environments into semantic categories using KinectTM data. By constructing fast DEFS (efficient feature selection algorithm) and support vector machine (SVM) classifiers, a fast scene classification algorithm is realized. Huang et al. (2020) studied the scene recognition method from 3D point cloud (or voxel) data and showed that its performance is much better than that based on 2D data.

3. METHODS

Based on the PointNet + + model, this paper builds a dual-task model where indoor scene recognition and semantic segmentation can be constructed at the same time, and the association rules of scenes and elements are constructed for the correction of scene recognition results. The process of this paper is shown in Figure 1, which is mainly divided into three steps: data loading, two-task model construction, and scene recognition result correction.

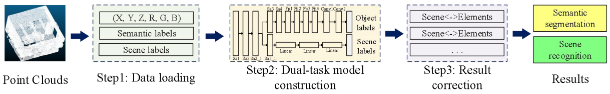


Figure 1. The process of proposed method

The data loading step is mainly to modify the information content output by the original data set loader. The new loader also outputs the object category and scene category in addition to the (x, y, z, r, g, b) information of the point cloud. There are two ways to return a scene category: the first is to assign a scene label to each point, and the second is to treat all the points as a whole with only one scene label. When the dual-task model is built, add a collection abstraction layer (Set Abstraction Layer) (Sa2_1) as the common layer of scene recognition and semantic segmentation. By modifying the dual-task Loss function and optimizing the model storage strategy, the dual-task state hyperparameters are set to improve the operation efficiency of the model. Since the scene is closely related to its internal elements, this paper analyzes the probability of 2191 scenes and 5386 types of indoor elements, obtains the elements related to each type of scene, and then formulates the correlation rules of the scene and elements to correct the scene recognition results.

3.1 Data loading

The traditional point cloud data set loader usually outputs the point cloud (XYZ format or XYZRGB format) and the point cloud labels corresponding to its tasks (e. g., semantic segmentation outputs the object category to which each point belongs). In order to meet the requirements of semantic segmentation and scene recognition, this paper adds two labels, object category and scene category, as output in the output part of the traditional data set. There are two methods for outputting labels. One is point-based loading, that is, assigning scene labels to each point; the other is scene-based loading, that is, viewing all points in the same scene as a whole and assigning only one scene label.

3.2 Dual-task model construction

The steps to build a dual-task model are as follows: (i) add new common feature layers for scene recognition and semantic segmentation, which are used to unify the initial features of the two tasks and ensure the complementary feature dimensions of the two tasks. (ii) change the original PointNet + + Loss function from Nll to the Cross Entropy function, which is more beneficial to multi-label classification tasks. Calculates the losses for scene recognition and semantic segmentation, respectively, and adds the values of the two losses for new models. (iii) optimizes the model storage strategy and adopts a single-task optimum strategy to improve the feature complementarity function of the dual-task model training. The structure of the dual-task model is shown in Figure 2, where Sa is the set abstraction layer, the purpose of Sa is to downsample the features of point clouds, and Fp is the feature propagation layer for the upsampling of point clouds.

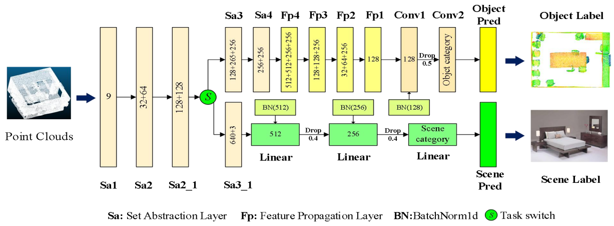


Figure 2. Dual-task model structure.

(1) Add a new common feature layer

The purpose of adding the common layer is to unify the original input features of semantic segmentation and scene recognition so as to facilitate the fusion and complementarity of subsequent dual-task results and deepen the feature depth of the acquisition. In the Sa 2_1 layer, the local search radius of feature extraction is 0.2, 0.4, and 0.8, respectively, and the number of points sampled by each search radius is 32, 64, and 128. The extracted features are processed by three MLP (multilayer perceptrons), and the parameters of the three sets of multi-layer perceptrons are [64, 64, 128], [128, 128, 256], and [128, 128, 256], respectively. The structure of the Sa 2_1 layer is shown in Figure 3.

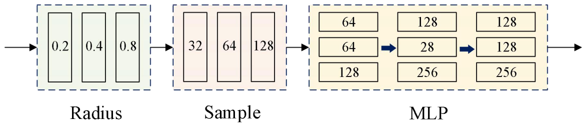


Figure 3. Sa 2_1 layer structure.

(2) Task state control

Add a dual-task state control hyperparameter, set the state of the two tasks of scene recognition and semantic segmentation when the mode initializes, and realize free task control in the model training and application stages. The main purpose is to improve the efficiency of the model. In addition, the model trained under dual tasks can be set to enable a single task in the application stage to further improve the flexibility and efficiency of the model application. When both tasks are turned off, the model will automatically open the scene recognition task by default to improve the model's stability.

(3) Loss of the dual-task model

(i) The loss function used by the original model is NLL_Loss. It is found that NLL_Loss is not applicable to the dual-task model (see Section 5.2 for detailed analysis). Therefore, the loss function of scene recognition and semantic segmentation is modified to a cross-entropy function. When the loss function adopted by the dual task is inconsistent, the final accuracy of the model is even lower.

(ii) The final model Loss is:

$$Loss = k1 \times Loss1 + k2 \times Loss2 \qquad （1）$$

In Equation (1), k1 and k2 in represent the weight of scene recognition and semantic segmentation loss values., and the value interval is [0,1]. For specific tasks, the value can be set independently according to the importance of the task. The default value is set to 1, meaning that the two tasks are equally important.

(4) Model storage strategy optimization

One of the purposes of constructing the dual-task model is to combine the global features and object features in scene recognition and semantic segmentation tasks and improve the effect of two tasks, so this paper optimizes the model parameter storage strategy using the single-task optimal strategy to save the model parameters (as shown in Figure 4).

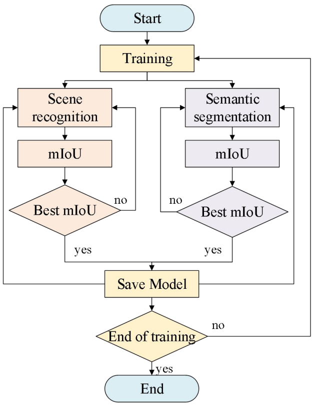


Figure 4. Single-task optimal model storage strategy.

In the training process, as long as the scene recognition and semantic segmentation tasks have a higher than the current optimal mIoU (the initial two tasks optimal mIoU are set to 0) to save the model parameters, the operation can make the scene recognition task use the optimal results of semantic segmentation and provide object-level local features, and the scene recognition task can also use the scene-level global features for the semantic segmentation task.

3.3 Correction of the scene recognition results

Due to the complexity of indoor scenes and the similarity of the layout, the characteristics of different scenes have similarities, so only using the deep learning model may lead to scene recognition errors. In general, the type of indoor scene has a strong correlation with the indoor elements contained (such as furniture, etc.). The elements contained in the room define the category of the scene. Therefore, this paper analyzes the probability of indoor scene elements, formulates the rules between scene and elements, assists scene recognition, and corrects the results. The process is shown in Figure 5.

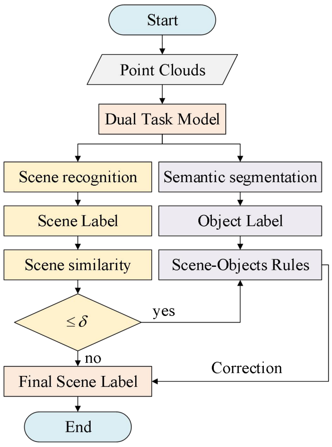


Figure 5. Correction Process of scene recognition results.

https://doi.org/10.5194/isprs-annals-X-1-W1-2023-469-2023 | © Author(s) 2023. CC BY 4.0 License.

(1) Construction of scene-element association rules

This paper determines the correlation between scenes and elements by counting the frequency of elements in common indoor scenes. The basic data used is the MatterPort 3D data set, and the occurrence probability of 2191 scenes and 5386 indoor elements is counted. According to the probability of common elements appearing in the scene, combined with the prior knowledge of human beings on the scene, the correlation rules of scenes and elements are constructed. The steps are as follows:

(i) Screening the elements with an occurrence probability greater than 85% in all scenes;

(ii) According to human cognition of the relationship between elements and scenes, simplify the results of (i) and build a sample of scene-element relationships;

(iii) The association rule algorithm mines the scene-element relationship sample built by (ii) and builds the scene-element association rules.

(2) Correction of the scene recognition results

The output of the dual-task model includes the probability of the scene to be identified and each scene category, and the correction of the scene recognition result is based on the similarity of the probability of the scene to be selected. The scene similarity threshold δ is set. When the difference in the probability of the scene to be identified is less than δ, the two scenes are considered to be similar, and the scene-element correlation rule constructed by (1) is used for similar scenes.

4. EXPERIMENTS AND ANALYSIS

4.1 Experimental data and the environment

(1) Experimental data

Experimental data for semantic segmentation and scene recognition are obtained from S3DIS(Armeni, I. et al, 2016) and Matterport3D(Chang, A. et al, 2017). The S3DIS dataset contains indoor point clouds of six regions, each with spatial coordinates, RGB information, and labels. Matterpot3D is a large indoor scene data set consisting of several types of annotation: color and depth images, camera poses, texture 3D mesh, building plans and area annotation, and object instance semantic annotation. Experimental data are presented in Table 1.

scene	dataset	training	validation
bedroom	Matterport3D	13	3
conference	S3DIS	8	3
hospital	Matterport3D	1	1
office	S3DIS	15	3
saloon	Matterport3D	10	3
parlor	Matterport3D	16	4

Table 1. Experimental dataset.

The preprocessing of the S3DIS and Matterport3D datasets is as follows:

(i) Reclassification of the point cloud categories into 15 categories: walls, ceilings, floors, doors, beams, boards, windows, beds, chairs, tables, bookshelves, pillows, cushions, sofas, and others.

(ii) Matterport3D Point clouds in the data set cannot be used directly for the proposed method, so this paper developed a tool to convert point clouds in ply format into (x, y, z, r, g, b) formats.

(iii) The true labels of the scenes and objects in the training set are obtained from the S3DIS and Matterport3D datasets, where the scene labels of the Matterport3D dataset are manually determined.

(2) Experimental environment

The experimental environment is equipped with two E5-2680 (2.4 GHz) CPUs and two Tesla T4 (16 GB) GPUs with 128 GB of memory. The experimental environment system is Ubuntu 20.04.6 LTS, and the deep learning framework is Pytorch 2.0.

4.2 Experimental details and evaluation indicators

(1) Experimental details

(i) Due to the large size of the point clouds, the original point clouds need to be sampled. In this paper, the sample size is 4096. During model training, the batch size was set to 24, training momentum was set to 0.9, learning rate was set to 0.001, decay rate was set to 0.0001, and the optimizer used in the model was Adam (Adaptive Moment Estimation, adaptive moment estimation), with a total of 50 epochs.

(ii) Save the model every five epochs, and the optimal model is set to be saved once when the mIoU of scene recognition and semantic segmentation tasks is larger than the respective current optimal mIoU.

(2) Evaluation indicators

In this paper, the overall accuracy (oAcc), (Acc) and (mIoU) are used to evaluate semantic segmentation and scene recognition results.

$$oAcc = \frac{\sum TP}{TN+FN+TP+FP} \tag{2}$$

$$Acc = \frac{TN+TP}{TN+FN+TP+FP} \tag{3}$$

$$mIoU = \frac{1}{n+1}\sum_{i=0}^{n}\frac{TP}{TN+FN+FP} \tag{4}$$

TP (True positive) represents the sample number of positive cases predicted by the model, and the actual value is also positive. FP (False positive) represents the number of samples that the model predicts are positive examples, but the actual value is negative examples. The TN (true negative) represents the model prediction as a negative example, and the actual value is also the sample number of negative examples. FN (False negative) represents the sample number of positive cases predicted by the model, but the actual value is negative cases.

4.3 Experimental analysis of indoor scene recognition

This section mainly analyzes the scene recognition results from three aspects: the influence on scene recognition accuracy based on point and scene, the role of the model storage optimization strategy, and the effect of the correction module.

(1) Analysis of scene recognition results

In this paper, the core concept of the scene recognition model is to make the model learn the global and object features of the scene at the same time. In addition, the semantic segmentation module also provides element-level semantic features for the scene recognition task to improve the effect of the scene recognition model. As shown in Figure 6, the change curve of the loss and the overall accuracy on the dual task with the increase in training times, the contribution of semantic segmentation to scene recognition is gradually enhanced. Meanwhile, the global features acquired by scene recognition also provide additional features for the semantic segmentation model, which has the effect of mutually promoting dual tasks.

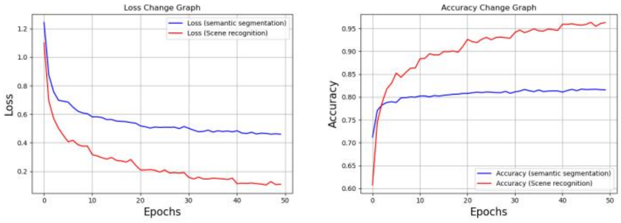


Figure 6. The change curve of loss and the accuracy of the proposed model.

In this paper, scene recognition is realized by fine-tuning PointNet and PointNet + + models, and it is compared with the method of this paper. The specific results are listed in Tables 2 and 3.

model	oAcc	Acc	mIoU
Ours	76.5	80.6	72.2
PointNet-cls	52.9	56.9	44.2
PointNet++-cls-msg	64.7	69.4	55.0

Table 2. Overall accuracy of indoor scene recognition (%).

scene	bedroom	parlor	hospital	restaurant	office	conference
criteria	Acc	Acc	Acc	Acc	Acc	Acc
Ours	100.0	66.7	100.0	50.0	100.0	66.7
PointNet-cls	73.7	30.6	100	61.0	89.1	43.8
PointNet++-cls-msg	66.7	33.3	100.0	50.0	100.0	66.7

Table 3. Single scene accuracy of indoor scene recognition (%).

Through the comparative analysis of Table 2 and Table 3, the overall scene recognition accuracy of this method is 76.5%, the average recognition accuracy is 80.6%, and the mIoU is 72.2%, which is higher than the comparison model, indicating that the proposed method can not only realize scene recognition based on point clouds but also has high scene recognition accuracy.

(2) The impact of point-based and scene-based data loading mode on the model

Tables 4 and 5 list the results of the model under point-based and scene-based data loading methods. By analyzing the numbers in Tables 4 and 5, the scene-based data loading method is better than the point-based.

model	oAcc	Acc	mIoU
Scene-based	76.5	80.6	72.2
Point-based	71.2	70.8	53.3

Table 4. Results of point-based and scene-based data loading modes on the overall accuracy (%).

scene	bedroom	parlor	hospital	restaurant	office	conference
criteria	Acc	Acc	Acc	Acc	Acc	Acc
Scene-based	100	66.7	100	50	100	66.7
Point-based	74.5	43.1	99.5	60.7	88.0	59.1

Table 5. Results of point-based and scene-based data loading modes on the single scene accuracy (%).

(3) Optimization experiment of model storage strategy

This paper optimizes the storage strategy of the dual-task model to improve the overall effect of the model. In order to verify the effectiveness of the single-task optimal strategy, this paper compares the single-task optimal and full-task optimal strategies with the same parameters, environment, and data. The specific experimental results are listed in Tables 6 and 7.

model	oAcc	Acc	mIoU
Single task optimum	76.5	80.6	72.2
Full task optimum	70.6	75.0	61.1

Table 6. Results of different model storage strategies – overall accuracy (%).

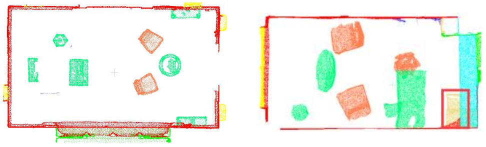
scene	bedroom	parlor	hospital	restaurant	office	conference
criteria	Acc	Acc	Acc	Acc	Acc	Acc
Single task optimum	100	66.7	100	50	100	66.7
Full task optimum	100	33.3	100	50	100	66.7

Table 7. Results of different model storage strategies – single scene accuracy (%).

As listed in Tables 6 and 7, the single-task optimal model storage strategy is better than the full-task optimal storage strategy in the overall accuracy, so the proposed method is effective.

(4) Correct the experimental results of the module

Through the analysis of the experimental results from (1) to (3), it can be found that among the six types of scenes, the parlor recognition accuracy is the lowest. After analysis, the main reason is that the elements in the parlor scene overlap with other scenes, and there are no exclusive key elements in the parlor, so there are mispoints. Figure 7 shows the top view of a parlor scene and an office scene. It can be seen that the elements in the parlor both exist in the office, but the elements in the office do not exist in the parlor (as shown in figure (b) at the red box). Therefore, the results directly identified by the model may have certain errors.



(a) Parlor (b) Office

Figure 7. Comparison of parlor and office scenes.

In order to solve the scene recognition errors caused by indoor scene similarity, this paper corrects the model scene recognition results by introducing scene-element association rules, and the corrected experimental results are listed in Tables 8 and 9.

model	oAcc	Acc	mIoU
Model only	76.5	80.6	72.2
Model and association rule	82.4	84.7	77.8

Table 8. Results of the correct module – overall accuracy (%).

scene	bedroom	parlor	hospital	restaurant	office	conference
criteria	Acc	Acc	Acc	Acc	Acc	Acc
Model only	100	66.7	100	50	100	66.7
Model and association rule	100	66.7	100	75.0	100	66.7

Table 9. Results of the correct module - single scene accuracy (%).

It can be seen from the data listed in Tables 9 and 10 that after the addition of the scene-element association rule, the overall accuracy of the revised model increases by 5.9% compared with the uncorrected model. Therefore, correcting the results of the scene recognition through the scene-element association rule can further improve the recognition accuracy of the scene.

4.4 Experimental analysis of semantic segmentation

The PointNet, PointNet + +, and PointNet-MRF (Li et al., 2019) and PointNet + +-MRF (Jiang et al., 2023) models were selected for comparative analysis. Table 10 lists the experimental results.

model	evaluation criterion		
	oAcc	Acc	mIoU
PointNet	75.9	56.3	45.3
PointNet++	80.1	59.4	50.7
PointNet-MRF	76.3	54.7	44.1
PointNet++-MRF	81.3	62.1	53.9
Ours	98.9	91.4	87.8

Table 10. Overall accuracy of point cloud semantic segmentation (%).

According to the data listed in Table 10, it can be seen that the dual-task model constructed in this paper significantly improves the semantic segmentation accuracy of point clouds, indicating that it is effective to apply the global scene features to the semantic segmentation task of point clouds. The results of the partial semantic segmentation of the model and the contrast model in this paper are shown in Figure 8.

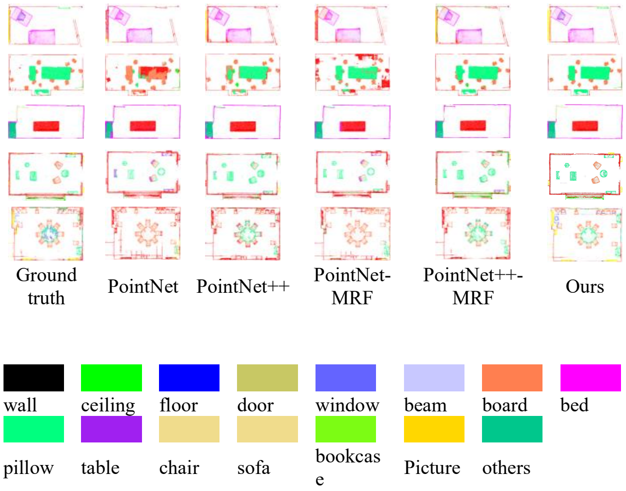


Figure 8. Results of point-cloud semantic segmentation.

5. DISCUSSION

5.1 Influence of a single task on the model

The dual-task model constructed in this paper can control the open state of the task by setting the hyperparameters. In order to verify the support of the features acquired by semantic segmentation for scene recognition, this paper constructs a comparison experiment in which only the task of scene recognition and the dual task are turned on. The experimental results are listed in Tables 11 and 12.

model	oAcc	Acc	mIoU
Scene recognition and semantic segmentation	76.5	80.6	72.2
Scene recognition only	70.6	72.2	61.1

Table 11. Results of single task on scene recognition – overall accuracy (%).

scene	bedroom	parlor	hospital	restaurant	office	conference
criteria	Acc	Acc	Acc	Acc	Acc	Acc
Scene recognition and semantic segmentation	100	66.7	100	50	100	66.7
Scene recognition only	33.3	33.3	100	100	100	66.7

Table 12. Results of single task on scene recognition - single scene accuracy (%).

As listed in tables 11 and 12, both overall accuracy and single scene recognition accuracy show that the double task open effect is better than single task scene recognition accuracy, and the double task scene's overall accuracy increased by 5.9%, while the average accuracy increased by 8.4%, indicating that this paper's approach of adding semantic segmentation features to the scene recognition accuracy is effective.

5.2 Influence of the Loss function on the model

The loss function adopted by the original PointNet++ is Nll loss. Since the cross entropy function is more suitable for classification tasks, this paper replaces the cross entropy function. In order to verify the effectiveness of the modification, this paper constructs a comparative experiment under different loss combinations, and the experimental results are listed in Tables 13 and 14.

model	oAcc	Acc	mIoU
Ours	76.5	80.6	72.2
Nll Loss	64.7	69.4	55.3
Cross Entropy + Nll	64.7	70.8	56.7

Table 13. Results of different Loss function – overall accuracy (%).

scene	bedroom	parlor	hospital	restaurant	office	conference
criteria	Acc	Acc	Acc	Acc	Acc	Acc
Ours	100	66.7	100	50	100	66.7
Nll Loss	66.7	33.3	100	50.0	100	66.7
Cross Entropy + Nll	66.7	66.7	100	25.0	100	66.7

Table 14. Results of different Loss function - single scene accuracy (%).

As can be seen from the data listed in Table 13 and 14, the scene is the best effect, so the loss function is effective in this paper.

5.3 Influence of the categories of indoor scene elements on the model

The type of a scene is generally determined by the elements contained inside it (for example, the space existing in the bathtub is generally defined as the bathroom scene). Due to the similarity of the indoor scenes, the same elements will appear in different scenes. When the elements existing in different scene categories are consistent, the scene recognition task can only use global features. In order to verify the effect of scene recognition, this paper constructs the elements category consistent scene recognition comparison test, with the specific experimental results shown in tables 15 and 16. The data is listed in 4.1, and only the elements common to different scenes are loaded during training.

Loading mode	oAcc	Acc	mIoU
Point-based	28.8	26.3	14.6
Scene-based	35.7	35.6	22.3

Table 15. scene recognition results when element categories are consistent-overall accuracy (%).

scene	bedroom	parlor	hospital	restaurant	office	conference
criteria	Acc	Acc	Acc	Acc	Acc	Acc
Point-based	49.5	34.5	15.3	46.2	5.4	7.0
Scene-based	80.0	20.0	40.0	40.0	0	33.3

Table 16. scene recognition results when element categories are consistent-single scene accuracy (%).

According to the results of Table 15 and 16, it can be found that when the internal element categories are consistent, the results of scene recognition are much lower than the scenes with different categories, and the scene-based loading mode is better than the point-based loading mode.

6. CONCLUSION

This paper constructs a dual-task model that can perform both scene recognition and semantic segmentation based on PointNet++, and draws the following conclusions through experimental analysis:

(1) The dual-task model constructed in this paper can realize the scene recognition and semantic segmentation tasks at the same time, and the accuracy is higher than that of a single task, and the accuracy also has great advantages compared with the latest model.

(2) The scene recognition results can be effectively corrected through the scene-element association rules, especially when the indoor scene category coincidence degree is high.

(3) In model training, the scene-based method of data loading is more accurate than the point-based method; the use of the cross entropy function is more conducive to the training of the dual-task model; and the saving of the single-task optimal strategy can improve the final accuracy of the model.

(4) The element categories contained in different indoor scenes have a great influence on the scene recognition results.

The dual-task model of scene recognition and semantic segmentation constructed in this paper provides a method for robotic indoor space understanding.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant Numbers 41872207 and 41961063), the National Key R&D Program of China (Grant Number 2019YFE0123300) and the Introduction Plan of High-End Foreign Experts (Grant Number G2021025006L).

REFERENCES

Afif, M., Ayachi, R., Said, Y., & Atri, M. (2020). Deep learning based application for indoor scene recognition. Neural Processing Letters, 51, 2827-2837.

Ali, H., Kabir, S., & Ullah, G. (2021). Indoor scene recognition using ResNet-18. International Journal of Research Publications, 69(1), 7.

Basu, A., Petropoulakis, L., Di Caterina, G., & Soraghan, J. (2020). Indoor home scene recognition using capsule neural networks. Procedia Computer Science, 167, 440-448.

Chen, G., Song, X., Zeng, H., & Jiang, S. (2020). Scene recognition with prototype-agnostic scene layout. IEEE Transactions on Image Processing, 29, 5877-5888.

Chen, L., Bo, K., Lee, F., & Chen, Q. (2020). Advanced feature fusion algorithm based on multiple convolutional neural network for scene recognition. Computer Modeling in Engineering & Sciences, 122(2), 505-523.

Du, D., Wang, L., Wang, H., Zhao, K., & Wu, G. (2019). Translate-to-recognize networks for rgb-d scene recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11836-11845).

Espinace, P., Kollar, T., Soto, A., & Roy, N. (2010, May). Indoor scene recognition through object detection. In 2010 IEEE International conference on robotics and automation (pp. 1406-1413). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Huang, S., Usvyatsov, M., & Schindler, K. (2020, October). Indoor scene recognition in 3D. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 8041-8048). IEEE.

Jiang, J., Li, F., Yang, J., Kang, Z., & Li, J. (2023). Construction of indoor obstacle element map based on scene-aware priori obstacle rules. ISPRS Journal of Photogrammetry and Remote Sensing, 195, 43-64.

Li, F., Wang, H., Akwensi, P. H., & Kang, Z. (2019). CONSTRUCTION OF OBSTACLE ELEMENT MAP BASED ON INDOOR SCENE RECOGNITION. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences.

López-Cifuentes, A., Escudero-Vinolo, M., Bescós, J., & García-Martín, Á. (2020). Semantic-aware scene recognition. Pattern Recognition, 102, 107256.

Miao, B., Zhou, L., Mian, A. S., Lam, T. L., & Xu, Y. (2021, September). Object-to-scene: Learning to transfer object knowledge to indoor scene recognition. In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 2069-2075). IEEE.

Mosella-Montoro, A., & Ruiz-Hidalgo, J. (2021). 2d – 3d geometric fusion network using multi-neighbourhood graph convolution for rgb-d indoor scene classification. Information Fusion, 76, 46-54.

Nascimento, G., Laranjeira, C., Braz, V., Lacerda, A., & Nascimento, E. R. (2018). A robust indoor scene recognition method based on sparse representation. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 22nd Iberoamerican Congress, CIARP 2017, Valparaíso, Chile, November 7–10, 2017, Proceedings 22 (pp. 408-415). Springer International Publishing.

Njoku, J. N., Amaizu, G. C., Lee, J. M., & Kim, D. S. (2022, February). Real-time deep learning-based scene recognition model for metaverse applications. In Proceedings of the KICS Winter Conference, Pyeongchang, South Korea (pp. 195-198).

Pal, A., Nieto-Granda, C., & Christensen, H. I. (2019, November). Deduce: Diverse scene detection methods in unseen challenging environments. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 4198-4204). IEEE.

Pereira, R., Gonçalves, N., Garrote, L., Barros, T., Lopes, A., & Nunes, U. J. (2020, April). Deep-learning based global and semantic feature fusion for indoor scene classification. In 2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC) (pp. 67-73). IEEE.

Qiao, Z., Yuan, X., Zhuang, C., & Meyarian, A. (2021, January). Attention pyramid module for scene recognition. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 7521-7528). IEEE.

Romero-Gonzalez, C., Martinez-Gomez, J., Garcia-Varea, I., & Rodriguez-Ruiz, L. (2016). 3D spatial pyramid: descriptors generation from point clouds for indoor scene classification. Machine Vision and Applications, 27, 263-273.

Seong, H., Hyun, J., & Kim, E. (2020). Fosnet: An end-to-end trainable deep neural network for scene recognition. IEEE Access, 8, 82066-82077.

Shi, L., Kodagoda, S., & Ranasinghe, R. (2011, December). Fast indoor scene classification using 3D point clouds. In Proceedings of the 2011 Australasian Conference on Robotics and Automation.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Song, X., Jiang, S., Wang, B., Chen, C., & Chen, G. (2019). Image representations with spatial object-to-object relations for RGB-D scene recognition. IEEE Transactions on Image Processing, 29, 525-537.

Wang, C., Peng, G., & De Baets, B. (2020). Deep feature fusion through adaptive discriminative metric learning for scene recognition. Information Fusion, 63, 1-12.

Xia, S., Zeng, J., Leng, L., & Fu, X. (2019). WS-AM: weakly supervised attention map for scene recognition. Electronics, 8(10), 1072.

Xiong, Z., Yuan, Y., & Wang, Q. (2019). RGB-D scene recognition via spatial-related multi-modal feature learning. IEEE Access, 7, 106739-106747.

Xiong, Z., Yuan, Y., & Wang, Q. (2020). MSN: Modality separation networks for RGB-D scene recognition. Neurocomputing, 373, 81-89.

Xiong, Z., Yuan, Y., & Wang, Q. (2021). ASK: Adaptively selecting key local features for RGB-D scene recognition. IEEE Transactions on Image Processing, 30, 2722-2733.

Zhu, H., Weibel, J. B., & Lu, S. (2016). Discriminative multi-modal feature fusion for rgbd indoor scene recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2969-2976).