# SEMANTIC SEGMENTATION OF INDOOR 3D POINT CLOUD WITH SLENET

Youli Ding[1], Xianwei Zheng[1,*], Hanjiang Xiong[1], Yi Zhang[2]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, WuHan University, Hubei, Wuhan - (whu_dyl, zhengxw, xionghanjiang)@whu.edu.cn
[2] School of Mathematics and Statistics, Wuhan University, Hubei, Wuhan - 201492025@qq.com

**KEY WORDS:** Indoor Mapping, 2D-3D Semantic Label Propagation, Semantic Segmentation, 3D Point Cloud

**ABSTRACT:**

With the rapid development of new indoor sensors and acquisition techniques, the amount of indoor three dimensional (3D) point cloud models was significantly increased. However, these massive "blind" point clouds are difficult to satisfy the demand of many location-based indoor applications and GIS analysis. The robust semantic segmentation of 3D point clouds remains a challenge. In this paper, a segmentation with layout estimation network (SLENet)-based 2D-3D semantic transfer method is proposed for robust segmentation of image-based indoor 3D point clouds. Firstly, a SLENet is devised to simultaneously achieve the semantic labels and indoor spatial layout estimation from 2D images. A pixel labeling pool is then constructed to incorporate the visual graphical model to realize the efficient 2D-3D semantic transfer for 3D point clouds, which avoids the time-consuming pixel-wise label transfer and the reprojection error. Finally, a 3D-contextual refinement, which explores the extra-image consistency with 3D constraints is developed to suppress the labeling contradiction caused by multi-superpixel aggregation. The experiments were conducted on an open dataset (NYUDv2 indoor dataset) and a local dataset. In comparison with the state-of-the-art methods in terms of 2D semantic segmentation, SLENet can both learn discriminative enough features for inter-class segmentation while preserving clear boundaries for intra-class segmentation. Based on the excellence of SLENet, the final 3D semantic segmentation tested on the point cloud created from the local image dataset can reach a total accuracy of 89.97%, with the object semantics and indoor structural information both expressed.

## 1. INTRODUCTION

In recent years, the location-based services (LBS) and GIS applications have been extended from outdoor to indoor environments (Zhou et al., 2017), which also induces the rapid development of indoor data acquisition methodologies and sensors. The indoor 3D scenes created by light detection and ranging (LiDAR) surveying, consumer-level RGB-D camera collection, or structure from motion (SFM) technique are now widely available (Dimitrov , Mani, 2015). These 3D scenes are usually represented as semantically unkown 3D point clouds (Hermans et al., 2014). However, for many indoor applications, such as 3D object tracking and retrieval, robot object manipulation, autonomous navigation and augmented reality (Wang et al., 2015), which requires not only the precise representation of the data, but also the semantic describing of the full indoor 3D models. Therefore, the semantic segmentation of indoor 3D point clouds is vital for semantic mapping of indoor scene (Tchapmi et al., 2017).

The semantic 3D point cloud segmentation, which aims to label the cluttered scene into meaningful semantic objects (Lu et al., 2016), has been an active topic touches diverse research fields (Liu et al., 2017). In the previous works, one of the common approach for semantic 3D point cloud modeling is to directly draw on the ideas of 2D image semantic classification, that is, train point-by-point feature classifier based on 3D geometric features (Koppula et al., 2011), then obtain the semantic classification results by a optimization with spatial context constraints (Munoz et al., 2010). With the development of deep learning in 2D image classification, these methods have been progressively extended to 3D point cloud classification, becoming an important approach for 3D point cloud semantic annotation (Boulch et al., 2018) (Su et al., 2015). However, the design of effective 3D features still exists large difficulty, due to the problem of occlusions, data noise, holes and non-uniform distribution of 3D point clouds (Zhou et al., 2009).

In the field of 2D image recognition and segmentation, the construction of large-scale 2D label training datasets has been studied for decades (Deng et al., 2009) (Rakelly et al., 2018) (Xiao et al., 2010) (Russell et al., 2008), forming a large number of standard datasets, such as LabelMe (Russell et al., 2008), ImageNet (Deng et al., 2009), ImageNet-segment (Kuettel et al., 2012), etc. Considering the rich image datasets in 2D side, W. Yan (Wang et al., 2013) proposed an exemplar SVM-based method which propagates the label information from imageNet to 3D point clouds. II fouad (Fouad et al., 2017) also adopted a 2D-3D transfer to realize the 3D point segmentation for indoor scenes from RGB-D images. Currently, deep learning algorithms have bloomed and show impressive performance in different application fields, such as object segmentation (Rakelly et al., 2018), object recognition and 3D scene understanding (Zhao et al., 2017). However, their results are restricted to the segmentation results in 2D images, due to the challenge in unconstrained spatial layouts, large variability of both object and indoor scene types, and illumination variance. Moreover, the spatial consistency and scene context between images did not fully considered in these methods. While, indoor spaces often have strong structural features such as vertical horizontal structures of wall, ceilings, and floors. Therefore, the indoor semantic 3D model not only needs to express object information, but also requires to express the structural information.

To summarize, the direct segmentation of point clouds in 3D space remains very hard, using the 2D-3D transfer to

*Corresponding author

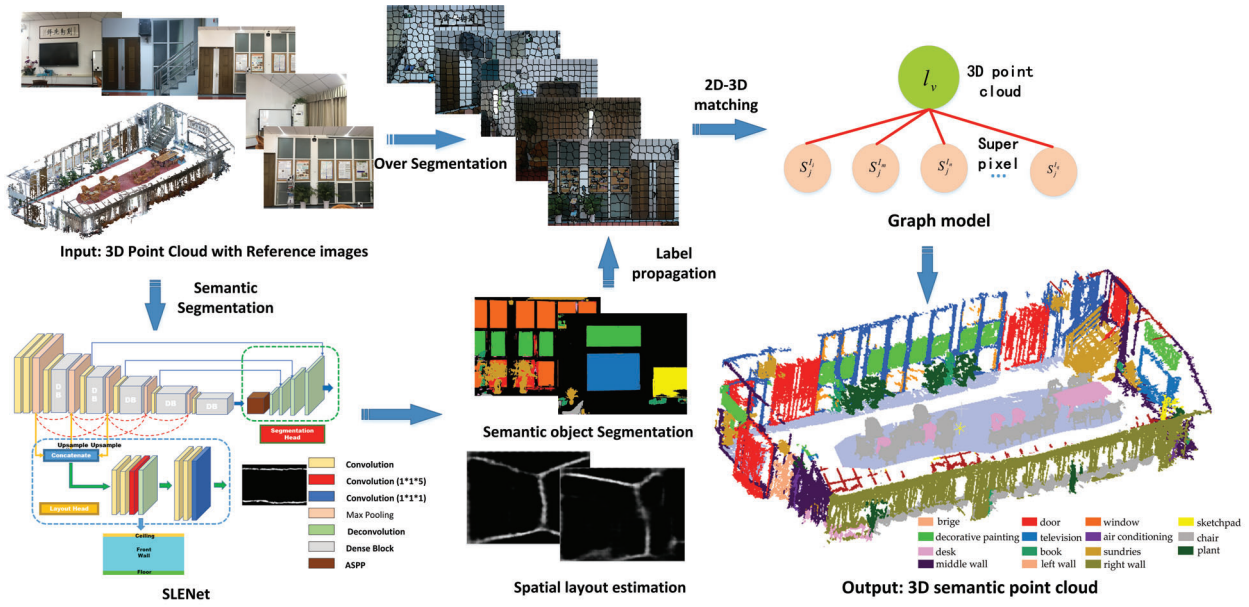



Figure 1. The framework of labeling transfer from 2D images to 3D point clouds

realize the semantic segmentation of indoor 3D point clouds is currently a more practical and efficient way, especially for 3D scenes created from image collections with SfM systems. To achieve this goal, a segmentation with layout estimation network (SLENet)–based 2D-3D semantic transfer method is proposed for robust segmentation of indoor 3D point clouds. In the proposed method, we divided into three separate running processes, namely "2D labeling and spatial layout estimation based on SLENet", "2D-3D labeling Transfer" and "3D contextual refinement" (see Fig. 1). We first devised a SLENet to combine the tasks of indoor semantic segmentation and space layout estimation in an integrated network. Specifically, SLENet can both learn discriminative enough features for inter-class segmentation while preserving clear boundaries for intra-class segmentation. The 2D-3D semantic transfer which incorporates a pixel labeling pool and a graphical visibility model is then utilized to efficiently propagate the 2D labels and spatial information to 3D point clouds. Finally, a label prediction approach is developed to realize the 3D-contextual refinement, which explores the extra-image consistency with 3D constraints to suppress the labeling contradiction caused by multi-superpixel label fusion.

## 2. METHODOLOGY

### 2.1 Segmentation and Spatial Layout Estimation Based On SLENet

Although object semantics and spatial layout are both essential to describe an indoor scene, within many works based on deep learning, none of them tackle these two tasks simultaneously. To this end, this paper proposes a new network, called SLENet (Segmentation with Layout Estimation Network), to combine indoor semantic segmentation with space layout estimation. This new network involves three components: the Backbone Network, the Segmentation Head and the Layout Head, as Figure 2 illustrates.

The Backbone Network can be divided into six stages according to the size of feature maps, and it is worth noting that low-stage

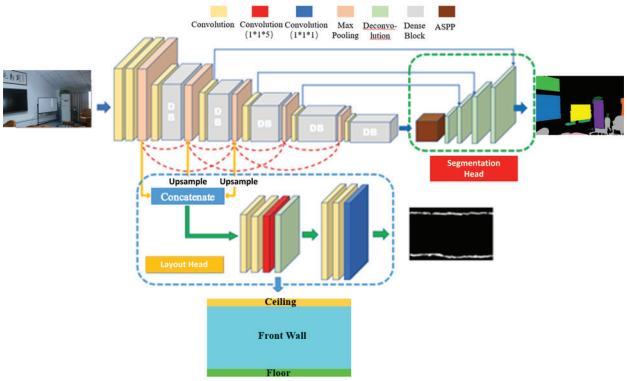



Figure 2. The architecture of SLENet

features contain more spatial information while the higher ones contain more semantic information. Therefore, we introduce a new architecture called Large-scale Residual Connection, as the dotted lines in Figure 2 and Figure 3(a) show, to transmit spatial information to high stages for layout estimation and semantic segmentation. Unlike residual structure in ResNet (He et al., 2016) that only connects adjacent layers, large-scale residual connection connects different stages rather than layers in a dense manner, in which case it can utilize special features from previous stages to a more considerable extent while preserving the ability of residual structure that facilitates gradient backpropagation and prevents vanishing gradient problem. Moreover, from the second stage, each flowing stage in the Backbone Network contains a solid block(Figure 3(b)), which is proposed in DenseNet (Li , Vu, 2018), to improve information flow between layers and fully exploit features extracted previously. Semantic Segmentation can be regarded as a pixel-level classification that includes object classification and localization, so the design of the Segmentation Head should take these two tasks into account. To fully utilize high-level semantics for classification, the Segmentation Head is attached to the top of the backbone network, and adopts the Atrous Spatial Pyramid Pooling(ASPP) (Chen et al.,

2018) of Deeplab v3 to extract and merge multi-scale features(Figure 3(c)). Also, segmentation head also leverages the encoder-decoder architecture to recover high-resolution result from low-resolution feature maps. However, different from Deeplab v3, we add the correspondent features from previous stages into every deconvolution stage by concatenation instead of recovering directly from the features linearly produced by the encoder, which ignores the spatial information contained in previous layers. Therefore, while the output of the backbone network provides semantic features for object classification, the extra spatial information in the decoding process promotes more accurate localization prediction.

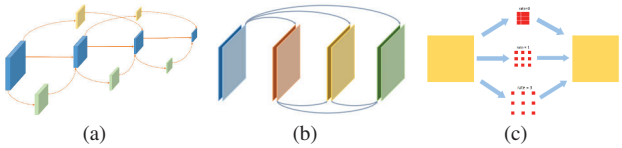



Figure 3. Illustrations of Large-scale Residual Connection, Dense Block and ASPP. (a) Large-scale Residual Connection. The blue block means a stage in the backbone network; others refer to convolutional layers. (b) The structure of Dense Block. A square here refers to a convolutional layer. (c) Atrous Spatial Pyramid.

As for the layout estimation, it should be noted that layout estimation can be formulated as dividing indoor space into five categories: left wall, middle wall, right wall, floor and ceiling, where spatial features play a much more important role than the semantic ones. This observation guides us to attach the layout head to lower layers to utilize the spatial information directly. Since different stages produce multi-size feature maps, we upsample smaller maps to fuse these multi-level features. After the fusing block, the fused features will go through a specifical design block to yield surface semantic labels, and the box layout (Figure 2). The box layout then acts as a rough estimation, combined with the extracted lines and vanishing points of the indoor scene to get a series of candidate spatial layouts. Finally, compared with the ground-truth, we choose the layout with the highest score as the final result.

## 2.2 2D-3D Labeling Transfer

In this section, we describe how to achieve the label transfer from 2D images to 3D point clouds. We first oversegment the 2D images into superpixels and assign the corresponding semantic label to each superpixel according to the segmentation result from SLENet. The label propagation can then be transmitted in units of superpixels, to avoid the time-consuming pairwise 2D-3D semantic transfer and reinforce the robustness to mismatching error occurring in 3D reconstruction from images. The specific 2D-3D transfer is formulated as a visibility graph model, which is constructed based on the view relationship between the camera and the 3D point in the SfM system.

**2.2.1 Pixel Label Pool Construction** Assuming that the 3D point cloud model is defined as $P = \{p_i\}$, each point is described by 3D coordinates and RGB colors $\{x_i, y_i, z_i, R_i, G_i, B_i\}$. The point cloud model is created with the hierarchical SfM-PMVS algorithm with R real images. Since the spatial 3D point is inversely calculated from the feature point matching sequence in multiple images during the 3D reconstruction of SfM, the mapping relationship between the 2D feature points and the corresponding 3D point can be established. However, SfM can only get sparse 3D point clouds, while the dense 3D point clouds are obtained through the patch-based region growing algorithm (PMVS). To semantically label the dense point cloud constructed from SfM-PMVS, we first use the Simple Line Interface Method (SLIC) method (Noh , Woodward, 1976) to segment an image in units of superpixels. Then we construct the region block of the feature points so that the category of a superpixel is represented by the categories of the feature points in it. Through the visibility model built from the SfM system, the superpixels can be further back-projected into the 3D space, and the 2D-3D label propagation for dense point clouds can finally be realized with the superpixel label pool.

The pixel label pool is a collection of semantic labels of all superpixels in a 2D image, which can be expressed as $S = \{S_i, l_{s_i}\}$ , where $i \in N$ indicates the superpixel number (N is the number of superpixels),$S_i$ represents the $ith$ superpixel, $l_{S_i} \in L$ signifies the semantic label value of the superpixel $S_i$. The specific construction process of the semantic label pool is shown in Figure 4. And the label of each 2D superpixel is directly delivered by the segmentation result of SLENet. The goal of 2D-3D semantic transfer is to assign a corresponding semantic label to each point in the point cloud according to the pixel label in the superpixel pool. Using the units of superpixels to realize 2D-3D label transfer is a robust solution for 3D dense point cloud segmentation, which can also avoid the reprojection error by establishing a superpixel buffer for 2D feature points and the corresponding 3D point.

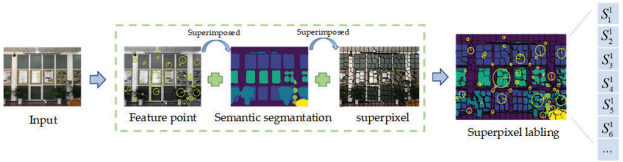
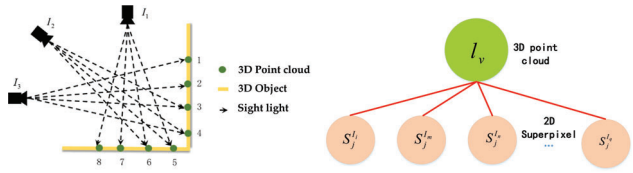



Figure 4. The process of superpixel label pool construction

**2.2.2 Graph Model Construction** After getting the semantic label of the 2D image in units of superpixels, we need to establish a bridge between 3D point cloud and the corresponding 2D superpixels to realize the semantic transfer from the 2D superpixels to 3D points. Here, we define 3D point cloud $P = \{p_i\}$ and superpixels $S = \{S_i\}$ as nodes $V = \{p_i\} \cup \{S_i\}$, and define the links of superpixels and 3D point as edges $\varepsilon$. Thus, a graph model $G = \{v, \varepsilon\}$ can be constructed based on Markov Random Field (MRF) (Li, 1994). According to the definition of the MRF graph model, each node is only associated with itself and its neighbors, which is independent of other nodes. Since the label of each 2D superpixel is delivered by the segmentation result of SLENet, there is no need to bridge edges between 2D superpixel nodes, which is different from the work of Wang (Wang et al., 2013). Hence, we only need to consider the relationship between every 3D point and its associated 2D superpixels. Recall that the projection relationship between 3D point cloud and superpixels is constructed with the visibility model created from the SfM system. The visibility model connects the camera to the point cloud and can tell that a 3D point is reconstructed from which 2D images. In other words, a 3D point node is connected with

multiple 2D superpixel nodes via the SfM-encoded visibility model. Thus its semantic label distribution can be determined by all connected superpixel labels. As shown in Figure 5(a), it can be seen that point 3 is visible from camera 2 and camera 3, the visible images are therefore imaged 2 and image 3. This means the semantic label distribution of point 3 is determined by the semantic of two superpixels (one is in image 2, and the other is in image 3). Based on the visibility model, a 3D point can be projected to the corresponding image space with the transform matrices (i.e., camera parameter), thus connected with its associated superpixels. The edges between a 3D point and its associated superpixel nodes are then created as the red links drawn in Figure 5(b). Finally, the desired graph model is constructed.




(a) (b)

Figure 5. Construction of the graph model from the visibility model. (a) This graph shows the projection relationship between cameras and 3D point cloud in the SfM-encoded visibility model. (b) is the construction of the graph model, which express the corresponding relationship from 2D to 3D, where $I_*$represent the $*th$ image, $S_*^{I_*}$ indicate the superpixel in the image $I_*$.

### 2.3 3D Contextual Refinement

Given the semantic superpixels and the graph model, we can fuse the 2D superpixels in multiple images to achieve the propagation of semantic labels from 2D superpixels to 3D points. Since for a 3D point $p_*$, the 2D superpixels in set $\{S_i \mid (S_i, p_i) \in \varepsilon\}$ may have different label categories, the label distribution of the point can then be defined by the occurrence probability of every label. We, therefore, formulate the problem of assigning a 3D point with the correct label as a multi-category labeling problem. The potential energy function is constructed as the following equation :

$$E = \sum_{p_i \in P} \Psi_d(l(p_i)) \tag{1}$$

In this quation, E is the dependent variable of the potential energy function, $p_i$ is the set of 3D points, $l(p_i)$ indicates the semantic label of 3D point $p_i$, and $\Psi_d$ denotes a data term. Since the superpixels associated with the same 3D point can define a distribution function for the point, label propagation can be conceived as labeling the 3D point with the label with maximum occurrence probability. Therefore, if the label distribution function of the 3D point is $P_{p_*}$, the data term definition can be defined as:

$$\Psi_d(l(p_i)) = -P_{p_i}(l(p_i)) \tag{2}$$

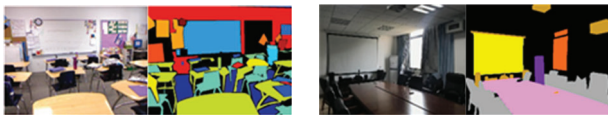
Finally, setting L as the semantic annotations of the 3D point cloud, we can then minimize the potential energy function E by graph cut algorithm to get the label mapping $l(.)$:

$$l(.) = \arg\min_{l \in L} E = \arg\min_{l \in L} \sum_{p_i \in P} \Psi_d(l(p_i)) \tag{3}$$

## 3. EXPERIMENTS

### 3.1 Experimental setup

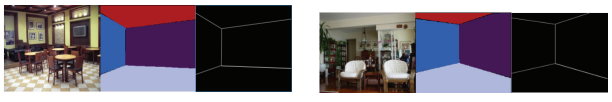
**3.1.1 Dataset** For the semantic segmentation training, we use the NYUDv2 RGBD indoor dataset (Silberman et al., 2012) and a local dataset collected from a large meeting room in LIESMARS (State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing) of Wuhan University. From NYUDv2 RGBD, we select 769 images with 30 classes where 600 for training, 100 for validation and 69 for the test. Compared to scenes in NYUDv2 RGBD, the 3D scene we aim to reconstruct contains some objects that possess some unique features, e.g., the cane chair to be reconstructed later is different from any chair in NYUDv2 RGBD. Therefore, we construct a local dataset including such objects to fine-tune the net. The additional dataset is composed of 261 images with 39 densely annotated. The additional dataset includes 12 categories, such as a door, window, sketchpad, decorative painting, television, air conditioning, chair, desk, book, bridge, plant, and sundries, which are related to the indoor semantics to be reconstructed later.



(a) (b)

Figure 6. Example of semantic segmentation training dataset for SLENet.

For the layout estimation, we train the layout head on the dataset published by Hedau et al. (Hedau , Hoiem, 2009), which consists of 313 images, and test it on the local dataset mentioned in segmentation training. The spatial layout can be described by a 3D box or a surface layout. As shown in Figure 7, a box layout separates the ceiling, the wall and the ground by lines extracted from corresponding borders, and a surface layout segments the whole scene into five classes: the ceiling, the middle wall, the left wall, the right wall and the ground, in which all occlusions in the interior space are ignored.



(a) (b)

Figure 7. Example of spatial layout training dataset.

The 261 images collected from LIESMARS (local dataset) were used to create the 3D scene for the final semantic segmentation test of the indoor 3D point cloud. As shown in figure 8, the test 3D scene is a large-scale complete indoor scene.

**3.1.2 Multi-stage Training Strategy**

1. As we hope the backbone network can extract rich and discriminative features, it is natural to train the network for

(a) (b)

Figure 8. Dataset of meeting room.(a) shows the RGB images we collected;(b) shows the corresponding 3D point cloud.

segmentation first since it is more complicated than layout estimation.

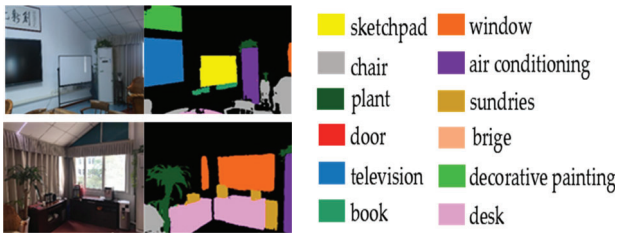
2. In the first stage, only the backbone network and segmentation head are trained on the NYUDv2 RGBD indoor dataset and then are fine-tuned on the additional collected images.
3. In the second stage, only the head for layout estimation will be trained. In detail, we jointly train the layout structure lines and semantic surface labels to alleviate the issue caused by occluded boundaries in the cluttered room. The datasets used here are the same as the ones used in the first stage.
4. In the third stage, all layers in the whole network are jointly optimized on the same two datasets.

### 3.2 Results and Analysis

**3.2.1 Segmentation and Layout Estimation** Since SLENet adopts some structures from Deeplab v3+ and DenseNet, e.g., dense block and atrous spatial pyramid pooling, we compare the results of Deeplab v3+, DenseNet121, and SLENet on the publicly available NYUDv2 RGBD dataset to validate the advantage of SLENet. We implement the three models in Keras with Tensorflow as a backend, and employ Adam optimizer with 1e-4 as initial learning rate. The models are trained for 10K iterations on one NVIDIA GTX 1080Ti.

Figure 9 shows the segmentation results of some sample images. From the results in the first and second columns, it is evident that SLENet can balance better between preserving structure details and learning discriminative features than DenseNet121 and Deeplab v3+. More specifically, although DenseNet121 and Deeplab v3+ respectively perform well in terms of preserving structure details, e.g. object edge, and learning different categories, DenseNet121 tends to mistake objects sharing similar features (the light green refers to shelf while the red denotes bookshelf), and Deeplab v3+ is more inclined to over smooth edges and even obscure boundaries between objects belonging to the same class. Different from the two nets, SLENet retains sufficient structure details by the spatial information from large-scale residual connection and solid blocks, and succeeds in learning discriminative features with the multi-scale features integrated by ASPP, in which case, SLENet circumvents the shortcomings found in DenseNet121 and Deeplab v3+.

Figure 10 shows the segmentation results of SLENet after fine-tuning on the local dataset. The segmentation accuracy on local dataset reaches 96%. Figure 11 shows the results of layout estimation. The pictures with black background are the rough outputs of SLENet, indicating that SLENet successfully acquires the ability to estimate the spatial layout of an indoor scene, and the red boxes in images denote the final layout refined from the estimation of SLENet.




(a) (b)

Figure 10. The semantic segmentation result with our proposed SLENet. (a) shows the input image and semantic segmentation results of our network. (b) shows the segmentation legend of SLENet.

**3.2.2 3D point cloud Segmentation** Within the extracted semantic segmentation and layout estimation results, we then oversegment each of the 261 images used for point cloud reconstruction into superpixels and assign the semantic labels to each superpixel to form the superpixel label pool. In this work, we employ the SLIC algorithm to perform the superpixel oversegmentation, and the region size is set to 10, while the regular term coefficient is set to 1.

A quantitative evaluation of the point cloud segmentation result was also made as a reference. The manually labeled validation semantic point cloud was used as the reference data. The total number of testing 3D points is 13917738, while the corrected labeled points are 12521788. Thus, the average segmentation accuracy for the entire point cloud is 89.97%.

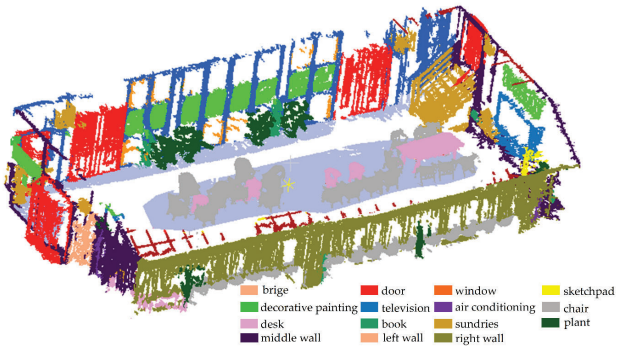



Figure 12. The semantic segmentation result of the test 3D point cloud.

## 4. CONCLUSION

In this paper, we presented a 2D-3D semantic transfer method for robust segmentation of indoor 3D point clouds to tackle the difficulties of lack of robust 3D features and adequate indoor 3D training data in the point cloud semantic segmentation process. In order to semantic modeling of the indoor scene with both individual object and spatial structure information, a SLENet is devised to simultaneously tackle these two tasks. Except for the combination of semantic segmentation with layout estimation, the proposed SLENet also adopt some network structure for feature extraction and fusion to form a new architecture, which can robustly solve the notorious problem, such as uneven illumination, complicated texture, and high-occluded situations. In comparison with the current popular DenseNet

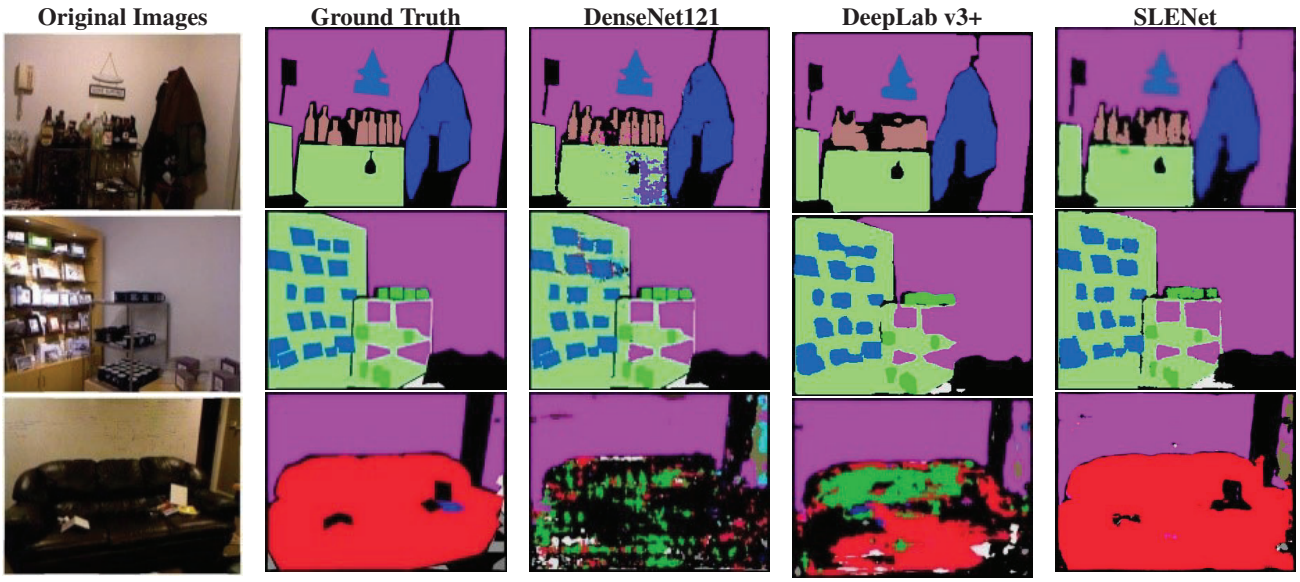



Figure 9. Qualitative comparisons among DenseNet121 model, Deeplab v3 model and our SLENet model on NYUDv2 RGBD dataset.

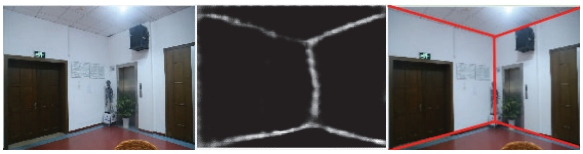



Figure 11. Indoor spatial layout estimation results. (a) shows the input images of SLENet. (b) shows the rough spatial layout estimation results with our Network. (c) shows the detail spatial layout extraction results.

and DeepLab v3+, the experimental results in both benchmark dataset and local dataset also showed the superiority of the proposed SLENet. Since the semantic labels of 3D point cloud are transferred from 2D images in our solution, the excellence of SLENet thus guarantees the accuracy of the final 3D point cloud segmentation.

Moreover, the final testing on the 3D scene created with local image dataset further verified the effectiveness of the proposed method. It is worth noting that the proposed 2D-3D semantic label propagation method is mainly suitable for the SfM modeling point cloud, to some extent, our method has some limitation in the processing of some other segmentation situations. As we know, the 3D geometry features play an essential role in the 3D semantic annotation process. Therefore, we will explore the effective combination of the extracted 2D annotation and 3D geometry features in the future work, leading to better label assignment of the 3D point cloud model.

## ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China under Grants 41871361 and 41701445.

## REFERENCES

Boulch, A., Guerry, J., Bertrand, L., Audebert, N., 2018. SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. *Computers & Graphics*, 71, 189–198.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A., 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40, 834–848.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F., 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 248–255.

Dimitrov, A., Mani, G., 2015. Segmentation of building point cloud models including detailed architectural/structural features and MEP systems. *Automation in Construction*, 51, 32-45.

Fouad, I, Rady, S., Mostafa, M., 2017. Efficient image segmentation of rgb-d images. 353–358.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hedau, V., Hoiem, D.and Forsyth, D., 2009. Recovering the spatial layout of cluttered rooms. *2009 IEEE 12th international conference on computer vision*, IEEE, 1849–1856.

Hermans, A., Floros, G., Leibe, B., 2014. Dense 3d semantic mapping of indoor scenes from rgb-d images. *2014 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2631–2638.

Koppula, H., Anand, A., Joachims, T., Saxena, A., 2011. Semantic labeling of 3d point clouds for indoor scenes. *Advances in neural information processing systems*, 244–252.

Kuettel, D., Guillaumin, M., Ferrari, V., 2012. Segmentation propagation in imagenet. *European Conference on Computer Vision*, Springer, 459–473.

Li, C., Vu, N., 2018. Densely connected convolutional networks for speech recognition. *Speech Communication; 13th ITG-Symposium*, VDE, 1–5.

Li, S., 1994. Markov random field models in computer vision. *European conference on computer vision*, Springer, 361–370.

Liu, M., Guo, Y., Wang, J., 2017. Indoor scene modeling from a single image using normal inference and edge features. *The Visual Computer*, 33, 1227–1240.

Lu, X., Yao, J., Tu, J., Li, K., Li, L., Liu, Y., 2016. PAIRWISE LINKAGE FOR POINT CLOUD SEGMENTATION. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 3.

Munoz, D., Bagnell, J., Hebert, M., 2010. Stacked hierarchical labeling. *European Conference on Computer Vision*, Springer, 57–70.

Noh, W., Woodward, P., 1976. Slic (simple line interface calculation). *Proceedings of the fifth international conference on numerical methods in fluid dynamics June 28–July 2, 1976 Twente University, Enschede*, Springer, 330–340.

Rakelly, K., Shelhamer, E., Darrell, T., Efros, A., Levine, S., 2018. Few-Shot Segmentation Propagation with Guided Networks. *arXiv preprint arXiv:1806.07373*.

Russell, B., Torralba, A., Murphy, K., Freeman, W., 2008. LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, 77, 157–173.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from rgbd images. *European conference on computer vision*, Springer, 746–760.

Su, H., Maji, S., Kalogerakis, E., Erik, L., 2015. Multi-view convolutional neural networks for 3d shape recognition. *Proceedings of the IEEE international conference on computer vision*, 945–953.

Tchapmi, L., Choy, C., Armeni, I., Gwak, J., Savarese, S., 2017. Segcloud: Semantic segmentation of 3d point clouds. *2017 International Conference on 3D Vision (3DV)*, IEEE, 537–547.

Wang, Y., Ji, R., Chang, S., 2013. Label propagation from imagenet to 3d point clouds. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3135–3142.

Wang, Z., Zhang, L., Fang, T., Mathiopoulos, P., Tong, X., Qu, H., Xiao, Z., Li, F., Chen, D., 2015. A multiscale and hierarchical feature extraction method for terrestrial laser scanning point cloud classification. *IEEE Transactions on Geoscience and Remote Sensing*, 53, 2409–2425.

Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A., 2010. Sun database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 3485–3492.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zhou, H., Yuan, Y., Shi, C., 2009. Object tracking using SIFT features and mean shift. *Computer vision and image understanding*, 113, 345–352.

Zhou, Y., Zheng, X., Xiong, H., Chen, R., 2017. Robust Indoor Mobile Localization with a Semantic Augmented Route Network Graph. *ISPRS International Journal of Geo-Information*, 6, 221.