

Envisioning the Unseen: Revolutionizing Indoor Spaces with Deep Learning-Enhanced 3D Semantic Segmentation

MUHAMMAD ARIF
Wuhan University, China
arifmuhmand@gmail.com

AINAZ JAMSHIDI
University of Maryland, Baltimore County
ainzj1@ada.rs.umbc.edu

Abstract

In recent years, advancements in indoor sensor technology and 3D model acquisition methods have led to a significant increase in the volume of indoor three-dimensional (3D) point cloud models. However, these extensive, "blind" point clouds present substantial challenges for advanced indoor applications and GIS analysis due to their lack of semantic segmentation. Addressing this demand, our study explores the spatial dimensions of semantic segmentation through the application of convolutional neural networks (CNNs). Utilizing a structured dataset, we aim to predict point cloud characteristics and assess the accuracy of machine learning models. The dataset is divided into training and testing segments, both subjected to an extensive training process. The outcomes are then compared with current state-of-the-art benchmarks, and visualized to demonstrate our model's efficacy. Historically, the segmentation of 3D point clouds has been hindered by the absence of robust 3D features, limited 3D training data, and the complexities inherent to indoor environments, such as high occlusion, uneven lighting, and diverse objects. To overcome these challenges, we propose an effective algorithm for transferring semantic labels from 2D semantic images to raw 3D point clouds. This method establishes a foundation for 3D semantic point cloud models that effectively resolve the issues related to object semantics and unclear spatial structures. Leveraging the SfM point cloud model's attributes and utilizing extensive 2D image databases, our algorithm estimates the structural layout and semantic labels of images, transferring this information as semantic labels to 3D point cloud data. This approach simplifies the extraction of structure and semantics from 3D point clouds. To demonstrate the algorithm's performance in complex indoor settings, we introduce a new architecture, the Large-scale Residual Connection, which transmits spatial information from lower to higher levels. Additionally, we incorporate the Atrous Spatial Pyramid Pooling (ASPP) of

DeepLabv3+ and the DenseBlock structure of DenseNet, along with a multi-stage training strategy to address the challenges posed by indoor environments' occlusions and complexity. Our methodology, aimed at robust semantic segmentation of indoor 3D point clouds, comprises two major innovations. First, a novel Combined Network is designed to label 2D images and estimate indoor spatial layouts, enhancing classification capabilities. Second, we implement a 2D-3D label propagation based on a graphical model, facilitating label transfer from 2D to 3D and constructing contextual consistency between images. Notably, our approach does not require any 3D scene training data, yet it achieves remarkable segmentation results in complex indoor scenes with an accuracy of 87%. Our experiments, conducted on the public NYUDv2 indoor dataset and a proprietary local dataset, demonstrate that compared to leading-edge techniques in 2D semantic segmentation, DeepLabv3+ adeptly learns discriminative features for inter-class segmentation while preserving clear boundaries for intra-class distinctions.

1. Introduction

1.1. Research Background and Significance

The migration of GIS applications and location-based services to indoor environments, fueled by advances in sensor technology and data acquisition methods, marks a pivotal shift in GIS research. Technologies like laser radar scanning and RGB-D cameras now enable the creation of extensive indoor 3D point cloud data, setting new benchmarks for indoor modeling precision. This evolution underscores the necessity for models that go beyond geometric accuracy, catering to the nuanced needs of augmented reality and autonomous navigation by offering detailed semantic understanding. The inadequacy of traditional "blind" point clouds, lacking in actionable contextual information, significantly restricts their utility in higher-level intelligent applications, highlighting the

need for semantic enrichment [1, 2, 3, 4]

This paradigm shift towards semantic-aware models seeks to transform mere geometric reconstructions into richly annotated datasets that are meaningful for a wide array of applications. The inherent complexity of indoor environments, characterized by occlusions, variable lighting, and textural diversity, poses unique challenges for 3D scene semantic segmentation. Addressing these challenges, the study leverages Structure from Motion (SfM) alongside dense reconstruction technologies, offering a scalable and cost-effective methodology for developing detailed indoor 3D models [5, 6].

1.2. Research Questions and Objectives

At the heart of this study lies the exploration of semantic segmentation for indoor 3D point clouds, a critical component in realizing intelligent, location-oriented services and GIS systems within indoor spaces. The research interrogates the application of deep learning for enhancing the segmentation of 2D images and, by extension, its potential to revolutionize the segmentation of 3D point clouds. The investigation encompasses three core areas: overcoming indoor segmentation challenges to refine 2D image segmentation; bridging the gap between 2D image semantics and 3D point cloud segmentation; and establishing a framework for semantic consistency across dimensions, ensuring the accuracy and integrity of the semantic segmentation process.

1.3. State of the Art

The semantic segmentation of indoor 3D point clouds is gaining traction, driven by breakthroughs in sensing and 3D modeling. This burgeoning research domain traditionally focused on extracting point cloud features directly—a method now supplemented by the advent of deep learning algorithms. Despite such advancements, the scarcity of 3D training datasets remains a significant hurdle. The recent advent of 2D-3D semantic transfer methodologies promises a novel direction, leveraging the robust framework of 2D image segmentation techniques against the complex backdrop of 3D point cloud segmentation. This synthesis of cross-domain methodologies underscores a transformative approach to semantic annotation, facilitating richer, more actionable 3D models.

1.4. Research Domain

The complexity of indoor environments demands a nuanced approach to semantic segmentation, one that considers the intricate interplay of spatial layout and object semantics. This study posits the Deeplabv3+ neural network model as a solution, aiming to meld spatial layout estimation with object segmentation to foster a

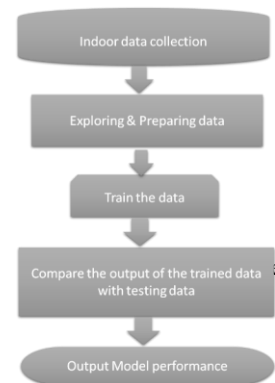
comprehensive understanding of indoor scenes. This dual approach not only enriches the semantic segmentation process but also paves the way for intelligent indoor space management, enhancing the utility and application potential of indoor 3D models.

1.5. Methodology

Leveraging Python's robust ecosystem for deep learning, this research adopts a methodical approach to neural network development, emphasizing data collection, preprocessing, model training, and evaluation. This detailed methodology underpins the model's capacity to navigate the complexities of indoor 3D semantic segmentation, offering insights into both the challenges and solutions inherent in this innovative research domain.

1.6. Technical Approach

The model's development follows a structured approach, from initial data acquisition through to comprehensive model evaluation. The detailed exploration of model architecture, optimized through rigorous training,



illustrates the study's commitment to addressing the outlined research questions. This meticulous process not only underscores the study's innovative contributions but also highlights the potential of deep learning in revolutionizing the semantic segmentation of indoor 3D point clouds.

2. Methodology

In our research, we employ an integrative approach leveraging the synergy of Structure from Motion (SfM) algorithms and Convolutional Neural Networks (CNNs) to advance the field of indoor 3D modeling through precise

semantic segmentation. This methodology encompasses a series of meticulously designed experimental and analytical steps, underpinned by robust data collection and processing techniques, to address the challenges posed by indoor environments' complexity.

2.1. Feature Point Extraction and Matching

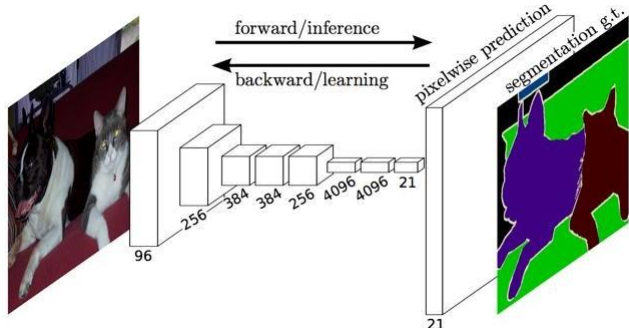
At the core of our 3D reconstruction process is the SfM algorithm, which plays a crucial role in determining the spatial arrangement of cameras and the subsequent modeling of 3D scenes. A critical step in this procedure involves the identification and matching of feature points across images. For this purpose, we employ the Scale-Invariant Feature Transform (SIFT) operator, renowned for its effectiveness in maintaining stable feature points across varying conditions. The process involves two main phases: the detection of keypoints via a multi-scale image analysis facilitated by the Gaussian blur algorithm, and the generation of a 128-dimensional descriptor for each keypoint. This approach ensures robust feature extraction, capable of withstanding the intricacies of indoor settings marked by obstacles, changing illumination, and diverse objects.

2.2. Dual-View Geometry and Bundle Adjustment

To enhance the accuracy of 3D spatial reconstructions, our study delves into dual-view geometry, examining the geometric constraints that emerge when observing a scene from two distinct perspectives. This examination aids in the consistent representation of space and feature matching across images. Further precision in the 3D modeling process is achieved through bundle adjustment, a method that refines the camera projection matrix by minimizing discrepancies between projected and actual image points. This sophisticated calibration method significantly enhances the fidelity of the 3D point estimations.

2.3. Convolutional Neural Networks (CNNs)

Central to our semantic segmentation process are CNNs, which extract features from input images through convolutional layers, reducing dimensionality while

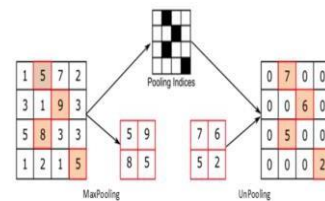


preserving critical information via pooling layers. Our research notably advances this area by implementing Full Convolutional Neural Networks (FCNs), which excel in semantic segmentation. FCNs transform the feature map matrix into a fully stacked vector, crucial for model creation. The resultant feature vector feeds into activation

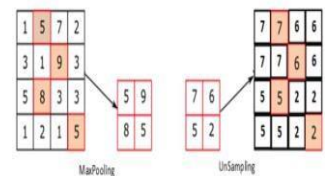
Figure2: Principle of image-by-pixel semantic segmentation of full convolutional neural networks

functions—softmax for multi-classification and sigmoid for binary classification—to classify outputs. FCNs, an advancement over CNNs, excel in semantic segmentation by enabling pixel-level image segmentation results for images of any size. Through deconvolution layers, FCNs upsample to match the original image size for pixel-wise semantic predictions, efficiently performing semantic segmentation on the feature map (Figure 2).

Upsampling can be achieved via unpooling, unsampling, or transpose convolution. Unpooling redistributes local maximum values to a higher resolution feature map, maintaining position information, while unsampling directly replicates local features for expansion. Deconvolution, mimicking convolution, is a popular upscaling method, offering a detailed approach to feature map expansion (Figure 3).



(a) Unpooling (Maxpooling)



(b) Unsampling

Figure3: Two ways to expand the feature map

FCNs allow for pixel-level segmentation across images of any size, employing deconvolution layers for upsampling and achieving high-resolution semantic predictions. This model's ability to learn discriminative features and maintain clear boundaries between different

classes is particularly effective for indoor scenes.

2.4. Innovative Architectural Elements

Our approach incorporates state-of-the-art architectural innovations, such as the Atrous Spatial Pyramid Pooling (ASPP) from DeepLabv3+ and the DenseBlock structure from DenseNet, alongside a Large-scale Residual Connection mechanism. These elements are instrumental in transmitting spatial information efficiently across network layers, enhancing the classification and segmentation capabilities of our model.

2.5. Multi-Stage Training Strategy

Addressing the complexities of indoor environments requires a tailored training strategy. Our method involves an initial focus on complex semantic segmentation, leveraging the multi-level connectivity of the backbone network for nuanced feature extraction. This process is followed by fine-tuning with targeted scene data, ensuring the model's adaptability to specific indoor settings. In summary, our research methodology represents a comprehensive fusion of advanced algorithms and neural network architectures, aimed at overcoming the inherent challenges of indoor 3D model semantic segmentation. By bridging the gap between 2D image semantics and 3D spatial modeling, our work lays a solid foundation for future advancements in indoor mapping, navigation, and intelligent space management systems.

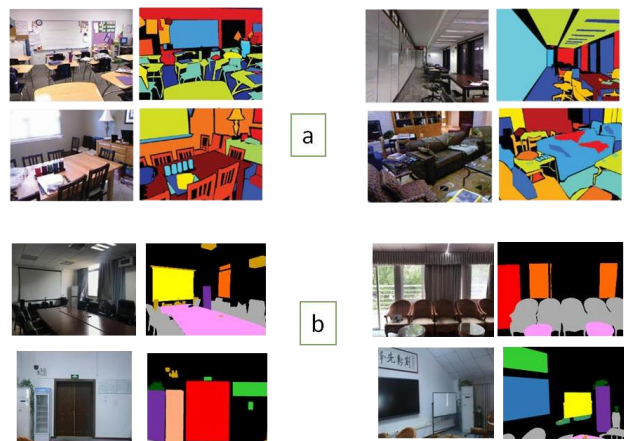
3. Experiment and Analysis

Our research into indoor 3D model semantic segmentation underscores the transformative impact of Full Convolutional Neural Networks (FCNs), demonstrating their superiority over traditional Convolutional Neural Networks (CNNs) in achieving precise, pixel-level segmentation across various image sizes. Key findings reveal the remarkable efficacy of FCNs in handling complex segmentation tasks, notably through advanced upsampling techniques such as unpooling, unsampling, and transpose convolution, which significantly enhance the quality and resolution of semantic segmentation. These techniques ensure the retention of essential spatial information, facilitating detailed and accurate analysis of indoor scenes. The integration of FCNs marks a significant advancement in semantic segmentation performance, offering unparalleled precision and efficiency in parsing complex indoor environments. This breakthrough paves the way for the development of sophisticated applications in indoor mapping, navigation, and intelligent

environmental interaction, establishing FCNs as a pivotal technology in the semantic analysis of indoor 3D models.

3.1. Experimental Data

The evaluation of the Deeplabv3+ neural network for indoor object semantic segmentation and spatial layout estimation was conducted using the NYUDv2 RGBD dataset. This dataset consists of 47 camera-captured images of a conference room setting. For this study, 49 images encompassing 30 types of indoor objects were selected: 49 for training, 100 for validation, and the remaining 47 for testing. To address the presence of objects with unique shape features in our target scenes, additional data comprising 261 conference room images were collected, with 39 marked using LabelMe for fine-tuning the network. These images augment the original dataset, ensuring comprehensive training coverage (see Figure 4).



Additionally, the RoomNet dataset consisting of 313 labeled images, was utilized for training the spatial layout estimation sub-network [7, 8]. This choice is predicated on the critical role of 3D spatial layout in understanding indoor scenes, marked by the regular horizontal and



vertical structures commonly found indoors (see Figure 5).

3.2. Network Stratification Training Strategy

The Deeplabv3+ network employs a layered training strategy to tackle both indoor spatial layout estimation and object semantic segmentation:

1. Initial training focuses on complex semantic segmentation leveraging the multi-level connectivity of the backbone network for feature extraction.
2. The first phase involves training on the NYUDv2 RGBD dataset, followed by fine-tuning with targeted scene data.
3. Subsequently, only the spatial layout estimation sub-network undergoes training.
4. The final phase integrates and optimizes data from the previous phases across all network layers.

3.3. Experimental Results and Analysis

Deeplabv3+ demonstrated notable improvement in semantic segmentation after fine-tuning with targeted scene datasets, as illustrated in Figures 3.10 and 3.11. The network accurately estimates indoor space layouts, evident from the segmentation and layout estimation results, showcasing Deeplabv3+'s advanced capability in accurately parsing indoor scenes.

Comparative analysis with DeepLabv3+ and DenseNet121 on the NYUDv2 RGBD dataset validates Deeplabv3+'s superior performance. Utilizing TensorFlow and Keras with an Adam optimizer, experiments across network models were iterated 10K times on an NVIDIA GTX 1080Ti. Deeplabv3+ outperforms in balancing object detail retention and identifiable features, surpassing DenseNet121 and DeepLabv3+ in edge feature processing and object characteristic differentiation.

Despite occasional segmentation anomalies attributed to similar texture features or significant shape variations in training versus test datasets, Deeplabv3+ demonstrates robust segmentation capabilities. This robustness, coupled with its innovative feature fusion structure and comprehensive training strategy, underscores Deeplabv3+'s effectiveness and reliability in indoor semantic segmentation tasks (see Figure 6).

Original image			
Tag true value			

Figure 4: 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.



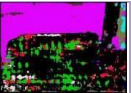


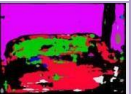



DenseNet 121			
SLENet			
Deeplabv 3+			

Figure 6: DenseNet121 model, Deeplabv3+ model and the semantic segmentation test

Results of the Deeplabv3+ model proposed in this paper on NYUDv2 RGBD indoor dataset

4. Conclusion

4.1. Synopsis of Research Achievements

In an era where the precision of indoor location services is increasingly critical, the transition from conventional 2D mapping to more immersive 3D point cloud representations has unveiled new dimensions in spatial data's utility and application. This transformation, however, is not without its challenges. The initial acquisition of 3D point clouds, often through sophisticated methodologies like laser scanning, RGBD imaging, and photogrammetry, yields data sets that are rich in detail but devoid of semantic context—rendering them unsuitable for direct applications in navigation, virtual reality, and intelligent object interaction. Addressing this gap, our research leverages the advancements in 2D image semantic segmentation, employing deep learning techniques to imbue 3D point clouds with meaningful annotations. By synthesizing the strengths of 2D semantic segmentation with the structural insights afforded by 3D reconstruction techniques, we introduce a novel framework for the semantic enrichment of indoor 3D spaces. This work is anchored on three pivotal innovations:

- The deployment of the Deeplabv3+ neural network, engineered to navigate the intricate semantics of indoor environments. This architecture not only excels in recognizing and delineating indoor spatial layouts and object semantics but also incorporates a hierarchical training approach, large-scale residual

connections, and dense feature transfer mechanisms to refine the accuracy of its semantic interpretations.

- The formulation of a 2D to 3D semantic transfer protocol, predicated on a superpixel segmentation strategy that optimizes the accuracy and efficiency of semantic annotations across dimensions.
- The development of a novel viewable model, underpinned by semantic consistency constraints between images, to mitigate the inaccuracies born from missegmentations and enhance the fidelity of the semantic transfer process.

4.2. Contributions and Innovations

This research transcends the conventional boundaries of "blind point clouds," transforming them into semantically enriched canvases that can support a wide array of practical applications, from sophisticated indoor navigation solutions to interactive virtual reality experiences. The core contributions of this study are twofold:

1. The creation of the Deeplabv3+ network, a nuanced approach to the semantic segmentation of complex indoor scenes, leveraging state-of-the-art structural and algorithmic advancements to ensure unparalleled precision in object identification and spatial layout estimation.
2. The introduction of an innovative semantic transfer methodology, which harmoniously bridges 2D image segmentation results with 3D point cloud models, marking a significant leap forward in the semantic modeling of indoor spaces.

4.3. Future Directions

While this study marks a significant advancement in the semantic segmentation of indoor 3D point clouds, it also opens several avenues for future research:

1. The exploration of more dynamic segmentation techniques that minimize the reliance on extensive training datasets, potentially incorporating active-learning strategies to streamline the training process and enhance the granularity of segmentation results.
2. The integration of inherent 3D point cloud features—geometrical, textural, and structural—into the segmentation framework to provide a more rounded and robust analysis of indoor spaces.
3. The expansion of the semantic segmentation methodology to accommodate point clouds derived from a broader spectrum of sources beyond those generated by SfM algorithms, aiming for a universal segmentation approach.

References

1. Hermans A., Floros G., Leibe B. (2014). Dense 3D semantic mapping of indoor scenes from RGB-D images. *Proceedings*, pp. 2631-2638.
2. Lu G. (2016). From coarse to fine: Quickly and accurately obtaining indoor image-based localization under various illuminations. University of Delaware.
3. Dai A., Chang A.X., Savva M., et al. (2017). ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes.
4. Tchapmi Shengjun. (2017). Multi-view image enhancement RGB-D indoor high-precision 3D mapping method. Wuhan University.
5. Hedman P., Ritschel T., Drettakis G., et al. (2016). Scalable inside-out image-based rendering. *ACM Transactions on Graphics*, 35, Article 231.
6. Lu G. (2016). From coarse to fine: Quickly and accurately obtaining indoor image-based localization under various illuminations. University of Delaware. (Note: This entry appears to be a duplicate of reference 2.)
7. He K., Zhang X., Ren S., et al. (2016). Deep Residual Learning for Image Recognition. *Proceedings*, pp. 770-778.
8. Lee C.Y., Badrinarayanan V., Malisiewicz T., et al. (2017). Roomnet: End-to-end room layout estimation. *Proceedings*, pp. 4865-4874.