Abstract

With the rapid development of new indoor sensors and 3D model securing techniques, the amount of indoor three dimensional (3D) point cloud models was incredibly expanded. Be that as it may, these huge "dazzle" point clouds are hard to apply for some advanced indoor applications and GIS analysis. It therefore, has a large demand for semantic segmentation of indoor 3D point clouds. In this work, the spatial aspects of semantic segmentation is done using convolution neural network (CNN). An existing structured data set is used to predict the point cloud and machine learning model accuracy. The data splits into training, testing data. Both the testing and training undergo the training process. The resultant training, testing data compare with the state of the art accuracy count, and it isvisualized.

Earlier it was difficult to handle the troubles of lacking of hearty 3D highlights and 3D training information in 3D point clouds segmentation preparation, as well as the high impediment, the lopsided light and complex objects in the indoor condition. Then the compelling calculation for marking spread from 2D semantic pictures to crude 3D point clouds is proposed right now the foundation of 3D semantic point cloud, which can viably address the issue of object semantic lacking and indistinct spatial structure in indoor 3D point cloud model. In view of the modeling attributes of SfM point cloud model, our Algorithm utilizes a huge number of existing 2D picture databases and develop picture semantic order calculations to gauge the structure design and semantic marks of the pictures, and transmit the data as semantic names to 3D point cloud information, subsequently, reduce the trouble of structure and semantic Extraction of 3D point cloud. So as to show the powerful semantic segmentation perform Ance of complex indoor picture, we make a new design called Large-scale Residual Connection to transmit spatial data from low-stage to the higher ones, in addition, the Atrous Spatial Pyramid Pooling (ASPP) of DeepLabv3+ and the DenseBlock structure of DenseNet Furthermore, a multi-organize training system is contrived to beat the issue of the high impediment and complex objects in the indoor condition.

Our proposed strategy, which focuses on the strong semantic segmentation of indoor 3D point cloud, is fundamental made out of two significant developments. Initially, a novel Combined Network has demonstrated to mark 2D pictures and gauge indoor spatial design, which not exclusively can improve the characterization Secondly, 2D-3D name engendering dependent on a chart model, which executes the name spread from 2D to 3D, and afterward develop relevant consistency between pictures consistency to understands the right in particular, we don't require any training information from 3D scene, and we can get acceptable 3D point cloud arrangement brings about increasingly complex indoor scenes and achieved 87% accuracy. The examinations were directed on an open dataset (NYUDv2 indoor dataset) and a neighborhood dataset. In assessment with the top tier procedures to the extent 2D semantic segmentation, Deeplabv3+ can both learn discriminative enough Features for between class segmentation while sparing clear cutoff points for intra-class.

Key words: semantic 3D point cloud, semantic label transfer, semantic segmentation, structure extraction, graph model, object semantic label, CNN,

Table of Contents

List of Figure

1. Introduction

1.1 Research background and meaning

With the expansion of location-based services and GIS applications from the outdoor environment to the indoor environment (Zhou et al., 2017), the precision indoor 3D model supported by indoor data acquisition technology and sensors has become a research hotspot in the current GIS field. At present, a large amount of indoor 3D point cloud data can be obtained through laser radar scanning technology, RGB-D camera and SfM. However, the development of intelligent devices has placed higher demands on current indoor modeling technologies. The original 3D modeling technology was designed to restore a geometrically complete indoor 3D point cloud model for a good visual experience. However, with the development of augmented reality, service robots and unmanned vehicles, smart devices have become one of the target users of modeling. Their zero-precursive understanding of real-world scenes makes only 3D-dimensional point clouds with geometric characteristics. Apply again. Therefore, the “blind point cloud” with missing objectified information (Hermans et al., 2014; Lu et al., 2016; Dai et al., 2017; Tchapmi et al., 2017) cannot be applied to higher levels of intelligence Service.

Compared to only focus on the recovery point cloud model geometry (Furukawa et Al, 2010a; Newcombe et Al, 2011; Wu, 2011) , the indoor semantic model based on the full 3D point cloud point cloud model, meaningless " cloud blind spots " semantically segmented object- oriented, making it a 3D-dimensional model semantic meaningful (and Yong Wang,, 2015) . Compared with the outdoor environment with open view and classification rules, the indoor space is prone to high occlusion, uneven illumination (glare, highlight), and some weak texture areas. Moreover, the indoor objects are complicated and complex, which causes the semantic segmentation of the indoor 3D scene. A series of difficulties (Tang Shengjun , 2017) . However, indoor spaces mostly exhibit horizontal and vertical structural features.

Compared with the 2D-dimensional plane map, the indoor 3D point cloud model not only has the advantages of high geometric precision and strong visual reality, but also provides people with rich scene knowledge and semantic information, and can support the retrieval and tracking of indoor objects. The large demand for building indoor 3D models has facilitated the development of various indoor 3D modeling techniques. Compared to lidar scanners and RGB-

D cameras (Kinect , Tangle, etc.), SfM (Motion Structure Restoration) modeling technology maps unordered indoor 2D images into 3D space by solving camera parameters (Agarwal et al , 2009) , and get a fine room through PMVS (dense reconstruction) technology.

The internal 3D-dimensional point cloud model (Furukawa et al., 2010b) has low cost and high popularity, and can quickly and automatically construct a large-scale indoor 3D point cloud model (Hedman et al., 2016; Lu, 2016) . Therefore, this study uses SfM-PMVS modeling technology to construct the basic point cloud data of point cloud segmentation experiments.

For the semantic classification of indoor 3D point clouds, most of the methods for point cloud classification by directly extracting point cloud features are difficult to be widely applied due to the lack of robust 3D feature operators and sufficient 3D training data sets. In recent years, deep learning algorithms have brought about technological innovations for 2D image segmentation (Chiu et al., 2019; Fu et al., 2019; Liu et al., 2019), using neural network models for mobile phones, cameras, etc. Narrow dimensional interior image extracting mobile device photographed reality semantic information, and multi-view images modeling principle (Furukawa et al., 2015) is formed in a 2D-dimensional semantic " cross- domain " channels transferred to a 3D-dimensional indoor The implementation of semantic segmentation of point clouds provides another solution. At the same time, deep learning technology has been in the field of 2D-dimensional image segmentation for decades, and has formed a large number of large-scale classification categories of standard indoor 2D training data sets, providing sufficient space for designing indoor 2D image semantic segmentation networks.

In summary, the popularity of mobile smart devices and the development of visually oriented indoor location-oriented services and indoor 3D GIS systems urgently require us to more intelligently understand the semantics of indoor 3D scenes, thus making indoor 3D point clouds practical. Play more value in the app. However, due to the lack of robust 3D feature operators and 3D training data, the method of directly using 3D features for point cloud classification has certain limitations. In addition, the complexity of the indoor environment poses a significant challenge to the precise semantic classification of indoor 3D point clouds (Choi et al., 2015). The image data collected by the convenient and usable mobile devices (mobile phones, cameras) provides a low-cost, rich and usable data foundation for building indoor 3D scenes; the cross-development of multidisciplinary theory (computer science.

1.2 Research questions and meaning

Semantic segmentation of indoor 3D point cloud models is an important part of building intelligent location-oriented services and indoor GIS systems. The multidimensional nature of 3D point cloud data and the complexity of the indoor environment pose a series of challenges for the accurate semantic segmentation of indoor 3D point cloud models. The increasingly mature deep learning algorithm provides a new solution to the semantic semantic segmentation of 2D-dimensional images. How to design a deep learning network model for accurate semantic segmentation of indoor 2D images and how to use the semantic features of indoor 2D images to assist the semantic segmentation of indoor 3D point clouds is the main scientific problem that this paper needs to solve. Focusing on this scientific research topic, this paper discusses the following three parts: (1) Compared with the outdoor environment, indoor space often has high occlusion, uneven illumination (glare, highlight), and some weak texture areas. The complexity of indoor objects brings a series of challenges to the fine semantic segmentation of indoor 2D images. How to design a deep network to extract complex indoor object features, how to combine indoor object semantics and indoor spatial layout structure to alleviate the impact of indoor complex conditions on point cloud segmentation and obtain finer indoor 2D image segmentation results? This is the basis for implementing 2D-dimensional semantic transfer. (2) How to use the semantic information of 2D-dimensional images to assist the 3D-dimensional point cloud for semantic segmentation. How to establish a "bridge "for 2D-dimensional semantic transfer is the core issue of this research. (3) For the semantic inconsistency of multiple 2D- dimensional pixels corresponding to the same 3D-dimensional point in the 2D-dimensional semantic transfer process, how to establish the context semantic consistency constraint to improve the accuracy of semantic transfer. This provides a further guarantee for the accuracy of the final 3D point cloud semantic segmentation.

Based on the deep foundation of the current deep learning algorithm in the field of 2D image semantic segmentation, and the large-scale rich available indoor 2D image open training dataset, this paper studies the 3D point cloud segmentation method based on 2D and 3D semantic transfer, firstly for indoor complex The scene design deep learning neural network performs indoor 2D image objectification and structured semantic extraction, and then builds a bridge of 2D and 3D semantics based on the guideline of multi-see picture modeling.

⑴ Provides an answer for the semantic segmentation of 3D point cloud, utilizing the critical focal points of the present profound learning algorithm in the field of 2D picture semantic segmentation, assisting with illuminating the confusion and turn of 3D point cloud comparative with 2D data. The issue of semantic segmentation brought about by sex gives a data establishment to the application of indoor 3D point cloud model in actual creation and life.

⑵ By developing the Deeplabv3+ profound neural network, utilizing the indoor spatial design structure to enhance the current indoor scene getting innovation, somewhat alleviate the troublesome issue of semantic segmentation brought about by complex scenes in indoor space, and the intelligent semantics of indoor environment Perception, the intelligent administration of indoor space offers a specific technical help.

⑶ Semantic comment of the indoor 3D "dazzle point cloud" by methods for 2D- dimensional semantic exchange, which has improved the application value of the good for nothing "daze point cloud" which is hard to interface with the actual application to a limited degree, enhancing the point cloud data. The application mode, and utilize the minimal effort 2D pictures obtained by the portable terminal, establishes a technical framework for the essential 2D-dimensional data to give a more elevated level of intelligent value-included services.

1.3 Research status at home and abroad

With the rapid development of new sensors and indoor 3D model securing innovation, the number of indoor 3D point cloud models has developed rapidly. What's more, the gigantic futile “daze point cloud “is hard to interface with practical applications. Denoting an enormous number of straightforwardly acquired good for nothing, 3D points as semantic 3D point clouds is beneficial to the huge scale application of point clouds in real creation and life. In this manner, the semantic segmentation of indoor 3D point clouds has become an exploration hotspot in numerous fields (Hermans et al., 2014; Husain et al., 2016; Liu et al., 2017). At present, the methods of 3D point cloud semantic segmentation can be generally isolated into three sorts: 3D point cloud based on 3D geometric features, 3D local shape features and surface features; nonetheless, it is hard to develop an adequately powerful 3D Based on the component administrator, based on the immediate extraction of 3D features for point cloud characterization, an enormous number of researchers utilize the profound learning algorithm to get familiar with the 3D point cloud model features based on the 3D model training data set, and afterward play out the semantic segmentation of the 3D scene. In any case, the 3D points the absence of cloud model training data blocks the development of point cloud segmentation network models. In the modeling procedure, the 2D-dimensional projection relationship gives a new answer for the semantic segmentation of the 3D point cloud model. Subsequently, based on the profound establishment of the present profundity learning algorithm in the field of 2D picture semantic segmentation, through the 2D-dimensional perceivability relationship “cross-space " conveyance of 2D-dimensional semantics is conceivable.

1.3.1 Point cloud semantic classification based on classical 3D point feature

Currently, one common practice in the 3D point cloud reference semantics division is a 2D-dimensional image research ideas semantic segmentation based on image characteristics, namely direct 3D-dimensional point cloud feature extraction and selection, and then based on 3D feature-by-feature points classification training (Xiong et al., 2011; Koppula et al., 2011; Weinmann et al., 2015; Armeni et al., 2017; Hackel et al., 2017) , and then obtained 3D point cloud semantic classification results (Munoz et al., 2010; Zhou et al., 2016) . Among them, the extraction of 3D features is the key to affect the segmentation results. Commonly used 3D features include: 3D geometric features, 3D local shape features, texture features, and a variety of statistical features. Lichti et al. (2006) assessed the local ebb and flow of a point cloud based on covariance analysis of eigenvalues and eigenvectors and gathered the limit data of the point cloud by the congruity of bend to perform point cloud segmentation; on this premise, Mitr et al. (2003)) by changing the examining point is calculated bend neighborhood point cloud data, the thickness circulation of the example and to extricate a 3D-dimensional component point is normal; Chehata et al. (2009) to extricate different multi-reverberation 3D-dimensional laser point cloud and full waveform highlight and utilize the arbitrary timberland classifier to acquire the semantic point cloud; simultaneously, on the grounds that the indoor environment generally presents horizontal and vertical normal structures, for example, Manhattan (Manhatan) structure, Planar (Planar) structure, etc., it very well may be built by RANSAC shape estimation algorithm. Parametric models to fit different geometric shapes in a point cloud, for example, planar shapes (Schnabel et al., 2010; Holzmann et al., 2017) , blended shapes (Cura et al., 2018; Lafarge et al., 2010) , And describe their comparing object features. Nonetheless, these methods can just speak to 3D-dimensional point clouds with straightforward geometric shapes, and it is hard to speak to complex scenes with high impediment in the room.

To this end, the utilization of different 3D features for semantic segmentation of 3D point clouds has become a significant method for point cloud segmentation and 3D scene order (Kalogerakis et al., 2010; Song and Xiao, 2016; Dai et al., 2017; Georgakis et Al., 2017) . For complex scenes with high impediment in an indoor environment, Gould (2009) builds up an arrangement model based on areas, induces the interrelationship between objects, and automatically orders scene semantics by joining local surface features and scene geometry features. Koppula et al. (2011) separated the multi-dimensional features of point clouds and combined the contextual connections of point clouds to perform semantic segmentation of 3D point clouds. Tests show that the different features of point clouds and the combination of between relationship features between objects are somewhat it can improve the vigor of 3D eigenvectors. On this premise, Munoz et al. (2010) utilized recursive methods to fragment 3D scenes while consistently including neighborhood squares of a similar classification to accomplish object fitting. The above examinations show that in high-impediment indoor environments, the trouble of semantic extraction is alleviated by consolidating visual features and 3D shape features. In addition, to tackle the commotion in point, cloud data and the tedious issue brought about by point-by-point semantic classification, Aijazi et al. (2013) isolate the point cloud into super-voxels as per geometric and surface qualities, and use the geometric attributes of super-voxels. Also, local descriptors complete the segmentation of point clouds; Valentin et al. (2013) perform up close and personal classifier training on the geometry and setting features of 3D surface models, lastly limit the Markov irregular field issue to acquire the final 3D model semantics. Imprint; Ni et al. (2017) removed the 2D-dimensional re-anticipated picture features based on point cloud segmentation, a progression of eigenvalues and rise features and utilized the arbitrary backwoods classifier to perform semantic classification of point cloud segmentation; Kang(2018) Semantic segmentation of Lidar point cloud is performed by extricating the geometry and radiation qualities based on super-voxels and developing a semantic likelihood chart model utilizing the word pack model .

1.3.2 Point cloud semantic segmentation with 3D neural network

Lately, deep taking in algorithms have been reached out from the field of 2D picture segmentation to 3D point cloud segmentation investigate, turning into a significant algorithm for 3D point cloud semantic comment (Su et al., 2015; Boulch et al., 2017) . Wu et al. (2015) assembled a 3D ShapeNet convolutional neural network to perform 3D shape reclamation based on feature learning of object voxels. Zhou et al. (2018) suggested that VelNet consolidates the forecast consequences of 3D feature extraction and 3D object bouncing box into a solitary stage start to finish trainable deep network, in this manner staying away from the wasteful activity of manual extraction of 3D point cloud features. Charles et al. planned the point cloud segmentation neural network PointNet (Qi et al., 2017a) to acquire the global point cloud feature by learning the spatial coding comparing to each point in the information point cloud. Nonetheless, the local structural irregularity brought about by the metric space confines the exact ID of complex objects in the PointNet network model. To this end, PointNet++ (Qi et al., 2017b) learns the local features of indoor scenes through a hierarchical structure, accordingly improving the generalization capacity of the network model for indoor complex scene acknowledgment. Be that as it may, the above methods have solid reliance on 3D training data, coming about in even the most developed point cloud segmentation innovation can just portion point cloud data in great and controllable environments under comparative training environments and test conditions. In this manner, these methods are not universal in the field of point cloud segmentation. Likewise, the point cloud training data is inadequate, the training data learning time is long, and the memory misfortune in the calculation procedure is also a key factor influencing the exactness and productivity of the point cloud segmentation.

1.3.3 Point cloud semantic segmentation via 2D-3D semantic transfer

So as to alleviate the dependence on 3D training data, a huge number of develop 2D-dimensional picture semantic segmentation methods are utilized to aid the classification of 3D point clouds (Russell et al., 2008; Deng et al., 2009; Xiao et al., 2010; Kuettel et al., 2012) . At present, the profundity learning algorithm has an extraordinary preferred position in the field of 2D-dimensional picture semantic segmentation, and has built a huge scale openly accessible indoor 2D marker training data set, for example, ImageNet (Deng et al., 2009) and LabelMe (Russell et al. , 2008) , InteriorNet dataset (Li et al., 2018), etc. At present, numerous researchers actualize semantic markup of indoor 3D scenes through semantic exchange methods. Dish et al. (2010) get comparing semantic data by moving semantic data of source data to various scenes of a similar measurement, for example, object segmentation (Rakelly et al., 2018) , object acknowledgment and 3D scene understanding (Liu et al ., 2016) . Nan et al. (2012) utilized online integrated data for semantic label delivery. The method finally gets the target semantic 3D point cloud by distinguishing the object with higher likeness in the reference data set and afterward spreading its semantics to the target inquiry object. In any case, traditional semantic label delivery is for the most part done in a similar measurement. Along these lines of semantic explanation is hard to meet our requirements for "cross-space " semantic delivery. To this end, Su et al. (2015) venture the point cloud data from alternate points of view to get a 2D- dimensional feature image, and use it as part of the convolutional neural network training data set to train the network to perform semantic recognition of the 3D-dimensional object. In this operation, compared to training a 3D-dimensional model directly in the network, the training data is changed from 3D-dimensional to 2D-dimensional, which has great advantages in terms of maturity and training speed. The GIFT algorithm also uses the same idea to process and identify 3D-dimensional objects by processing 2D-dimensional image information (Bai et al., 2016). Audeber T, etc. (2016) using SegNet (Badrinarayanan et al, 2017.) Coder - decoder semantically segmented image frame, and by introducing multiple cores to perform convolution in the network to identify and correct the residual, and finally the segmented 2D-dimensional pixel label is backprojected into the 3D-dimensional space to complete the semantic annotation of the 3D-dimensional point cloud. Caltagirone et al. (2017) simplified the semantic segmentation of point clouds by reducing the dimensionality of point cloud data by generating unstructured point cloud data with top view images of several basic statistics such as average elevation and density. Boulch etc. (2018) using the principle of 2D-dimensional modeling to give a 3D-dimensional point cloud model semantic first obtained the point cloud corresponding to the mapping relationship between the cloud point and the multi-view images corresponding to the RGB image and depth map, and then to RGB image is performed CNN by The pixel semantics are estimated and projected into the 3D space to implement semantic category annotation of the point cloud. Compared with direct semantic annotation of 3D point cloud data, this method avoids constructing 3D feature operator and training data set, and realizes semantic object acquisition of 3D point cloud. However, this method is not transmitted in semantic “cross- domain”. The consistency of context semantics is not considered in the process.

In summary, the direct extraction of 3D point features for semantic segmentation of indoor 3D point cloud models makes it difficult to construct robust 3D eigenvectors that can maintain scale, rotation, and photometric stability ; using a deep learning network model to directly perform 3D point clouds Semantic segmentation, lack of large-scale marker 3D point cloud training data; and the use of mature deep neural network for 2D image semantic segmentation to assist 3D point cloud for semantic segmentation can solve the above problems. Therefore, for the above study and shortcomings, this research work based on the current 2D-dimensional image.

1.4 Research Area

Contrasted to the outdoor environment, the indoor space will in general have high impediment and lopsided brightening. Moreover, the indoor space objects are complex and convoluted, which brings a progression of challenges for the fine semantic segmentation of indoor picture objects. In the field of picture semantic segmentation, even the as of now existing moderately propelled deep learning algorithms are hard to create hearty object segmentation results for some complex indoor environments. Since the indoor space usually presents a horizontal and vertical normal structure, the inside space understanding is helped by the indoor space layout, which can improve the strength of the semantic segmentation of the neural network. Along these lines, this paper plans (Deeplabv3+) neural network that can at the same time perform indoor object segmentation and indoor space l ayout estimation for the high impediment and complex complexity of the inside, and the spatial structure features and indoor object semantics. The features supplement one another and combine with one another to deliver an increasingly exact comprehension of the indoor scene. Among them, the principle look into substance include: Deeplabv3+ network structure plan and capacity of each structure, Deeplabv3+ neural network training rules and Deeplabv3+ and current standard semantic segmentation network segmentation results.

(2) 2D-dimensional picture semantic features and spatial structure features helped 3D-dimensional point cloud for semantic segmentation The demand for location-based services and GIS applications in the indoor environment has encouraged the development of indoor 3D data obtaining innovation and sensor refreshes. As of now, 3D point cloud data that can be directly gotten is usually spoken to by a progression of inane "daze point clouds”. The issue of absence of structure and objectification makes it hard to interface with real applications. Denoting the "dazzle point cloud” as a semantic point cloud with practical essentialness is of extraordinary importance in all real life fields. At present, in the field of 2D-dimensional picture segmentation, the deep learning algorithm has accomplished moderately propelled outcomes. In this way, based on the semantic segmentation and spatial layout extraction consequences of indoor 2D picture objects got by Deeplabv3+ neural network in (1) , this paper proposes a 2D-dimensional semantic " cross-area " move method based on the semantic consistency requirement between sees.

1.5 Research route

The technical roadmap for this study is as follows:

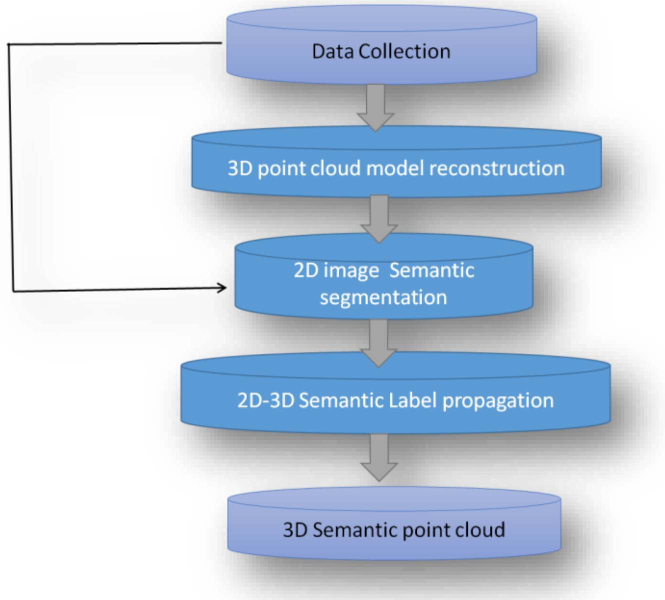

Data Collection
3D point cloud model reconstruction
2D image Semantic segmentation
2D-3D Semantic Label propagation
3D Semantic point cloud

Figure1. 1: Flowchart

(1) The indoor picture data is gathered by common mobile gadgets, such as mobile telephones and cameras , and the indoor 3D point cloud model is constructed by utilizing SfM modeling devices such as VisualSfM , Bundler+PMVS , and the procured picture data.

(2) Based on the structural focal points of the present driving Deeplabv3+ and DenseNet networks in the field of 2D picture segmentation, a Deeplabv3+ neural network that can at the same time perform indoor 2D picture semantic segmentation and spatial layout extraction is constructed.

(3) Constructing a super-pixel semantic markup pool, based on the perceivability connection between 2D-dimensional superpixels and 3D-dimensional points in the SfM modeling process, build up a diagram model of 2D-dimensional semantic exchange.

(4) In the process of 2D-dimensional semantic transfer, due to the one-to-many visual relationship between 3D points and 2D superpixels, it is necessary to construct context semantic consistency constraints to alleviate the semantic confusion caused by the miss-segmentation semantics of 2D images. High-precision 3D point cloud model semantic annotation.

(5) Experimental results and discussion and analysis.

1.6 Methodology Design

I have implemented a new neural network using the free packages available in Python programming language. Python is a programming language and free software environment popular among statisticians and data miners for its ease-of-use, as well as its sophisticated visualizations and analyses. The support for deep learning in Python has grown ever since, with an increasing number of packages becoming available.

Step 1 –Collecting Own Data

Step 2 – Exploring and Preparing the Data

Step 3 – Training a Model on the Data

Specifying the model:

My first model will have 3 nodes in the hidden layer, with a relatively low learning rate of 0.05, sigmoid activation function, and optimized over 600 epochs:

The typical neural network-learning algorithm applies error propagation from outputs to inputs, and gradually fine-tunes the network weights to minimize the sum of error. Each cycle through this learning process is called an epoch

Step 4 – Evaluating Model Performance

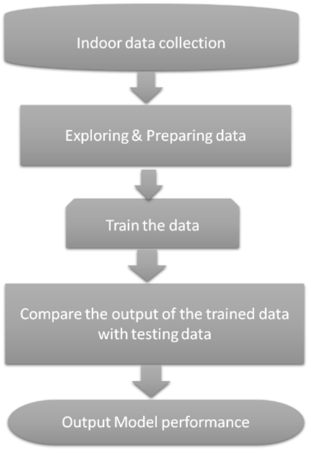

Indoor data collection
Exploring & Preparing data
Train the data
Compare the output of the trained data with testing data
Output Model performance

Figure1.2**: Technical route**

2. Indoor multi-view image semantic modeling

Indoor multi-see picture modeling, based on the standard of SfM utilizes the indoor disarranged images gathered by savvy cell phones to compute the camera parameter network one by one and reestablish the modeling procedure of indoor 3D-dimensional scenes, including: 2D-dimensional picture includes The three essential procedures of point extraction and coordinating, estimation of camera parameters, and rebuilding of spatial 3D arranges. Semantic segmentation based on 3D-dimensional models of indoor multi-see images is a procedure of utilizing 2D-dimensional picture semantics to check the 3D-dimensional point semantics one by one. Based on the profound learning calculation, which is as of now in front of different techniques in the field of 2D picture semantic segmentation, the indoor 2D picture semantics with high segmentation accuracy can be gotten. What's more, SfM modeling a comparing connection between the 2D-dimensional pixel visual 3D-dimensional point cloud, 2D and 3D task was to construct a semantic significance “connect “. To this end, this chapter mainly presents the hypothetical and specialized premise of 3D-dimensional modeling of indoor multi-see images and semantic segmentation of 2D-dimensional images based on profound learning calculations. Segment 2.1 presents the essential theoretical data on indoor multi-see picture modeling; Section
2.2 presents the specialized premise of semantic segmentation neural networks.

2.1 Multi-view image modeling theory

2.1.1 Feature point extraction and matching

The SfM algorithm recreates a 3D-dimensional scene by reestablishing the posture of every camera one by one. Among them, the camera present parameter is unraveled based on the feature matching pair between images. Therefore, it is first important to compute the feature points of the info image and the feature matching sets between the two images. Normally utilized image feature matching algorithms are Moravec administrator (Moravec, 1980), Harris administrator (Harris and Stephens, 1988), SIFT administrator (Ng and Henikoff, 2003) , HOG administrator (Dalal and Triggs, 2005).Since the feature extraction object of this examination is the image taken in the indoor condition, the indoor space will in general have some high impediment, lopsided brightening, and many entangled items. Therefore, this investigation removes the bearing and scale by SIFT administrator., brightening, pivot, and feature administrators that protect the invariance of relative changes and point of view changes to distinguish nearby features in the image. The SIFT feature point identification image is self-evident, and can keep up stable corners, edge points, and so forth for nearby alters, for example, course, scale, light, In addition, relative change. The SIFT feature points are created by two key advances: 1) discovering key points in the image; 2) producing a 128 - dimensional portrayal administrator on the key points.

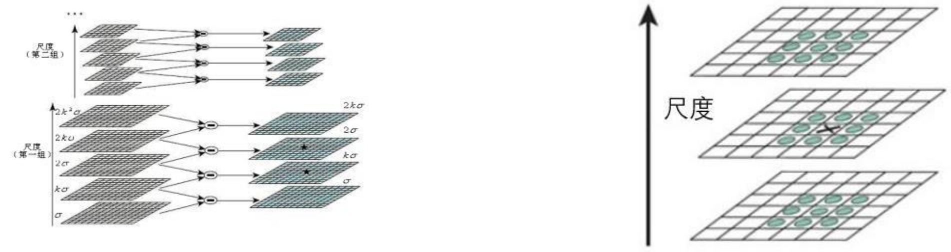
The key points of the SIFT feature administrator are produced in the examination of various scale spaces. Therefore, the SIFT feature points can keep up dependability in various scale spaces. The multi-scale space of a similar image is determined by Gaussian fluffy algorithm. Gaussian fluffy algorithm is an ordinarily utilized image-separating algorithm. It utilizes a specific size of window layout (fluffy format) to convolute with the original image to scale the original image. The Gaussian fuzzy template is calculated by the following 2D-dimensional Gaussian kernel function:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

In the above formula, the image coordinates of the pixels in the original image are shown, indicating the position of each pixel in the original image.

In order to efficiently generate sufficiently stable image features, the adjacent images of the same size and different scales in the scale space are subtracted (Equation 2.3) to obtain Gaussian difference images (Fig. 2.1(a)), and then each is compared in Gaussian difference space. The size of a gradient of 18 local adjacent points corresponding to 8 adjacent points in the 3*3 neighborhood of the same scale and the adjacent upper and lower scales (see Figure 2.1(b)) and the local extremism is selected.



(a) Gaussian differential scale space (b) Detection of extreme points in Gaussian

Figure2.1: Extreme Point Detection

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned}$$

The curve is fitted to the Gaussian difference function of the scale space (Equation 2.4) and iteratively solved to obtain the exact position of the true feature point.

$$D(X) = D + \frac{\partial D^T}{\partial X} X + \frac{1}{2} X^T \frac{\partial^2 D}{\partial X^2}$$

Obtaining the feature point direction mainly through the accompanying two stages: 1) calculating the gradient distribution value of the local pixel; 2) developing the gradient histogram.

$$\begin{aligned} m(x, y) &= \sqrt{\left(L(x+1, y) - L(x-1, y)\right)^2 + \left(L(x, y+1) - L(x, y-1)\right)^2} \\ \theta(x, y) &= \tan^{-1}\left(\left(L(x, y+1) - L(x, y-1)\right) / \left(L(x+1, y) - L(x-1, y)\right)\right) \end{aligned}$$

So as to keep up the invariance of SIFT feature points in image pivot, solid and powerless illumination, and so forth., the accompanying changes are made to the frontal gradient direction of the feature points: First, the 7*7 neighborhood of the feature points is turned to the feature focusing on the feature points. Point the fundamental direction (Fig. 2.2), and afterward utilize the 4*4 seed points in the neighborhood of the feature point to describe the feature points, and create a 128 - dimensional feature vector to describe the feature points (Fig. 2.3). At long last, the 128 - dimensional feature vector is standardized to decrease the impact of image illumination changes on the soundness of feature points.

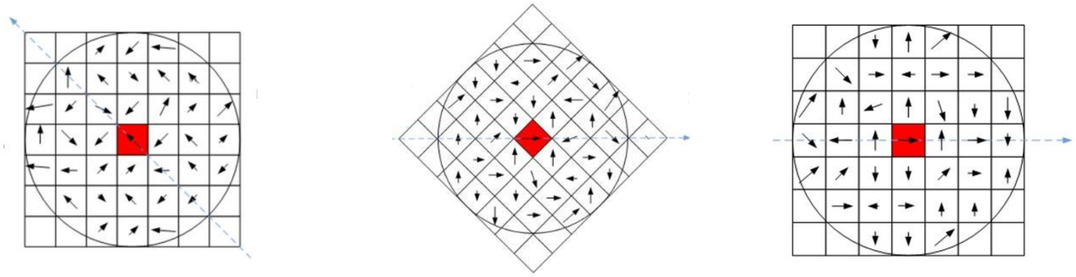


Figure2. 2: Adjustment of the gradient direction

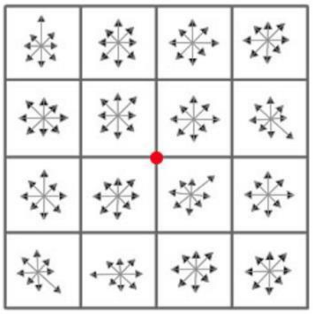


Figure2. 3: 128 -dimensional SIFT feature vector

After obtaining the SIFT feature points, it is generally considered that the closer the descriptors are in the vector space, the higher the comparability of the feature points. Therefore, the closest neighbor-matching algorithm performs the matching of feature points between images.

2.1.2 Dual view geometry

As per the comparing highlight points in numerous perspectives on a similar scene, the relative positional connection between the cameras and the relative places of the camera and the 3D-dimensional space points, some geometric limitations of the element coordinating sets between the pictures can be acquired. Taking one of the cameras in the multiview as the arrange cause, the primary camera parameter matrix is recorded as the principal camera, and the second camera can be viewed as the camera position dependent on the main camera for the rotational interpretation change, communicated as, among them, The inside reference matrix of the reference camera and the second camera, individually, speak to the pivot matrix and interpretation vector of the second camera relative to the reference camera. At that point a similar 3D-dimensional point in the space is anticipated on the principal camera and the second camera, and the projection points x and x' delivered by the main camera and the second camera fulfill the polar geometric requirement (Fig. 2.4). The projection point of the 3D-dimensional space point x in the left view is x. At that point, the projection point of the correct view must be on the crossing point line l' between the left camera C and the correct camera C' and the plane shaped by the point and the correct photograph. The plane shaped by the three points is known as the atomic plane, and the straight line l' speaks to the atomic line.

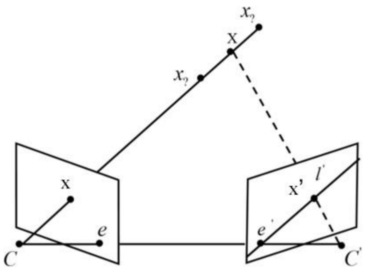


Figure2. 4: polar geometry

Equation 2.6), and the relative posture relationship between the two cameras (Equation 2.7), the correspondence between the basic matrix F and the matching pairs in the picture can be inferred. (Equation 2.8). At the same time, F can be represented by the parameter matrix of the two cameras (Equation 2.9), so that the corresponding camera parameter matrix can be calculated as long as the base matrix is solved.

$$X\,(\lambda) = P^{+}x + \lambda C$$

Among them, is the inverse matrix, C represents the position of the camera center, and represents the distance parameter between the 3D-dimensional point and the Camera. When =0, X represents the position of the 3D-dimensional space point; when, X represents the camera center, X=C.

$$x^{'}Fx = 0 \qquad (2.8)$$

$$F = [K_2 t]\times K_2 R K_1^{-1} \qquad (2.9)$$

In general, the use of eight algorithm fundamental matrix F. Remember, in the form of a matrix, the formula 2.10 is expressed as:

$$\begin{bmatrix} u & v & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = 0 \tag{2.10}$$

The accompanying homogeneous (Equation 2.11) is set up by a set of n (n=8) pairs of matching points. The coefficient matrix An is tackled according to the corresponding point set, and the base solution of An is obtained as the solution of the basic matrix f by singular worth deterioration So as to improve the vigor and precision of estimation calculations regularly singular worth decay for discharging standardized, so the particular activity, in the standardization procedure of the procedure, and it is made out of picture interpretation Normalize the change matrix. The at last obtained matrix F is the basic matrix corresponding to the pair of highlight matching points.

$$Af = \begin{bmatrix} u_1'u_1 & u_1'v_1 & u_1' & v_1'u_1 & v_1'v_1 & v_1' & u_1 & v_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_n'u_n & u_n'v_n & u_n' & v_n'u_n & v_n'v_n & v_n' & u_n & v_n & 1 \end{bmatrix} f = 0 \tag{2.11}$$

2.1.3 Beam method adjustment

Assume that the nearness of 3D-dimensional space, and X-J point the camera O I noticeable, the camera O I projection matrix is P. In a perfect case, the pixel point of the P j point anticipated on the corresponding 2D-dimensional image by the camera O I is x ij (UV). In any case, because of image contortion, camera parameters and such, the light of the 2D-dimensional pixel point x ij and its corresponding matching point x ij ' on the matching image anticipated through the camera place in the 3D-dimensional space can't be met in the space X j point. Hence, the pillar projection can be utilized to improve the camera projection matrix P to alter the light radiated by the camera, with the goal that a point X j in the 3D-dimensional space can be back projected to the image's 2D-dimensional pixel organize point x ij ' by the advanced camera parameters. Above, preferably, |x ij '- x ij |=0. Be that as it may, there is constantly a blunder in the genuine activity, and accordingly, the estimation of the 3D-dimensional point is made as precise as conceivable by limiting the separation between the re-projection point x ij ' of X j and the genuine estimation of the image point x ij .

Shaft change gives a genuine greatest probability gauge for the camera parameter matrix and 3D-dimensional arrange solution in the projection space by building the accompanying cost work (Equation 2.12).

$$\arg\min \sum \left\| x_{ij} - f_j \mathrm{P}(O_j (X_i - c_j)) \right\|^2 \tag{2.12}$$

Where P is the projection matrix. Speaks to a 3D-dimensional point in space, and N speaks to the number of 3D-dimensional points. Demonstrates the course, position and central length of the jth camera, individually.

This cost capacity can be seen as a nonlinear least squares issue and tackled utilizing the Levenberg-Marquardt calculation (Nocedal and Wright, 2006). During the time spent explaining the nearby least, it is essential to give an increasingly precise camera initialization parameter. Globalized camera parameter estimation often results in lower precision local minimum values. Therefore, this study uses an incremental method to add a camera to the reconstruction system one by one to obtain more accurate camera parameter initial values.

First, the picture with the most matching feature points among many pictures is used as the initial value of the initialization camera parameters. The initialization internal parameters of the camera can be obtained from the stored image format information document (EXIF tags) (Snavely et al., 2006). The camera is initialized as the coordinate origin, so its foreign parameter is set to, then the outer parameter of another picture with the largest number of matching pairs. Then, each time a picture is added to the reconstruction system and the parameters of the camera are optimized by parameter iteration. The principle of incremental optimization is to select a picture with the largest number of feature matching point pairs with the initial matching picture as the picture input successively. Finally, the camera parameters are optimized by the matching point pairs of the newly added pictures in the system. The program runs iteratively until all the images are used for 3D reconstruction.

2.2 Convolutional neural network related basic knowledge

.

2.2.1 Convolutional layer

It is the first and essential part of hidden layers which is used to extract features of input images. It preserves relationship between pixels using Small Square of input image. It convolves the filters of window size ($fh \times fw \times nc$) and input matrix of image ($h \times w \times nc$). The mathematical form as following:

$$(h \times w \times nc) * (fh \times fw \times nc) = (h - fh + 1)\ (w - fw \quad + 1) + 1 \qquad (2.1)$$

$if\ filter\ nc = 1$

If have two filters for horizontal and vertical edges then output will be given as:

$(h - fh + 1)(w - fw + 1) + 2.$

Convolutional neural networks extract imIn recent years; the efficient recognition algorithm of convolutional neural networks has developed rapidly in the field of image processing. Compared with the traditional image classification algorithm, the convolutional neural network can use the image as an input to realize end-to-end object recognition, thus avoiding the complex and time-consuming feature extraction and data recovery operations in the previous semantic object extraction algorithm.

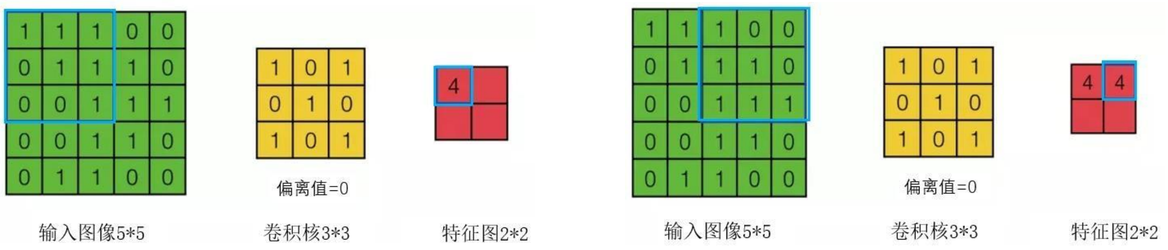
The specific process of convolution calculation is as follows:

Assuming that a 5*5 input image is convoluted with a 3*3 convolution template, a 3*3 feature map can be obtained. The specific steps of the convolution calculation are as follows: first, the image is numbered pixel by pixel; then the convolution kernel is numbered by weight (each value in the convolution kernel is called a weight), expressed as; the feature map calculated by convolution each element is numbered. Convolution calculation using the following formula,

$$a_{i,j} = f(\sum_{m=0}^{2}\sum_{n=0}^{2} w_{m,n} x_{i+m,j+n} + w_b) \qquad （2.13）$$

Wherein, the pixel representing, the i-th row and the j-th column in the input image; the convolution weight, of the mth row and the nth column in the convolution kernel; the, eigenvalue of the i-th row and the j-th column in the feature map; Adjustments are made, and offsets are added during the convolution process ; f is the activation function, which is designed to enhance the nonlinear expression of the network.

In the process of convolution calculation, the size of the output feature map can be changed by adjusting the sliding core in the input layer window stride. When the convolution step is 1, a feature map with a size of 3*3 is generated. When the convolution step is 2, the convolution calculation is as follows, and the convolution calculation result is a feature map of 2*2.



(a) The first convolution window is calculated (b) Convolution window sliding

Figure2. 5: When the convolution step is 2, the convolution calculation process is indicated

The size of the feature map calculated by image size, step size, and convolution satisfies the following relationship

$$W_2 = (W_1 - F + 2P) / S + 1$$
$$H_2 = (H_1 - F + 2P) / S + 1 \qquad (2.14)$$

Wherein represents a wide output characteristic diagram; wide input layer image; convolution kernel width; convolution sliding window successive steps; is the number of zero-padded input image edge layer convolution operation, if is 1, fills a circle at the edge of the input layer image; the height of the feature map after the convolution calculation; indicates the height of the input layer image before the convolution calculation.

The above convolution calculation is for a convolution layer with a depth of 1. If the depth of the convolution layer is D, the convolution calculation is performed as follows.

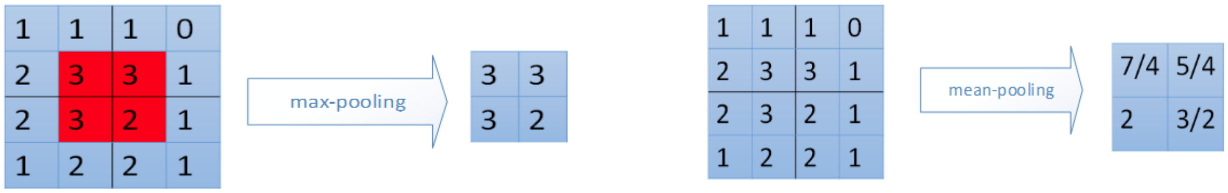
$$a_{i,j} = f\left(\sum_{d=0}^{D-1}\sum_{m=0}^{F-1}\sum_{n=0}^{F-1} w_{d,m,n}\, x_{d,i+m,j+n} + w_b\right) \qquad (2.15)$$

2.2.2 Pooling layer

It can reduce only the number of parameters during forward propagation when the input images are too large. Pooling layer is also called sub sampling or down sampling due to reduction of dimensions of each map and retains the important information. There are three popular pooling techniques; max pooling, average pooling and sum pooling and the most useful is max pooling. It does not use padding and number of input and output channels remain constant. Max pooling does not learn parameters because of fixed function which does not allow neural network to learn during back propagation of neural network from pooling. Max pooling takes largest element from window rectified feature map. While average and sum pooling take average and sum,

The pooling layer compresses the feature map obtained by the convolution layer to reduce the computational complexity of the network and reduce the impact of noise on the main features (Scherer et al., 2010); (Shen et al., 2014); (He et al., 2015).

In the forward propagation process of the feature, the downsampling operation is performed on each non-overlapping local region of size n*n in the feature map generated by the previous convolutional layer to achieve the purpose of dimension reduction. As shown in Figure 2.6(a), Max Pooling uses the largest value in each 2*2 local area as the output feature; Mean Pooling selects the average as the output feature of this local area (Figure 2.6(b)).



1	1	1	0
2	3	3	1
2	3	2	1
1	2	2	1

3	3
3	2

1	1	1	0
2	3	3	1
2	3	2	1
1	2	2	1

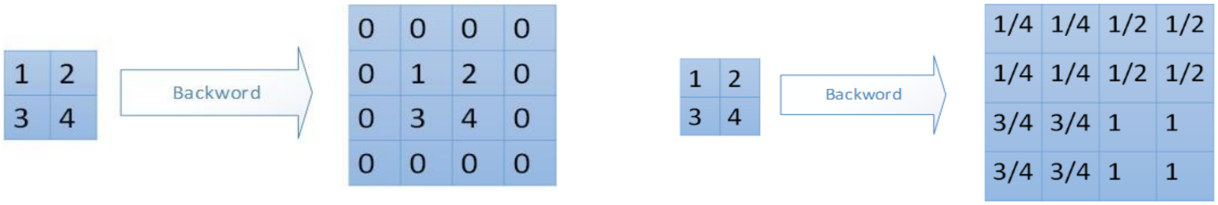
7/4	5/4
2	3/2

(a) Max Pooling

(b) Mean Pooling

Figure2. 6: Two ways of feature forward propagation

The backward propagation mechanism of the characteristics of the pooling layer is slightly different from the forward propagation (Fig. 2.7(a)). When the Mean Pooling method performs backward propagation, the residuals of the pooled layer are equally divided into n parts (n indicates that one residual corresponds to several pixels in the next convolutional layer), and the average value is correspondingly transmitted to the corresponding n parts. In the pixel (Figure 2.7(b)).



(a) Max Pooling

(b) Mean Pooling

Figure2. 7: Two ways of feature backward propagation

2.2.3 Full Convolutional Neural Network

In these layers, the feature map matrix is converted with a fully staked vector. Because this vector of features will be useful in creation of model. In last it will feed to activation functions such as softmax (for multi classification) and sigmoid (binary classification) to classify output.

The Full Convolutional Neural Network (FCN) is an upgraded semantic segmentation network for convolutional neural networks (CNN) (Wu, 2015). The full convolutional neural network is better than the convolutional neural network. Its focal points are principally reflected in the accompanying two angles: 1) The FCN neural network can obtain start to finish pixel-level image segmentation results; 2) The FCN input can be any size image. In view of the feature map extricated by each convolutional layer, the FCN utilizes the deconvolution layer to upsample the semantic segmentation consequence of the first image size, in this way performing pixel-by-pixel semantic forecast, lastly performing pixel-level semantics on the feature guide of the first image size. Split (Figure 2.9).

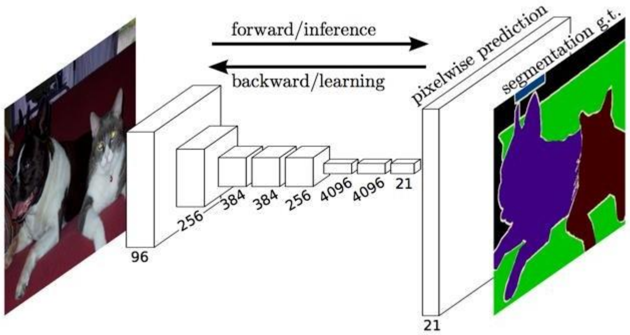
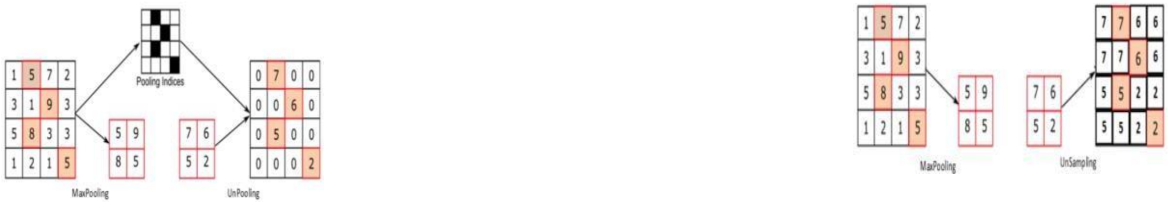


Figure2. 8: Principle of image-by-pixel semantic segmentation of full convolutional neural networks

As Up sampling can be accomplished in a few different ways. (1) Unpooling; (2) Un Sampling; (3) Transpose convolution. Corresponding to the local diminishing operation, the Pooling, Unpooling operation doles out the removed local feature esteems to the higher resolution feature map to accomplish the reason for up sampling. As appeared in Figure 2.10 (a), the feature map is expanded by utilizing the local greatest data obtained by Maxpooling, wherein the extreme worth holds its unique position data, and every single other position are loaded up with zeros.



(a) Unpooling (Maxpooling)

(b) Unsampling

Figure2. 9: Two ways to expand the feature map,

Not at all like Unpooling, has the Unsampling operation legitimately duplicated the extricated local features to a higher resolution feature map for feature map development. As appeared in Figure 2.10 (b),

In the up sampling strategy, the most generally utilized technique is the deconvolution technique. Not at all like Unpooling and up sampling, is the deconvolution like the convolution operation. In the picture. As appeared in the figure underneath, an input image of size 2*2 uses a 4*4 convolution portion to play out a deconvolution count with a stage size of 3, and yields a 7*7 feature map.

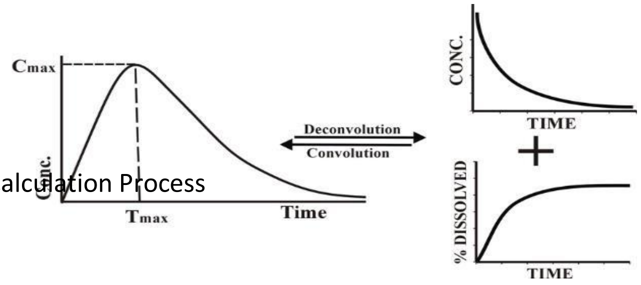


Figure2. 10: Deconvolution Calculation Process

2.3 Summary of this chapter

While recently deep learning achieved very high accuracy, which is an imaginary level for other tools and techniques. Wehave chosen deep learning among all research areas due to future direction and everything become internet base. As we explained that convolutional neural network is a superior part of supervised deep learning. It includes on different hidden layers, which connects with each other. These layers extract the features of input objects and pass through different stages of layers, which have been briefly explained in this chapter.

In order to realize the semantic segmentation of indoor 3D point cloud model with 2D- dimensional joints, this chapter introduces the algorithm foundation of indoor 3D model semantic markup through 2D-dimensional modeling from two aspects: multi-view 3D modeling algorithm and deep learning basic algorithm. It lays a theoretical foundation for the subsequent introduction of 2D-dimensional indoor scene understanding and 2D-dimensional3D-dimensional semantic transfer of indoor 3D-dimensional point cloud method.

3. Parsing Indoor Scene Based on Deeplabv3+ Neural Network

Dissimilar to Unpooling, the Unsampling operation legitimately duplicates the separated local features to a higher resolution feature map for feature map development. As appeared in Figure 2.10 (b), after the local greatest is obtained in the Maxpooling stage, Unsampling straightforwardly duplicates the most extreme value data into the feature template of the first image resolution size to obtain an up sampled feature map.

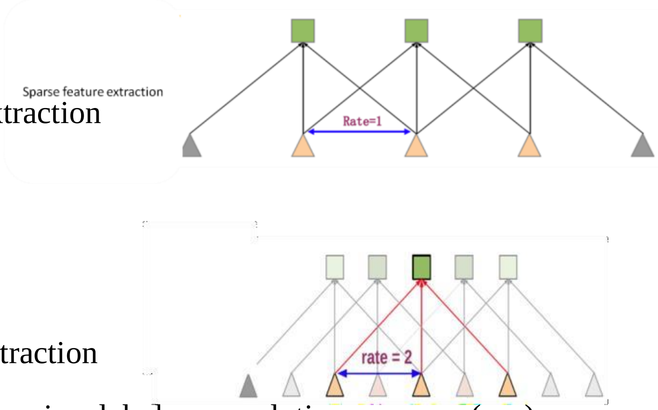
The up sampled result is obtained by setting the convolution portion parameters and learning. The convolution operation is to yield a solitary operation value at the corresponding situation of the feature map by tiling all the pixel values in the image with the template window. Conversely, deconvolution, by increasing a value in the downsampled low-resolution feature map by the weight value set in the convolution bit, obtains the weighted eigenvalue yield to a higher resolution feature. In the picture. As appeared in the figure underneath, an input image of size 2*2 uses a 4*4 convolution piece to play out a deconvolution computation with a stage size of 3, and yields a 7*7 feature map. With the rapid development of deep learning algorithms in the field of 2D-dimensional image semantic segmentation, pixel-level precise image semantic segmentation results assume an important job in expanded reality, self-administration robot route, and a progression of savvy wearable gadgets. The indoor help and application necessities advance the development of indoor object semantic segmentation. Simultaneously, an enormous number of open and accessible indoor standard training informational indexes are likewise framed. The present standard deep learning image semantic segmentation models are CNN (Zeiler and Fergus, 2014), FCN (Wu, 2015), DeepLab (Chen et al., 2016), SegNet (Badrinarayanan et al., 2017), DenseNet (Li and Vu, 2018), and so on., while DeepLab v3+ drives the other image semantic segmentation algorithms by bringing an opening convolution model into the encoder - decoder structure to meld multi-scale data. DenseNet diminishes the number of parameters of the network through multi-post feature associations, and upgrades the reuse of network features, in this way improving network execution. Be that as it may, compared with the open air condition, the indoor space is complex and different, and there are numerous objects. The utilization of object semantics to depict the indoor space has a progression of constraints, and the indoor space frequently presents a flat and vertical square structure. In this way, the indoor space layout the structure and indoor object semantic extraction results complement each other, which is helpful for the comprehension of indoor scenes. Be that as it may, until this point in time, not very many deep neural networks can join indoor spatial layout estimation and indoor object semantic extraction to comprehend indoor scenes. In this way, this paper proposes a new kind of neural network Deeplabv3+, which can at the same time incorporate indoor object semantic segmentation and spatial layout estimation into the network for training, lastly obtain a progressively exact indoor scene semantic getting impact.

3.1 Deeplabv3+ network design basis

3.1.1 Cavity convolution structure for dense feature extraction

The customary deep convolutional neural network incredibly advances the development of 2D-dimensional image semantic segmentation algorithm with its ground-breaking feature learning capacity and pixel-level semantic recognition capacity. Be that as it may, the rehashed superposition of different convolutional layers and pooled layers in the convolutional neural network altogether decreases the spatial resolution of the yield feature map. Presently, the feature map can be extended by the up sampling technique to reestablish the feature map of the first image size. The semantic segmentation aftereffect of the pixel, however this operation causes the loss of part of the data and the loss of time and memory.

To this end, numerous researchers have made improvements in convolutional neural networks. Chen et al. (2016) in view of the examination consequences of convolutional neural networks for a long time, proposed the deeplab neural network for the resolution of feature maps. Among them, deeplab utilizes an empty convolution structure to adequately separate multi-scale network thick features. The opening convolution structure extricates bigger scope features by amplifying the responsive area of the convolution part, and can alter the size of the depression convolution bit to adjust to various features. Diagram sampling rate to separate features of multi- scale objects (Figure 3.1). Such a network structure can adequately keep away from the securing of repetitive data by the network for various scale feature extraction, and simultaneously give more consideration to the relationship of features between objects.



(a) Sparse feature extraction

(b) Dense feature extraction

Figure3. 1:The one-dimensional hole convolution process (a) represents the sparse feature extracted by the low-resolution feature map after standard convolution; (b) the cavity convolution of the sampling rate r=2 is applied to the low-resolution feature map.

In the one-dimensional convolution structure, the whole convolution calculation process is as follows:

$$y[i] = \sum_{i=1}^{K} x[i + r * k]w[k] \qquad (3.1) \qquad \text{（3.1）}$$

The one-dimensional input signal indicates the yield result after the cavity convolution computation, and indicates the whole convolution bit, indicating the size of the convolution part, and indicating the sampling pace of the input signal. In standard convolution figuring's, convolution sampling is performed. The convolution consequence of one-dimensional whole convolution and standard convolution is appeared in Figure 3.1. It very well may be seen from the figure that the feature map under the impact of cavity convolution can separate increasingly thick image features.

In the 2D-dimensional convolution structure, the operation of the whole convolution is as appeared in the figure beneath, expecting a downsampling operation with a factor of 2 on the input image, and afterward playing out a standard convolution operation with a convolution piece size of 7, lastly just 1 is obtained. /4 The feature map of the first image size. On the off chance that this feature map is sampled to the first image size, a ton of detail information will be lost. Be that as it may, if the image feature is extricated utilizing the whole convolution with a sampling pace of 2, the definite information of the image can be better safeguarded while obtaining the first image size feature map. Also, the utilization of whole convolution structure decreases the multi-layer convolution operation, and doesn't add extra training parameters to the network while removing image escalated features, which quickens the training productivity of the network.

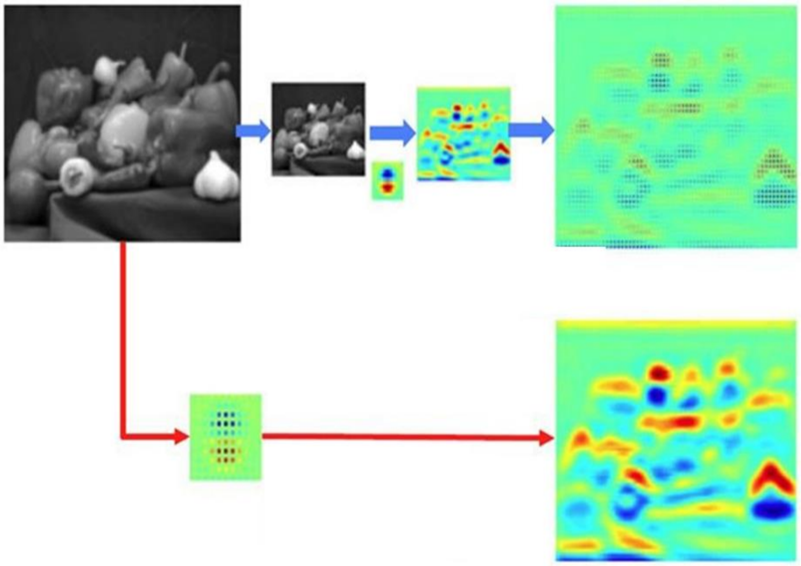


Figure3. 2: The cavity convolution acts on the 2D-dimensional image. The first row represents the sparse feature output of the low-resolution feature map under the standard convolution; the lower row represents the dense feature extracted after the Cavity convolution of the sampling rate r=2.

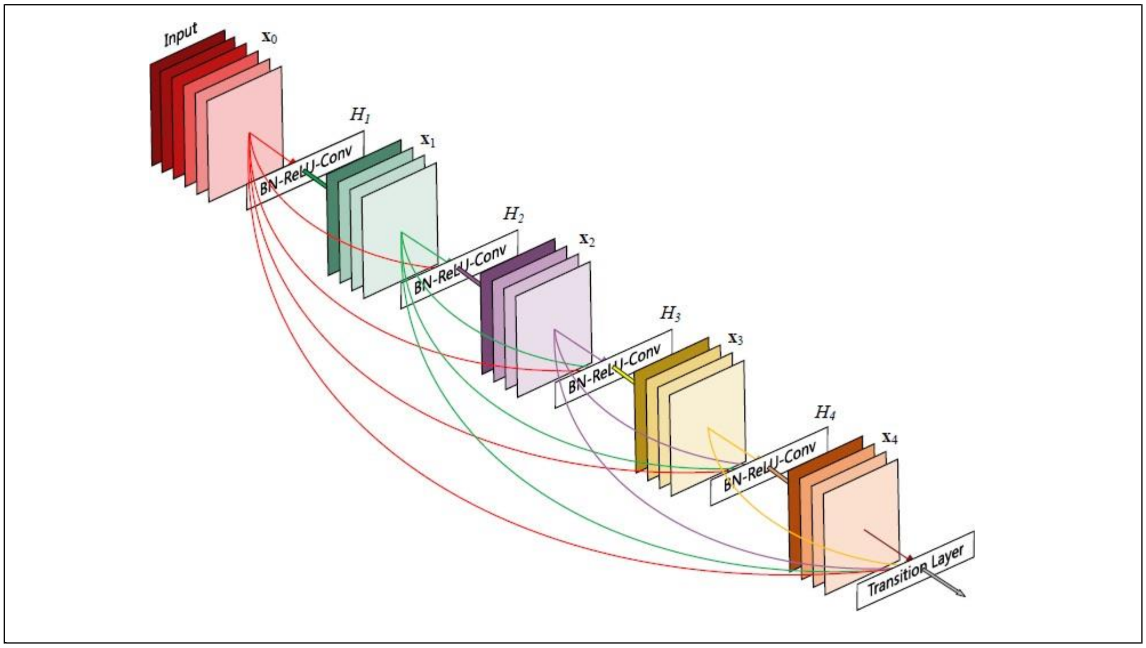
3.1.2 Tight compact structure for hierarchical feature joints

The traditional convolutional neural network uses the feature output of the previous layer as the input of the next layer to accomplish the purpose of feature forward propagation. On this basis, ResNet (Targ et al., 2016) uses a jump-connect structure with an identification function to perform backhaul of the gradient (Equation 3.2). However, the feature values output in this way are obtained by summing the features of the feature connection layer, which may hinder the transmission of information. In order to further improve the utilization of features between layers, Li et al. (Li and Vu, 2018) designed a denseblock structure to transfer features between layers. Figure 3.3 show the neural network connected using the denseblock structure. Suppose the network before l-1 characteristics are transmitted to the first layer l layer, the first l wherein layer represented by the formula 3.3:

$$x_l = H_l(x_{l-1}) + x_{l-1} \qquad (3.2)$$

$$x_l = H_l([x_0, x_1, \cdots, x_{l-1}]) \qquad (3.3)$$

Where in, it represents the connection of the feature map extracted from the 0th layer to the 1st -1th layer in the network. It represents l nonlinear transformation function
Layer, represents l wherein FIG layer obtained.



DenseNet neural network connected by multiple tightly dense structures

Figure3. 3: While changing the feature map size, the between layer association operation utilized in Equation 3.3 will be constrained. The downsampling structure in the convolutional neural network can change the size of the output feature map. In this way, it is important to expand the convolutional layer and the pooling layer between the firmly stuffed firmly pressed structures to change the size of the feature map. So as to encourage the downsampling operation in the DenseNet network model, a plurality of firmly associated compact dense structures can be utilized for the transmission of dense features. As appeared in FIG. 3.4, a part associating nearby compact dense structures is called a change layer. The progress layer generally incorporates convolution and pooling operations.

3.1.3 Network output based on cascaded multitasking structure

Traditional multitasking methods in deep convolutional neural networks are usually founded on shared features of numerous tasks, each task working freely. Nonetheless, the features separated by each task right now are reused for different subtasks, bringing about a complicated network structure, enormous computational load and memory misfortune. To this end, Dai et al. (2016) proposed a cascaded multitasking structure network, in which the subtasks are associated, and the features of the past stage subtask extraction can be applied to the last stage subtasks, in this manner simplifying the traditional The task network upgrades feature reuse, diminishes network load and improves network computing proficiency. The comparison between the traditional multitasking network and the cascaded multitasking structure network is appeared in Figure 3.5.

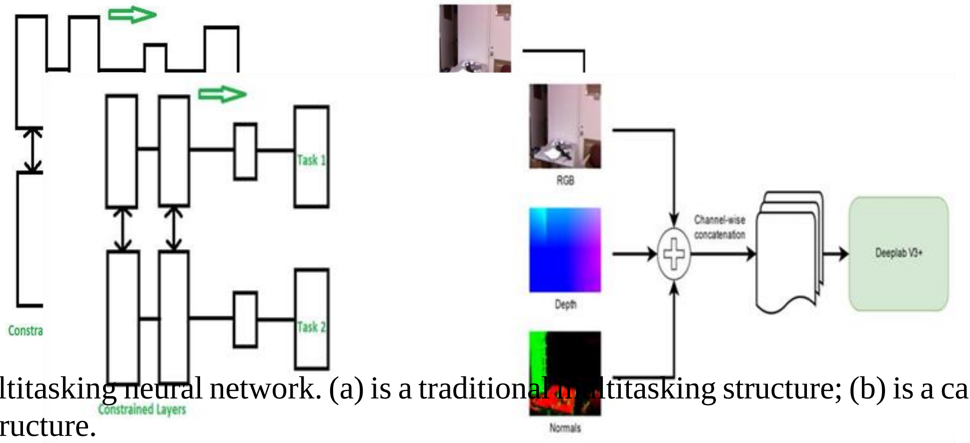


Figure 3.4: Multitasking neural network. (a) is a traditional multitasking structure; (b) is a cascaded multitasking structure.

3.2 Deeplabv3+ Network Architecture

Based on the great achievements of DeepLabv3+, DenseNet and cascading multi-task structure in the field of image semantic segmentation, this paper proposes Deeplabv3+ neural network. The Deeplabv3+ neural network mainly consists of three parts: the backbone network part, the semantic segmentation sub-network and the spatial layout segmentation sub-network, as shown in Figure 3.6.

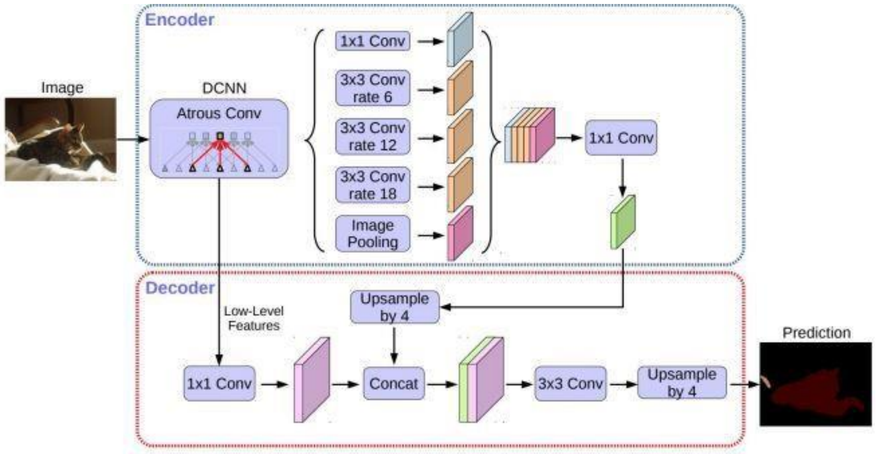


Figure3. 5: Organizational structure of Deeplabv3+ neural network

In order to solve the problem of high occlusion, uneven illumination and complex objects in indoor space, this paper proposes to use indoor space layout to assist indoor object segmentation to understand indoor scenes. Based on the structural advantages of cascaded multitasking networks, this paper proposes a Deeplabv3+ uses a cascaded multitasking network structure to segment indoor objects and spatial layout. It is estimated that the two subtasks of the network are output separately. Since the lower part of the network contains a large number of spatial structural features, and the high-level structure of the network contains more object semantic features, this paper designs the spatial layout estimation results of the low-level part of the Deeplabv3+ neural network and transmits the spatial structural features. Going to the upper level of the network not only simplifies the network results, but also achieves the purpose of feature reuse and feature fusion. Gradient back propagation and prevent the gradient from disappearing. Beginning from the second layer of the backbone network, a DenseBlock structure (Fig. 3.7(b)) is presented. The DenseBlock structure is a dense convolutional network structure planned by Liu in DenseNet (Li and Vu, 2018). The skip association structure empowers each layer of the network to use the features separated by the past layer, along these lines improving the information scattering and reusability of features between the layers of the network, while significantly lessening the parameters. The number and the issue of the gradient disappearing. Were utilized backbone network.

5-.the DenseBlock structure coupled to each other to accomplish the full-preferred position of the qualities of each layer object.

The semantic segmentation subnetwork is a pixel-level semantic segmentation network for object recognition and object localization. So as to utilize the semantic information of the elevated level structure of the network, the semantic extraction sub-network is associated with the most elevated layer of the backbone network through the enormous scale residual structure, and uses the hole convolution structure (ASPP) (Chen et al., 2016) to extricate the layers. Multi-scale features are consolidated to obtain more elevated level semantic information (Figure 3.2(c)). The structure separates convolution features at various scale levels by altering the open field of the convolution channel. Also, the semantic segmentation sub-network utilizes the encoder - decoder design to recuperate the extricated low-resolution feature maps to obtain high- resolution semantic segmentation results. DeepLab v3 (Chen et al., 2017) neural network legitimately performs feature recuperation through direct encoder. Compared with DeepLab v3, Deeplabv3+ neural network incorporates the layers separated from the front finish of the network into the semantic segmentation subnet through the fully associated network. The pooling layer utilizes the spatial feature information of the lower layer of the network. In this way, in the process of the backbone network giving semantic feature information to the object semantic segmentation task, the additional spatial structure feature also gives increasingly exact feature area expectation information in the decoder organize.

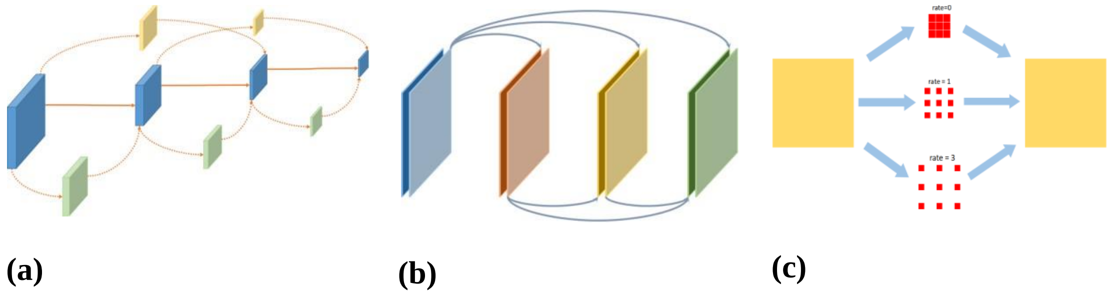


Figure3. 6: Main functional structures in the Deeplabv3+ neural network. (a) represents a large-scale residual structure, in which a blue square block represents each layer in the backbone

network, and other smaller square layers connected to the backbone layer represent convolutional layers; (b) indicates compactness intensive Structure (DenseBlock), the cube in the figure represents the convolution layer in the network; (c) represents the hole convolution structure, the red small squares represent convolution templates of different scales, and the red squares of the second and third layers represent Convolution template with holes.

3.3 Optimization of spatial layout estimation results

Assuming that the input size is w*h*3, the rough estimate of the spatial layout obtained by Deeplabv3+ in 3.2 is expressed as a probability function.

$$P^{(k)} = \Pr(L_{ij} = k \mid I), \forall k \in \{0,1\}, i \in [1,\ldots,h], j \in [1,\ldots w] \tag{3.4}$$

Where L is the position of each pixel in the input image of size w*h, and L ij is the spatial layout of the output to estimate the label value of the pixel in the image, and its value ranges from 0 to 1, L ij = 0 indicates that the pixel is the background, whereas L ij =1 indicates that the pixel is an estimate of the spatial layout structure.

However, the spatial layout obtained by Deeplabv3+ has a rough estimate of the structure lines that are noisier and less accurate (Ren et al. 2016). There are two reasons for this: First, because the multi-layer convolution and pooling operations in the neural network lose some of the spatial structure information, the spatial layout structure of the Deeplabv3+ output is thin and not straight. In addition, the highly occluded environment in the room poses a significant challenge to the accurate estimation of the spatial layout structure. Therefore, this paper uses the Manhattan world hypothesis based on indoor geometric information (Coughlan and Yuille, 1999) to optimize the rough estimate of the spatial layout obtained by Deeplabv3+ segmentation. The optimization process is mainly carried out in two steps: 1) construction of a candidate set of spatial layout structure; 2) selection of an optimal spatial layout structure.

3.3.1 Construction of candidate sets for spatial layout structure

In order to estimate the spatial layout of the indoor height occlusion environment, this paper divides the spatial layout structure candidate set into two steps. First, Deeplabv3+ rough estimation result obtained is the spatial layout of 4 mask pixels operate to establish an estimated buffer layout, indicated as C. Then, the vanishing line in the original image is extracted (Gupta et al., 2010), and the vanishing line falling in the mask area of the spatial layout rough estimation result is taken as the candidate set of the layout structure, expressed as l i (original) , as shown in Figure 3.8 . , c, d, e divide the indoor space structure into ceilings, walls, floors and solid lines.

Due to the complex indoor environment, there is often a highly occluded area. As shown in Figure 3.9 (b), the structural lines of the floor and the wall may be partially or completely obscured. If the structural line is partially occluded, the structural line of the occluded portion can be obtained by extending the remaining structural lines. If the structural line is completely occluded, it is difficult to accurately restore the position of the structural line, and at this time, the approximate position of the occluding portion structural line can be inferred by geometric rules. The occluded spatial layout structure line estimation result is denoted as l $_{i\ (occluded)}$.

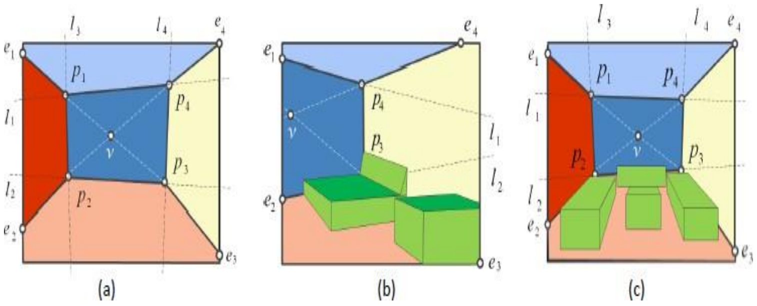


Figure3. 7: Example interior layout structure models, Layout = (L . 1 , L 2 , L . 3 , L . 4 , V) is the spatial layout parameters four vanishing line and vanishing point set consisting :(A) represents the indoor space . 5 th The layout is completely Visible; (b) indicates the occlusion in the indoor space.

In summary, the final spatial layout estimation result candidate set contains the estimation results of the above two cases, expressed as:

$$l_{critical} = l_{i\,(original)} \cup l_{i\,(occluded)} \qquad (3.5)$$

Here, the final indoor space structure line estimation candidate set is shown, indicating the vanishing line set of the original image in the rough layout estimation result mask area; and the structural line indicating the indoor occlusion area recovery.

3.3.2 Selection of optimal space layout structure

In order to obtain the optimal spatial layout estimation results, this paper uses the coarse layout estimation probability map Obtained by Deeplabv3+ as the mask weight to score the structural lines in the layout estimation candidate set. The scoring function looks like this:

$$S(L\mid P)=\frac{1}{N}\sum_{ij}P_{i,j},\ \forall L_{i,j}=1 \qquad (3.6)$$

Wherein, P represents a coarse layout estimation probability map of the Deeplabv3+ network output; represents a probability that the pixel of the i-th row and the jth column of the image belongs to the layout structure line pixel; L represents a candidate set of the layout hypothesis estimation; and represents that the pixel in the candidate set is the layout structure line Pixel; N is a normalization factor equal to the total number of layout pixels in L. Finally, the maximum value of the solution, that is, the layout with the highest selection score is assumed to be the optimal layout estimation result.

$$L^{*}=\arg\max_{L}S(L\mid P) \qquad (3.7)$$

3.3 Experiment and analysis

3.3.1 Experimental data

The Deeplabv3+ neural network experiments on indoor object semantic segmentation and indoor space layout estimation were performed using the indoor public data set NYUDv2 RGBD of 47 images taken by camera of Conference room.

In this experiment, 49 images were selected from the NYUDv2 RGBD dataset, including 30 types of indoor scene objects, of which 49 images were used as training data, 100 sheets were used as verification data, and the remaining 47 images were used as test data. Compared with the indoor image scene in the NYUDv2 RGBD dataset, the target scene selected in this paper contains some objects with special shape features. Therefore, this experiment additionally collected a small number of target scene data sets for fine-tuning the network. The data set of the target scene contains 261 conference room images, 39 of which are used as additional labeled training data, and are semantically marked with LabelMe . As shown in Table 3.1 , the labeled data set, into the interior scene 12 is training classes, which 12 is semantic category is passed to the semantic tags as a final 3D-dimensional point cloud. In addition, the semantic segmentation network training data set of this experiment is shown in Figure 3.10.

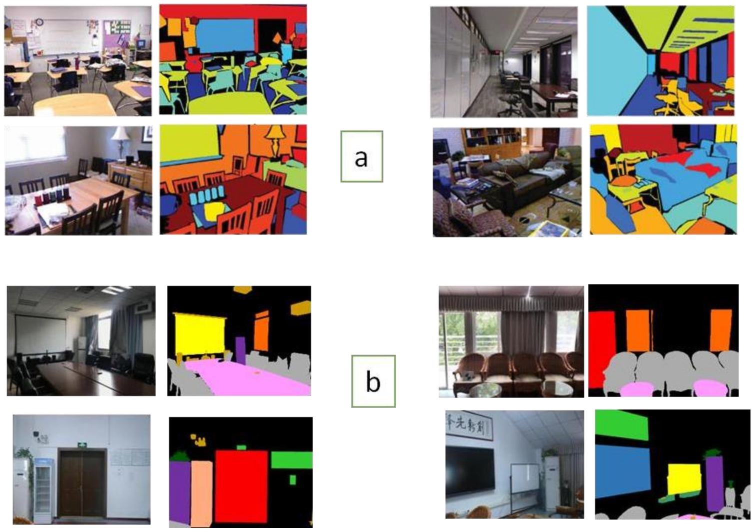


Figure3. 8: Example of a training data set for Deeplabv3+ neural network semantic segmentation. (a) Represents the training data set for semantic segmentation in the NYUDv2 RGBD indoor data set; (b) represents the conference room set for network fine-tuning.

In addition, because the indoor 3D-dimensional spatial layout structure plays a vital role in the semantic expression of indoor scenes and the indoor spatial structure often appears as a horizontal and vertical regular structure. Therefore, Deeplabv3+ indoor spatial layout estimation sub-network uses the training data set published by Hedau in RoomNet (Lee et al., 2017) for network training, which contains 313 indoor spatial layout marker images and uses 3.4.1. The conference room data set mentioned in the section tests the network layout estimation results. The indoor space layout can be expressed as a 3D-dimensional box model. As shown in Figure 3.11, the indoor space layout structure is marked with a straight line at the junction of the ceiling, the wall and the floor. Thus, the entire indoor space is divided into a ceiling, a floor, and a left. The wall, the right wall and the middle wall are divided into five parts. During the marking process, all occlusions in the interior space are ignored for surface-based spatial layout marking.

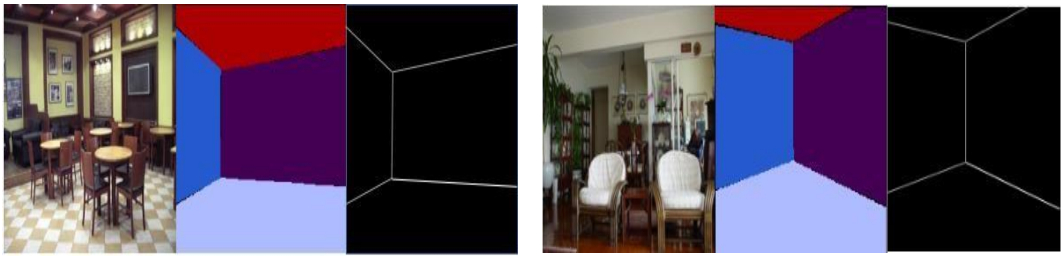


Figure3. 9: Spatial layout estimated network training data

3.3.2 Network Stratification Training Strategy

The Deeplabv3+ neural network proposed in this paper can simultaneously perform indoor spatial layout estimation and indoor object semantic segmentation. This research work proposes the following network layered training strategies:

(1) Since the backbone network of multi-level connection can extract rich identifiable features in the image, the network first performs complex semantic segmentation training, and then performs spatial layout estimation training.

(2) In the first phase of network training, the NYUDv2 RGBD indoor public data set is trained only on the backbone network and the semantic segmentation sub-network, and then the network data is fine-tuned with the additional collected target scene training data.

(3) In the second phase of network training, only the spatial layout estimation sub-network is trained. (4) In the third phase of network training, the data sets of the first two phases will be jointly optimized for all layers in the Deeplabv3+ network.

3.4.3 Experimental results and analysis

Figure 3.12 shows the semantic segmentation results of the Deeplabv3+ neural network after a small number of target scene training data sets participate in network fine- tuning. . Figure 3.13 shows the estimation results of the indoor space layout. The white frame with black as the background is the rough estimate of the indoor space obtained by the Deeplabv3+ neural network. From the layout estimation result graph, Deeplabv3+ can estimate the space of the indoor scene more accurately.

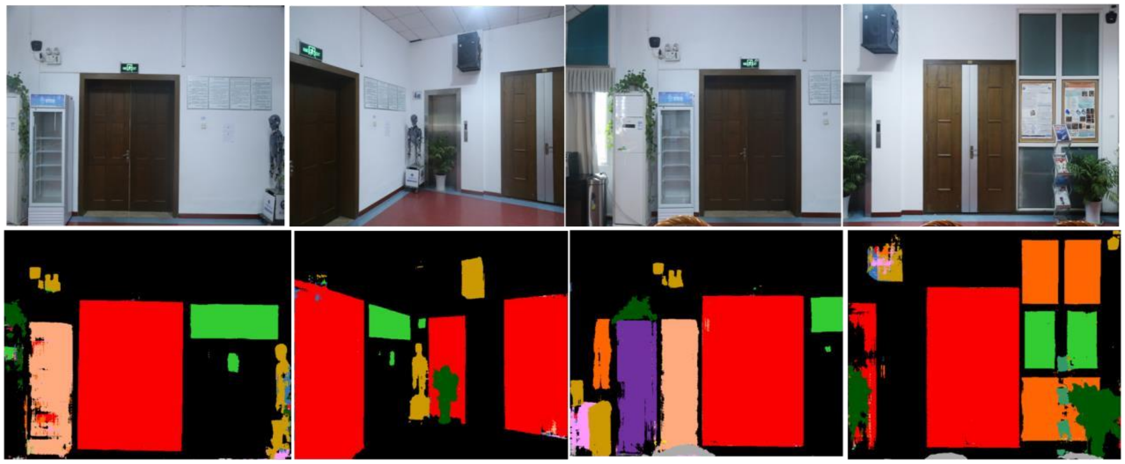


Figure3. 10: the semantic segmentation result of the indoor object obtained by the Deeplabv3+ neural network proposed in this work (a) Test pictures input for the Deeplabv3+ neural network; (b) Semantic segmentation results representing the Deeplabv3+ neural network;

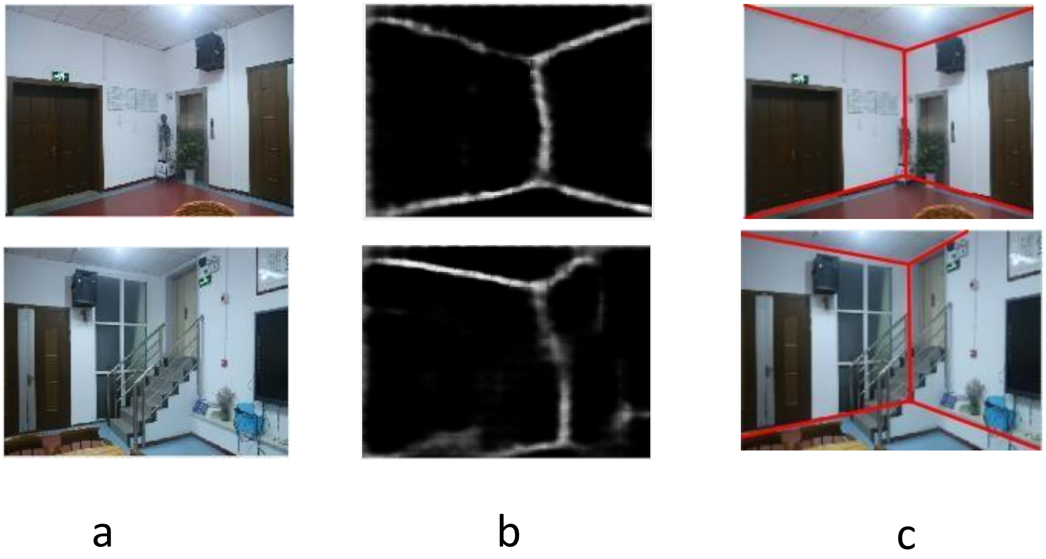


Figure3. 11: Estimated results of indoor space layout. (a) an input picture representing the layout estimation of the Deeplabv3+ neural network; (b) a rough estimation result indicating the indoor spatial layout of the Deeplabv3+ neural network; (c) a final fine spatial spatial structure line estimation result.

Since Deeplabv3+ using the design of the neural network DeepLab v3 + void convolution structure and DenseNet compactness dense structure, thus, herein by contrast DeepLabv3+ , DenseNet121 and proposed Deeplabv3+ in NYUDv2 RGBD semantic segmentation indoor disclosed datasets The results verify the superiority of the Deeplabv3+ neural network. This experiment uses Tensor flow as the back-end framework to perform semantic segmentation experiments of three network models in Keras, and uses the Adam optimizer with 1e-4 as the initial learning rate. These semantic segmentation network models are iterated 10K times on an NVIDIA GTX 1080Ti device to complete network training.

From the segmentation results in the first and second columns of the table, it can be seen that compared to DeepLab v3+ and DenseNet121, the two semantic segmentation models, Deeplabv3+ neural network can better balance the retention of indoor object detail features and object identifiable features. Learn these two tasks. Specifically, even DeepLab v3+ and DenseNet 121 can obtain relatively good processing results for the edge features of indoor objects. However, DenseNet121 tends to mix objects with similar characteristics, as shown in the second column of Table 3.14. Light green indicates the segmentation result of the shelf, red indicates the bookshelf, and the red bookshelf appears as noise in the green shelf area. The segmentation accuracy has a certain degree of influence, and the same misclassification occurs in the segmentation result of the first column; DeepLab v3+ is easy to produce smoothness at the edge of the image, and cannot even divide the edges between similar objects. It can be seen from the segmentation results in the table that Deeplabv3+ segmentation result is better than DenseNet121 and DeepLab v3+. It can extract spatial structure information through large-scale residual structure and tight-intensive structure, so it can effectively preserve the structural information of indoor objects. In addition, Deeplabv3+ multi-scale feature fusion structure based on whole convolution enables the network to effectively learn the identifiable features of indoor objects. Therefore, based on the structural advantages of the DenseNet and DeepLab v3+ networks, the Deeplabv3+ neural network can clearly balance the shortcomings of the two in the indoor semantic segmentation. The third column in the table is the result of abnormal segmentation. According to the analysis, there may be two reasons. One is that the texture features that are similar between the segmented object and the background cause the network to mix the sofa and the floor; the second is the training data center sofa. The shape differs greatly from the shape of the sofa object in the test image, resulting in a misclassification. However, in the case of segmentation anomalies, Deeplabv3+ can still exhibit a large degree of segmentation advantage, which indicates that Deeplabv3+ has certain robustness in indoor object semantic segmentation.

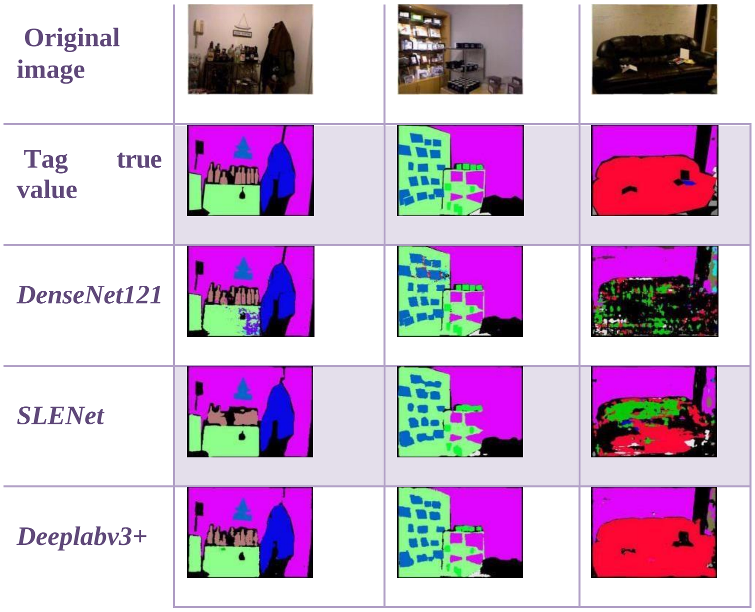


Figure3. 12: DenseNet121 model, Deeplabv3+ model and the semantic segmentation test Results of the Deeplabv3+ model proposed in this paper on NYUDv2 RGBD indoor dataset

4. Joint 2D-3D feature for indoor point cloud semantic segmentation

With the rapid growth of new sensors and indoor 3D model acquisition technology, the number of indoor 3D point cloud models has grown rapidly. However, 3D point cloud models reconstructed by traditional modeling methods and indoor 3D scanners are often a group of meaningless “blind point clouds” that can only be used for measurement and visualization, and are difficult to integrate with practical applications. At present, the construction of 3D semantic point cloud mainly draws on the idea of 2D image semantic segmentation, and semantically classifies 3D point cloud by extracting 3D feature operator or using a large number of 3D object model training data to learn in deep learning network. However, it is difficult to construct robust 3D eigenvectors and large-scale 3D training data sets that can maintain the invariance of illumination, rotation and scale, which hinder the process of point cloud semantic segmentation. Therefore, how to use effective methods to perform semantic markup of 3D point clouds makes a large number of indoor”blind point clouds” become semantic, perceptible, and applicable is an urgent problem to be solved. Based on the indoor 2D-dimensional image semantic segmentation results and indoor spatial layout estimation results obtained by the Deeplabv3+ neural network proposed in Chapter 3 , this paper proposes an indoor 3D point cloud semantic segmentation method based on 2D-dimensional joint to perform indoor 3D point cloud model. Semantic markup. Firstly, in order to improve the efficiency of indoor 3D point cloud semantic mark and reduce the impact of re projection error on semantic accuracy, this paper builds a 2D semantic mark pool based on SLIC super pixel segmentation method. Then, based on SfM modeling process The visibility relationship between 2D-dimensional pixel points and 3D-dimensional points, constructing a graph model to establish a"bridge" for 2D- dimensional semantic transfer . Finally, in the process of semantic transfer, in order to alleviate the one-to-many semantic confusion problem between 3D points and 2D pixels, the semantic consistency constraint between images is established to improve the accuracy of semantic transfer. Therefore, this chapter is organized as follows: Section 4.1 introduces the establishment of 2D-dimensional semantic markup pool; Section 4.2 introduces the construction of the " bridge " -graph model of the 2D-dimensional semantic transfer ; Section 4.3 introduces the semantic consistency constraint between images; Section 4.4 introduces the experimental part; 4.5 sections of this chapter summarized.

4.1 Establishment of a semantic label pool

The data set used in this study includes R indoor real-time images taken by mobile devices and the indoor 3D point cloud model reconstructed by SfM-PMV method. Since in the modeling process, through the camera projection, the 2D-dimensional matching pair sequence can be mapped into the 3D-dimensional space to obtain a 3D-dimensional point cloud. Therefore, the mapping relationship between the two dimensions can be established according to the visibility relationship between the 2D-dimensional pixel and the 3D-dimensional point. However, with SfM algorithm can only get sparse 3D point cloud, by PMVS can rebuild algorithm SfM sparse point cloud created encrypt obtain better visual effect of dense point clouds. In order to improve the efficiency of 3D dense point cloud model semantic markup, firstly, the SLIC (Simple Line Interface Method) algorithm (Noh and Woodward, 1976) is used to Superpixel segmentation of 2D images.

4.1.1 Principle of SLIC Superpixel Segmentation

The SLIC Superpixel segmentation algorithm iteratively clusters 2D-dimensional image pixels according to the similarity of features such as brightness and color between pixels, thereby segmenting the 2D-dimensional image into pixel block sets in superpixels (Noh and Woodward, 1976). Then construct a super pixel collection based on the key points (cluster centers), and represent the semantic categories of the super pixels with the semantic annotation of the key points. The specific steps for Superpixel segmentation are as follows:

(1) Initialize the cluster center

Assuming that the input image size is n*n, K pixels of the same size are pre-segmented, and the initial cluster center is obtained according to the following formula.

$$S = sqrt\,(n * n / K) \qquad (4.1)$$

(2) Update of the cluster center

The continuity of the gradient values explains the cohesion between the pixels to some extent. Therefore, the gradient values of all the pixels in the neighborhood of each cluster center in the input image are calculated, and the pixel with the smallest gradient is used as the updated cluster center.

(3) Pixel label allocation

In the input image, the search box of size 2s*2s is scanned, and the label of the cluster center closest to each pixel is assigned to the pixel.

(4) Distance measurement and iterative optimization

In the label assignment process, in order to select the cluster center closest to each pixel, the spatial distance and the color distance are used to measure the similarity between pixels. SLIC superpixel segmentation converts the RGB color of the input image into LAB color to compensate for the unevenness of the RGN color distribution. The pixel similarity measure is then performed according to the following equation.

$$\begin{aligned} d_c &= \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2} \\ d_s &= \sqrt{(x_i - x_i)^2 + (y_j - y_i)^2} \\ D' &= \sqrt{\left(\frac{d_c}{N_c}\right)^2 + \left(\frac{d_s}{N_s}\right)^2} \end{aligned} \qquad (4.2)$$

Wherein each represents a pixel i and j color space and the distance between the distances; represents the maximum spatial distance image class.

4.1.2 Process of semantic token pool construction.

Based on the 2D-dimensional visibility relationship in the SfM-PMVS modeling process, we can back project the 3D point cloud into the 2D image space through the camera imaging principle, so that each 3D point in the indoor 3D point cloud can be The super-pixel semantics with unique semantic values are marked to finally realize the semantic transfer of the 3D- dimensional point cloud in super pixel-based super-pixel semantic mark pool, thereby avoiding the time-consuming problem of pixel-by-pixel semantic transfer, and at the same time, it can alleviate the weight The effect of projection error on the accuracy of semantic transfer.

As shown in FIG. 4.1. Firstly, the SLIC super-pixel segmentation algorithm is used to iteratively segment the input indoor image. Then, based on the Deeplabv3+ deep convolutional neural network designed in Chapter 3, the indoor image semantic segmentation result and spatial layout estimation result are obtained in Superpixel units. Perform 2D-dimensional semantic markup. Among them, the K cluster centers in the super pixel segmentation are taken as the key points, and the semantic values of the pixels in which they are located are used as the semantic labels of the superpixels to form a series of semantic token pools in superpixels.

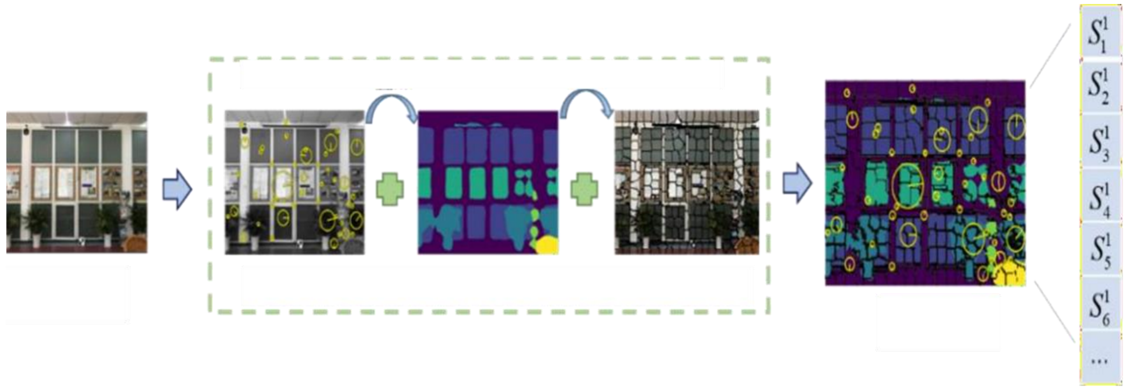


Figure4. 1: The construction process of the super pixel semantic mark pool

4.2 Construction of 2D-dimensional viewable model

After obtaining the semantic annotation of the 2D-dimensional image in superpixels, it is necessary to establish a”bridge” of semantic cross-domain transfer between the 2D-dimensional image and the 3D-dimensional point cloud to realize the transmission of the semantics of the 2D-dimensional object to the 3D-dimensional point semantics. To this end, based on the visibility relationship between 2D-dimensional image pixel points and 3D-dimensional point cloud in the multi-view modeling algorithm SfM-PMVS modeling process, this paper builds a viewable model of 2D-dimensional semantic cross-domain transfer to break through different dimensions.

4.2.1 2D-dimensional semantic transfer basis

In the SfM-PMVS modeling process, 2D-dimensional pixel points are projected into the 3D-dimensional space by the camera. This is the key to constructing a”bridge” for 2D-dimensional semantic transmission.

The basic principles are as follows:

SfM (Structure from Motion) reconstructs the relative position of the corresponding 3D points by restoring the pose of each camera. It is a low cost and high modeling efficiency modeling method in many 3D modeling methods (Agarwal et al., 2009)., Among them, the projection principle of 3D pixel generated by 2D pixel in SfM modeling process is small whole camera model, and the small hole imaging model approximates the imaging behavior of real camera without distortion. In the process of SfM recovering camera pose parameters, it needs to be it performs distortion correction to get more accurate camera parameters. Based on the projection matrix [R, T] of the small whole camera model, the transformation of the 3D points and 2D image coordinates in the indoor environment needs to be transformed by the following coordinate system (Fig. 4.2).

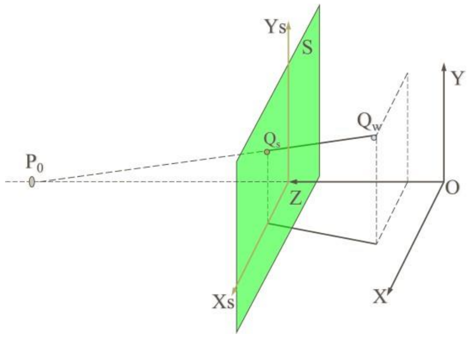


Figure4. 2: Camera Imaging Principle

The transformation of the 2D-dimensional coordinates includes the following projection transformation steps: first, the 3D-dimensional world coordinate system is converted to a 2D-dimensional camera coordinate system (formula 4.3), then the camera coordinate system is converted to the image physical coordinate system (formula 4.4), and finally, the image is The

2D-dimensional projection coordinates in the physical coordinate system are converted to the image pixel coordinate system (formula 4.5), and finally the 2D-dimensional image pixel coordinates obtained by projecting the 3D-dimensional space point coordinates in the world coordinate system on the corresponding 2D-dimensional image are obtained

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \sim \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{4.3}$$

Where R and T respectively represent a 3*3 rotation matrix and a 3*1 translation matrix of the world coordinate system transformed to the camera coordinate system.

$$x = f\frac{X_c}{Z_c}, y = f\frac{Y_c}{Z_c} \tag{4.4}$$

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{4.5}$$

Where f is the focal length of the camera, representing the projected coordinates of the 3D-dimensional space point in the 2D-dimensional physical coordinate system. The image physical coordinate system is transformed into the image pixel coordinate system by:

$$\begin{bmatrix} u \\ v \end{bmatrix} \sim K \begin{bmatrix} x \\ y \end{bmatrix} \tag{4.6}$$

$$K = \begin{bmatrix} \alpha_u & s & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 0 \end{bmatrix} \tag{4.7}$$

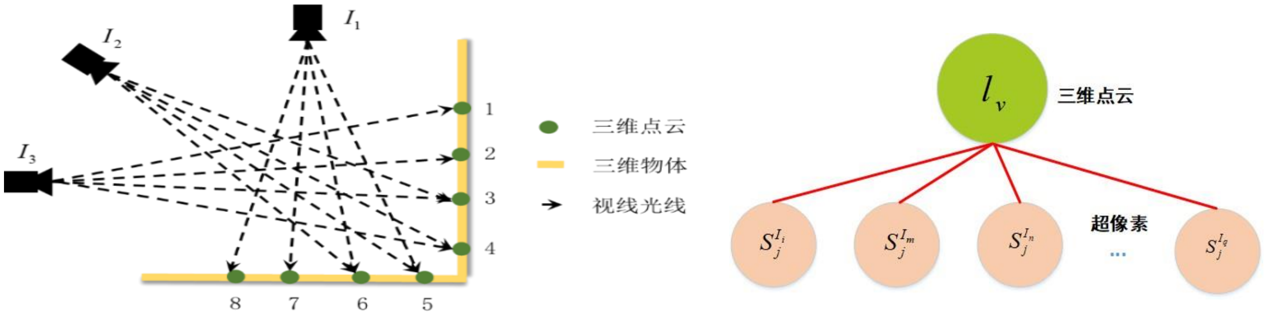
In the formula, K is a matrix camera internal reference, and are in the camera x –axis and y normalized focal axis, under normal circumstances, = ; S represents the camera axis tilt parameter, generally set 0; and represents a main image Point coordinates; and pixel coordinate system coordinates and image physical coordinate system coordinates respectively representing the projection points of the 3D-dimensional point .In summary, the 3D point cloud can be projected onto the 2D image according to the following formula to obtain the 2D image coordinates. The projection relationship between the 2D image pixel and the 3D point constitutes the basis of the 2D and 3D semantic transfer. The next section will introduce the construction of a 2D-dimensional viewable model based on 2D-dimensional projection relationships.

$$z\begin{bmatrix} u \\ v \end{bmatrix} \sim K\begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix}\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{4.8}$$

4.2.2 Construction of viewable models

In order to construct models, it is assumed that there is a 3D-dimensional point cloud, which represents the first 3D-dimensional point in the 3D- dimensional space, wherein the pixel point of the corresponding photo of the camera is calculated by triangulation, and therefore, the 3D-dimensional point has a visual relationship with the camera. A 2D-dimensional visibility relationship can be established by the projection relationship between the 3D-dimensional point and the camera. Based on the visual relationship, we can know which pictures are reconstructed from each 3D point. In other words, the 3D-dimensional point can be connected to its corresponding plurality of 2D-dimensional superpixels by the visibility relationship between the camera and the 3D-dimensional point. FIG. 4.2 (A) represented by a simple visual FIG relationship between the camera and the 3D-dimensional points.

According to the above-mentioned correspondence between 3D points and 2D Superpixel visibility, this research work constructs a graph model of 2D-dimensional semantic transfer to complete the semantic markup of 3D point cloud. In this chapter, the 3D point and 2D Superpixel are used as the nodes of the visibility graph model. At the same time, the connection between each super pixel and its corresponding 3D point is taken as the edge of the graph model, which is expressed as the construction based on Markov. Visibility map model for the random field (MRF). The Markov random field model defines the correlation between variables. The Markov model indicates that the related features of the state features of a certain thing exist only in the neighborhood, but have nothing to do with the characteristics of other fields. According to the Markov property, each node of the graph model connecting the 3D-dimensional space and the 2D-dimensional image defined in this paper is only related to the semantics of the node itself and the nodes in its neighborhood, and has nothing to do with other nodes. From the visibility relationship between the SfM- based camera and the 3D point, a viewable model as shown in Fig. 4.3 (b) is constructed. Therefore, the transfer of 3D points and 2D Superpixel semantics can be achieved through a viewable model.



(b)

Figure4. 3: Construction of the model can view. (A) Represented based SfM visibility of the 3D-dimensional relationship corresponding to the dot code camera. (B) Shows the construction of 3D-dimensional point connection over a 2D-dimensional pixel corresponding thereto may view the model, represents The * picture, which indicates the * th super pixel in the first picture.

4.3 Multi-category label semantic consistency constraints

In SfM-PMVS , the same 3D-dimensional point may pose different camera images obtained by projecting the multi-view modeling principle, therefore, based on 4.2 visibility graph constructed two model 3D semantic measure of semantic transfer labeling process in the 3D-dimensional points, The same 3D point may correspond to multiple 2D Superpixel semantic tag values. In other words, the semantics of each 3D point is determined by the semantic tags of all the 2D superpixels it corresponds to. At the same time, the semantic tags of the 2D Superpixel depend on the semantic tags of the corresponding key points. Although the Deeplabv3+ neural network proposed in the third chapter has improved the semantic segmentation of indoor height occlusion, uneven illumination and many objects, it still inevitably has some misclassification. This also leads to the inconsistency of the semantics of the 2D-dimensional Superpixel corresponding to the same 3D-dimensional point, which leads to the”confusion” problem in the semantic transfer process. Therefore, for the multi-class semantic mark problem in the process of 2D-dimensional transmission, this paper proposes a multi- category mark energy function based on the viewable model to suppress the multi-class mark error caused by the mis-segmentation of the neural network, and then improve the 2D- dimensional semantics.

For 3D-dimensional point X- I, assuming a 2D-dimensional super-set of pixels corresponding to, S I exist in different semantic category label. In this paper, the distribution of 3D-dimensional point semantic markers is represented by the probability of occurrence of each semantic label in 3D-dimensional space. Therefore, this paper explores the consistency of the semantic labels of multiple superpixels corresponding to each 3D-dimensional point to predict the 3D-dimensional point semantic value, thus alleviating the multi-category semantic transfer error caused by Deeplabv3+ mis-segmentation. Construct the following energy function with the semantic tag of the 3D point as a variable:

$$E = \sum_{X_i \in \mathrm{C}} \psi_d \big(l \, (\, X_i \,) \big) \tag{4.9}$$

Represents a 3D-dimensional point set that represents the semantic mark value of a 3D-dimensional point as a data item. Since the semantic label distribution of a 3D-dimensional point is affected by the semantics of multiple superpixels associated with it, assuming that the semantic distribution function of the 3D-dimensional point is, the data item in the above energy function can be defined as:

$$\psi_{\mathrm{d}} (l(p_i)) = \text{-}P_{i \; p} (l(p_i)) \qquad \text{（4.10）}$$

Finally, the semantic mark value of each 3D point in the 3D point cloud is obtained by solving the minimum value of the energy function E. This paper uses the graph cut algorithm to solve the minimum energy function:

$$l(\cdot) = \arg \min_{l \in L} E = \arg \min_{l \in L} \sum_{p_i \in P} \psi_d (l(p_i))$$

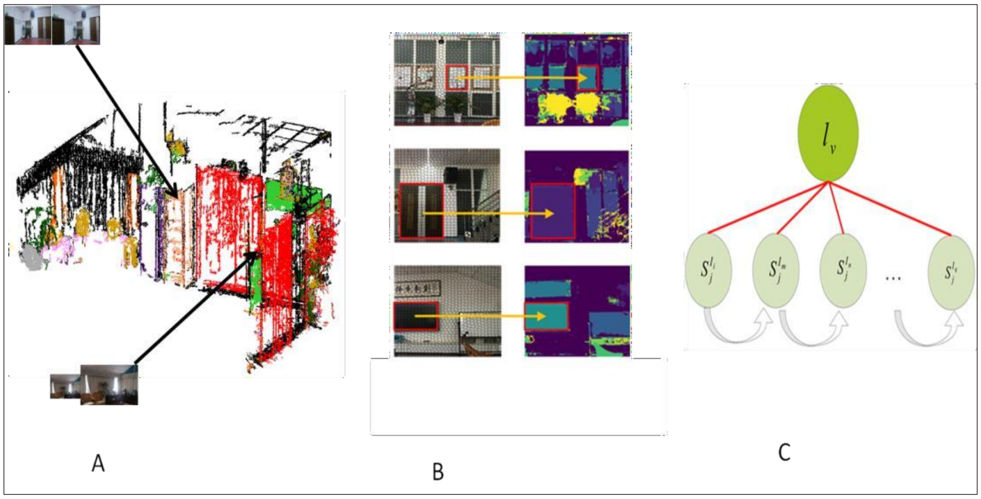


Figure4. 4: the composition of the data item. (A) shows a correspondence point cloud with 3D-dimensional point 2D-dimensional pixel-to-many; (B) indicates a 2D-dimensional super-cell Superpixel semantic tags pixel Deeplabv3+ corresponding relationship between pixels in semantic segmentation result; (C) is a multi- A graph model that classifies semantic consistency constraints.

4.4 Experiment and analysis

In order to verify the effectiveness of the indoor 3D point cloud semantic markup method proposed by the 2D-dimensional joint method, this paper builds a 3D point cloud model based on the SfM-PMVS modeling principle, using the conference room dataset collected by the mobile camera in Chapter 3. The experimental results of semantic segmentation and spatial layout estimation of conference room data in Deeplabv3+ neural network in Chapter 3 are used to transmit the corresponding 3D point cloud semantics. Finally, a 3D semantic point cloud model with high semantic precision is obtained.

4.4.1 Experimental data

In order to verify the robustness of the proposed method to the semantic segmentation of point cloud models in complex indoor scenes, the conference room dataset is used for different time periods. The data set of the same conference room photographed by the mobile camera; the 3D-dimensional point cloud corresponding to the conference room image set obtained by the SfM-PMVS algorithm as shown in Fig. 4.5 (b), which is included in the reconstructed 3D-dimensional scene. More objects, is a closed indoor scene with a relatively large scale.

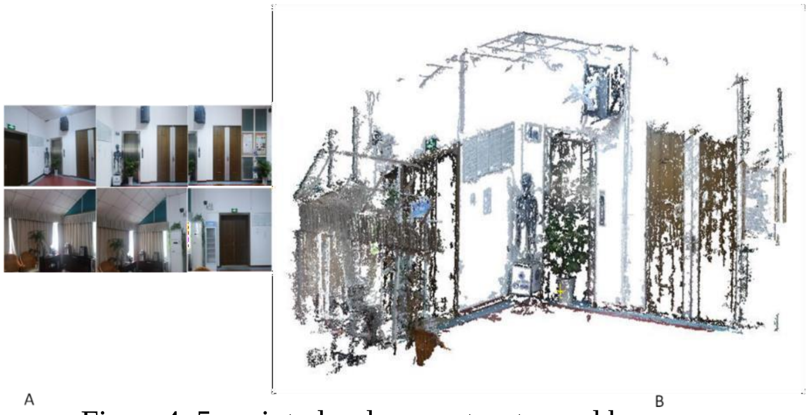


Figure4. 5: point cloud reconstruct a and b

4.4.2 Superpixel Segmentation Experiment

Based on the indoor semantic segmentation results and layout estimation results extracted by Deeplabv3+ neural network, this paper proposes an indoor 3D point cloud semantic segmentation method based on 2D-dimensional joints to perform semantic tagging of indoor 3D point clouds. In order to improve the efficiency of semantic markup of 3D dense point cloud model, SLIC Superpixel segmentation algorithm is used to Superpixel segmentation of 49 conference room images. Then, the K cluster centers in the super pixel segmentation are taken as the key points, and the semantic tag values are used as the semantic tags of the superpixels to form a 2D-dimensional semantic mark pool in superpixels. In this experiment, the SLIC Superpixel segmentation algorithm has a segmentation block size of 10 and a regular term coefficient of 1, and the resulting Superpixel segmentation result is shown in Figure 4.6.

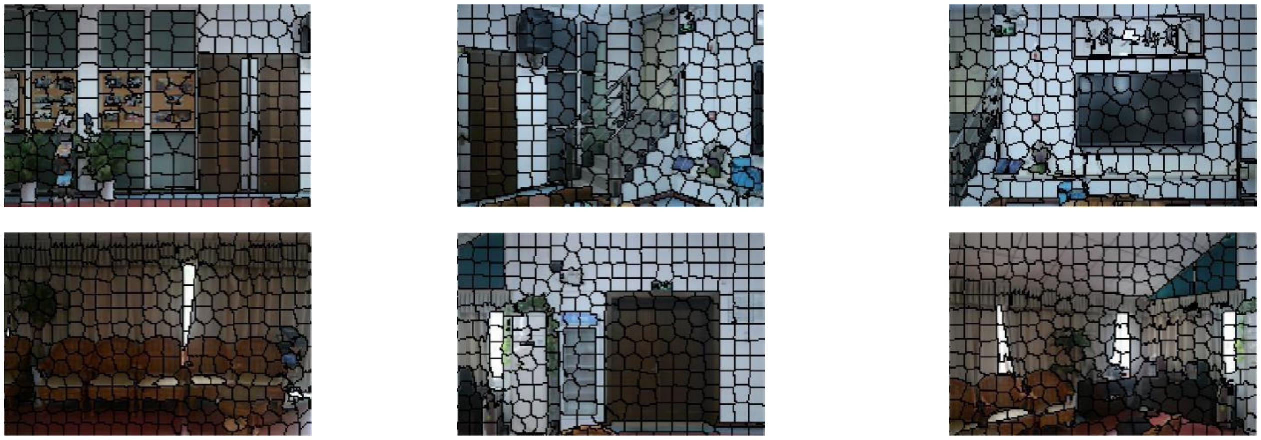


Figure4. 6: Example of Superpixel segmentation results

4.4.3 3D-dimensional point cloud semantic segmentation experiment

To understand the visibility relationship between 2D-dimensional pixels and 3D-dimensional points in the SfM modeling process, a 2D-dimensional super-pixel and 3D-dimensional point-based viewable model is established to transfer the 2D-dimensional semantics. In order to alleviate the semantic confusion between 3D points and 2D superpixels caused by Deeplabv3+ mis-segmentation, the semantic transfer-based graph model establishes semantic consistency constraints between images to improve the accuracy of semantic transfer. Based on the semantic segmentation and spatial layout estimation results obtained by Deeplabv3+ neural network, this experiment uses 2D-dimensional semantic transfer in superpixels to obtain the 3D-dimensional semantic point cloud model as shown in Figure 4.7 .

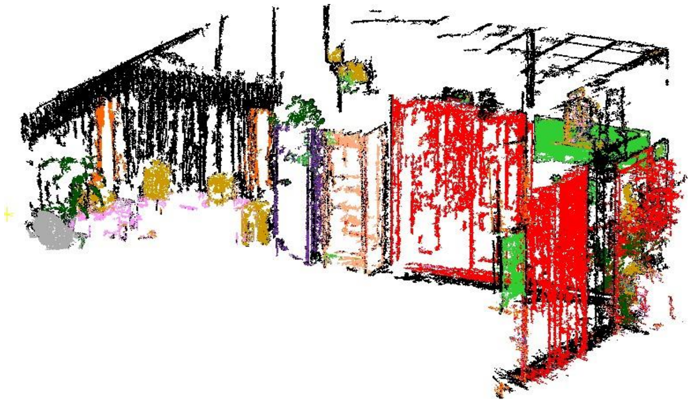


Figure4. 7: 3D point cloud semantic segmentation results

In order to evaluate the segmentation accuracy of the 3D semantic point cloud obtained by the 2D-dimensional semantic transfer method proposed in this paper, this experiment uses the 3D point cloud used by manual marking as the true value (Figure 4.8) to calculate the segmentation precision. The 3D test point cloud used in the experiment has a total of 5,581,443 points. According to statistics, the number of points correctly classified is 12,521,788.

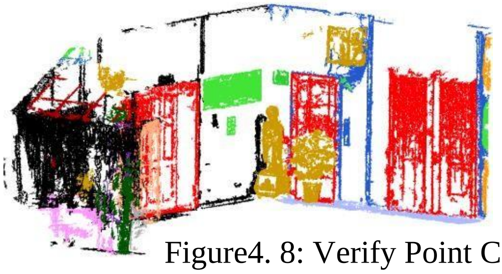
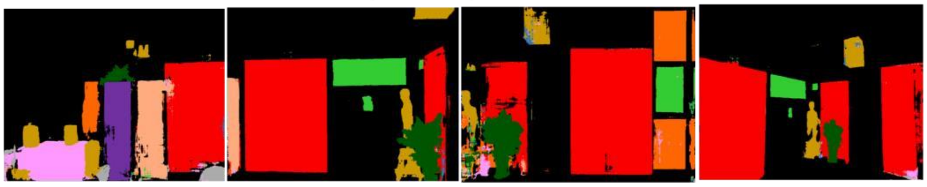


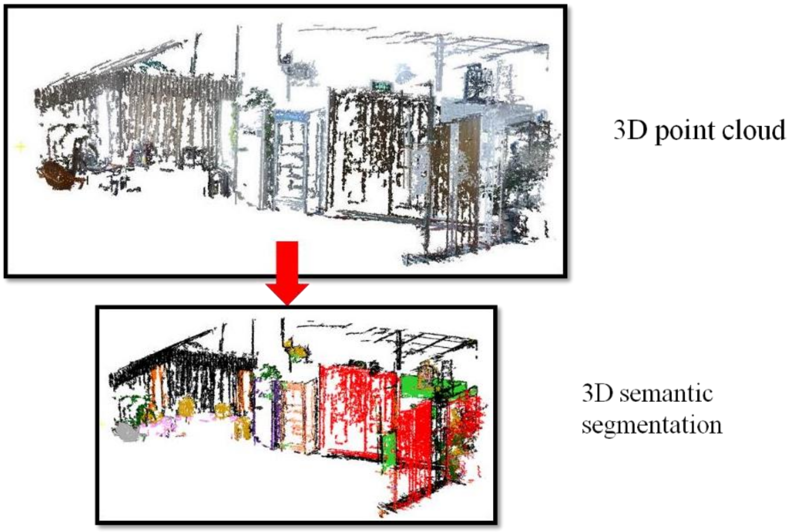
Figure4. 8: Verify Point Cloud

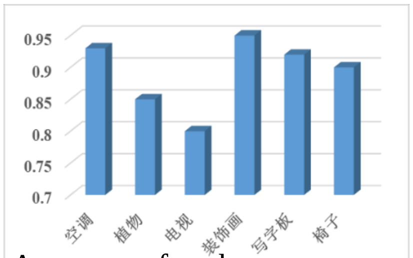
4.4.4 Analysis of experimental results

(a) original image



(b) Deeplabv3+ image object semantic segmentation and spatial layout estimation results





(f) Accuracy of each scene category

Figure4. 9: 2D-dimensional semantic transfer accuracy analysis

In order to validate the 3D-dimensional point cloud segmentation method proposed in this paper, four more complex scenes in conference room point cloud data are selected for accuracy analysis. As shown in Figure 4.9, scene 1 contains six categories of air conditioners, plants, televisions, decorative paintings, writing boards and chairs. The overall classification accuracy is 89.17%. Among them, the accuracy of air conditioners, decorative paintings and writing boards are all above 90% .

In summary, according to the comparative analysis of the classification accuracy of each category in the indoor environment, the objects in the conference room data set used in this paper can basically be correctly classified, but inevitably there are some special cases where the segmentation accuracy is low. The main reasons are divided into the following three points: (1) Due to the 2D-dimensional joint 3D point cloud semantic segmentation method proposed in this paper, the semantic segmentation precision of 3D point cloud depends to a large extent on the semantic segmentation accuracy of 2D image objects. Therefore, like the plant, there are some miscellaneous objects and other shape features are more complex, difficult to segment the object, the segmentation accuracy is often lower than the other more regular object segmentation accuracy; In addition, the target conference room scene used in this paper is mainly large Object- based, relatively trivial objects have less training data, so the recognition of small objects such as small decorations on the wall is not ideal.

(2) The source data used in this work is the point cloud data generated by SfM- PMVS algorithm and its corresponding 2D-dimensional image. The semantic segmentation result is obtained by semantic separation of 2D image semantics with 3D point cloud. Therefore, the effect of the modeled 3D point cloud model will also affect the accuracy of the 2D and 3D semantic transfer of the point cloud to some extent. (3) The indoor environment is complex, and there are often a large number of occlusion phenomena. Even if the spatial layout estimation proposed in this paper can alleviate some occlusion errors, it is still difficult to separate the semantics of multiple objects with multiple occlusions.

At present, the commonly used 3D point cloud classification methods are divided into direct classification and indirect classification. The point cloud classification can be directly performed by extracting 3D point features or using deep learning algorithms. However, due to the lack of robust 3D feature operators and sufficient The indoor 3D point cloud training data set makes it difficult to directly classify the 3D point cloud. Therefore, for the complex situations that may occur indoors, this chapter proposes an indoor 3D point cloud semantic segmentation method based on 2D-dimensional joints based on the image semantic segmentation results and spatial layout estimation results obtained by the Deeplabv3+ neural network designed in Chapter 3 . The research results are as follows:

(1) Based on the visibility relationship between 2D-dimensional pixels and 3D-dimensional points in the SfM-PMVS modeling process, a 2D-dimensional viewable model is constructed for 2D-dimensional semantic transfer.

(2) In order to solve the problem of semantic confusion between 3D points and 2D superpixels due to 2D-dimensional mis-segmentation, this paper constructs semantic consistency constraints between images to ensure the accuracy of 3D point cloud semantic segmentation.

5. Conclusion

5.1 Research Summary

With the rapid development of location-based services and GIS applications in outdoor environments, indoor spaces are increasingly demanding precision location services. Compared with the 2D-dimensional map, the indoor 3D-dimensional point cloud model has the advantages of strong sense of reality and rich information. At present, indoor 3D point cloud models can be acquired by laser scanners, RGBD cameras and image-based modeling methods. However, the directly obtained 3D point cloud model is a series of meaningless “blind point clouds”. It lacks objectified and structured information and can only be used for distance measurement and visualization, but not with indoor navigation, 3D object retrieval and tracking. Practical applications such as virtual reality interaction are in line. The 3D semantic point cloud model assigns semantic features to the 3D point cloud model point by point, so that the 3D point cloud model can be analyzed and applied in actual production life. At present, the semantic segmentation of 3D point cloud can be divided into two ways. One is to classify the 3D point cloud by directly extracting the features of 3D points; the other is based on the depth learning algorithm in the field of 2D image segmentation. Semantic recognition ability, indirectly semantic segmentation of 3D point cloud. Among them, due to the lack of robust 3D feature operators and sufficient 3D training datasets, the method of semantically classifying 3D point clouds directly is not universal. The current hot learning algorithm has been in the field of 2D-dimensional image semantic segmentation for many years, and has formed a relatively complete semantic recognition system. In addition, a large number of well-distributed and publicly available indoor training data sets make 2D-dimensional image semantic segmentation much simpler than semantic extraction in 3D-dimensional space. At the same time, the visual projection relationship between the 3D point cloud model based on multi-view image and its corresponding 2D image provides the possibility of 2D-dimensional semantic”cross- domain " transfer. Therefore, based on the deep foundation of deep learning algorithm in the field of 2D image semantic segmentation, combined with the modeling principle of multi-view image 3D reconstruction, this paper studies the image semantic segmentation and spatial layout extraction neural network for indoor complex environment and extracts it. The semantics are passed to the 3D point cloud model. Specifically, the main research content of this paper can be summarized as the following three aspects:

(1) Indoor scene semantic segmentation and spatial layout extraction based on Deeplabv3+ deep network

Compared with the outdoor environment, the indoor space tends to have some high occlusion, uneven illumination, and the indoor objects are complicated, which makes the indoor image semantic segmentation difficult. The indoor environment usually presents a horizontal and vertical regular structure. Therefore, based on the NYUDv2 RGBD indoor public data set, this paper designs a Deeplabv3+ neural network that can simultaneously perform indoor spatial layout estimation and object semantic segmentation. Among them, the network is divided into three layers, including: backbone network, semantic segmentation sub-network and spatial layout extraction sub-network. In the training process of the network, the strategy of layered training is adopted, and the more complex semantic features are prioritized, and then the network training of spatial layout is carried out. Finally, the mutual integration of the features of each layer is carried out and jointly optimized. In the process of training, since there are some objects in the target scene that have the same category as the NYUDv2 RGBD training data set but the shape features are significantly different, the target scene marker data is added to the training data set for network fine-tuning to improve the recognition of the target scene object by the network.

Robustness. In addition, large-scale residual structure is adopted in the network to realize the transfer of low-level spatial structure features to high-level semantic features. At the same time, the tight-density structure is used in the backbone network to transfer and reuse features between layers. In the network, the spatial convolution structure is used to extract and fuse multi-scale features to obtain higher-level semantic information. For the estimation of indoor spatial layout, the spatial layout rough estimation results are obtained by combining the multi-scale features extracted by the low-level network, and the layout structure is further refined through indoor straight line extraction and vanishing point estimation. Finally, the semantic segmentation results and spatial layout estimation results of indoor objects with higher precision are obtained. In this paper, the robustness of Deeplabv3+ neural network is verified by comparing the semantic segmentation effect of Deeplabv3+ neural network with the current mainstream DeepLab v3+ and DenseNet121 on the same dataset.

(2) 3D point cloud semantic segmentation based on 2D-dimensional semantic transfer

Based on the one-to-one mapping relationship between the 2D-dimensional image and the 3D-dimensional point in the multi-view image 3D-dimensional reconstruction process, a “bridge “for the 2D-dimensional semantic cross-domain transmission can be established. In order to improve the efficiency of semantic transfer and alleviate the semantic transfer error caused by re-projection error, this paper builds a super-pixel semantic mark pool based on the indoor object semantic segmentation and spatial layout estimation results obtained by Deeplabv3+ neural network, and performs semantic transfer in super-pixel units. This avoids time-consuming operations of pixel-by-pixel semantic delivery. Based on the visibility relationship between 3D points and 2D pixels in SfM modeling process, the 3D points and their corresponding 2D superpixels are used to establish the visibility graph model for the transmission of 2D and 3D semantics.

(3) Viewable model optimization of semantic consistency constraints between images

Due to the complexity of the indoor environment, the result of semantic segmentation of the Deeplabv3+ neural network often leads to some inevitable mis-segmentation. In the multi- view image modeling process, a 3D-dimensional point is usually reconstructed by multiple cameras, that is, one 3D-dimensional point corresponds to multiple super-pixels, and 2D- dimensional mis-segmentation semantics causes multiple super-pixels corresponding to the same

3D-dimensional point. The problem of inconsistent semantic tags. In order to alleviate the one- to-many semantic confusion problem between 3D points and 2D superpixels, this paper improves the accuracy of 2D and 3D semantic transmission by establishing semantic consistency constraints between images.

5.2 Innovation points

The meaningless " blind point cloud " directly obtained by 3D laser scanners, RGB- D cameras or image-based modeling algorithms cannot be applied to actual production and life. Semantic marking of 3D point cloud model helps 3D point cloud data to provide meaningful information for indoor navigation, 3D target retrieval and tracking, and virtual reality interaction. Aiming at the complex ambiguity, uneven illumination and many objects in the indoor environment, this paper proposes a kind of second based on the strong recognition ability of deep learning algorithm in the field of 2D-dimensional image semantic segmentation, combined with the principle of multi-view image modeling. Three joint indoor 3D point cloud semantic segmentation method. The experimental results show that the indoor object semantic segmentation and spatial layout extraction network proposed in this paper - Deeplabv3+ neural network can robustly identify and extract indoor objects. At the same time, the 2D-dimensional semantic transfer method proposed in this paper can also obtain high precision. 3D semantic point cloud model. In general, the main work and innovations of this article can be summarized as the following two points:

(1) In this paper, a neural network- Deeplabv3+ neural network capable of simultaneous image object semantic segmentation and spatial layout extraction is designed for complex and variable indoor scenes. The network combines large-scale residual structure, tight-intensive structure and The cavity convolution structure is used to extract and fuse multi-scale features, and adopts multi-level training strategy to obtain more robust indoor object semantic segmentation results and spatial layout rough estimation results. The Deeplabv3+ network is designed to provide a new solution to the indoor image semantic segmentation algorithm.

(2) Based on the Deeplabv3+ neural network, the indoor object semantic segmentation results and spatial layout estimation results, combined with the principle of multi-view image modeling, this paper proposes a 2D-dimensional semantic transfer method for viewable models based on semantic consistency constraints between images for 3D-dimensional Semantic

segmentation of point clouds. This method provides a complementary method for 3D point cloud segmentation by making full use of the advantages of deep learning algorithms in the field of image semantic segmentation.

5.3 Discussion and outlook

In this paper, based on the application of 3D semantic point cloud model in actual production and life, this paper studies an indoor 3D point cloud segmentation method based on 2D-dimensional joint. Among them, for the indoor environment is a huge complex challenges posed by indoor objects semantic segmentation task, this paper designed a multi-task output can be extracted for indoor object semantics division and spatial layout structure Deeplabv3+ neural network. According to the SfM modeling principle, the graph model of semantic consistency constraint between images is constructed to realize the semantic transfer from 2D image space to 3D point cloud space to complete the semantic segmentation of indoor 3D point cloud. Finally, according to the indoor herein, two 3D-dimensional data acquisition , a series of semantic segmentation experiments , demonstrate the effectiveness of the proposed method. However, there are still many areas in this study that need further improvement and improvement. Specifically, future research work focuses on the following three aspects:

(1) In terms of 2D-dimensional image segmentation, this paper designs Deeplabv3+ neural network to simultaneously perform indoor object segmentation and spatial layout extraction to alleviate the semantic segmentation problem caused by complex indoor environment. Semantic segmentation of neural networks often requires a large amount of network training data, which has high requirements for equipment and manpower. In addition, a large amount of training data will also generate relatively large memory loss for the device, hindering the rapid development of fine segmentation of image semantics. Accordingly, reference Zhou et (Zhou et Al. , 2018) active-learning network training pattern proposed, which can reduce the training mode only the training data, and can provide more targeted training difficult segmented object, the present study The network and its training mode will be further improved to make the segmentation more efficient and detailed.

(2) In terms of 3D point cloud semantic segmentation, this paper explores a viewable model with semantic consistency constraints between images for the transfer of 2D-dimensional semantics. However, this semantic markup method has a strong dependence on the 2D-dimensional image segmentation results. In other words, the semantic segmentation result of the

2D-dimensional image determines the semantic segmentation precision of the 3D-dimensional point cloud to some extent. The 3D-dimensional point cloud's own 3D-dimensional features, such as geometric features, shape structure features, etc., can be used as an important feature of 3D point cloud semantic segmentation to assist the 3D point cloud for more robust semantic segmentation. Therefore, the future segmentation work will combine the 2D-dimensional image object semantic features and the 3D-dimensional point cloud features to perform the fine semantic segmentation of the 3D-dimensional point cloud.

(3) The indoor 3D point cloud semantic segmentation method based on the 2D- dimensional joint method proposed in this paper is only applicable to the 3D point cloud model constructed by SfM algorithm, and cannot provide effective semantic segmentation for the point cloud model obtained by other methods. Therefore, in the next study, a more universal 3D point cloud segmentation method will be sought to cope with the semantic segmentation of 3D point cloud models obtained in various ways.

References

AIJAZI A, CHECCHIN P, TRASSOUDAINE LJRS 2013. Segmentation based classification of 3D urban point clouds: A super-voxel based approach with evaluation. 5: 1624-1650.

AUDEBERT N, LE SAUX B, LEFèVRE S 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks [C] //; Springer; 180-196.

BADRINARAYANAN V, KENDALL A, CIPOLLA RJITOPA, et al. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. 39: 2481-2495.

BAI S, BAI X, ZHOU Z, et al. 2016. Gift: A real-time and scalable 3d shape search engine [C] //; 5023-5032.

BELTON D, LICHTI DDJIAPRSSIS 2006. Classification and segmentation of terrestrial laser scanner point clouds using local variance information. 36: 44-49.

BOULCH A, GUERRY J, LE SAUX B, et al. 2018. SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. 71: 189-198.

71. BOULCH A, GUERRY J, SAUX BL, et al. 2017. SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks.

CALTAGIRONE L, SCHEIDEGGER S, SVENSSON L, et al. 2017. Fast LIDAR-based road detection using fully convolutional neural networks [C] //, IEEE; 1019-1024.

CHEHATA N, GUO L, MALLET CJIAOP, REMOTE SENSING, et al. 2009. Airborne lidar feature selection for urban classification using random forests. 38: W8.

CHEN LC, PAPANDREOU G, SCHROFF F, et al. 2017. Rethinking atrous convolution for semantic image segmentation.

CHEN LC, PAPANDREOU G, KOKKINOS I, et al. 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. PP: 834-848.

CHIU HP, SAMARASEKERA S, KUMAR R, et al. 2019. Augmenting reality using semantic segmentation [M]. Google Patents.

CHOI W, CHAO YW, PANTOFARU C, et al. 2015. Indoor scene understanding with geometric and semantic contexts. 112: 204-220.

CURA R, PERRET J, PAPARODITIS NJAPA 2018. A state of the art of urban reconstruction: street, street network, vegetation, urban feature.

DAI A, CHANG AX, SAVVA M, et al. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes.

DAI J, HE K, SUN J 2016. Instance-aware semantic segmentation via multi-task network cascades [C] //; 3150-3158.

DALAL N, TRIGGS B 2005. Histograms of oriented gradients for human detection [C] //, IEEE Computer Society; 886--893.

DIMITROV A, GOLPARVAR-FARD MJAIC 2015. Segmentation of building point cloud models including detailed architectural/structural features and MEP systems. 51: 32-45.

. DENG J, DONG W, SOCHER R, et al 2009. ImageNet: A large-scale hierarchical image database [C] //; 248-255.

FU J, LIU J, WANG Y, et al. 2019. Stacked deconvolutional network for semantic segmentation.

FURUKAWA Y, CURLESS B, SEITZ SM, et al. 2010. Towards internet-scale multi-view stereo [C] //, IEEE; 1434-1441.

FURUKAWA Y, HERNáNDEZ CJF, GRAPHICS TIC, et al. 2015. Multi-view stereo: A tutorial. 9: 1-148.

FURUKAWA Y, PONCE JJITOPA, INTELLIGENCE M 2010. Accurate, dense, and robust multiview stereopsis. 32: 1362-1376.

GEORGAKIS G, MOUSAVIAN A, BERG AC, et al. 2017. Synthesizing Training Data for Object Detection in Indoor Scenes.

GIUSTI A, CIREŞAN DC, MASCI J, et al. 2013. Fast image scanning with deep max-pooling convolutional neural networks [C] //, IEEE; 4034-4038.

GOULD S, FULTON R, KOLLER D 2009. Decomposing a scene into geometric and semantically consistent regions [C] //; 1-8.

GUPTA A, HEBERT M, KANADE T, et al. 2010. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces [C] //; 1288-1296.

HACKEL T, WEGNER JD, SCHINDLER KJIJOP, et al. 2017. Joint classification and contour extraction of large 3D point clouds. 130: 231-245.

HARRIS CG, STEPHENS M 1988. A combined corner and edge detector [C] //, Citeseer; 10-5244.

HERMANS A, FLOROS G, LEIBE B 2014. Dense 3D semantic mapping of indoor scenes from RGB-D images [C] //; 2631-2638.

HEDMAN P, RITSCHEL T, DRETTAKIS G, et al. 2016. Scalable inside-out image-based rendering. 35: 231.

HE K, ZHANG X, REN S, et al. 2015. Deep Residual Learning for Image Recognition. 770-778.

HOLZMANN T, OSWALD MR, POLLEFEYS M, et al. 2017. Plane-based Surface Regularization for Urban 3D Reconstruction [C] //.

HUSAIN F, SCHULZ H, DELLEN B, et al. 2016. Combining Semantic and Geometric Features for Object Class Segmentation of Indoor Scenes. 2: 49-55.

KALOGERAKIS E, HERTZMANN A, SINGH K 2010. Learning 3D mesh segmentation and labeling [C] //; 1-12.

KANG Z, YANG JJIJOP, SENSING R 2018. A probabilistic graphical model for the classification of mobile LiDAR point clouds. 143: 108-123.

LEE CY, BADRINARAYANAN V, MALISIEWICZ T, et al. 2017. Roomnet: End-to-end room layout estimation [C] //; 4865-4874.

LI CY, VU NT 2018. Densely Connected Convolutional Networks for Speech Recognition.

LI W, SAEEDI S, MCCORMAC J, et al. 2018. InteriorNet: Mega-scale multi-sensor photo-realistic indoor scenes dataset.

LIU C, CHEN LC, SCHROFF F, et al 2019. Auto-deeplab:. Hierarchical neural architecture search for semantic image segmentation.

LIU C, YUEN J, TORRALBA A 2016. Nonparametric scene parsing via label transfer [M], Dense Image Correspondences for Computer Vision. Springer: 207-236.

LIU M, GUO Y, WANG JJVC 2017. Indoor scene modeling from a single image using normal inference and edge features. 33: 1-14.

LU G 2016. From coarse to fine: quickly and accurately obtaining indoor image-based localization under various illuminations [M]. University of Delaware.

LU X, YAO J, TU J, et al. 2016. PAIRWISE LINKAGE FOR POINT CLOUD SEGMENTATION. III-3: 201-208.

MITRA NJ, NGUYEN A 2003. Estimating surface normals in noisy point cloud data [C] //, ACM; 322-328.

MORAVEC HP 1980. Obstacle avoidance and navigation in the real world by a seeing robot rover [M]. STANFORD UNIV CA DEPT OF COMPUTER SCIENCE.

MUNOZ D, BAGNELL JA, HEBERT M 2010. Stacked Hierarchical Labeling [C] //; 57-70.

NAN L, XIE K, SHARF AJATOG 2012. A search-classify approach for cluttered indoor scene understanding. 31: 137.

NEWCOMBE RA, IZADI S, HILLIGES O, et al. 2011. Kinectfusion: Real-time dense surface mapping and tracking [C] //; 127-136.

NG PC, HENIKOFF SJNAR 2003. SIFT: Predicting amino acid changes that affect protein function. 31: 3812-3814.

NI H, LIN X, ZHANG JJRS 2017. Classification of ALS point cloud with improved point cloud segmentation and random forests. 9: 288.

NOCEDAL J, WRIGHT S 2006. Numerical optimization [M]. Springer Science & Business Media.

NOH WF, WOODWARD P 1976. SLIC (Simple Line Interface Calculation) [M]. Springer Berlin Heidelberg.

point sets in a metric space [C] //; 5099-5108.

RAKELLY K, SHELHAMER E, DARRELL T, et al. 2018. Few-Shot Segmentation Propagation with Guided Networks.

REN Y, LI S, CHEN C, et al. 2016. A coarse-to-fine indoor layout estimation (cfile) method [C] //, Springer; 36-51.

ROSKA T, CHUA LOJITOC, ANALOG SI, et al. 1993. The CNN universal machine: an analogic array computer. 40: 163-173.

RUSSELL BC, TORRALBA A, MURPHY KP, et al. 2008. LabelMe: A Database and Web-Based Tool for Image Annotation. 77: 157-173.

SCHERER D, MüLLER A, BEHNKE S 2010. Evaluation of pooling operations in convolutional architectures for object recognition [C] //, Springer; 92-101.

SCHNABEL R, WAHL R, KLEIN RJCGF 2010. Efficient RANSAC for Point-Cloud Shape Detection. 26: 214-226.

SENGUPTA S, VALENTIN J, WARRELL J, et al. 2013. Mesh based semantic modelling for indoor and outdoor scenes [C] //; 6.2.

SHEN Y, HE X, GAO J, et al. 2014. A latent semantic model with convolutional-pooling structure for information retrieval [C] //, ACM; 101-110.

WANG C, YONG KCJAIC 2015. Smart scanning and near real-time 3D surface modeling of dynamic construction equipment from a point cloud. 49: 239-249.

WEINMANN M, JUTZI B, HINZ S, et al. 2015. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. 105: 286-304.

WU CJHWCWEHCV 2011. VisualSFM: A visual structure from motion system.

WU XJCS 2015. Fully convolutional networks for semantic segmentation.

WU Z, SONG S, KHOSLA A, et al. 2015. 3D ShapeNets: A deep representation for volumetric shapes [C] //; 1912-1920.

XIAO J, HAYS J, EHINGER KA, et al. 2010. SUN database: Large-scale scene recognition from abbey to zoo [C] //; 3485-3492.

XIONG X, MUNOZ D, BAGNELL JA, et al. 2011. 3-D scene analysis via sequenced predictions over points and regions [C] //; 2609-2616.

Tang Shengjun 2017. Multi-view image enhancement RGB-D indoor high-precision 3D mapping method [M]. Wuhan University .

Xiong Hanjiang , Zheng Xianwei , Ding Youli , et al. 2018. Semantic segmentation of indoor 3D point cloud model based on 2D-3D semantic transfer . 43: 2303-2309.

April 2020