3D Semantic Explorer: Deep Learning for Complex Semantic Labeling and Occupancy Prediction

Syfullah Mohammad[1*] **Nishanth Ravula**[2*] **Uday Tej Togiti**[3*]

[1] The State University of New York at Buffalo
syfullah@buffalo.edu,
nravula@buffalo.edu,
udaytejt@buffalo.edu

Abstract

Humans have the ability to formulate a 3D mental map of the entire surrounding present around them. While this ability to form a semantic understanding of the environment might be taken for granted, it is in practice a challenging task as it required extraction and processing of multiple scene information. While existing solution making use of 3D Simultaneous Localization and Mapping (SLAM) propose reliable robotics-based solutions to the above problem, incorporating recent advancements in the field of AI (visual transformers) might provide the versatile solutions researchers have been looking for. Towards this we propose 3D Semantic Explorer, a transformer based 3D mapping model that analyzes, extract semantic information from various view points, and performs complex semantic labelling and occupancy prediction given an environment. To achieve this we use voxels to develop 3D semantic scene completion followed by using CNNs to predict occupancy of semantic labels for each voxel in 3D space. By proper consideration of Sparsity and class imbalance, the proposed model performs significantly well on the semantic kitty dataset. The code is available with instructions on our ***Github repository***.

Introduction

Complete 3D understanding of the 3D environment is a very important task in the domain of autonomous vehicles. It is particularly of utmost importance because 3D understanding directly effects the subsequent tasks such as planning and 3D map construction. However, unlike human brain that captures information from multiple modes and clearly formulates a 3D understanding of the surrounding complex environments, machines cannot form a 3D map with such ease. Theses challenges can be due to lack of resolution while sensing and incomplete observation due to occlusions present in the complex environments. Humans have the ability to see beyond occlusions and use prior experience to for a mental map of the environment even in the presence of occlusions in the environment. This ability in humans is very difficult to replicate in machines.

In order to tackle the above stated challenges during 3D reconstruction (mapping) of complex environments, Semantic Scene Completion (**cite**) was proposed. The aim of the work is to jointly gather information of the entire scene geometry and semantic information from limited information provided to the algorithm. This method tries to address two challenges at the same time:

- Scene Reconstruction (for areas that are clearly visible)
- Scene Hallucination (for areas that are partially visible or occluded)

The model draws inference from the above stated fact that humans can naturally reason about the complex scene geometry even with incomplete or occluded objets present in the complex environment. This naturally leads to the point that humans have far superior ability to form complex 3D maps of the environment and there is a significant gap to be bridged between Semantic Scence Completion and human comprehension of a given 3D complex environment.

Data for such a semantic scene mapping algorithm comes in multiple forms. Many of the models that try to form a semantic map of the environment use LiDAR data to form accurate 3D geometric measurement. However, the downside to using LiDAR data is that it is quite expensive and the portability of LiDAR data is on the lower side. This brings us to the use of cameras owing to the fact that cameras are cheaper and provide richer visual information for the given environment using driving data taken from a car. This has given a rise in the use of camera-based SSC models that use 2D image frames to form a 3D reconstruction of the entire environment by performing stereo vision and depth matching. One such work is MonoScene Monoscene performs 2D image inputs to 3D using dense feature projection. However, during the MonoScene calculation, the algorithm assigns empty values to the scenes that are occluded, making the algorithm perform suboptimally. This leads to several ambiguities in the environment representation especially due to occlusions in the environment.

The main contributions of this work include:

- We use the latest developments in the field of AI, including Transformer architecture to perform 3D-2D representation of the environment using cross-attention mechanism.

[*] These authors contributed equally.

- Before reconstructing locations that are partially visible, we reconstruct fully visible regions to improve the performance of the model.
- Using sparse representation, we perform efficient and scalable 3D reconstruction, this is owing to the hypothesis that majority of the regions in the environment are unoccupied.
- A novel two-stage methodology to reconstruct 3D environment using visual transformers (voxels).
- Image depth map creation for enhanced depth estimation for the objects in the environment.
- A novel Masked Encoder model (MAE) that outputs complete 3D reconstruction of the environment.
- We present our results on the SemanticKITTI dataset that provides image data in the form of odometry data. We then perform occupancy prediction to enhance our results and yield better results that the existing state of the art models.

The entire architecture is divided into two parts. The first part of the model uses Convolutional Neural Networks to perform scene depth estimation and scene reconstruction of the given geometry. The output of the first part of the model is a sparse set of voxels from predefined learnable visual tokens. The second part of the model is the transformer architecture that uses self-attention to perform extensive semantic segmentation on the received visual tokens and completes the scene representation.

Related Works

3D reconstruction is the task that is aimed at performing 3D map of the given geometry using multiple 2D images from multiple view points. As this problem is deeply grounded in Computer Vision, traditional CV algorithms try to solve the problem. This problem can be broke down into 2 stages. 1) Single-view reconstruction by learning shape priors from massive data. (Li et al. 2023), and stereo view reconstruction using images from multiple cameras and multiple view points. Due to occlusions present in the environment, 3D reconstructions requires the model to perform hallucinations, that is closely related to the imaginative ability of humans to perform unseen object formation given an environment.

The semanticKITTI dataset (Behley et al. 2019) is the dataset that is considered in this work. This is preferred by many researchers because it provides a cheaper and computationally less intensive alternative to the LiDAR based odometry datasets. The paper that introduces the semanticKITTI dataset states the need for large scale datasets to propel research on laser-based semantic segmentation. The authors of the semanticKITTI dataset have annotated all the sequences of the KITTI dataset (Behley et al. 2019) odometry benchmark and provide dense point-wise annotations for the complete 360-degree field-of-view of the employes automotice LiDAR. The paper proposes 3 benchmark tasks including 1) Semantic segmentation of the point clouds using a single scan, 2) Semantic segmentation of point clouds using multiple scans, 3) Semantic scene completion which requires reconstruction and anticipation including occlusions in the environment. This task of semantic scene completion is the one that we use for our task. We use the dataset to perform 3D recpnstruction of the environment and use the existing models to perform benchmark on our model.

(Agrawal, Nakazawa, and Takemura 2009) Proposes a method to accurately segment and classify 3D ranging data, into the objects present in the image frames. The object classification in the work is defined from the application perspective of robot navigation and automation. The main focus of the work is with respect to the planning using the image features present in the datasets. The paper proposes an idea of object features which represent the entire distribution of neighbouring pixels to better perform the object classification tasks. The paper proposes a method to construct polygons from the point data in contrast to using and processing raw points present in the dataset. Finally the paper refines Markov Random Field (MRF) calculation providing more emphasis on ”related points” to perform classification of the detected objects in the image data.

(Anguelov et al. 2005) Addressed the problem of segmenting the image using 3D scan data. A subclass of Markov Random Fields (MRFs) that supports effective graph-cut inference serves as the foundation for our segmentation framework. The MRF models enforce the preference that neighboring scan points have the same categorization label and integrate a broad range of unique attributes. The authors discriminatively train the model using a collection of labeled scans using a recently established maximum-margin framework, and as a result, authors automatically determine the relative relevance of the features for the segmentation task. The trained MRF can subsequently be utilized to segment new scenes very effectively by performing graph-cut inference. We put our method to the test on three sizable datasets generated by several 3D sensors, demonstrating its adaptability to both indoor and outdoor contexts with a variety of items.

(Armeni et al. 2017) propose a dataset of large-scale interior settings that includes instance-level semantic and geometric annotations together with a range of mutually registered modalities from the 2D, 2.5D, and 3D domains. The collection, which spans over 6,000m2, includes over 70,000 RGB photos, depth data, surface normals, semantic annotations, global XYZ images (both regular and 360-degree equirectangular views), and camera data. Additionally, it contains registered raw 3D meshes and point clouds that have been semantically tagged. Using the regularities seen in large-scale interior areas, the dataset enables the construction of joint and cross-modal learning models as well as possibly unsupervised approaches.

(Behley et al. 2010) A newer and more active area of robotics study is range data segmentation into semantic

categories. In this study, the authors argue that this task should be seen as a problem of quick, extensive retrieval. Given a dataset of millions of labeled scan points and their surrounding areas, the authors can logically search for points that are similar in the datasets and use the labels of those points to forecast the labels of novel points using a local prediction model, such as majority vote or logistic regression. However, in order to really implement this, extremely effective methods are needed for (1) keeping millions of scan points in memory and (2) quickly locating scan points that are comparable to a target scan point. In this work, we suggest utilizing to address both concerns the most recent spectral hashing by Weiss et al. It uses a compact binary code to represent each item in a database, making sure that things with comparable properties will have similar binary code words. Likewise, close neighbors of the code for the query have codes that are only a little Hamming distance apart. Then, using the same binary code word across all points, the authors locally learn a logistic regression model. Our tests on actual 3D scans demonstrate that the resulting method, known as spectrally hashed logistic regression, may be extremely quick at prediction time and surpasses cutting-edge methods like closest neighbor and logistic regression.

(Anguelov et al. 2005) The majority of autonomous robots require accurate and trustworthy localization and mapping as a fundamental building ingredient. We present a unique, dense laser-based mapping method that uses three-dimensional point clouds to achieve this from revolving laser sensors, which. The authors create a surfel-based map and use the projective data link between the current scan and a rendered model view from that surfel map to predict changes in the robot's position. The authors use the map representation to create a virtual image of the map prior to a potential loop closure for the purposes of loop closure detection and verification. This method enables a more reliable identification even when there is little overlap between the scan and the areas that have already been mapped. Our method provides real-time registration and is effective.

(Behley, Steinhage, and Cremers 2012) For the categorization of three-dimensional laser range data, choosing the right features and their parameters is essential for producing high-quality results. In this study, the authors examine three urban datasets recorded with distinct sensors—sweeping SICK lasers, tilting SICK lasers, and a Velodyne 3D laser range scanner—the performance of various histogram descriptors and their parameters. These descriptors are 1D, 2D, and 3D histograms capturing the distribution of normals or points around a query point. The authors also suggest a brand-new histogram descriptor that makes use of spectral values at various scales. The authors contend that selecting a larger support radius and a global reference frame/axis based on the z-axis can significantly improve the performance of all types of investigated classification models.

(Boulch et al. 2018) In this paper, the authors present a novel, all-encompassing, and effective approach to unstructured point cloud labeling. We offer a framework that employs CNNs on several 2D image views (or snapshots) of the point cloud because the subject of effectively applying deep Convolutional Neural Networks (CNNs) on 3D data is still open. Three main concepts make up the strategy. (i) We pick numerous acceptable snapshots of the point cloud. The authors produce two different kinds of images: a depth composite view with geometric features and a Red-Green-Blue (RGB) view. Then, using fully convolutional networks, we label each pair of 2D snapshots pixel-by-pixel. To accomplish a lucrative fusion of our heterogeneous inputs, various architectures are tested.

(Chen et al. 2016) In this work the authors address the task of semantic image segmentation with Deep Learning and make three main contributions that are experimentally shown to have substantial practical merit. First, we highlight convolution with upsampled filters, or 'atrous convolution', as a powerful tool in dense prediction tasks. Atrous convolution allows us to explicitly control the resolution at which feature responses are computed within Deep Convolutional Neural Networks. It also allows the authors to effectively enlarge the field of view of filters to incorporate larger context without increasing the number of parameters or the amount of computation. Second, the authors propose atrous spatial pyramid pooling (ASPP) to robustly segment objects at multiple scales. ASPP probes an incoming convolutional feature layer with filters at multiple sampling rates and effective fields-of-views, thus capturing objects as well as image context at multiple scales. Third, we improve the localization of object boundaries by combining methods from DCNNs and probabilistic graphical models. The commonly deployed combination of max-pooling and downsampling in DCNNs achieves invariance but has a toll on localization accuracy. We overcome this by combining the responses at the final DCNN layer with a fully connected Conditional Random Field (CRF), which is shown both qualitatively and quantitatively to improve localization performance. Our proposed "DeepLab" system sets the new state-of-art at the PASCAL VOC-2012 semantic image segmentation task, reaching 79.7% mIOU in the test set, and advances the results on three other datasets: PASCAL-Context, PASCAL-Person-Part, and Cityscapes.

(Cordts et al. 2016) Visual comprehension of intricate urban street sceneries is a necessary component for a variety of applications. Large-scale datasets have greatly aided object detection, particularly when used to deep learning. However, no existing dataset completely reflects the complexity of actual urban environments for semantic scene interpretation. To combat this, the authors present Cityscapes, a benchmark collection and sizable dataset for training and testing methods for instance- and pixel-level semantic tagging. Cityscapes is made up of a sizable and varied collection of stereo video sequences shot in the streets of 50 various cities. For approaches that make use of huge amounts of weakly labeled data, 5000 of these images

have high quality pixel-level annotations, while 20000 additional images have coarse annotations. Importantly, the authors' effort outperforms prior efforts in terms of dataset size, richness of annotation, scene variety, and complexity. The dataset characteristics are further examined in our companion empirical study, which also assesses the performance of a number of cutting-edge methods using our benchmark.

(Dai et al. 2017a) The availability of big, labeled datasets is a crucial necessity for utilizing supervised deep learning techniques. Unfortunately, there is very little data available in the area of RGB-D scene interpretation. Current datasets only cover a restricted variety of scene views and have a tiny number of semantic annotations. ScanNet, an RGB-D video dataset with 2.5M views in 1513 scenes annotated with 3D camera poses, surface reconstructions, and semantic segmentations, is our solution to this problem. The authors created a user-friendly, scalable RGB-D capturing system with automatic surface reconstruction and crowdsourcing semantic annotation to gather this data. The authors demonstrate that utilizing this data makes it possible to perform at the cutting edge on a number of 3D scene understanding tasks, such as 3D object categorization, semantic voxel labeling, and CAD model retrieval.

(Dai et al. 2017b) offer ScanComplete, a unique data-driven method for predicting a complete 3D model together with per-voxel semantic labels from an imperfect 3D scan of a scene. The main benefit of our approach is that it can manage big scenes with different spatial extents, controlling the cubic expansion in data size as scene size rises. In order to do this, the authors create a fully-convolutional generative 3D CNN model whose filter kernels are independent of the size of the entire picture. The model can be tested on any size of scene, but it can be trained on scene subvolumes. Additionally, the authors provide a coarse-to-fine inference approach to take advantage of huge input context sizes and generate high-resolution output. The authors carefully assess various model design decisions in a large set of tests, taking into account both deterministic and probabilistic models for completion and semantic inference. Our findings demonstrate that we significantly beat other methods in terms of completion quality and semantic segmentation performance, as well as in terms of the size of the environments handled and processing efficiency.

(Deng et al. 2009) An increase in sophisticated and reliable models and algorithms for indexing, retrieving, organizing, and interacting with photos and multimedia data may result from the explosion of image data on the Internet. But the precise method for organizing and utilizing such data continues to be a major challenge. Here, we introduce "ImageNet," a brand-new database that is a sizable ontology of images constructed on top of the WordNet structure. The majority of WordNet's 80,000 synsets will be filled with an average of 500–1000 crisp, full-resolution images thanks to ImageNet. The outcome will be tens of millions of annotated photographs arranged according to WordNet's semantic hierarchy. This paper provides a thorough evaluation of ImageNet as it stands today: 5247 synsets and 3.2 million in 12 subtrees in total. They demonstrate that compared to the present image databases, ImageNet is significantly more accurate, diverse, and big in scale. Building a database of this size is a difficult endeavor. We outline the method of data collecting using Amazon Mechanical Turk. Finally, we present three straightforward applications in automatic object clustering, image classification, and object identification to demonstrate the utility of ImageNet. We anticipate that ImageNet's size, precision, diversity, and hierarchical structure will present academics working in the field of computer vision and beyond with unmatched opportunities.

(Engelmann et al. 2018) In this research, researchers describe a deep learning architecture that deals with the issue of unstructured point cloud 3D semantic segmentation. We offer grouping algorithms that establish point neighborhoods in the initial world space and the learnt feature space, in contrast to prior work. Neighborhoods are significant because, depending on their spatial extent, they enable the computation of local or global point characteristics. The pairwise distance loss and the centroid loss are additional dedicated loss functions that the authors include to further structure the learned point feature space. We demonstrate how to use these processes for the job of 3D semantic point cloud segmentation and present cutting-edge results on indoor and outdoor datasets.

(Everingham et al. 2014) The Pascal Visual Object Classes (VOC) challenge has two parts: (i) an annual competition and workshop; and (ii) a publicly accessible collection of photos with ground truth annotation and standardized evaluation software. Classification, detection, segmentation, action classification, and person layout are the five difficulties. The authors review the challenge from 2008 to 2012 in this essay. The paper is intended for two audiences: challenge designers who want to see what we as the organizers have learned from the process and our recommendations for the organization of future challenges; and algorithm designers, researchers who want to see what the state of the art is, as measured by performance on the VOC datasets, along with the limitations and weak points of the current generation of algorithms. The authors introduce several novel evaluation methods to analyze the performance of submitted algorithms on the VOC datasets, including a bootstrapping method to determine whether or not differences between two algorithms' performances are significant; a normalized average precision to compare performance across classes with different proportions of positive instances; and a clustering method to visualize the performance across multiple algorithms so that the hard and e In order to detect the different kinds of errors that occur, we also examine the community's development over time using the techniques proposed by Hoiem et al. in the Proceedings of the 2012 European Conference on Computer Vision.

(Firman et al. 2016) Building a complete 3D model

of a scene, given only a single depth image, is underconstrained. To gain a full volumetric model, one needs either multiple views, or a single view together with a library of unambiguous 3D models that will fit the shape of each individual object in the scene. We hypothesize that objects of dissimilar semantic classes often share similar 3D shape components, enabling a limited dataset to model the shape of a wide range of objects, and hence estimate their hidden geometry. Exploring this notion, we provide an approach that can complete the unobserved geometry of tabletop-sized objects, based on a supervised model trained on currently accessible volumetric elements.

(Gaidon et al. 2016) Modern computer vision algorithms typically require expensive data acquisition and accurate manual labeling. In this work, the authors instead leverage the recent progress in computer graphics to generate fully labeled, dynamic, and photo-realistic proxy virtual worlds. We propose an efficient real-to-virtual world cloning method, and validate our approach by building and publicly releasing a new video dataset, called Virtual KITTI (see this http URL), automatically labeled with accurate ground truth for object detection, tracking, scene and instance segmentation, depth, and optical flow. We provide quantitative experimental evidence suggesting that (i) modern deep learning algorithms pre-trained on real data behave similarly in real and virtual worlds, and (ii) pre-training on virtual data improves performance. As the gap between real and virtual worlds is small, virtual worlds enable measuring the impact of various weather and imaging conditions on recognition performance, all other things being equal. We show these factors may affect drastically otherwise high-performing deep models for tracking.

(Garbade et al. 2018) Inferring the 3D geometry and the semantic meaning of surfaces, which are occluded, is a very challenging task. Recently, a first end-to-end learning approach has been proposed that completes a scene from a single depth image. The approach voxelizes the scene and predicts for each voxel if it is occupied and, if it is occupied, the semantic class label. In this work, we propose a two stream approach that leverages depth information and semantic information, which is inferred from the RGB image, for this task. The approach constructs an incomplete 3D semantic tensor, which uses a compact three-channel encoding for the inferred semantic information, and uses a 3D CNN to infer the complete 3D semantic tensor. In our experimental evaluation, the authors show that the proposed two stream approach substantially outperforms the state-of-the-art for semantic scene completion.

(Vaswani et al. 2017) Is a seminal paper that introduces the transformer archicture that currently is widely used in almost every deep learning application.

After the above research study is performed, we have understood the importance of a novel algorithm required for the task of 3D environment scene completion. This pushed our motivation to complete the project and contribute this field of research in a better way.

Proposed Methodology

The aim of the work is to perform 3D scene mapping from RGB images of the environment from multiple view points. Specifically, we use the images in the form of a time series data.

$$\mathbf{I}_t = \{I_t, I_{t-1}, ...\}$$

The steps for the implementation of the project can be broken down into the following categories:

- Data Preprocessing
- Camera Calibration
- Visual Odometry
- 3D Reconstruction
- Semantic Segmentation
- Occupancy Prediction
- Evaluation

Data Preprocessing

As discussed above the data from LiDAR sensors is very expensive and is very hard port the data from one source to another source. For this reasons, we take data from SemanticKITTI dataset. The dataset is received in the form of image frames. We perform data augmentation using **dataaug** library. The library provides tools to remove/add noise to input images. We also have the ability to perform augmentations such as:

- Random Rotate
- Random Translate
- Random Shear
- White Noise addition
- Random horizontal Flip
- Random Vertical Flip

We also split the data required for training, testing and validation. The images are then written into pickle files for proper reading and storing the data for future use.

Camera Calibration

To perform 2D to 3D image reconstruction using image data, we need to first calibrate the cameras. This is to make sure that the points in one camera frame correspond to the points at the same location in the other camera frame. This is the most important step to progress further because without calibration, we cannot get a estimation of the points relative location in the given frame. We make use of features such as corners using Harris corner detector, SIFT features, and scale invariant gradient features to find features in the images and then we use RANSAC to match features from one camera image to another. We then estimate the Essential and Fundamental matrices to find the feature correspondence between cameras.

Visual Odometry

Using the images in the SemanticKITTI dataset, we need to estimate the relative motion of the car. We use Simultaneous Localization and Mapping Algorithms and Stereo Vision Odometry to perform odometry analysis of the dataset. As a result we get the relative pose of the camera locations at each step.

Semantic Segmentation

Once the feature map is attained after the odometry analysis is done, one important issue that arises is the way to deal with occlusions present in the environment. We address this issue by performing Semantic Segmentation. This is done using Deep Learning models including PointRCNN to perform Semantic Segmentation.

The main classes that we use to perform semantic segmentation come from the common attributes that can be seen on road including the signal posts, vehicles, pedestrains, bicycles, houses, road and so on. Once semantic segmentation is done, each data point in the point cloud data comes with an atttributed class label which is one of the above mentioned classes.

Occupancy Prediction

As discussed earlier occupancy prediction is an important step to perform optimal 3D reconstruction of the environment. This is to predict the object that is associated with each pixel in the input image. To perform this, we use 3-Dimensional Convolutional Neural Networks, to predict the occupancy of each point in the point cloud data that is output from semantic segmentation. We then split the received data into train, validation and test data to perform inference and evaluation.

Evaluation

Once the occupancy prediction is done for every data point in the point cloud, we generate a 3D map of the environment. We test the prediction against the ground truth present in the SemanticKITTI dataset and use metrics such as Precision, Recall and F1 score to find the number of points that have been correctly classified.

Implementation in detail

Upon completion of camera calibration and visual odometry data analysis, the cameras will be calibrated in such a way that any point on the left camera will be mapped to a point on the right camera. This means the feature matching step between the cameras becomes a linear search and the image features can be properly mapped.

Once calibration is performed, Semantic Segmentation should be performed. The 3D point cloud segmentation module takes as input a 3D point cloud and performs semantic segmentation to identify different objects in the point cloud. The module utilizes a PointNet++ network that takes as input a set of points and outputs a segmentation label for each point. PointNet++ is a deep learning network that can operate on unordered point sets and is invariant to the permutation of points. PointNet++ uses a hierarchical network architecture to process the point cloud data at different scales and captures both local and global features of the point cloud.

For the proposed model the novelty lies in the fact that we use PointNet++ network to perform 3D point cloud semantic segmentation. As the PointNet++ model architecture has proven to produce state of the art results on tasks such as segmentation, object detection and classification. To get high-quality semantic annotations of the cloud point data, we train PointNet++ on the SemanticKITTI dataset.

With semantic segmentation done, we need to perform feature extraction on the segemented locations in the images, the input to the feature extractor is a segmented point cloud data and the output is the extracted features fromt the model. The proposed method utilizes the hierarchical network architecture of the PointNet++ network to extract both local and global features of the point cloud data. The extracted features can capture the geometry and semantics of the point cloud data and can be used for 3D reconstruction.

With the feature extraction completed, we now finally perform 3D reconstruction using the features extracted. This model takes inputs as the features extracted and provides a 3D map of the complex environment as the output. We make use of PointSDF network. The PointSDF network takes the set of features extracted as inputs, and provides a signed distance field representation of the 3D environment. We train the PointSDF network on the semanticKITTI dataset to reconstruct the 3D environment from the features extracted in the above step.

As discussed earlier occupancy prediction is an important step to perform optimal 3D reconstruction of the environment. This is to predict the object that is associated with each pixel in the input image. To perform this, we use 3-Dimensional Convolutional Neural Networks, to predict the occupancy of each point in the point cloud data that is output from semantic segmentation. We then split the received data into train, validation and test data to perform inference and evaluation.

The novelty of our proposed method lies in the use of the PointSDF network to reconstruct the 3D environment. PointSDF is a deep learning network that can reconstruct the 3D environment from point cloud data without the need for explicit mesh generation. The proposed method utilizes the features extracted from the segmented point cloud data to reconstruct the 3D environment using the PointSDF network. The reconstructed 3D environment can capture both the geometry and semantics of the point cloud data and can be used for various applications such as robotics and virtual reality.

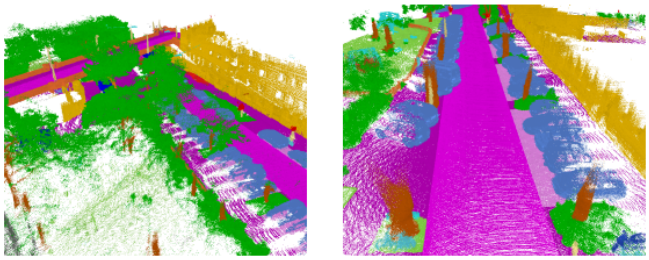

Figure 1: semanticKITTI data visualization from the dataset.

Loss Functions

We take inspiration from the Voxel loss proposed in **cite voxformer**. The loss function proposed in this work is a weighted cross-entropy loss. The ground truth contains data in the form of class representation of team time frame in the form of a voxel grid. The loss can be formulated as,

$$\mathcal{L} = \sum_{k=1}^{K} \sum_{c=c_0}^{c_M} w_c \hat{y}_{k,c} log(\frac{e^{y_{k,c}}}{\sum_c e^{y_{k,c}}})$$

where k is the voxel index and K is the total number of voxels.

Implementation details

Dataset: We obtain data from the SemanticKITTI dataset which provides dense semantic annotations for each LiDAR sweep from the KITTI Odometry Benchmark composed of 22 outdoor driving scenarios. The volume head of the semanticKITTI dataset is $51.2m$ ahead of the car, on the left side the volume head is $25.6m$ and on the right side it is $25.6m$. The voxelization is done using 3D voxel grids with a dimension of $256 \times 256 \times 32$ since each voxel has a size of $0.2m \times 0.2m \times 0.2m$. There are a total of 20 classes in the dataset. 19 of these classes are semantic and 1 is a free class. Regarding the target output, ground truth voxel information is provided the semantic voxel grid information and the sparse input to the SSC model, it can either be a single voxelized LiDAR model or an RGB image.

Implementation: Stage-1 of the project is to perform Depth Estimation using the input images. As the images come from multiple views of the vehicle, the stereo vision analysis is to be performed and the depth of any object in the image frame from the car location is to be estimated. While estimating the depth accurately, the baseline distance between cameras and rectification error between cameras must be taken into consideration. We use conventional Computer Vision and Image Processing techniques to estimate the depth using the disparity parameter. We employ a light-weight Convolutional Neural Networks architecture to estimate the depth using Deep Learning frameworks. The neural network utilized has $23M$ parameters, and to make the model light-weight, we train the depth estimation model on $3.7M$ parameters for faster prediction and training. This gives us the pointcloud of the data depth points as output and then we move on to the later stages of the implementation. We then build a transformer based architecture and PointRCNN based architectures to test the performance of each architecture.

Evaluation Metrics

: As the task includes both Object Detection and classification, we use Intersection over Union (IoU) as the evaluation metric for the implementation. For occupancy prediction, the problem becomes a binary cross entropy problem and we use accuracy for direct performance evaluation. Although there are several variations of IoU such as high mIoU, mIoU, we utilize the naive version of IoU to report our performance of the model. The ranges of IoU and mIoU for the proposed model are inside the volume of $12.8m \times 12.8m \times 6.4m$.

Baselines Methods: We compare our implementaton for the 3D environment reconstruction using deep learning to many of the existing methods including the traditional approaches and existing transformer based approaches. We test the performance with the publically avaiable camera-based models and the deep learning architectures like LMSCNet and SSC-Net and so on.

Performance and Results:

For the stage-1 of the project, we compare the performance of the model with camera-based systems. Our 2D to 3D depth estimation we outperform camera-based methods bu over 18%, the existing benchmark for 2D-3D depth estimate include models such as MonoScene that report 36.80 accuracy. This improvement in the performance is owing to the fact that, using depth estimation and image correction, we eliminate a lot of empty regions in the image. This brings down the required regions to test to a very small subset over which the performance calculation is performed. In comparison to our model, MonoScene associates all the empty spaces to false features negating the impact of all these features and reducing the importance, while many of these features are required to attain a semantic information of the environment.

Performance comparison w.r.t existing methods		
Method	IoU (%)	mIoU (%)
MonoScence	37.60	12.11
Proposed Method	57.69	20.42
SSCNet	55.22	19.68
LMSCNet	55.22	21.50
JS3CNet	55.09	28.12

The proposed model performs in a superior was for two cases. 1) For short-range areas and the second is with respect to the detection of smaller objects.

In addition to the cloud point depth estimation, we also perform occupancy prediction and we report our performance as compared to the existing state of the art models that perform the same task.

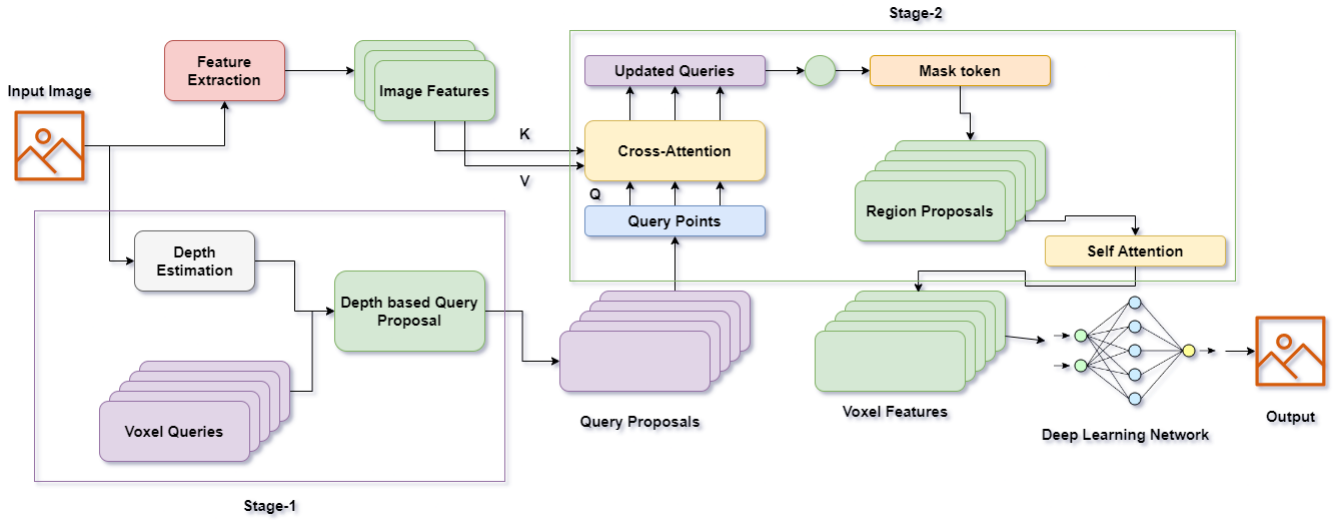

Figure 2: Proposed Architecture

Ablation Studies

For the task of depth estimation we try various combinations of monocular and depth images received from the SemanticKITTI dataset. In general we have observed that the stereo vision based depth estimation provides better accuracy and performance as compared to monocular camera based depth estimation. This can be naturally understood because the error in estimation is reduced due to the presence of multiple views of the same object present in the frame. Using stereo vision odometry data and the disparity, baseline parameters in the given image frame, the estimation of the depth happens in a more accurate fashion. The improvement in performance is also attributed to the fact that stereo vision based depth estimation relies on the epipolar geometry in comparison the monocular camera based estimation that relies purely on pattern recognition. This is why our model with stereo depth estimation performs the best.

For the task of queying the voxel point in the image, we use two types of querying techniques. One is the direct dense query technique. We have found that dense querying is not only inefficient in terms of memory consumption but also performs worse than the second type of querying we use named, random querying. This includes a randomly proposed query region of interest, of dimensions 128128×16 voxel. This allows us to obtain the best trade-off between the memory consumed and the performance time of the model. With the best of both worlds, we perform query mechanism in an efficient fashion.

Input to the model is a temporal series of frames from the cameras, this means there is a semantic relation between the frames that come as the input to our model, we utilize this information given to the model by assigning more weight to the immediate frames in the input and assigning less weights to the older frames of the input. We derive this discounting from the discounting done in Reinforcement Learning where the discounted return given to the agent helps the agent to take intelligent sequence of decisions over time.

The image features in the model are captured using several deep learning transformer-based models. While using IoU and mIoU as the performance indicators, we extract information from the image with a neural network layer that is of size $1/12$ of the input image. During ablation studies, we found that this sizing of the feature extractor with respect to the input image provides the best performance.

For depth estimation, semantic segmentation and occupancy prediction, we perform proper ablation studies with respect to the number of layers in the model, number of nodes per each layer in the model, utilization of self-attention as compared to cross-attention, enabling voxel-to-image or voxel-to-voxel interactions.

Inferences and future work

In this project we have utilized the semanticKITTI dataset to perform 3D environment reconstruction. We utilized Deep Learning methods for 2D-3D depth estimation. We study the performance of several models for the same and provide a clear ablation study for the model working. We then perform semantic segmentation of the point cloud data that is output by the depth estimation to improve the performance of the model thereby reducing all the points that are irrelavant to the prediction from the model.

The proposed architecture has several novelties. Firstly, it utilizes PointNet++ for 3D point cloud segmentation, which has been shown to achieve state-of-the-art results in various point cloud related tasks. Secondly, it utilizes PointNet++ for feature extraction, which can capture both local and global features of the point cloud data. Finally, it utilizes

PointSDF for 3D reconstruction, which can reconstruct the 3D environment without the need for explicit mesh generation.

The proposed architecture has several applications in various fields such as robotics, virtual reality, and autonomous driving. The reconstructed 3D environment can be used to train autonomous robots to navigate in complex environments. It can also be used to create immersive virtual reality experiences and simulate real-world scenarios for training purposes. Additionally, it can be used to develop and test autonomous driving algorithms in a simulated environment.

In conclusion, the proposed architecture presents a novel method for 3D environment reconstruction using the SemanticKITTI dataset. The architecture utilizes various deep learning techniques such as PointNet++ and PointSDF to achieve state-of-the-art results in 3D environment reconstruction. The reconstructed 3D environment has several applications in various fields and can be used to train autonomous robots, create immersive virtual reality experiences, and develop and test autonomous driving algorithms.

We would like to extend the current model to better the prediction performance of the model. In particular, we believe that given a more accurate semantic segmentation model, we would arrive at a richer set of feature cloud point points and using those points for 3D reconstruction of the model would lead to better results. We would also like to enhance the work by using more sophisticated architectures that perform better than vanilla transformer based networks. Using these models can enhance the performance of the overall architecture. Using dynamic feature extraction models, automatic segmentation models, the performance of the proposed model can be further improved.

Conclusion

In conclusion, our work illustrates the use of deep learning architecture that incorporates our expertise in the field of Computer Vision and Image Processing for Depth Estimation. We then use a Tranformer based architecture to perform Semantic Segmentation and that is followed by Occupancy Prediction. Finally we propose our 3D semantic explorer to perform 3D reconstruction of a given complex environment. We have provided clear illustration of our proposed architecture in the form a 2D figure. We then perform extensive ablation studies on various parameters in the model architecture. This provides us with the validation to the proposed model and how it beats the exisiting benchmarks. We make use of PointNet++ for semantic segmentation and tune the model by providing our novelty in the work. Finally we look forward to extend our work to other datasets and other extensions to the model. We are yet to incorporate any physics into the model that will adjust based on the physical movement of cameras and the objects present within the view of the model.

References

Agrawal, A.; Nakazawa, A.; and Takemura, H. 2009. MMM-Classification of 3D Range Data. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation*, ICRA'09, 2269–2274. IEEE Press. ISBN 9781424427888.

Anguelov, D.; Taskar, B.; Chatalbashev, V.; Koller, D.; Gupta, D.; Heitz, G.; and Ng, A. 2005. Discriminative Learning of Markov Random Fields for Segmentation of 3D Scan Data. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, 169–176. USA: IEEE Computer Society. ISBN 0769523722.

Armeni, I.; Sax, S.; Zamir, A. R.; and Savarese, S. 2017. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *CoRR*, abs/1702.01105.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Behley, J.; Kersting, K.; Schulz, D.; Steinhage, V.; and Cremers, A. B. 2010. Learning to hash logistic regression for fast 3D scan point classification. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5960–5965.

Behley, J.; Steinhage, V.; and Cremers, A. B. 2012. Performance of histogram descriptors for the classification of 3D laser range data in urban environments. In *2012 IEEE International Conference on Robotics and Automation*, 4391–4398.

Boulch, A.; Guerry, J.; Le Saux, B.; and 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.; and Yuille, A. L. 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *CoRR*, abs/1606.00915.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. *CoRR*, abs/1604.01685.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T. A.; and Nießner, M. 2017a. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. *CoRR*, abs/1702.04405.

Dai, A.; Ritchie, D.; Bokeloh, M.; Reed, S.; Sturm, J.; and Nießner, M. 2017b. ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans. *CoRR*, abs/1712.10215.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Engelmann, F.; Kontogianni, T.; Schult, J.; and Leibe, B. 2018. Know What Your Neighbors Do: 3D Semantic Segmentation of Point Clouds. *CoRR*, abs/1810.01151.

Everingham, M.; Eslami, S. M. A.; Gool, L. V.; Williams, C. K. I.; Winn, J. M.; and Zisserman, A. 2014. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111: 98–136.

Firman, M.; Aodha, O. M.; Julier, S.; and Brostow, G. J. 2016. Structured Prediction of Unobserved Voxels from a Single Depth Image. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5431–5440.

Gaidon, A.; Wang, Q.; Cabon, Y.; and Vig, E. 2016. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. *CoRR*, abs/1605.06457.

Garbade, M.; Sawatzky, J.; Richard, A.; and Gall, J. 2018. Two Stream 3D Semantic Scene Completion. *CoRR*, abs/1804.03550.

Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023. VoxFormer: Sparse Voxel Transformer for Camera-based 3D Semantic Scene Completion. arXiv:2302.12251.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. *CoRR*, abs/1706.03762.