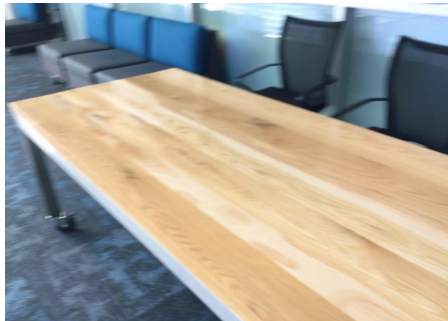


MSc thesis in Geomatics for the Built Environment

Indoor 3D Reconstruction from a Single Image

Chirag Garg
2020



INDOOR 3D RECONSTRUCTION FROM A SINGLE IMAGE

A thesis submitted to the Delft University of Technology in partial
fulfillment
of the requirements for the degree of

Master of Science in Geomatics for the Built Environment

by
Chirag Garg
July 2020

Chirag Garg: *Indoor 3D Reconstruction from a Single Image* (2020)
© ⓘ This work is licensed under a Creative Commons Attribution 4.0
International License. To view a copy of this license, visit
<http://creativecommons.org/licenses/by/4.0/>.

The work in this thesis was made in the:



3D geoinformation group
Department of Urbanism
Faculty of Architecture & the Built Environment
Delft University of Technology

Supervisors: Dr. Liangliang Nan
Dr. Jan van Gemert
Dr. Seyran Khademi

ABSTRACT

3D indoor reconstruction has been an important research area in the field of computer vision and photogrammetry. While the initial techniques developed for this purpose use sensor devices and multiple images for data acquisition and extracting 3D information and representation of the scene, with the advent of deep learning techniques, there has been a good progress in extracting 3D information of an indoor scene reconstruction using a single image. This has potential in minimizing user efforts and cost for data acquisition. The current state of the art method involves two main components, the global depth map and plane instances. After investigating the current state of the art methods, it is observed that there is inconsistency in reconstructed surface boundaries and depth estimation over the curvature and edges of the objects present in the scene, despite having good 3D representation in the surrounding regions. We devise a loss function for optimizing depth estimation during supervision of the neural network by providing geometric awareness to the pixels at local level based on its neighborhood properties defined by spatial compatibility and color similarity. A similar function is used during 3D reconstruction for orientation consistency of normals in the point cloud. Based on the quantitative and qualitative analysis, it is observed that the proposed approach helps in improving the 3D reconstruction of objects in the indoor environment.

ACKNOWLEDGEMENTS

I would like to begin by expressing my gratitude to my first supervisor, Dr. Liangliang for providing constant scientific and personal support during the project. I am also thankful to my second supervisor, Dr. Jan van Gemert for providing critical feedback on the project and assistance in setting up project drive. Finally, many thanks to Dr. Seyran Khademi for helpful suggestions during the project.

I am very grateful to my family and friends for their endless support throughout my masters journey, who also kept me motivated and determined during the difficult times. Lastly, I would like to thank the scientific community for providing open access tools for conducting research.

CONTENTS

1	INTRODUCTION	1
1.1	Background	1
1.2	Research Questions	5
1.3	Thesis Outline	6
2	RELATED WORK	7
2.1	Conventional 3D Reconstruction Approach	7
2.2	3D Reconstruction using single images	8
2.2.1	Object level 3D Reconstruction	9
2.2.2	Scene Level 3D Reconstruction	9
2.3	Observations	14
3	METHODOLOGY	19
3.1	Overview	19
3.2	Neural Network Architecture	19
3.3	Geometry Aware depth loss	23
3.4	3D Reconstruction	25
3.5	Evaluation	27
3.5.1	Geometric Accuracy	28
3.5.2	Plane Detection Accuracy	28
3.5.3	Visual Analysis	29
4	IMPLEMENTATION	31
4.1	Datasets	31
4.2	Programming Environment	31
4.3	Geometry Aware Depth Loss function	32
4.4	Experiment Setup	33
4.4.1	Training Specifications	33
4.4.2	Testing Specifications	33
4.4.3	Model Specifications	35
5	RESULTS AND EVALUATION	37
5.1	Results	37
5.1.1	Experiment with mean and centroid representation	37
5.1.2	Experiment with different weights of loss function terms	40
5.1.3	Final Results	40
5.2	Evaluation	49
5.3	Limitations	60
6	CONCLUSIONS	63
6.1	Discussion	63
6.2	Future Work	65

A	REPRODUCIBILITY SELF-ASSESSMENT	71
B	DATA OF QUANTITATIVE EVALUATION	73

LIST OF FIGURES

Figure 1.1	Application of 3D Indoor Reconstruction	2
Figure 1.2	Understanding of indoor environment	2
Figure 2.1	Multi-view Stereo	8
Figure 2.2	Object level 3D Mesh Reconstruction example	9
Figure 2.3	Mesh R-CNN : 3D Reconstruction Model	10
Figure 2.4	Eigen model for depth prediction	10
Figure 2.5	Planenet: 3D Reconstruction Model based on Dilated Residual Networks (DRN)	11
Figure 2.6	PlanarReconstruction: proposal free instance based planar reconstruction	12
Figure 2.7	PlaneRCNN: 3D Reconstruction Model based on MaskRCNN	12
Figure 2.8	Total3DUnderstanding	13
Figure 2.9	Comparison of plane segmentation in different models	14
Figure 2.10	Observations from current approach 1	15
Figure 2.11	Observations from current approach 2	15
Figure 2.12	Observations from current approach 3	16
Figure 2.13	Observations from current approach 4	16
Figure 3.1	Methodology : Overview	20
Figure 3.2	Neural Network Architecture	21
Figure 3.3	Backbone of Neural Network	22
Figure 3.4	PlaneRCNN : Normal Estimation	23
Figure 3.5	Geometry Aware Concept	24
Figure 3.6	3D Reconstruction	26
Figure 3.7	Error Visualization	29
Figure 4.1	Superpixels of Image	34
Figure 4.2	Experiment Setup	35
Figure 4.3	Experiment Setup	36
Figure 5.1	Experiment 1 : Comparison of predicted depth	38
Figure 5.2	Experiment 1: Performance Assessment	39
Figure 5.3	Experiment 2 : Comparison of predicted depth	41
Figure 5.4	Experiment 2 : Performance Assessment	42
Figure 5.5	Full Pipeline Example	44
Figure 5.6	Depth Estimation : Result 1	45
Figure 5.7	Depth Estimation : Result 2	46
Figure 5.8	3D Reconstruction : Result 1	47
Figure 5.9	3D Reconstruction : Result 2	48
Figure 5.10	Quantitative Evaluation : 1	50
Figure 5.11	Quantitative Evaluation : 2	51

Figure 5.12	Qualitative Comparison of predicted depth : 1	53
Figure 5.13	Qualitative Comparison of predicted depth : 2	54
Figure 5.14	Qualitative Comparison of predicted depth : 3	55
Figure 5.15	Qualitative Comparison of predicted depth : 4	56
Figure 5.16	Comparison of 3D models	57
Figure 5.17	Comparison of 3D models	58
Figure 5.18	Comparison of 3D models	59
Figure 5.19	Limitations	59
Figure A.1	Reproducibility criteria to be assessed.	71
Figure B.1	Quantitative Evaluation of reconstructed depth maps on Scannet Dataset	73
Figure B.2	Quantitative Evaluation of planar segmentation and reconstruction on Scannet Dataset	74
Figure B.3	Quantitative Evaluation of reconstructed depth maps on NYU Dataset	75

ACRONYMS

DRN	Dilated Residual Networks	xi
GPS	Global Position System	1
IMU	Inertial measurement unit	1
CNN	Convolutional Neural Network	3
SIFT	Scale-invariant feature Transform	7
SFM	Structure From Motion	7
FCN	Fully Convolution Network	11
FPN	Feature Pyramid Network	19
ROI	Region of Interest	22
HSV	Hue Saturation Value	25
SLIC	Simple Linear Iterative Clustering	32
IoU	Intersection Over Union	29
RMSE	Root Mean Square Error	42

1 | INTRODUCTION

For a human being, it takes a single glance at a room to understand the indoor built environment. A person understands both its semantic and geometric details. For example, there are table, walls, doors, windows and furniture present in the room and the door is at the right side of table or there is a visible walkable path to the door. Processing this information through a machine is a very challenging task and has been an important area of research in the field of computer vision. Getting 3D information has many applications. For example, it can be used for home or work assistance robots for indoor environment to understand various elements in the indoor space and take desired actions. A user can create a virtual model of the house or office which can further be used for redesigning by the real estate company. Similarly, 3D indoor environment is useful for infrastructure management, energy simulations and emergency services[Zlatanova and Isikdag, 2015]. This has a lot of applications in indoor navigation where a 3D model can be reconstructed and used as database for localisation. Few of the applications have been depicted in [Figure 1.1](#) and [Figure 1.2](#).

1.1 BACKGROUND

3D reconstruction has evolved significantly over the years. To get 3D information such as depth or planar surfaces from indoor space, a combination of various sensors such as using laser scanning device with Global Position System (GPS) device, Inertial measurement unit (IMU) and wifi access points, can provide 3D point clouds of indoor scene [Choi et al., 2015]. However, due to expensive setup and expertise, using multi-view stereo reconstruction proposed in [Sinha et al., 2009] and [Furukawa et al., 2009] is convenient than sensor-based approach. In this, multiple images having a minimum overlap to reconstruct the geometric primitives like vanishing points, planes, lines, and local features such as corners, blobs, which are grouped together into planar or surface patches[Gallup et al., 2010]. However, these techniques still face many difficulties: 1) there is occlusion present in image, thus only limited observation about objects is present, 2) the variation of light and texture hinders the feature extrac-

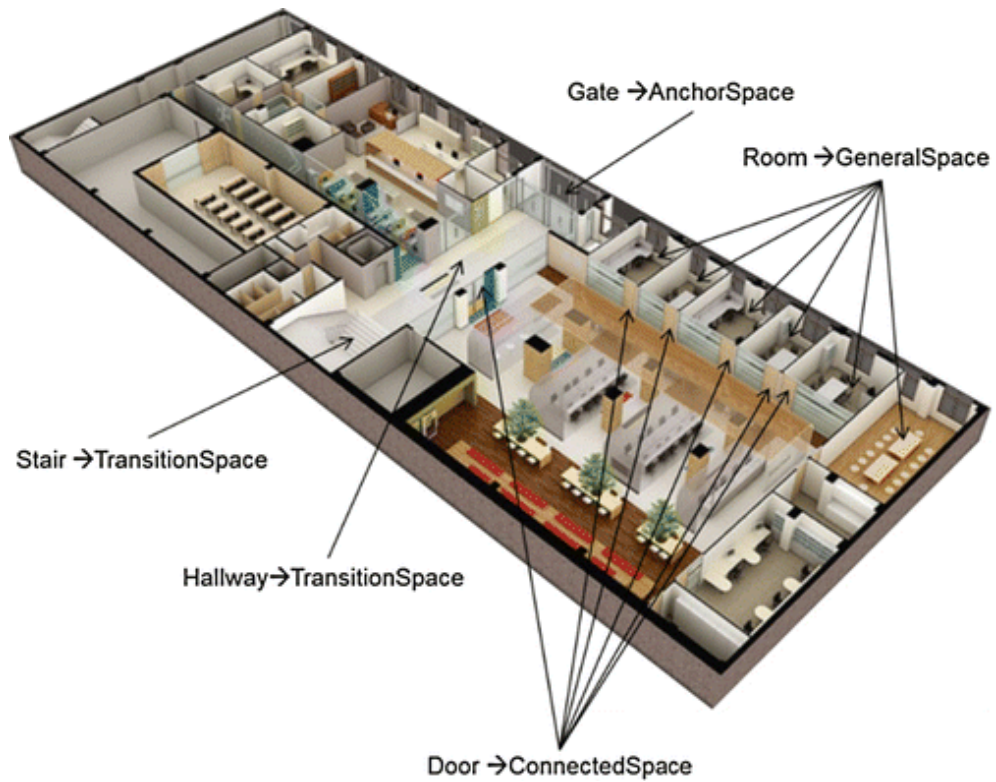


Figure 1.1: An example of 3D indoor model of a building which can be used for navigation, infrastructure development, virtual reality applications for cultural heritage. [Zlatanova and Isikdag, 2017]

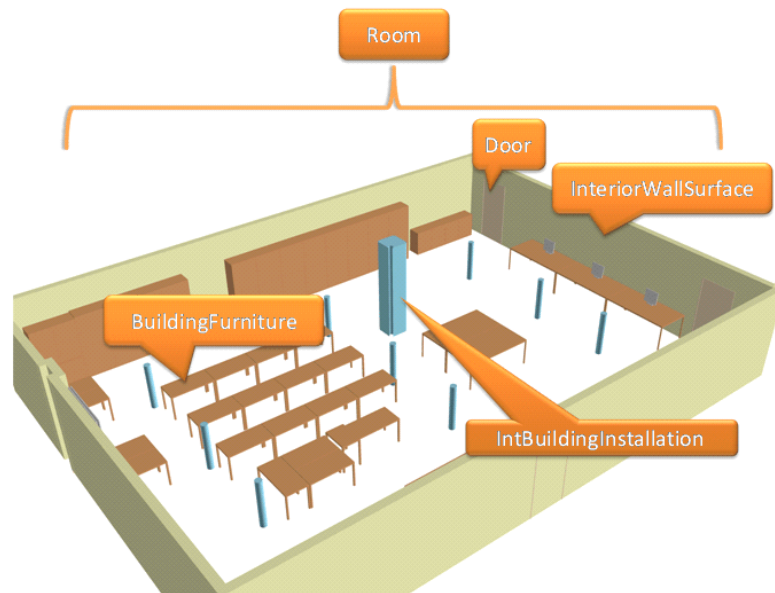


Figure 1.2: An example of understanding an indoor space using 3D model [Donaubauer et al., 2010]

tion algorithms for feature reconstruction, and 3) the complexity of placement of various objects challenges the Manhattan world norms. The neural networks [Liu et al., 2015] try to tackle these challenges by looking at an image from a holistic perspective. Keeping this in mind, extracting 3D information using a single image to extract 3D information can make the data acquisition process easier and can be critical when it is difficult to use traditional techniques. Getting maximum information from one image processing becomes crucial for many applications of real time 3D reconstruction. If reliable and good models can be generated from a single image, it will be helpful in minimizing the user efforts in post processing and the financial cost of data acquisition.

With the evolution of the deep learning techniques, the Convolution Neural networks Convolutional Neural Network (CNN) have been utilized to infer information such as depth maps, surface normals and meshes from a single image [Kang et al., 2020]. Using supervised learning techniques, ground truth information per pixel for an image is used to train a model and infer semantic labels, their location in an image and reconstructed depth [Mousavian et al., 2016]. Recently, new models have been developed which perform these tasks using networks designed for segmentation to reconstruct depth-map from single image, [Liu et al., 2018], [Yang and Zhou, 2018], [Yu et al., 2019]. Among these, PlaneRCNN [Liu et al., 2019] is the state-of-the-art method that outperforms the others in piecewise planar 3D reconstruction in indoor environment. It uses MaskRCNN, [He et al., 2017] as the convolutional backbone network and make improvements for extracting planar surfaces and global depth map from single image. In the basic model, 3D reconstruction involves two main components, one is the global depth map and the other is the plane segmentation. The detected plane instances and global depth information are combined to reconstruct the final piece-wise planar model. Through the analysis of the reconstruction of current techniques, it is observed that there is relatively higher error around the curvature and edges of the objects present in the indoor environment specially non-planar surfaces. The depth is not consistent over the parts of the object despite having good information available in nearby areas. An example of this is shown in Figure 1.3, wherein, we can see that the table top surface in the planar model is not accurately reconstructed when compared to ground truth. Similarly, the 3D point cloud generated does not distinguish boundaries between objects clearly and does not provide good understanding of the scene .

We propose a new energy function to enforce the spatial compatibility of depth and color information based on the local context and maintain depth consistency in accordance with surrounding neighborhood. The aim of this research project is to develop a geometry aware

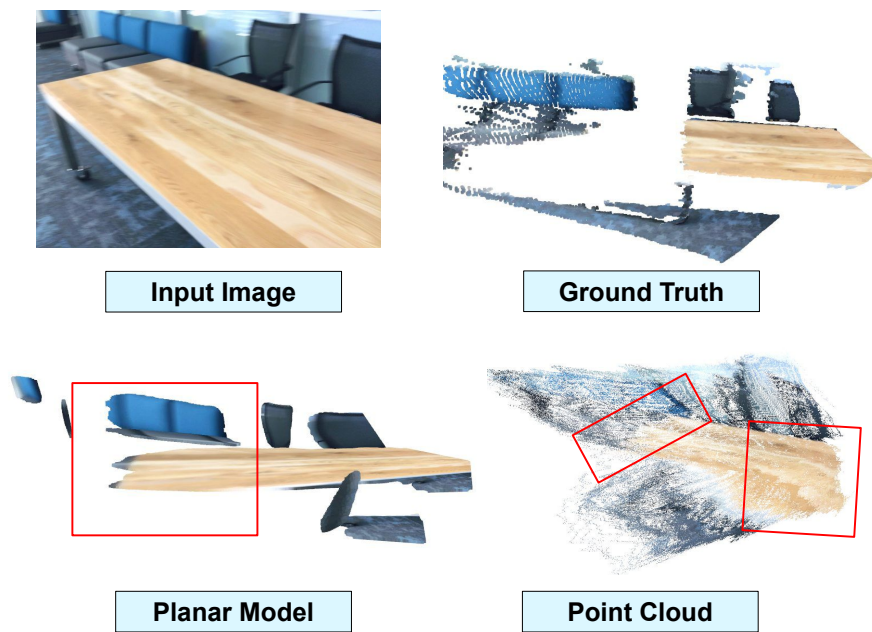


Figure 1.3: An example of 3D reconstruction from a single image using model of [Liu et al., 2019]. In side view of the planar model, it can be observed that the table surface is wrongly reconstructed and side view of the 3D point cloud, the boundaries of objects are not clearly distinguishable and has potential for improvement.

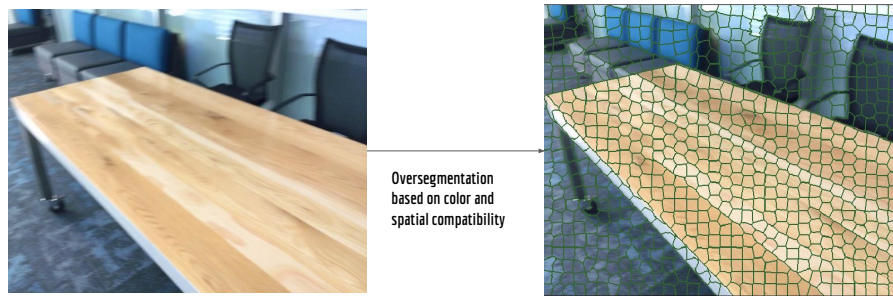


Figure 1.4: Superpixel representation of image helps in handling pixel neighboring region of pixels based on color and spatial proximity

optimization approach that can help in improving the 3D reconstruction process by providing supervision on depth map during training process of the neural network. The motivation is to influence the 3D Reconstruction process using a single image and provide insights into the whole procedure and investigate the role of global depth estimation as well.

1.2 RESEARCH QUESTIONS

Based on the information available at hand, our main research question is :

“Can optimization based on the spatial and color compatibility of pixels within image, help in the improvement of 3D reconstruction from a single image?”

By spatial and color compatibility of pixels, we refer to over-segmentation of an image that creates superpixels from image based on k-means algorithm using color information and pixel positions [Achanta et al., 2010] depicted in Figure 1.4. To support our main research objective, a sub question has also been formulated as following :

- How does the optimization approach influence the process of 3D Reconstruction and depth estimation in an indoor environment ?

Scope of the thesis

To focus on particular aspects of the main research question, we will only consider the following things :

- Only indoor scenes will be used for research. Hence, no outdoor scenes or buildings will be considered.
- Only a single image will be used as data input. Thus, no multiple images are utilized.
- Only those models are considered for research which deal with 3D Reconstruction. Stand alone depth estimation techniques are not considered for research.
- Our main focus will be on improving 3D reconstruction and investigating the effect of our optimization approach, although there are potential related research areas in this project which cannot be explored due to limitation of time. These have been provided in the later sections

1.3 THESIS OUTLINE

[Chapter 2](#) provides an overview of conventional approaches for 3D reconstruction, followed by the current state of the art methods that use single image as input. Lastly, observations based on visual analysis of results is provided for understanding the problems in the current approach. In [Chapter 3](#), the methodology is provided which first presents a brief overview of the pipeline used in the research, followed by detailed description of each component of the pipeline. Afterwards, [Chapter 4](#) lays down the practical details for implementing the methodology and conducting experiments. In [Chapter 5](#), the results and analysis of the conducted experiments are presented along with the evaluation. Lastly, [Chapter 6](#) provides the conclusion of the research answering the research questions and discussing future work.

2 | RELATED WORK

In this chapter, we will look at the current techniques available in the literature that use deep learning techniques for 3D reconstruction using single images. Firstly, an overview of traditional approaches will be provided followed by methods that use deep learning techniques for object and scene level 3D reconstruction. In the end, observations from current state of the art methods in piecewise planar reconstruction will be provided to support the motivation for the project.

2.1 CONVENTIONAL 3D RECONSTRUCTION APPROACH

To obtain 3D information from the real world, conventional approaches use either multiple images and information of camera trajectory or sensor based approach, in which different devices such as depth camera or laser scanners are used to directly obtain the depth or 3D coordinates [Kang et al., 2020] of the environment. A classic algorithm is Structure From Motion (SFM) technique which uses triangulation of feature matches among different images, and can also involve incrementally retrieving information using different sets of images and refining camera poses using bundle adjustment to minimize the re-projection error for a particular point in 3D space [Ullman, 1979]. Among many methods proposed to find the key points among pair of images, "Scale-invariant feature transform" Scale-invariant feature Transform (SIFT) is a benchmark , wherein, the features, invariant to the scale are detected using a image pyramid network, gradients at local level and normal orientation refinement [Lowe, 2004].

Many algorithms have been proposed that use a multi-view approach to first find correspondences between images and use epipolar geometry constraints to obtain 3D information [Goldlücke et al., 2014]. In [Furukawa and Ponce, 2009] and [Galliani et al., 2015], photometric and visibility constraints are used to produce a depth map, while in latter, normal information is also used to improve the performance. A depiction of multi-view stereo reconstruction is depicted in Figure 2.1 wherein, a dense point cloud is obtained using images

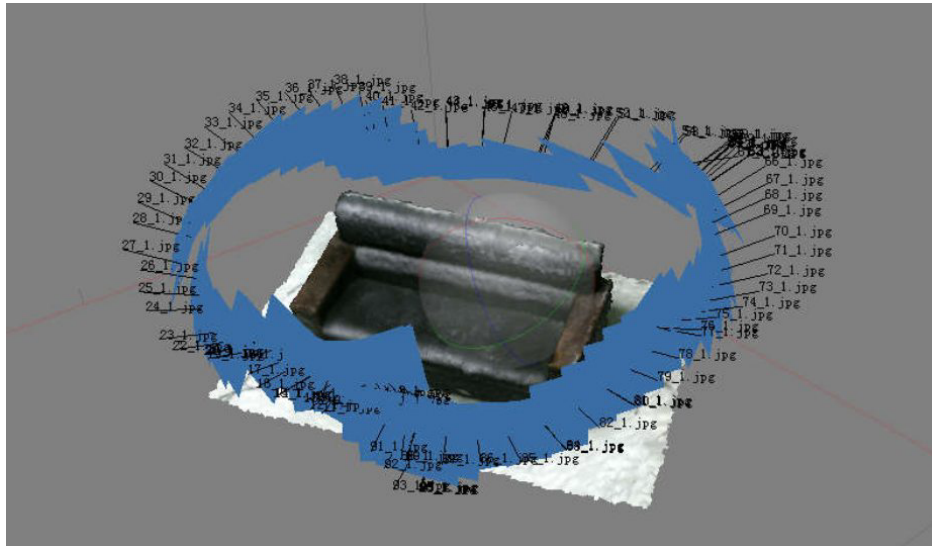


Figure 2.1: Images from different views of object are taken to retrieve point cloud with rich texture information[Kang et al., 2020]

from multiple views. In a sensor based approach, devices such as laser scanners and depth cameras are used to obtain depth or 3D point cloud representing the scanned scene. To recover the 3D information, common approaches use a type of structure from motion technique known as "Simultaneous Localization and Mapping" (SLAM) [Durrant-Whyte and Bailey, 2006] to combine information from different locations of device trajectory to recover the final point cloud. Once a dense point cloud is obtained, surface can be extracted by employing various algorithms depending on the user requirements. These can be smooth surface reconstruction, piecewise plane reconstruction [Gallup et al., 2010], or Poisson reconstruction [Kang et al., 2020]. These methods are often time-consuming and statistical optimization is required to achieve accurate results. This also restricts their utility in real-time applications.

2.2 3D RECONSTRUCTION USING SINGLE IMAGES

With the advent of deep learning techniques, the indoor 3D reconstruction has become popular research topic at the intersection of field of Deep Learning, Computer Vision and Photogrammetry. There is active research in generating object and scene level reconstruction from a single image.

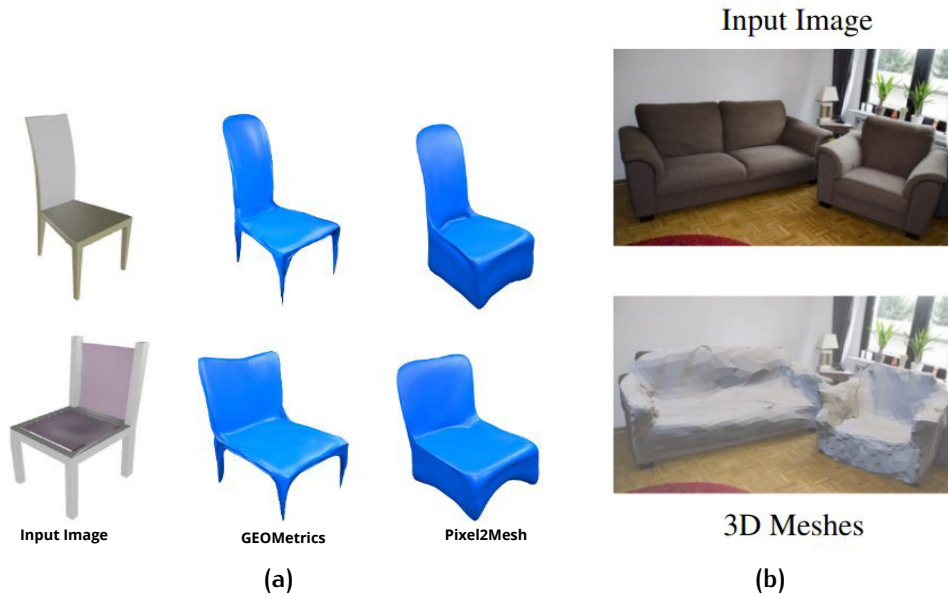


Figure 2.2: (a) A comparison of GEOMetrics and Pixel2Mesh models clipped from [Smith et al., 2019]. (b) An example of 3D reconstruction of indoor scene clipped from [Gkioxari et al., 2019]

2.2.1 Object level 3D Reconstruction

Many approaches proposed for object wise 3D reconstruction using deep learning networks provide representation of single or more objects from a single image. Initial methods conceptually follow an approach of deforming a given mesh into 3D structures. In [Wang et al., 2018], "Pixel2Mesh" is proposed which uses an ellipsoid as the input mesh while in [Smith et al., 2019], "GEOMetrics" model is proposed in which, an "adaptive face splitting" procedure is used to incorporate local context of vertices while mesh reconstruction and provide higher details of objects. A depiction of this is shown in Figure 2.2a. While these approaches focused on synthetic images, in [Gkioxari et al., 2019], a "MeshRCNN" is proposed which provides meshes of multiple objects from real world indoor scenes. They propose a voxel branch in the MaskRCNN model proposed in [He et al., 2017], a popular instance segmentation model and use new losses for supervision on mesh generated from sub-sampled point cloud of the objects during training. This is depicted in Figure 2.3.

2.2.2 Scene Level 3D Reconstruction

In scene level reconstruction, a single image is used to generate a 3D representation of the full scene. This representation can be of various types having different level of information. It provides an understanding of objects present in the room besides the room layout along with

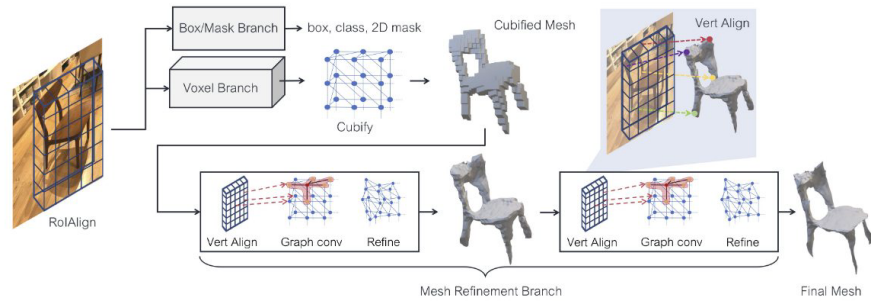


Figure 2.3: Depiction of overview of "Mesh R-CNN" from [Gkioxari et al., 2019], representing the voxel branch that provides a voxel representation of objects and a mesh refinement branch to deform the generated mesh and provide a detailed representation of detected objects in image.

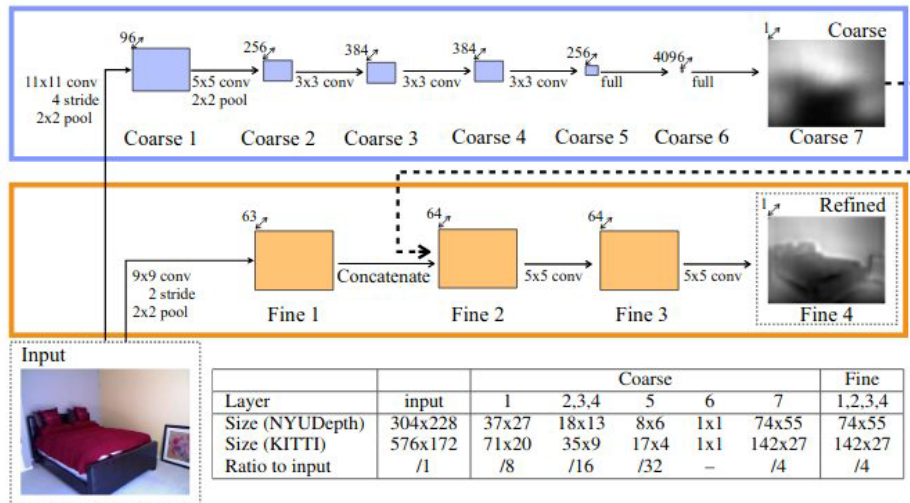


Figure 2.4: Two separate networks are used to improve the final depth prediction by using both fine and coarse level features of image[Eigen et al., 2014]

topological alignment of primitives in the 3D representation. One of the early pioneers in this field, [Saxena et al., 2006] infers depth from outdoor scenes using "markov random fields" to incorporate both global and local features of an image to refine depth prediction. With the advent of the deep neural networks, many convolutional neural network based techniques have been produced to infer depth maps or surface normals from single image [Li et al., 2015]. One well known approach was proposed by [Eigen et al., 2014], wherein, two networks are used to improve the final depth prediction by using both fine and coarse level features of image as depicted in Figure 2.4. But these methods do not provide planar segmentation or parameters which can help in inferring topological relationships among various elements in the scene.

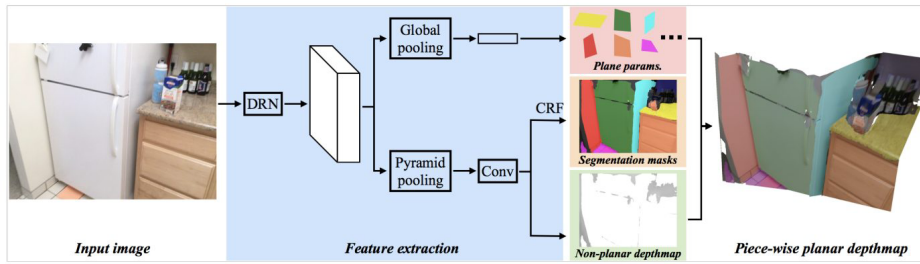


Figure 2.5: Planenet : provides a set of plane parameters, plane segmentation masks and a global depthmap using high resolution feature maps from a pretrained model and ground truth 3D planes [Liu et al., 2018]

Recently, a novel model, “Planenet”, was proposed by [Liu et al., 2018] to reconstruct a “piecewise depth map”, given a single RGB-image using end-to-end deep neural network built upon DRN proposed in [Yu et al., 2017]. As depicted in Figure 2.5, using high resolution feature maps at the end of DRN, three separate output branches are established. The network uses ground truth 3D planes for training to collectively provide a set of plane parameters, segmentation masks and a global depth map.[Liu et al., 2018]. In another approach, “PlaneRecover”, a Fully Convolution Network (FCN) based on DispNet, [Mayer et al., 2015], simultaneously predicts plane segmentation map and plane parameters, taking advantage of ground truth semantic labels, depth map and known camera pose, in outdoor RGB-D dataset, and categorising scene into planar and non-planar depending upon their semantic labels[Yang and Zhou, 2018]. The non-planar pixels are not considered in the depth prediction. It is important to note here that backbone networks used in above methods are flexible networks for image classification (global tasks) and semantic segmentation (pixel wise prediction tasks)[Yu et al., 2017]. Both Planenet and PlaneRecover provide limited number of planes(4-10) in the scene which generalises various small planes into one large plane, thus losing complexity in reconstructed 3D model.

The problem of generalisation of scene was recently resolved in [Yu et al., 2019], wherein, a encoder-decoder architecture is adopted to provide a proposal free instance level plane segmentation and plane parameters in a two stage process. The encoder is built upon Resnet-101 implemented by [Zhou et al., 2018], an established benchmark for semantic classification. In first stage, two decoders train CNN to infer plane segmentation and pixel level embedding which are further merged to provide instance level embedding. In second stage, these instance aware planar segmentation is combined with pixel-level plane parameters to provide final piece-wise planar 3D model[Yu et al., 2019].

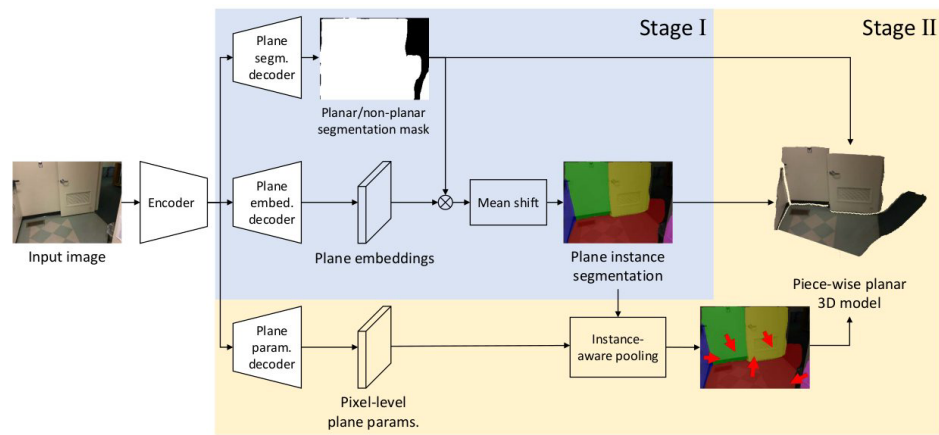


Figure 2.6: Associative embedding : Learns instance embedding to form proposals using mean shift clustering [Newell et al., 2016]

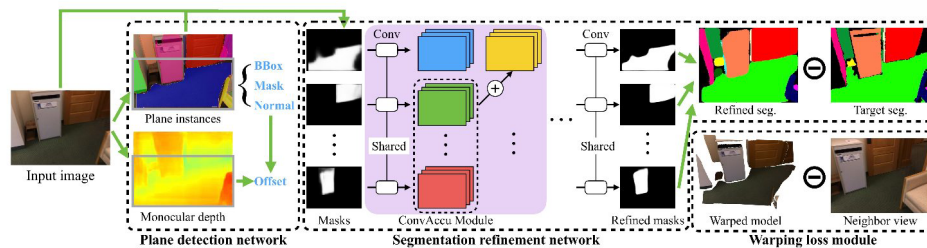


Figure 2.7: PlaneRCNN : Three separate networks are used to estimate plane instances, refine plane segmentation and incorporate warping loss module to boost performance [Liu et al., 2019]

In another method, a proposal-based method was adopted. “PlaneRCNN”, recently, made breakthrough in 3D planar reconstruction using single image by proposing a novel neural architecture in [Liu et al., 2019]. As depicted in Figure 2.7, it contains three networks: firstly, a plane detection network based on MaskRCNN, [He et al., 2017], [Kim, 2017] infers plane normals and offset information along with global depthmap to provide both instance level planar masks and global depth map. Secondly, a joint refinement network takes the output from previous stage to refine each planar instance mask and lastly a warping loss module is used to optimize the reconstructed 3D model from nearby view during training for performance boost. It provides significant improvement in planar reconstruction from all past methods. A visual comparison of some methods discussed in the literature so far, is shown in Figure 2.9, clipped from [Liu et al., 2019].

While piecewise planar representation provides a mid-level information for the scene, the curved surfaces are reconstructed as planar

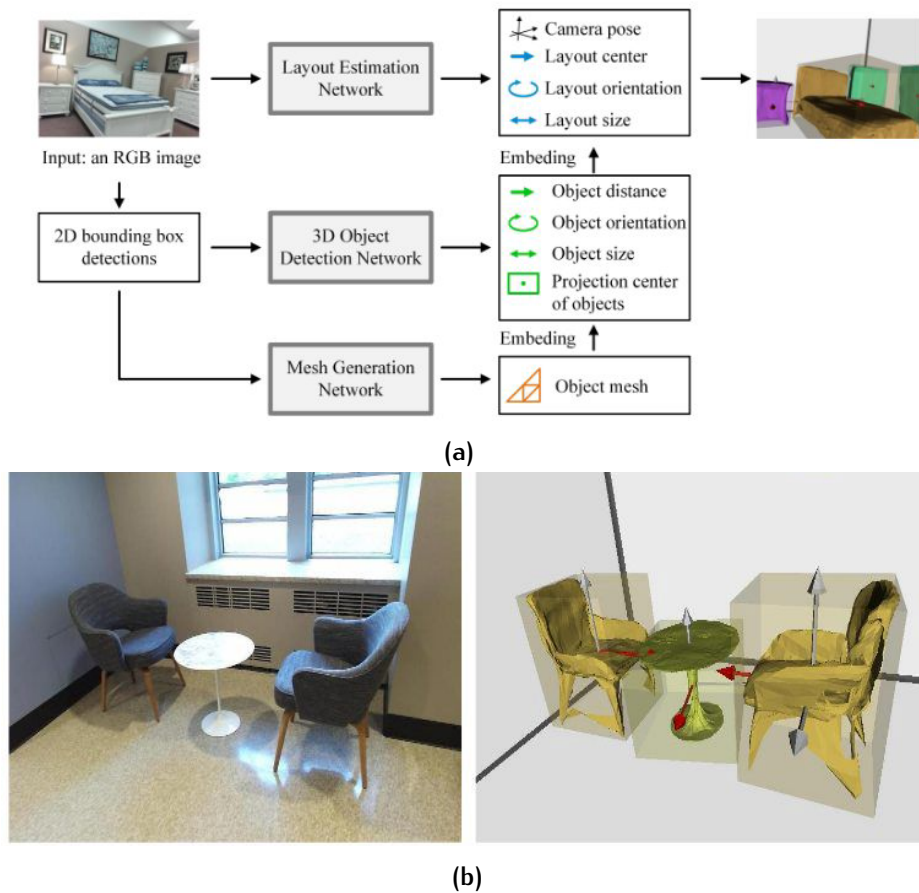


Figure 2.8: a) Total3D : Three separate networks are used to extract the room layout, object bounding boxes and meshes from a single image [Nie et al., 2020]; b)An example of scene level 3D reconstruction from [Nie et al., 2020]

surface and a rough room layout depending on the scene. Recently, a new approach proposed in [Nie et al., 2020] provides the bounding box of objects, the room layout and meshes using a single image as input, by combining 3 different modules. The “layout estimation module” provides supervision on the camera pose and room layout in the scene. The “object detection network” provides supervision on object level orientation and position with respect to the scene and camera. Finally, the “mesh generation network” is used to get objects level meshes present in the scene by generating a target shape prior from the image and using it to perform deforming of a sphere mesh [Nie et al., 2020]. An overview depiction of the network is shown in Figure 2.8a and an inference from indoor environment scene is shown in Figure 2.8b. The model uses combination of losses from different modules and jointly optimize them to get the final model.

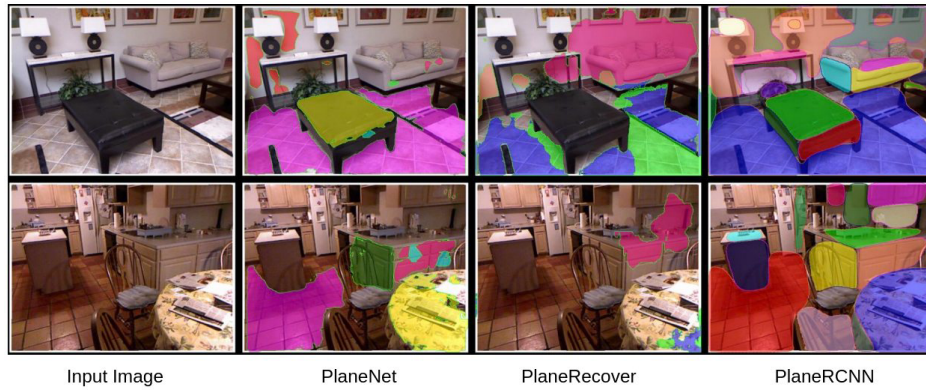


Figure 2.9: From left to right: input image, Planenet [Liu et al., 2018], PlaneRecover [Yang and Zhou, 2018], and PlaneRCNN [Liu et al., 2019]

2.3 OBSERVATIONS

In existing literature for neural networks in the 3D reconstruction, it is observed that the architectures are largely based on the image segmentation networks. A pre-trained model can be utilized to get the feature maps and new features are learnt by adding new modules on old architecture. The new weights are estimated by training only the new modules first and then end-to-end training is performed. This technique of transfer learning can be helpful in customizing state of the art methods for further research. For establishing a benchmark model, we look at both [Liu et al., 2019] and [Yu et al., 2019] for visual analysis. An example of an inference from an image is shown in Figure 2.10. From visual analysis, it becomes clear that the first approach by PlaneRCNN performs qualitatively better over second approach. It is globally and locally, better representation of indoor scene. In Figure 2.11, and Figure 2.12, both approaches do not maintain the orthogonality of planes at all places and their placement is also not consistent with nearby objects. However, PlaneRCNN has denser distribution of points and preserves the topology better than PlanarReconstruction where geometric complexity is not preserved.

The current approaches often fail in critical boundary conditions, resulting in misplaced planes or inconsistency in depth values. For example, in Figure 2.12 a table and rug in the image are together predicted as table while in the 3D model shown in Figure 2.12, major points of table are predicted between ground and table and only a small portion of table is appeared at a height in PlaneRCNN while a good part of table appears as planar surface in PlanarReconstruction. Similarly for pillows in Figure 2.11, there is no depth consistency maintained for single object and when the boundary is changing. More examples of piecewise planar model are provided in Fig-

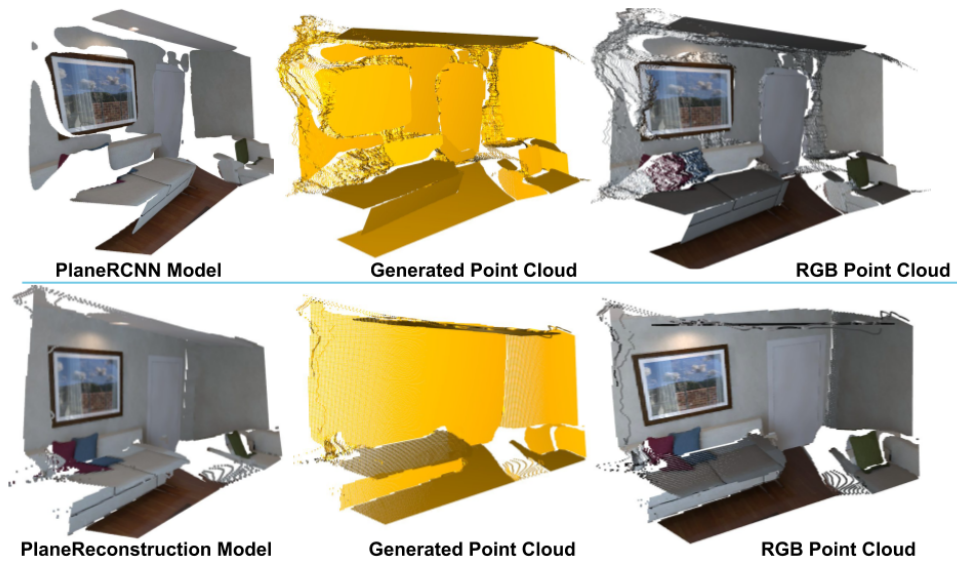


Figure 2.10: From left to right: 3D mesh reproduced from model, generated point cloud using predicted depth, and colored point cloud using original image rgb values

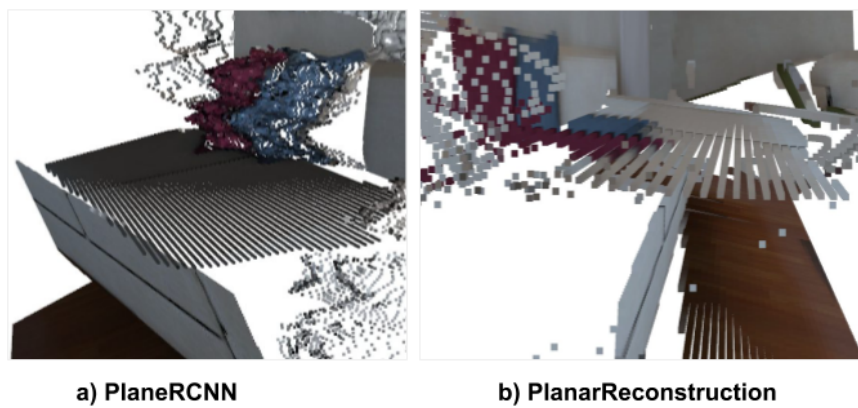


Figure 2.11: From left to right: point cloud generated from PlaneRCNN and PlanarReconstruction respectively and zoomed in on couch and pillows.

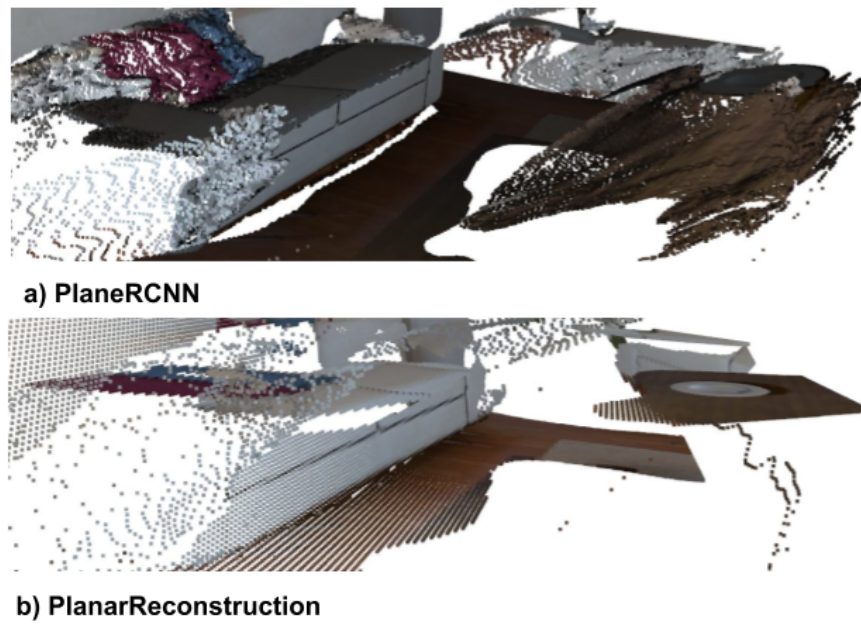


Figure 2.12: From left to right: point cloud generated from PlaneRCNN and PlanarReconstruction respectively and zoomed in on the table

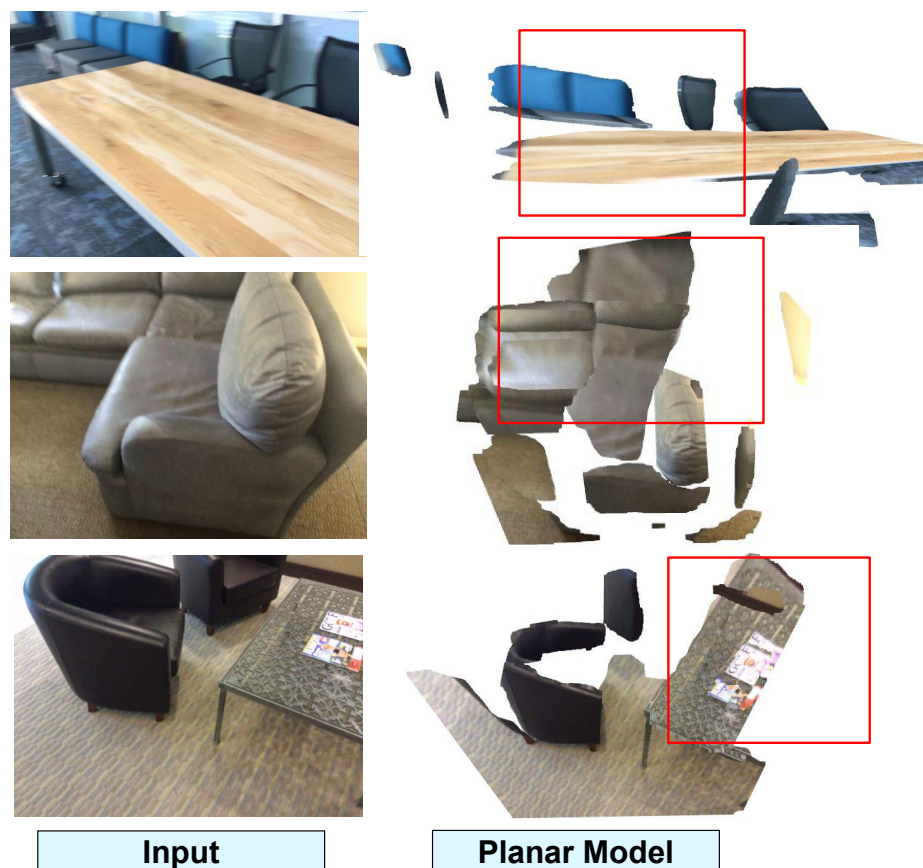


Figure 2.13: From left to right: input image and piecewise planar 3D model generated using [Liu et al., 2019]

ure 2.13. The red markings show the boundary and surface reconstruction problems in the basic version of [Liu et al., 2019]. Hence, adding a new energy function to enforce the relationship between depth and color information has a potential to improve the depth estimation and 3D reconstruction. In order to design a energy function that can provide geometric awareness during supervision of depth estimation based on color consistency and spatial proximity, the neural network needs to enforce the conditions in which the pixels in a spatially connected neighborhood should have consistent depth values. Thus, sudden depth changes should be penalised in a neighborhood of similar colors. We can formulate this as a loss function which has to be minimized by the neural network to provide geometric awareness for depth estimation. To further explain the conceptual idea, we will move to next chapter which defines the methodology in detailed manner for each step.

3 | METHODOLOGY

This chapter provides details of the methodology adopted for the research. Firstly, an overview is presented providing the full pipeline. Afterwards, a detailed description of the proposed loss function and neural network architecture used in the research is provided followed by 3D reconstruction.

3.1 OVERVIEW

A schematic overview of the methodology is provided in [Figure 3.1](#). For 3D planar reconstruction using single image, we need to extract plane parameters and per pixel depth map. Firstly, data needs to be prepared and pre-processed to use it as input for training a neural network. For each iteration, a set of RGB image, pose information, depth image and plane annotation are needed. In a single iteration, a RGB image is passed into the neural network, where loss functions are used for supervised learning to extract plane segments and global depth-map which then are used to calculate piece-wise planar depth. With known camera intrinsic, a piece-planar model can be reconstructed using plane instance parameters and depth information or a point cloud can be obtained.

3.2 NEURAL NETWORK ARCHITECTURE

Keeping our objective in mind, we use the neural network architecture used in PlaneRCNN [[Liu et al., 2019](#)] for our research to investigate the role of depth estimation and plane segmentation in 3D reconstruction process. A depiction of neural network architecture is provided in [Figure 3.2](#). The basic model builds upon the MaskRCNN network based on Resnet-101 and Feature Pyramid Network (FPN) architecture. In the first stage, the input image is normalized by subtracting the mean pixel values from the image, and padding is provided if necessary to resize the image into (640X640) size. This is fed into a bottom-up pathway based on the Resnet-101 architecture consisting of five convolutional modules, each extracting the features at different scale reducing the spatial resolution as we move up by half at

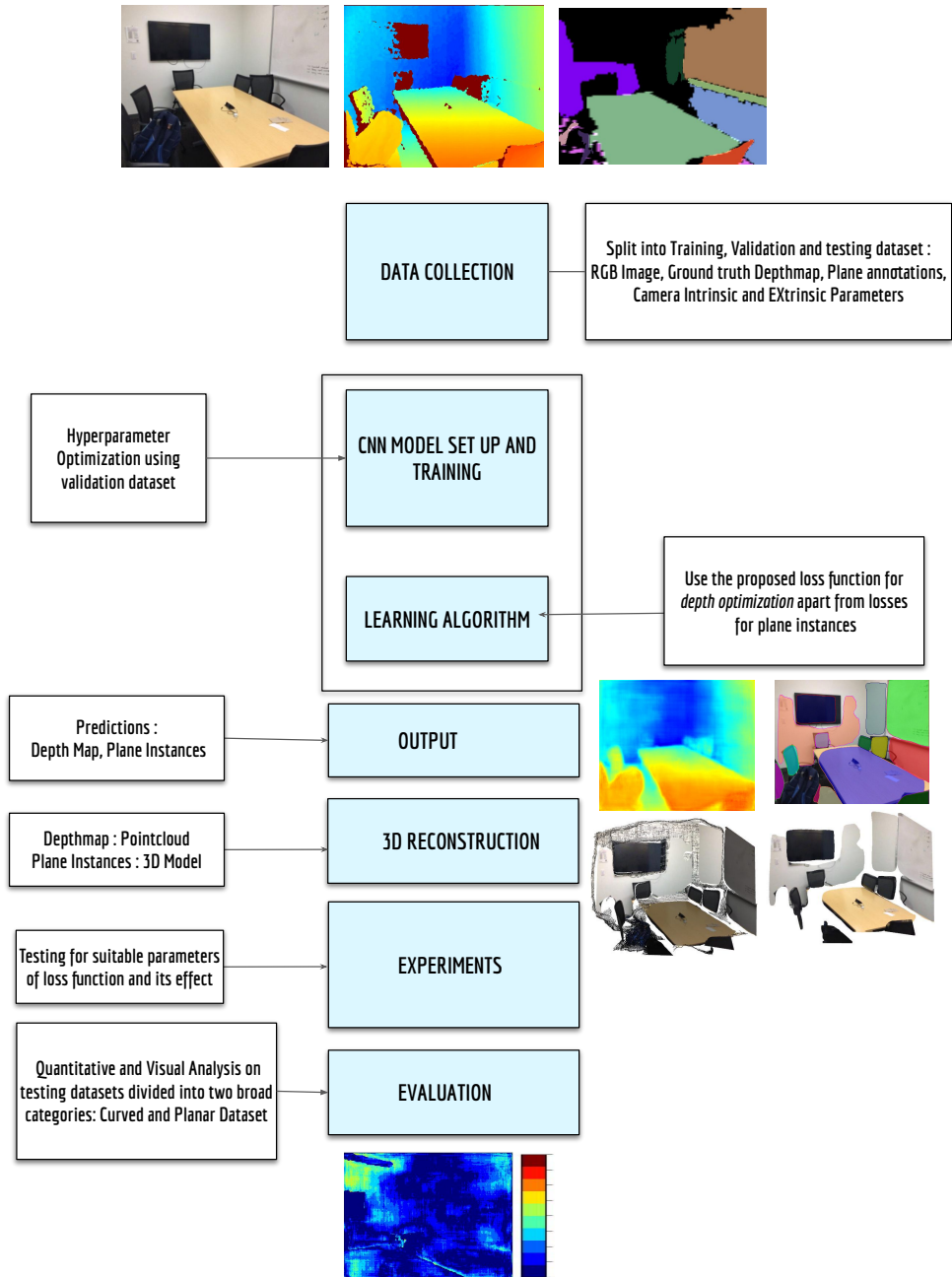


Figure 3.1: Full Pipeline of our methodology

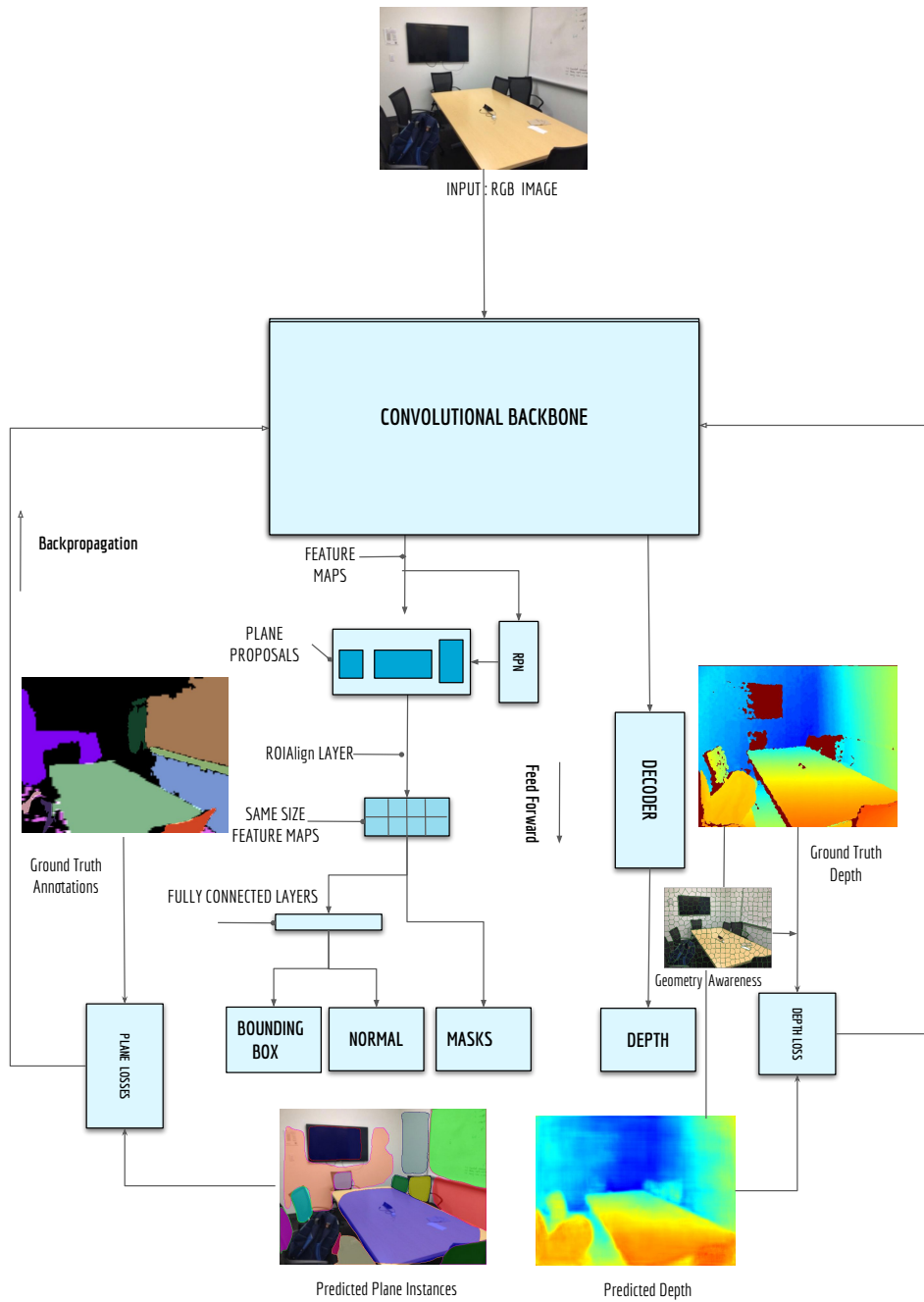


Figure 3.2: Neural network architecture for plane detection and depth map estimation

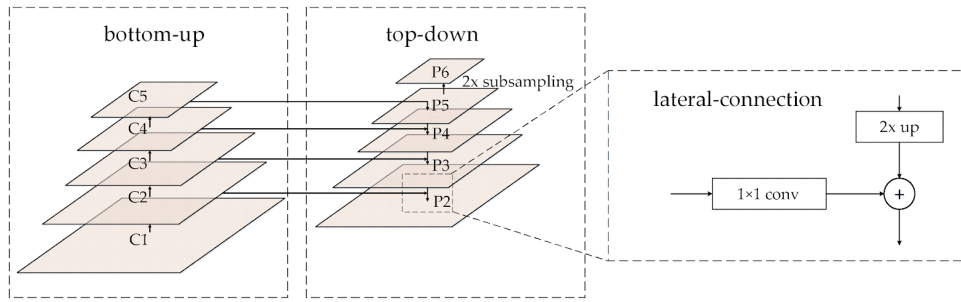


Figure 3.3: Depiction of backbone of the neural network based on Resnet 101-FPN clipped from [Chen et al., 2019]. a feature pyramid map is created by following a top-bottom pathway using the features from bottom-up pathway using lateral connections.

each level. Then, a feature pyramid map is created by following a top-bottom pathway using the features from bottom-up pathway using lateral connections. Each layer of top-bottom pathway is up-sampled, and is added element wise to each layer of bottom-up pathway after convolution. This is then merged together to generate a feature map at one level. This is done at all corresponding levels of the pathways to generate the final feature maps except the second layer of bottom-up pathway [He et al., 2017]. This has been depicted in Figure 3.3.

Plane Instances Estimation

Once, the feature maps from the backbone network are generated, the region proposal network is used as a sliding window at each level of FPN to predict plane instances and provide the normal predictions for the same. To obtain the plane normals, these anchor normals are defined on the basis of the ground truth plane normals. By using the K-means clustering algorithm on randomly sampled plane annotations in training dataset, the normals in equally distributed k directions are used to formulate cluster centers for each anchor normal [Liu et al., 2019]. A depiction of anchor normal and residual vector is provided in Figure 3.4. The anchor normals are then, selected using a cross entropy loss function. In the next stage, the regional proposals generated from first stage are resized to same dimensions using a ROIAlign method. In this method, for each Region of Interest (ROI), the resized features are estimated based on bilinear interpolation of the nearest cell feature values. The feature output of ROIAlign is passed on to the plane instance head comprising of fully connected layers to get the plane instance mask and 3D residual vector for normal estimation. To generate supervision on the residual vector, the nearest anchor normal is estimated and smooth L1 loss is used as proposed by [Liu et al., 2019]. A post processing is done to get the information according to input image and preserve the spatial compatibility of instances.

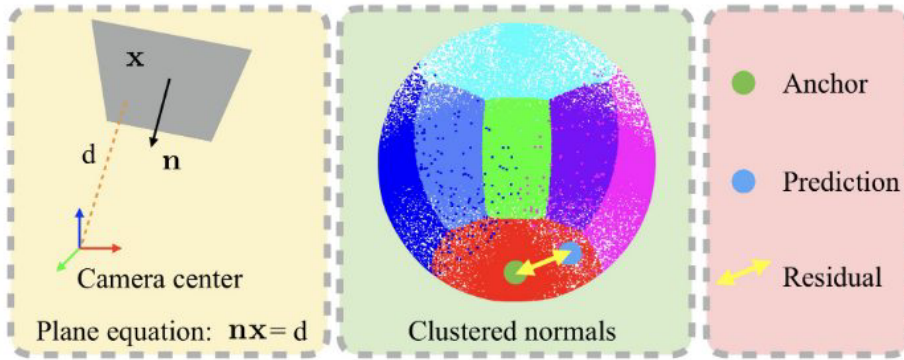


Figure 3.4: Normals of plane are estimated in the respective camera coordinate system using k anchor normals and finding residual vector by PlaneRCNN, [Liu et al., 2019]

Depth Estimation

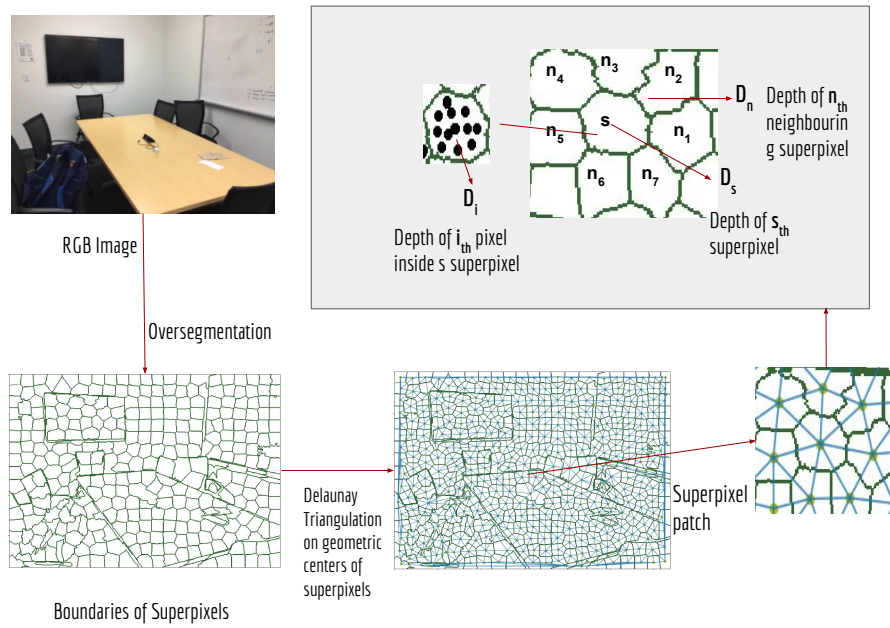
For global depth map, a decoder is added after the FPN, wherein, a convolutional block of five convolutional modules, each having convolutional layer, a batch normalization layer and a rectified linear activation unit layer with stride 1 and size 3 kernel is connected to a deconvolutional block with five respective deconvolutional modules each having a up-sampling layer with scale factor 2, a batch normalization layer and a a rectified linear activation unit layer is used. The features are then fed into the final convolutional layer to predict the global depth map in the dimension of (640 X 640)[Liu et al., 2019].

3.3 GEOMETRY AWARE DEPTH LOSS

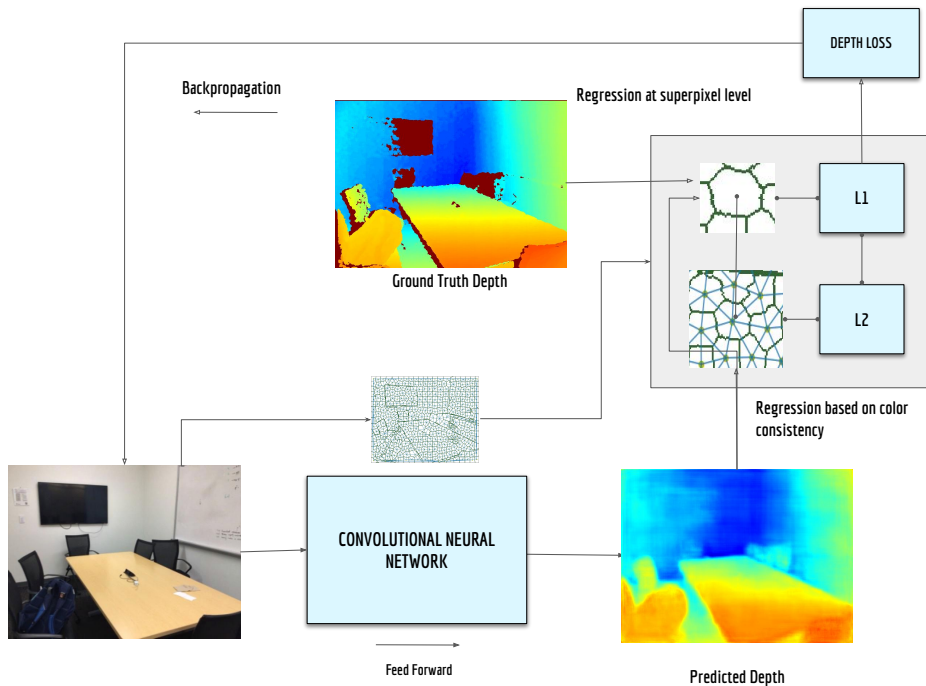
To optimize the depth values based on the local neighborhood geometric context, we formulate a loss function with two terms. The first term allows the pixels of the predicted depth to get geometric awareness on super-pixel level while the other term balances the error with respect to the nearby super-pixels to maintain depth consistency using a graph like structure within a spatial neighborhood. A depiction of the conceptual idea is provided in Figure 3.5. The overall loss function can be estimated by combining the loss from both terms :

$$L = (1 - w)L_1 + wL_2 \quad (3.1)$$

Let S be a set of superpixels in Image \mathbb{I} . For a given super-pixel $s \in S$, the pixels inside it are represented by set P , the centroid is represented by the mean of positions of all pixels within it. Using centers, we collect the neighbouring points using the Delaunay triangulation [Delaunay, 1934]. Doing this, we get a graph like structure where each superpixel assumed as a simplex is connected to neighboring super-pixels. The color information for each super-pixel is represented by its



(a) Depiction of how a graph network is created from an image for defining neighbors and pixels for each superpixel



(b) Overview of geometry aware depth optimization

Figure 3.5: Depiction of how the superpixels can be used provide geometric awareness to predicted depth

Hue Saturation Value (**HSV**) histogram represented by H . The depth information can be represented either by mean of depth of all valid pixels inside the super-pixel or pixels at the centroid of the super-pixel. Thus, for each superpixel $s \in S$, we have p, n, H, D representing its number of valid pixels, the number of neighbors, the histogram values and depth information representing the depth information.

For the first loss term, for each pixel, the predicted depth is compared against the ground truth depth of its representative super-pixel depth. This helps in constraining the pixel depth values based on the superpixel depth values to provide a small surface representation. The cumulative error is the weighted average of this regression for all super-pixels. It is depicted by equation below:

$$\mathbf{L}_1 = \frac{\sum_{s=1}^N \sum_{i=1}^{p_s} |D_i^{pred} - D_s^{gt}|}{\sum_{s=1}^N p_s} \quad (3.2)$$

where, D_s^{gt} is the ground truth depth of superpixel, D_i^{pred} is the predicted depth of the p_i pixel in the superpixel. p_s is the number of valid pixels belonging to the superpixel and N is the total number of superpixels in the image.

For the second term, firstly, only those neighbors are considered for each super-pixel which have high correlation whose value can be calculated by comparing the **HSV histograms** of the two super-pixels. The second term of the loss function is thus given by:

$$\mathbf{L}_2 = \frac{1}{N} \sum_{s=1}^N \frac{1}{n_s} \sum_{n=1}^{n_s} \text{corr}(H_s, H_n) \times |D_s^{pred} - D_n^{pred}| \quad (3.3)$$

In equation 3.1, the term $\text{corr}(a, b)$ is calculated by:

$$\text{Corr}(a, b) = \frac{\sum_{b=1}^k (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^k (a_i - \bar{a})^2 \sum_{i=1}^k (b_i - \bar{b})^2}} \quad (3.4)$$

The Pseudo-Code for the algorithm to calculate loss function is provided in [Algorithm 3.1](#)

3.4 3D RECONSTRUCTION

The global depth map estimated from the neural network and plane instances are used for estimating the plane offsets and point coordinates for 3D reconstruction. An overview of the process is provided

Algorithm 3.1: Geometric Aware Depth Optimization

Input: A RGB image \mathcal{I} , a ground truth depth image D_{gt} , a predicted depth image D_{pr} , number of segments N , compactness c

- 1 **for** each image I **do**
- 2 Get superpixels and metadata : histograms, neighbors, ground truth depth
- 3 **for** $s \leftarrow 0$ to N **do**
- 4 **if** *valid* (*superpixel*) **then**
- 5 // Calculate the L1 term
- 5 **for** each *valid pixel* **do**
- 6 error1 = regression at super pixel level;
- 7 $l_1 = \text{sum}(\text{error}_1)$
- 8 // Calculate the L2 term
- 8 **for** each *neighbor n of superpixel* **do**
- 9 $C = \text{Color similarity with current superpixel};$
- 10 error2 = proportionally regress the respective superpixel predicted depths;
- 11 $l_2 = \text{average}(\text{error}_2)$
- 12 $L_1 = \text{weighted average of } l_1;$
- 13 $L_2 = \text{average of } l_2;$
- 14 **return** $L = L_1 + L_2$

Output: L : Depth Loss value using geometric aware loss function

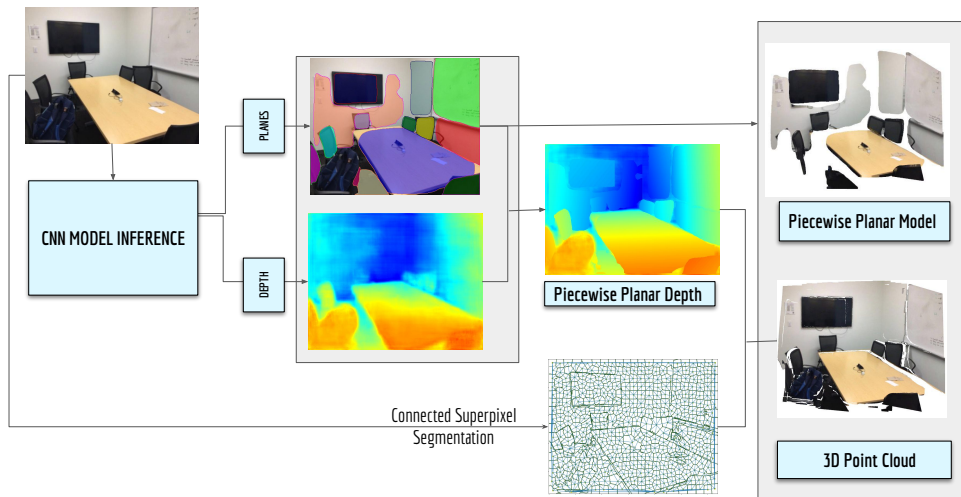


Figure 3.6: Overview of 3D reconstruction

in [Figure 3.6](#). Using the camera intrinsic parameters, a point cloud can be generated from the depth map. For a point P_c in the camera coordinate system, let $(x_c, y_c, z_c, 1)^t$ be the homogeneous coordinates for a pixel P_I in image with $(i, j, 1)^t$ as homogeneous coordinates. The pinhole camera equation provided by [Equation 3.5](#) gives us:

$$P_I = \pi(P_c) = \left(\frac{f_x x_c + c_x}{z(P_I)}, \frac{f_y y_c + c_y}{z(P_I)}, 1 \right)^t, \quad (3.5)$$

where f_x, f_y are focal lengths in x and y direction; c_x, c_y are respective principal point offsets; $z(p_I)$ is the depth value of 2d point p_I . If the depth is known for the 2D point, it can be projected back to a 3D point using inverse projection function provided by [Equation 3.6](#):

$$P_c = \pi^{-1}(P_I, z(P_I)) = z(P_I) \left(\frac{i - c_x}{f_x}, \frac{j - c_y}{f_y}, 1 \right)^t \quad (3.6)$$

For the pixel P_I , the predicted plane normal n is used to obtain the plane offset d using the equation below:

$$d = \frac{\sum_i m_i (n^\top (z(p_I) \pi^{-1} P_I))}{\sum_i m_i} \quad (3.7)$$

We use a normal consistency term to re-orient the normals of points based on the superpixel segmentation depicted in [Algorithm 3.1](#) and calculated in [Equation 3.3](#).

3.5 EVALUATION

For evaluating our designed loss function, we compare the performance with the basic version of [[Liu et al., 2019](#)] as baseline. The depth loss in the baseline model is a pixel level loss between ground truth and predicted depth. If d_i^{pr} represents the predicted depth and d_i^{st} represents the ground truth depth of a pixel i and N is the number of pixels in images to be tested, the baseline loss can be calculated by:

$$\frac{\sum_{i=1}^N |d_i^{pr} - d_i^{st}|}{N} \quad (3.8)$$

For all testing, only those pixels in the image are considered which have valid ground truth label for predicted depth and plane masks.

3.5.1 Geometric Accuracy

Following the previous works, the non planar global depth and piecewise planar depth estimation of the new model will be evaluated by using metrics adopted in [Eigen et al., 2014] and [Wang et al., 2015]. then the following errors will be calculated using their respective equations:

- mean relative error :

$$\frac{1}{N} \sum_{i=1}^N \frac{|d_i^{pr} - d_i^{gt}|}{d_i^{gt}} \quad (3.9)$$

- Root mean square error(rmse) :

$$\sqrt{\frac{1}{N} \sum_{i=1}^N (d_i^{pr} - d_i^{gt})^2} \quad (3.10)$$

- mean log 10 error :

$$\frac{1}{N} \sum_{i=1}^N \left\| \log_{10}(d_i^{pr}) - \log_{10}(d_i^{gt}) \right\| \quad (3.11)$$

- scale invariant rmse log error: rmse log error of normalized predicted and ground truth depth
- accuracy with respect to a certain threshold th , defined by equation 3.12 :

$$\max \left(\frac{d_i^{gt}}{d_i^{pr}}, \frac{d_i^{pr}}{d_i^{gt}} \right) = \delta < th \quad (th \in [1.25]) \quad (3.12)$$

3.5.2 Plane Detection Accuracy

- Random Index(RI): This measures the proportion of pixel pairs between the predicted and ground truth segmentation that are consistent [Arbelaez et al., 2010]. It ranges from 0 (for no intersection) to 1(for same clustering). If S_{pr} represents the predicted segmentation clusters and S_{gt} represents the ground truth segmentation cluster. Then for p_{mn} amount of points for m th cluster of S_{pr} and n th cluster of S_{gt} and N number of pixels in the image, the index can be estimated by :

$$RI(S_{gt}, S_{pr}) = \left\{ \binom{N}{2} - 1/2 \left\{ \sum_m (\sum_n p_{ij})^2 + \sum_n (\sum_m p_{mn})^2 - \sum \sum p_{mn}^2 \right\} \right\} \binom{N}{2} \quad (3.13)$$

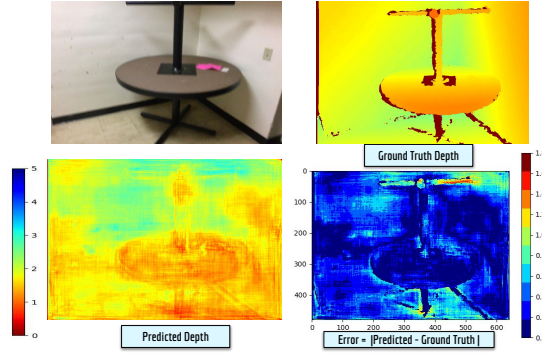


Figure 3.7: Visualization of the predicted and ground truth depth and the estimated error with the range in meters and respective colorbar for representing the information.

- Variation of Information (VOI): This measures the randomness in a image segmentation with respect to other. In [Meilă, 2005], it is calculated by using the entropy information, H and mutual information, I between two images using :

$$VOI(S_{gt}, S_{pr}) = H(S_{gt}) + H(S_{pr}) - 2I(S_{gr}, S_{pr}) \quad (3.14)$$

- Segmentation Covering (SC): This measures the overlap between the predicted region, R_{pr} and ground truth region R_{gt} segmentation, [Arbelaez et al., 2010] by using:

$$O(R_{pr}, R_{gt}) = \frac{|R_{pr} \cap R_{gt}|}{|R_{pr} \cup R_{gt}|} \quad (3.15)$$

- Average Precision(AP): is the index to measure the ration of True positives to the sum of True and False positives. It can be calculated for a particular depth error and Intersection Over Union (IoU) of ground truth and predicted plane instances.

For plane reconstruction accuracy of plane reconstruction, average precision of plane instance detection at three depth error thresholds is considered : 0.4 m, 0.6 m ,0.9 m.

3.5.3 Visual Analysis

To visually depict the depth map generated from the neural network, we use a continuous jet colormap from Hunter [2007] with maximum value being 5 meters, while for error visualization we use a discrete interval inverse version of the same colormap. The error values are clamped at 1.8 meters. This has been depicted in Figure 3.7

4 | IMPLEMENTATION

4.1 DATASETS

In order to conduct our research, we use the publicly available established benchmark data-sets which provide real world RGB-D ground truth with rich annotations at indoor level and toolbox to do pre-processing. For each iteration of experiment, we need a RGB image, ground truth depth image for supervision, plane annotations which provides the pixel wise anchor normal id, intrinsic and extrinsic parameters of camera and laser scanner used for data collection. For training and validation, we use the Scannet dataset while for final evaluation we use both NYUv2 and Scannet datasets. For plane annotations, we use the benchmark data provided by [Liu et al., 2019].

- **ScanNet** : Presented in [Dai et al., 2017], there are 1513 annotated scans available for 707 different spaces such as classrooms, apartments, offices, apartments. They have 1205 scans for training and other 312 scans for testing. We use the second version to create a dataset consisting of 7000 images for training, 1000 images for validation, 800 images for testing purposes.
- **NYU-Depth** : There are two versions of v1 [Silberman and Fergus, 2011] and v2 [Nathan Silberman and Fergus, 2012] introduced in 2011 and 2012, respectively. The first one has 64 indoor scenes with 2347 RGBD images available for training and testing at 60-40 ratio respectively. The second version has 1449 RGBD images with pixel level labelling for 26 scene types. There are 795 images for training set and 654 images for the testing set. We use the second version whole test dataset for evaluation of our models.

4.2 PROGRAMMING ENVIRONMENT

In order to conduct experiments, the following hardware and softwares were used for implementation. For programming and inference purposes, a device with Ubuntu 18.04 having graphics card, NVIDIA QUADRO P1000 with 4GB GDDR5 on-board memory is used. For

conducting training and testing, High Power Computing cluster, TU Delft is used. For each experiment, a certain virtual environment using Conda or venv is created to load CUDA modules. Then, for deep learning and basic operations, Pytorch¹ and Python² are used. Some of the other dependencies such as Numpy³, Scipy⁴, OpenCV⁵ and Sklearn⁶ are also used. Open3d⁷ is used for processing and rendering 3D models and point clouds. We adopt the PlaneRCNN repository⁸ for our research framework and add further modules to implement our methods.

4.3 GEOMETRY AWARE DEPTH LOSS FUNCTION

To implement the loss function mentioned in Section 3.3, we first perform over-segmentation using Simple Linear Iterative Clustering (SLIC) algorithm [Achanta et al., 2010]. We choose this algorithm because the computation time is very less given its performance compares to other segmentation techniques [Achanta et al., 2012]. For segmentation algorithm, the two major parameters involved are the *number of segments* and *compactness* parameter. The number of segments controls the approximate amount of superpixels to be generated in the image while the compactness parameters maintains the balance between the color similarity and spatial proximity within regions [Achanta et al., 2010]. A depiction is provided in Figure 4.1 giving superpixels with different values of parameters to show the difference between segmentation. The segmentation connect the pixels in a particular superpixel to a segment id. We find the geometric centers of each superpixel and compute delaunay triangulation using them to create a graph network of superpixels. Then neighbors indices are extracted for each segment. A representation of the graph like structure of superpixels with neighbors is depicted in figure below. For choosing only those neighbors which have high color similarity we keep a restriction of 0.85 on the correlation value between the histograms. We choose correlation measure as it provides range from 1 to -1 with 1 indicating high similarity within histograms which allows us to control compatibility within the corresponding neighbors. For estimating depth values at superpixel level, we choose two representations to test. One is mean representation computed by using the

¹ <https://pytorch.org>

² <https://www.python.org>

³ <https://www.numpy.org>

⁴ <https://www.scipy.org>

⁵ <https://opencv.org>

⁶ <https://scikit-learn.org>

⁷ <http://www.open3d.org>

⁸ <https://github.com/NVlabs/planercnn>

average of depth values present inside the superpixel. Other is the depth value represented by the geometric center of the bounding box of the superpixel.

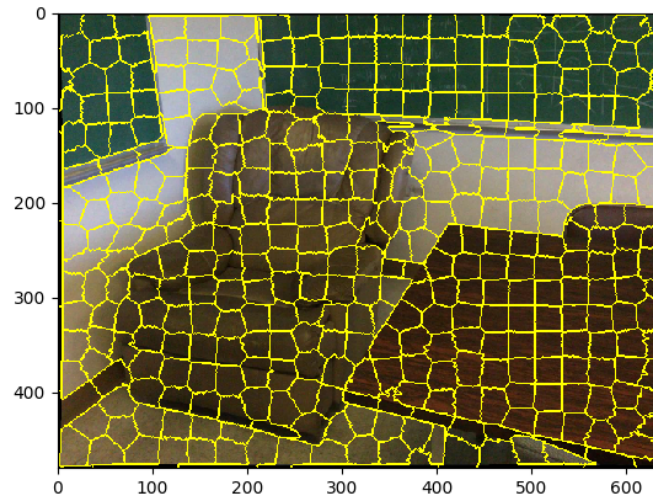
4.4 EXPERIMENT SETUP

4.4.1 Training Specifications

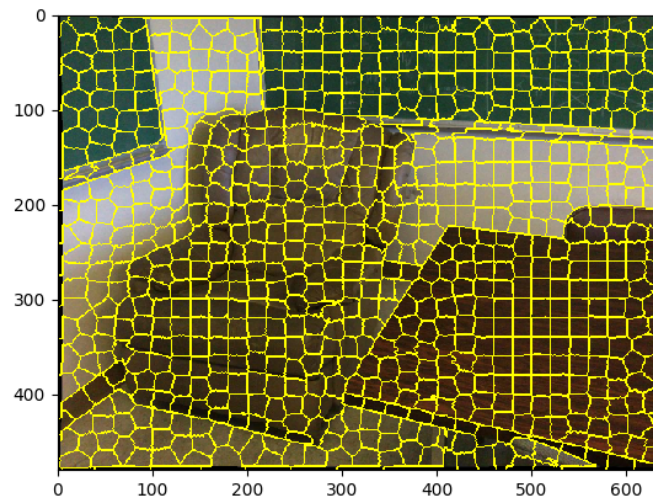
For training, we use the technique of transfer learning to load the pre-trained MaskRCNN weights to extract features from. Afterwards, we train all the layers of the model using randomly sampled 7000 Scannet training images with learning rate of 0.00001, momentum of 0.9, weight decay of 0.0001. A mini-batch of 15 images is used for faster training and stochastic gradient descent optimizer is used for achieving the convergence. The hyper-parameters are fixed using the performance on the validation dataset and convergence with respect to the training dataset. We fix the hyper-parameters and keep the procedure same for all our experiments and models for a fair comparison and do not fine-tune individual models. The configurations of all other variables involved in the neural network layers and convolution process are kept as set up in [Liu et al., 2019] and [He et al., 2017]. The set of available training images are kept same for all experiments with equal distribution from all type of indoor scenes such as living room, conference room, classroom and lounge.

4.4.2 Testing Specifications

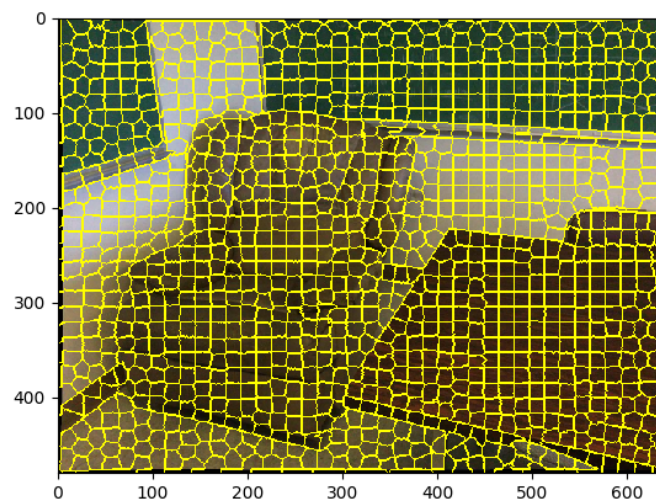
We evaluate the performance of our loss function, both quantitatively and qualitatively. The quantitative analysis will be done on the global depth, piece-wise planar depth reconstruction and plane instances. The qualitative analysis involves comparison of results using human eye which in ideal situation requires unbiased opinion from various users. But due to the limitation of time, we tackle this issue for comparing depth images by visualising the error difference between predicted and ground truth depth as shown in [Figure 3.7](#). This provides a better idea on how the models perform in comparison to each other. We also render the piece-wise planar model and point cloud composed of planar model coordinates along with non-planar points not present in the polygon surface model to investigate the individual effect of each term of loss function on different type of indoor scenes. We broadly create two types of datasets from Scannet dataset, defined as, curved and planar with set of 400 images each in the testing phase. The curved dataset has curved objects dominating the image while planar dataset has planar surfaces in most part of the scene. Apart



(a) segments=400



(b) segments=800



(c) segments=1200

Figure 4.1: Depiction of over-segmentation of image with different number of segments and same compactness

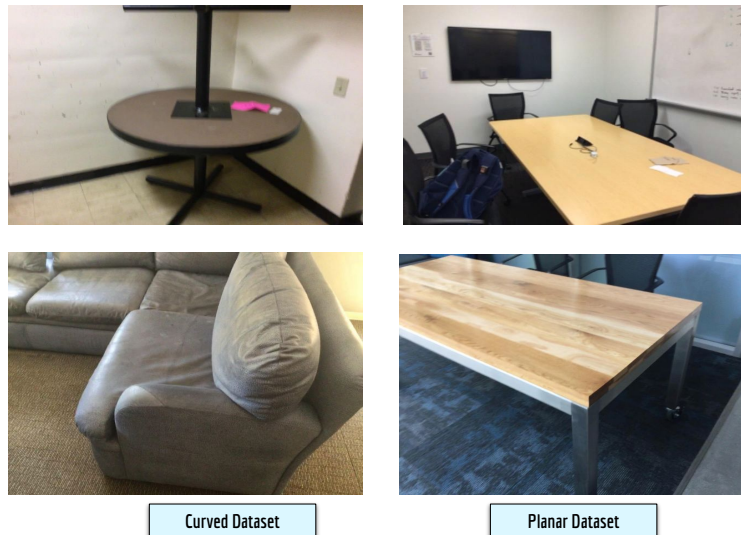














Figure 4.2: Testing dataset divided into two categories : curved and planar

from that we also test on NYU test dataset comprising of 645 images for the final evaluation although it is expected that the accuracy will be below due to low resolution of dataset and noisy ground truth values.







4.4.3 Model Specifications

To see the effect of mean and centroid representation of depth in loss terms, we conducted an initial experiment to determine the best representation for our terms. We test in total 7 models including the baseline. For both mean and centroid representation, three configurations are used : a) number of segments is 1200 and the regression is of type 1 b) number of segments is 1600 and regression is of type 1 c) the number of segments is 1200 and the regression is of type 2. To find out the right range of weight value in loss function we increase the weight as a step function from 0 to 1. This also helps us in seeing the effect of each term on different type of datasets. To tackle the issue of removing chance bias of model, we train the baseline and our model 3 times for final comparison. A full overview of the experiments is shown in Figure 4.3. For baseline, we choose the basic version of Liu et al. [2019] without the refinement and warping loss module and is represented by *b*. Further details of each experiment and results will be discussed in next section.

EXPERIMENT 1 : Test the options for regression at superpixel level to choose suitable representation for first and second term of loss function based on the potential for providing geometric awareness

Model Name	Predicted Depth	Ground Truth Depth	Description for representation for first term of loss function (L1)
<i>m</i>			PR-GT : Mean superpixel (N = 1200)
<i>c</i>			PR- GT : Centroid superpixel depth (N = 1200)
<i>m1</i>			PR-GT : Mean superpixel depth (N = 1600)
<i>c1</i>			PR- GT : Centroid superpixel depth (N = 1600)
<i>m2</i>			PR pixel depth - Mean GT Superpixel depth (N = 1200)
<i>c2</i>			PR pixel depth - Centroid GT Superpixel depth (N = 1200)

EXPERIMENT 2 : Test the effect of first and second term of loss function based on the weight balance with m2 representation for first term and both mean and centroid for second term
 $L = (1-w) L1 + w (L2)$

Model Name	L1		L2	Weight (k is representative of weight)
	Predicted Depth	Ground Truth Depth	Predicted Depth	
<i>w_k</i>				k= 0- 0.9
<i>wc_k</i>				k = 0, 0.1, 0.2

Legend for Representation




SYMBOL	Depth Value at Superpixel Level
	Mean of depth values of valid pixels inside superpixel
	Depth value of centroid of superpixel
	Depth value of pixels within the superpixel

Figure 4.3: Description of specifications of models used in the experiment

5

RESULTS AND EVALUATION

This Chapter presents the results and analysis of the conducted experiments during the research. Firstly, the results and analysis from each step in the process are provided to show the motivation for particular decision-making regarding the choice of particular parameters in our loss function. Then, the results from our best model are presented as compared to baseline. Afterwards, an evaluation of our method is done with respect to the evaluation metrics and visual comparison method provided in [Section 3.5](#).

5.1 RESULTS

5.1.1 Experiment with mean and centroid representation

In our first experiment, we trained the models with different representations mentioned in [Section 4.4](#) and depicted in [Figure 4.3](#). The objective of this experiment was two fold : a) choosing a suitable number of superpixels b) understanding the effect of each type of representation for picking suitable candidate for the first and second terms of our loss function. In [Figure 5.1](#), it can be observed that the mean representation for the predicted depth gives the essence of the structure with lot of edges while the centroid representation resulted in high smoothness. As the number of segments increase, both the mean and centroid models loose their influence on the structure of the chair depicted by *m_1*. Also, as the number of segments increase, the time required for pre-processing also increases. The 'c_2' model with second type of centroid representation has over-smoothing on the structure and it does not provide the outer curvature of the chair with sharp features. The assessment of the quality of the depth maps is also supported by the quantitative analysis shown in [Figure 5.2](#). In the curved dataset, the m_2 model has better accuracy along with the c_2 model, however, in the planar dataset, m_2 surpasses the later one in performance. It should also be noted that the curved dataset showed more sensitivity towards the different representations than the planar dataset.

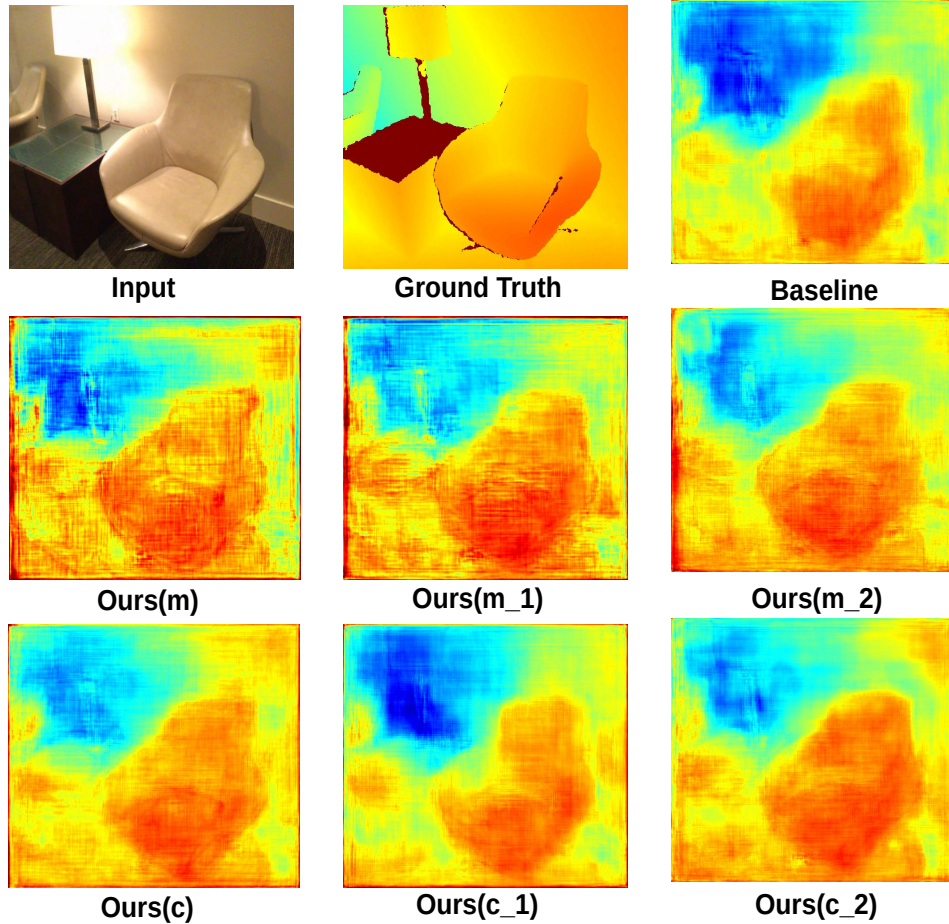
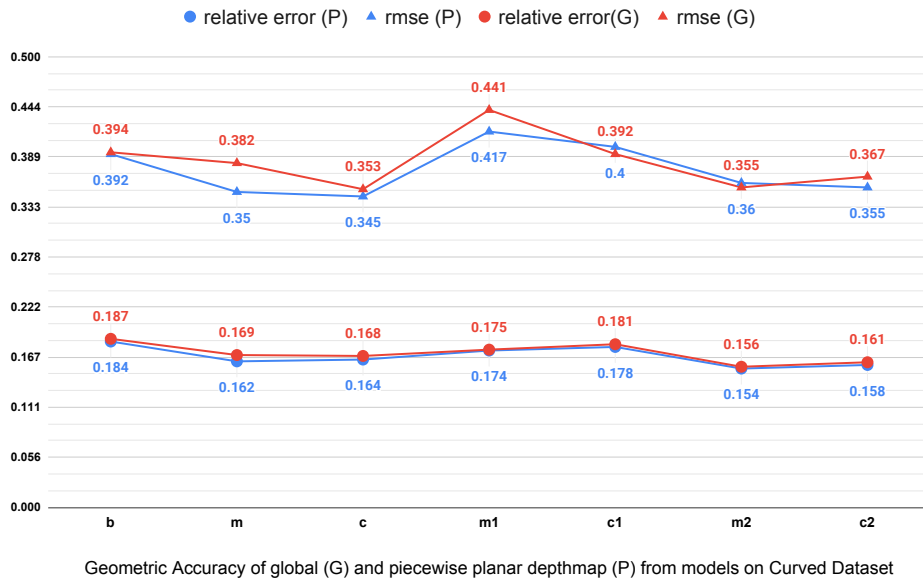
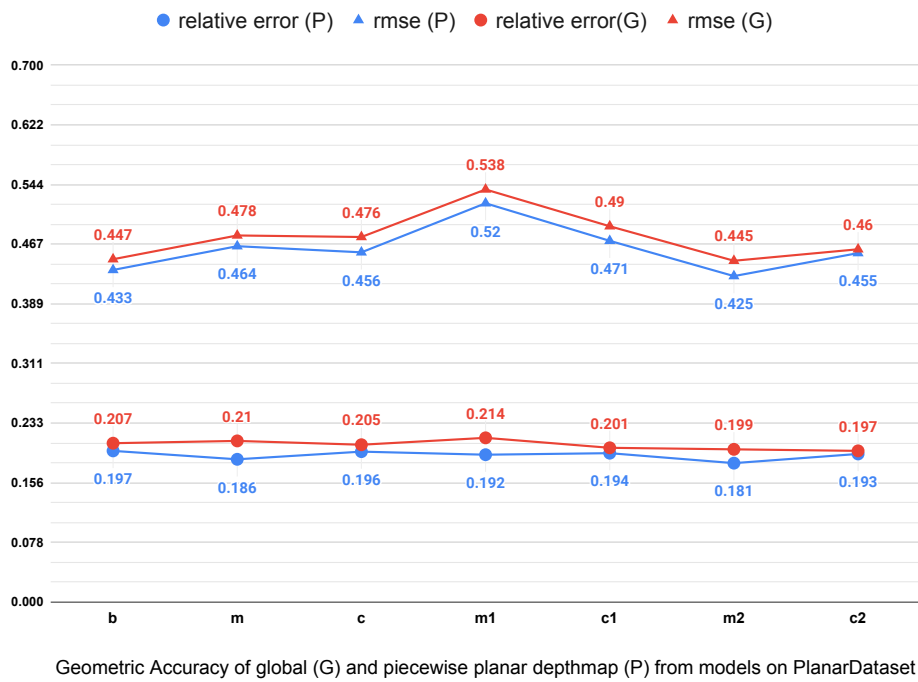


Figure 5.1: Comparison of predicted depth using different representation in our depth loss function.

From this experiment, we can observe that quantitatively and qualitatively, overall the 'm_2' model representing the depth loss is suitable for our first term. Visually, it maintains the balance between preserving edges and shape of the objects. It provides both the smoothness at surface level and sharpness around the boundary of the objects. For our second term, we need to provide the geometry awareness to the superpixel with respect to the local neighborhood and care for boundaries of the objects. This makes the mean representation, 'm' more suitable for providing depth consistency. In the next experiment, we fix the first term based on the analysis of this experiment and test for finding the right balance between the two terms by choosing a particular value of weight in our loss function.



(a)



(b)

Figure 5.2: Experiment 1 : Analysis of baseline and ours with different representation for first term. Lower values indicate better performance for all metrics

5.1.2 Experiment with different weights of loss function terms

After fixing our first term of loss function in Equation 3.1, we tested the effect of our second term of the loss function on depth estimation to find out a suitable value of weight parameter based on the geometric accuracy. We increase the weight from 0 to 0.9 for our loss term defined in Equation 3.3 and evaluate the performance both qualitatively and quantitatively. An example of the error between predicted and ground truth depth is shown in Figure 5.3. We can see that with only the first term, the error is slightly reduced on the sofa arm. As the second term with weight 0.1 is used, the error further decreases around the sofa surface as well as in the second sofa behind the first sofa. As we further increase the weight, the error starts to increase on the sofa surface as well as on the floor, becoming similar to baseline method. This indicates the high influence of the second term on the depth reconstruction.

From Figure 5.4, we observe that, both in the planar and curved datasets, the relative error is least when weight is 0.1. For the curved dataset, the root mean square error is reduced both in piecewise planar and global depth as compared to the baseline. The accuracy of the model reduces as the weight increases from 0.1 to 0.9. For the planar dataset, this reduction is less as compared to the curved dataset. For the planar dataset, the gap between the piecewise planar and global depth is higher for planar dataset than for the curved one. This shows that the planar segmentation further improved the depth estimation in the planar dataset. Since the curved objects have less planar instances the depth is also less affected by the segmentation. After finding out this behaviour, we confirm the second term by training the model with centroid representation in second term with weights 0.1 and 0.2 and compare it to the mean representation. It can be seen in Figure 5.4 that the performance decreases with respect to the baseline method and mean representation at 0.1 and 0.2 in the curved and planar datasets.

Keeping in mind the prior observations and results, we fix our first term with 'm_2' representation, and second term with 'm' representation and then, train the model with 0 and 0.1 weight, three times for comparison.

5.1.3 Final Results

We use the model trained with our geometry aware loss function having weight 0.1 to generate the depth maps and plane instances. Using plane parameters, masks and depth information, a piecewise planar model is reconstructed along with the point clouds using equations

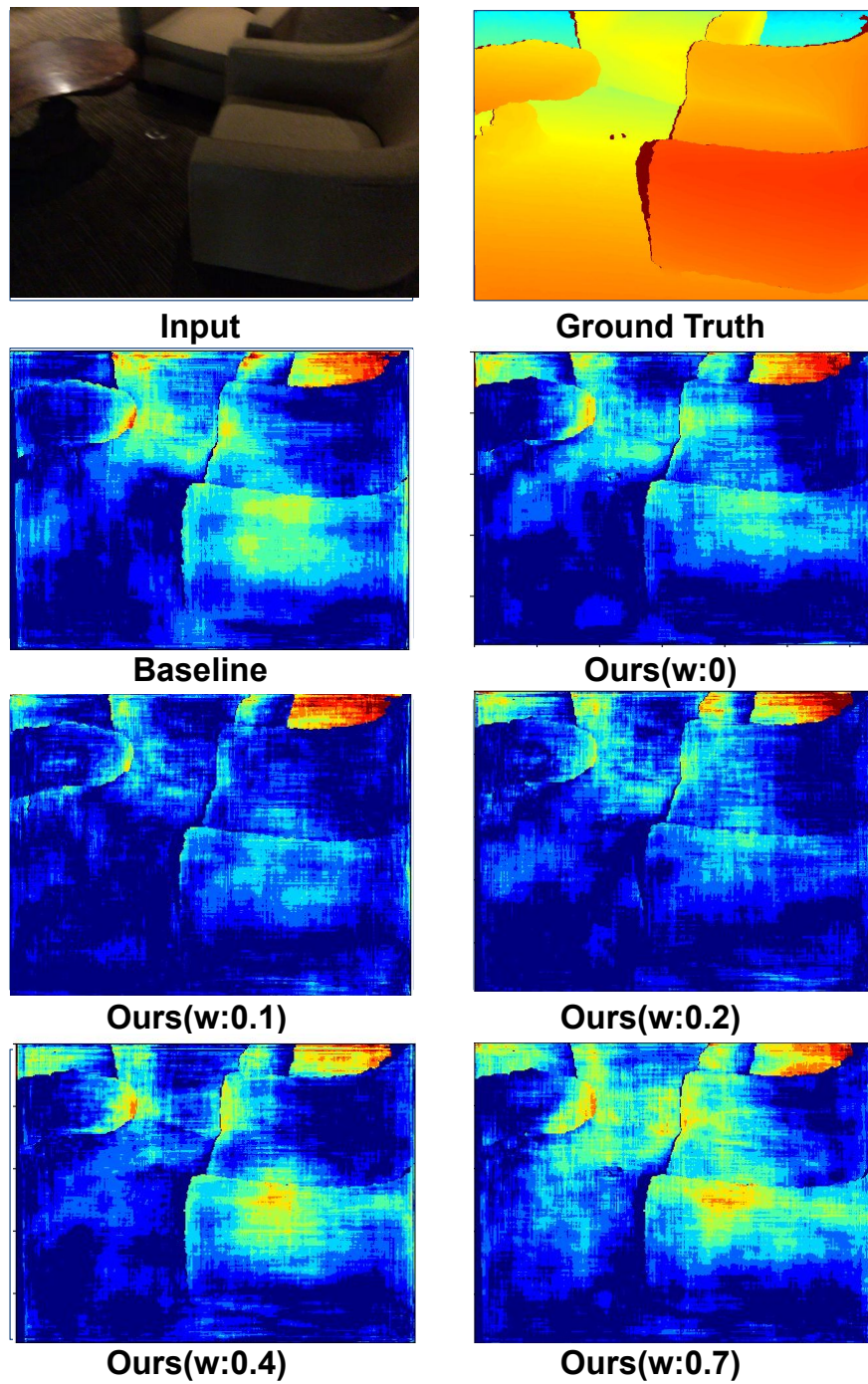
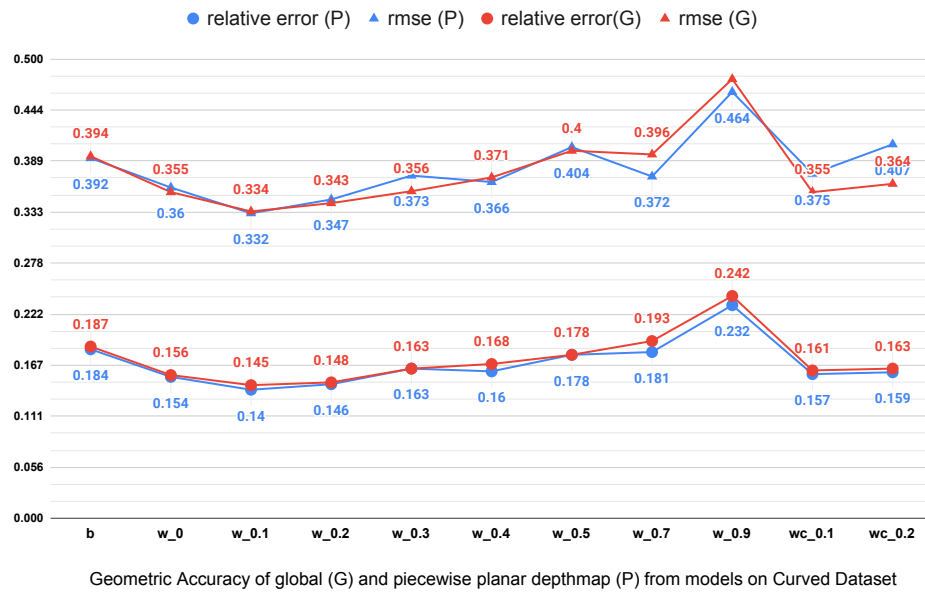
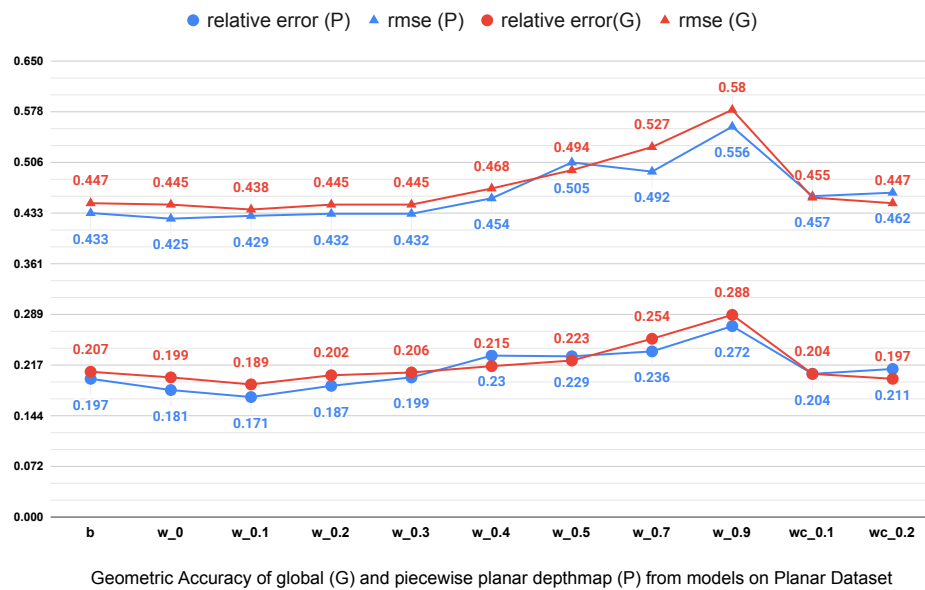


Figure 5.3: Comparison of predicted depth using different weights in our depth loss function.



(a)



(b)

Figure 5.4: Experiment 2: Analysis of baseline and ours with increasing weight in energy function. The relative error and Root Mean Square Error (RMSE) should be lower for better performance. (a) Model with weight 0.1 performs best on curved dataset with 0.2 close to it (b) Model with weight 0.1 performs best but is very close to baseline.

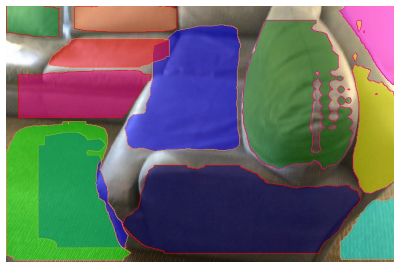
mentioned in [Section 3.4](#). An example of the full pipeline is depicted in [Figure 5.5](#). It can be observed that the piecewise planar model is highly influenced by the plane instances of the sofa and does not provide good representation of the curved surface of the sofa. The 3D point cloud instead provides better representation of the surface of sofa. Also, it can be seen that the reconstruction is noisy and does not provide the best surface representation. Still, the 3D point cloud combined from plane instances and non-planar depth provides balance between highly planar structure and full non-planar point cloud for the given image.

Few more examples of the depth maps estimated from our model have been shown in [Figure 5.6](#) and [Figure 5.19](#). In the first figure, it can be observed from the left column that the circular curvature of the table is better estimated in our model. The baseline method provides a limited representation of the table missing the round curvature as well as the cylindrical stand on the table. In the second image, the sofa has better outer curvature and representation of the arms. The baseline in this case misses the joint connection between arm and head of sofa. In figure [Figure 5.19](#), the left column result shows the depth estimation for a table. It can be observed that there is a slight improvement on the edges of the table and the leg of the table. From right column, it can be seen that the chairs and table surfaces are better represented in our case wherein, the structure is reconstructed around the thin legs of the chair.

The piece-wise planar surface models and point clouds for few images are shown in [Figure 5.8](#) and [Figure 5.9](#). From [Figure 5.8](#), it can be observed from the 3D models that the curved curvature of the table is broken and goes inside the wall due to high error, while in our case, it is reconstructed as an individual curved curvature. Similarly, in [Figure 5.9](#), it can be observed that in the 3D models, the table in baseline is reconstructed wrongly and goes inside the chairs while in our case, there is a clear boundary and good orientation of planar surface. From 3D point clouds, it can be seen that the non-planar region is quite noisy and does not provide good understanding of the scene while in our case after using our consistency term during 3D Reconstruction, patches have been reconstructed along with the original planar model.



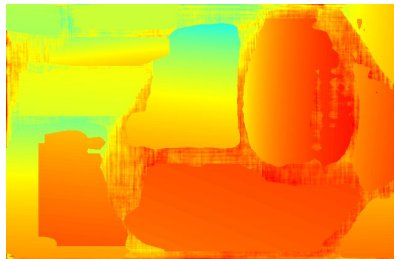
Input



Plane Instances



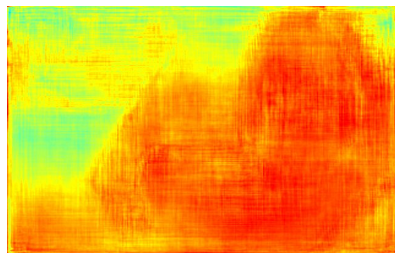
Piecewise planar Model



Piecewise planar depth



3D Point Cloud



Global Depth



3D Point Cloud

Figure 5.5: A depiction of 3D reconstruction pipeline with an example

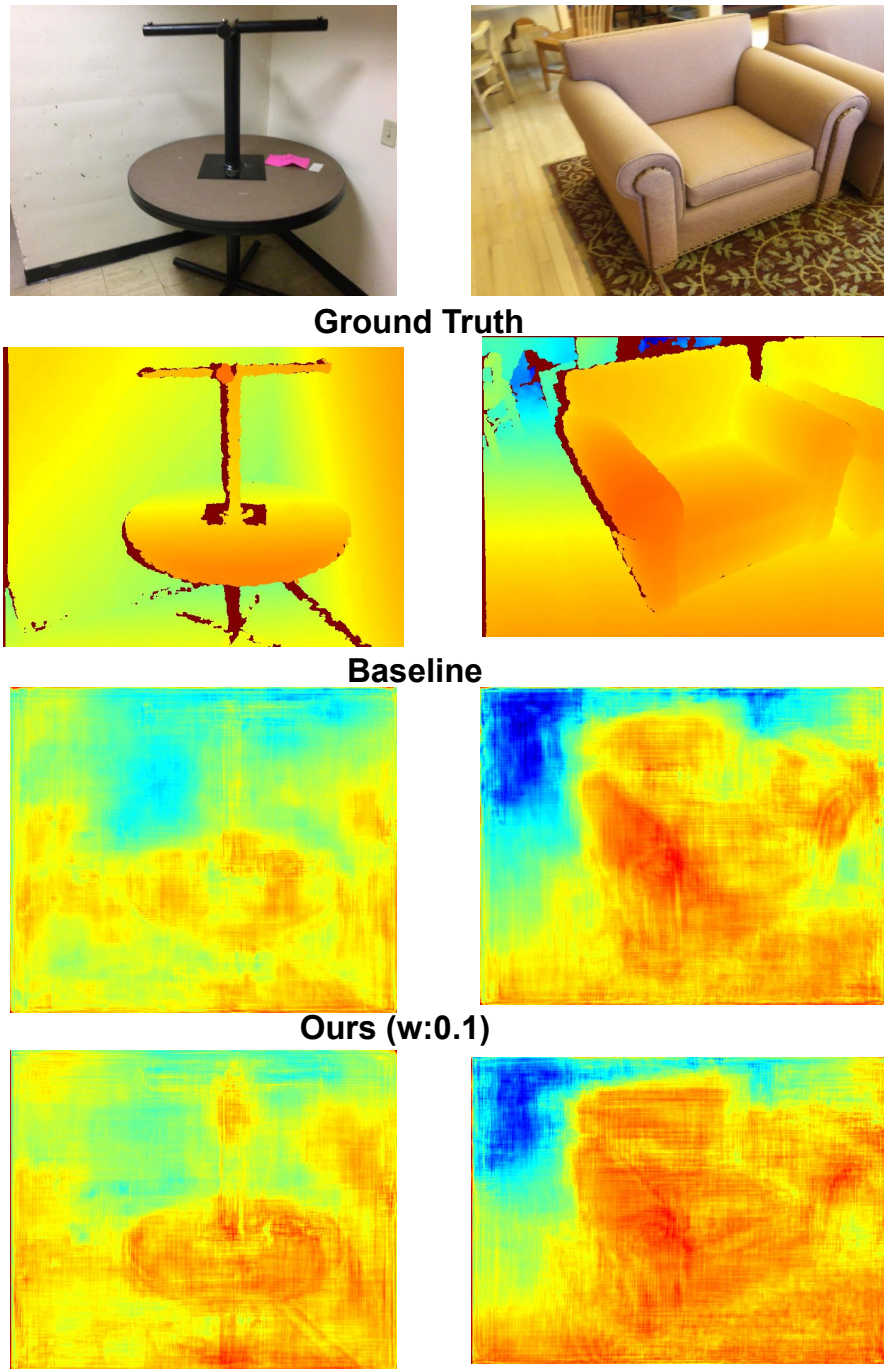


Figure 5.6: Comparison of depth map estimated from our model and baseline on curved dataset. There is more robustness in our model, when it comes to completing the curvature of the objects during reconstruction, as compared to the baseline.

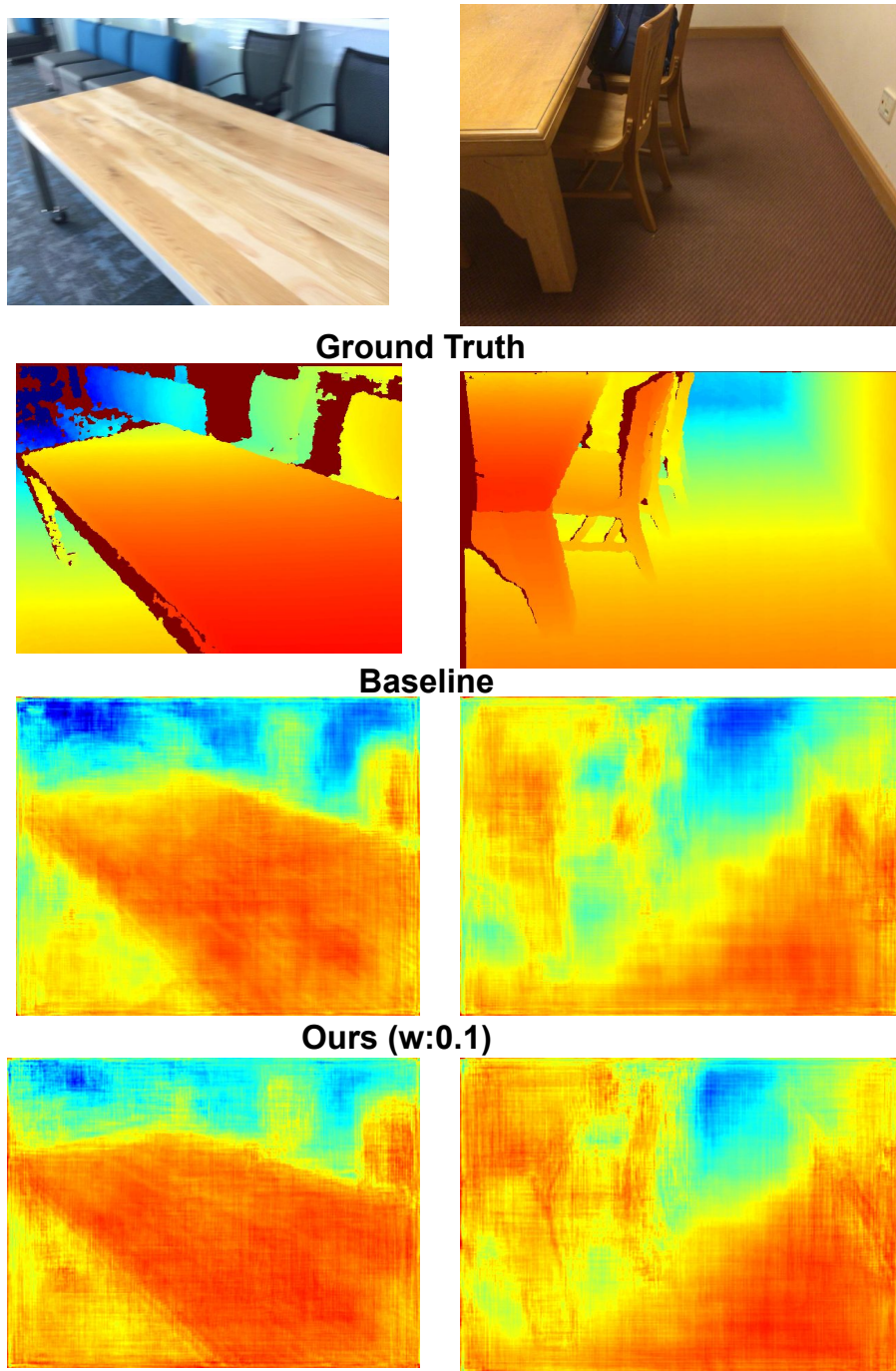


Figure 5.7: Comparison of depthmap estimated from our model and baseline on planar dataset. There is better sharpness in bringing out the edges of the planar surfaces from our model as compared to baseline.

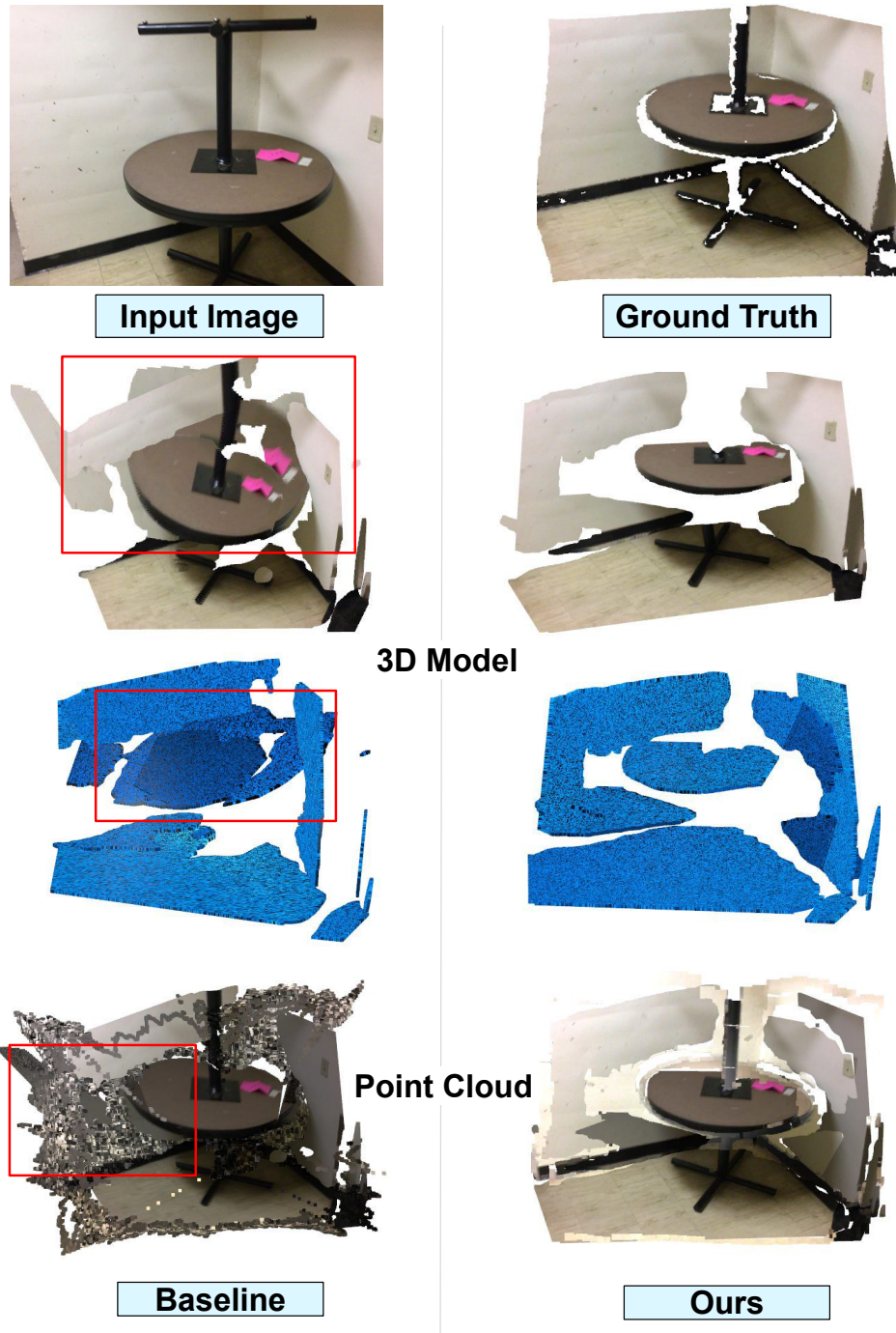


Figure 5.8: Comparison of piecewise planar model and point clouds from our model and baseline. It can be observed from the 3D models that the curved curvature of the table is broken and goes inside the wall due to high error, while in our case, it is reconstructed as an individual curved curvature



Figure 5.9: Comparison of piecewise planar model and point clouds from our model and baseline. It can be observed that in the 3D models, the table in baseline is reconstructed wrongly and goes inside the chairs while in our case, there is a clear boundary and good orientation of planar surface.

5.2 EVALUATION

We tested on the created Scannet test dataset and NYU dataset, mentioned in the [Section 4.1](#) and using metrics defined in [Section 3.5](#). We tested the geometric accuracy and plane reconstruction quality depending on the dataset quality.

Quantitative

The statistical comparison of the model with our loss function with respect to the baseline is shown in [Figure 5.10](#). Detailed statistics are shown in [Appendix B](#). For the Scannet curved dataset, our model reduces the relative error by 17% and [RMSE](#) by 10% with respect to the baseline method for piecewise planar depth. For global depth estimation, this reduction is 20% for relative error while 13% for [RMSE](#). There is overall gain of 9% in accuracy of global depth map with respect to a depth threshold(1.25). For the planar dataset, the reduction in relative error is approximately 16% for both global and piecewise planar depth. However, the [RMSE](#) is very close to the baseline with only 2% reduction for both the depth maps. There is gain of 10% in the accuracy of piecewise planar depth maps. On NYU dataset, our loss function leads to reduction in the relative error in piecewise planar depth by 11% with both our terms and 6% with only first term. The accuracy in general for NYU dataset is less for all models as compared to the Scannet dataset. This can be attributed to the low resolution of NYU dataset as well as it not being used during the training procedure.

For plane instance detection, the metrics comparison can be observed from [Figure 5.11](#). The segmentation quality is similar to the baseline method. The average precision of the plane reconstruction at depth threshold of 0.4 meter increases by 33% in curved dataset while 20% in planar dataset by using weight 0.1. This increment reduces as the error threshold increases to 0.9 meter. This shows that the reconstruction accuracy improves as the error threshold reduces.

Overall, it is observed that there is greater effect of proposed loss function on the non-planar regions than the planar regions. The first term is more effective alone on the curved dataset, while second term gives improvement irrespective of the dataset. However, the quantitative evaluation does not necessarily indicate the quality of 3D reconstruction.

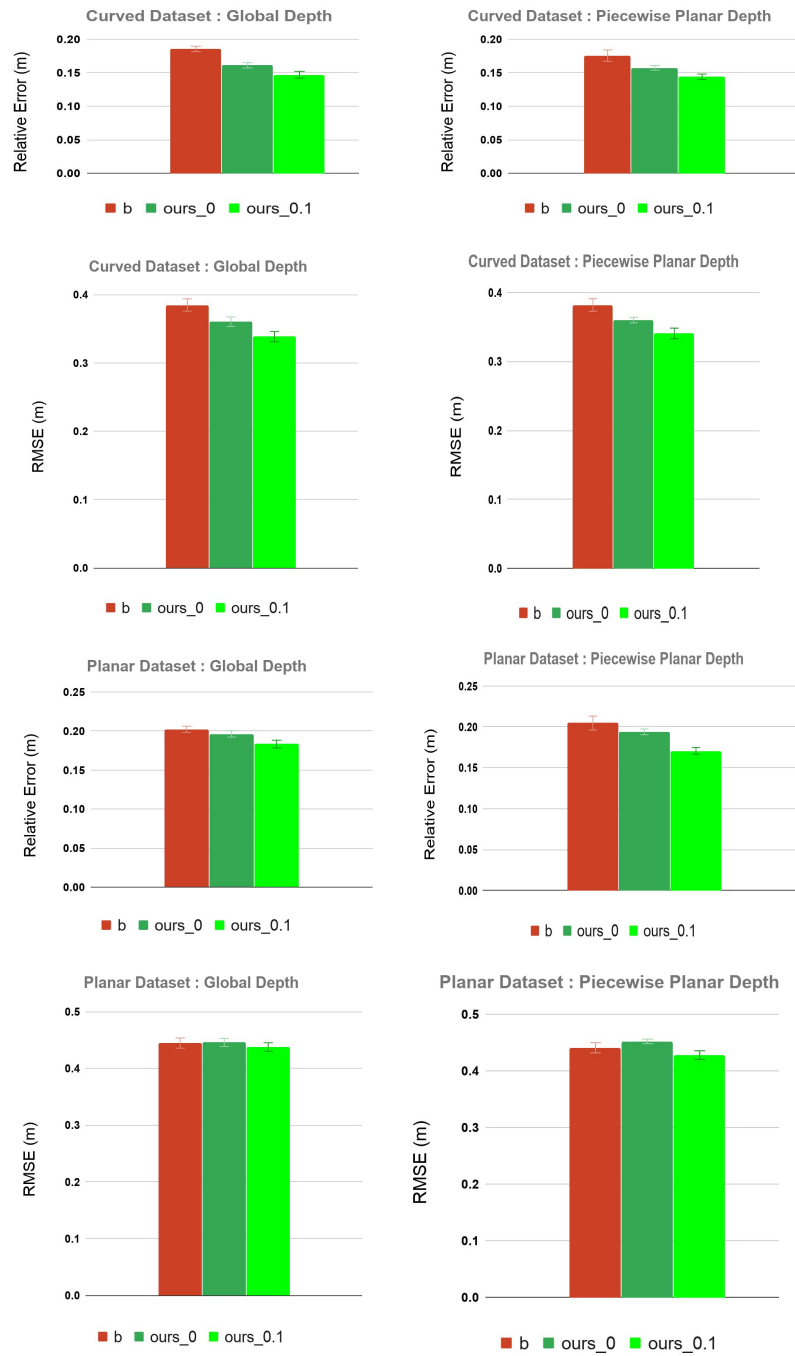
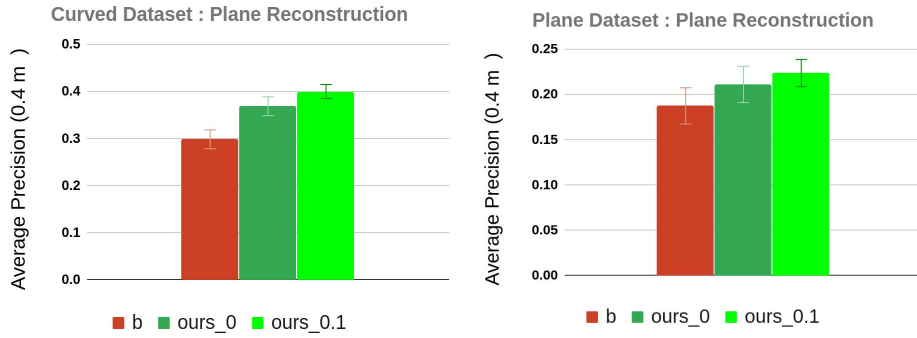


Figure 5.10: Geometric Accuracy Metrics for baseline and ours with 0.1 weight. Mean with standard deviation of the three times run experiment with created Scannet test datasets are reported



Scannet Curved Dataset : Mean values of 3 times run experiment						
Models	Plane Segmentation			Plane Reconstruction at certain depth error threshold		
	RI (high)	VOI↓(low)	SC(high)	AP0.4m(high)	AP0.6m(high)	AP0.9m(high)
b	0.812	2.07	0.578	0.298	0.468	0.484
ours_0	0.806	2.109	0.575	0.369	0.471	0.478
ours_0.1	0.819	2.022	0.589	0.4	0.484	0.487

Scannet Planar Dataset : Mean values of 3 times run experiment						
Models	Plane Segmentation			Plane Reconstruction at certain depth error threshold		
	RI (high)	VOI↓(low)	SC(high)	AP0.4m(high)	AP0.6m(high)	AP0.9m(high)
b	0.813	2.253	0.547	0.187	0.304	0.328
ours_0	0.81	2.254	0.546	0.211	0.293	0.319
ours_0.1	0.814	2.21	0.553	0.224	0.286	0.308

NYU Dataset : Mean values of 3 times run experiment													
Model s	Piecewise Planar Depth					Global Depth					Plane Segmentation		
	rel (low)	rel_sq rt(low)	rmse (low)	rmse_log	a1 (high)	rel (low)	rel_sq rt(low)	rmse (low)	rmse_log	a1 (high)	RI (high)	VOI↓ (low)	SC (high)
b	0.33	0.406	0.949	0.349	0.497	0.334	0.416	0.961	0.349	0.491	0.309	2.428	0.405
ours_0	0.307	0.363	0.935	0.36	0.491	0.312	0.373	0.949	0.369	0.485	0.336	2.334	0.435
ours_0.1	0.291	0.335	0.944	0.342	0.473	0.295	0.344	0.943	0.341	0.469	0.319	2.407	0.415

Figure 5.11: Geometric Accuracy and Plane Detection Metrics for baseline and ours with 0.1 weight. Mean with standard deviation of the three times run experiment on created Scannet test dataset and full NYU test dataset are reported

Qualitative

A visual comparison of global and piecewise planar depth in various cases has been shown in [Figure 5.12](#), [Figure 5.13](#), [Figure 5.14](#) and [Figure 5.15](#). From error analysis, it can be observed how both the terms affect the depth reconstruction in curved and planar scenes. The first term helps in reducing the error at global level while the second term acts at local level. In [Figure 5.12](#), there is high error around the curvature of sofa cushion as well as on the floor. With first term, the error on cushion is diminished while second term reduces the error on the floor and lower curvature of sofa. The piecewise planar depth is affected proportionally by this reduction in error. This reflects the effect of the global depth estimation on the reconstructed piecewise planar depth. Similar behaviour is seen in other cases.

A visual comparison of 3D Reconstruction is shown in [Figure 5.16](#) and [Figure 5.17](#). It can be seen in [Figure 5.16](#) that using only pixel level loss, the baseline model creates an inaccurate reconstruction of the table surface. By using only the first term of our loss function, there is improvement in the extent of the table surface, however, the floor and table surfaces are stitched together and the orientation of table surface with respect to the sofa in the behind is wrong. By using both terms in loss function, there is improvement in both orientation and extent of the table surface. Furthermore, it can be observed from the red markings in [Figure 5.17](#) that the curvature of the round table is better represented with both terms of our loss function. With only the first term, there is improvement in the round curvature but there is a false surface nearby due to high error around the edges, while in the baseline model, the whole table is broken and going inside the wall. A comparison of the point clouds generated by combining planar coordinates from the model and non-planar depth for other regions is shown in [Figure 5.18](#). It can be observed that there is lot of noise present in the reconstructed point cloud of baseline and overall structure is not maintained. This issue is resolved by using the consistency term to reorient the point normals based on generated superpixels during the 3D reconstruction.

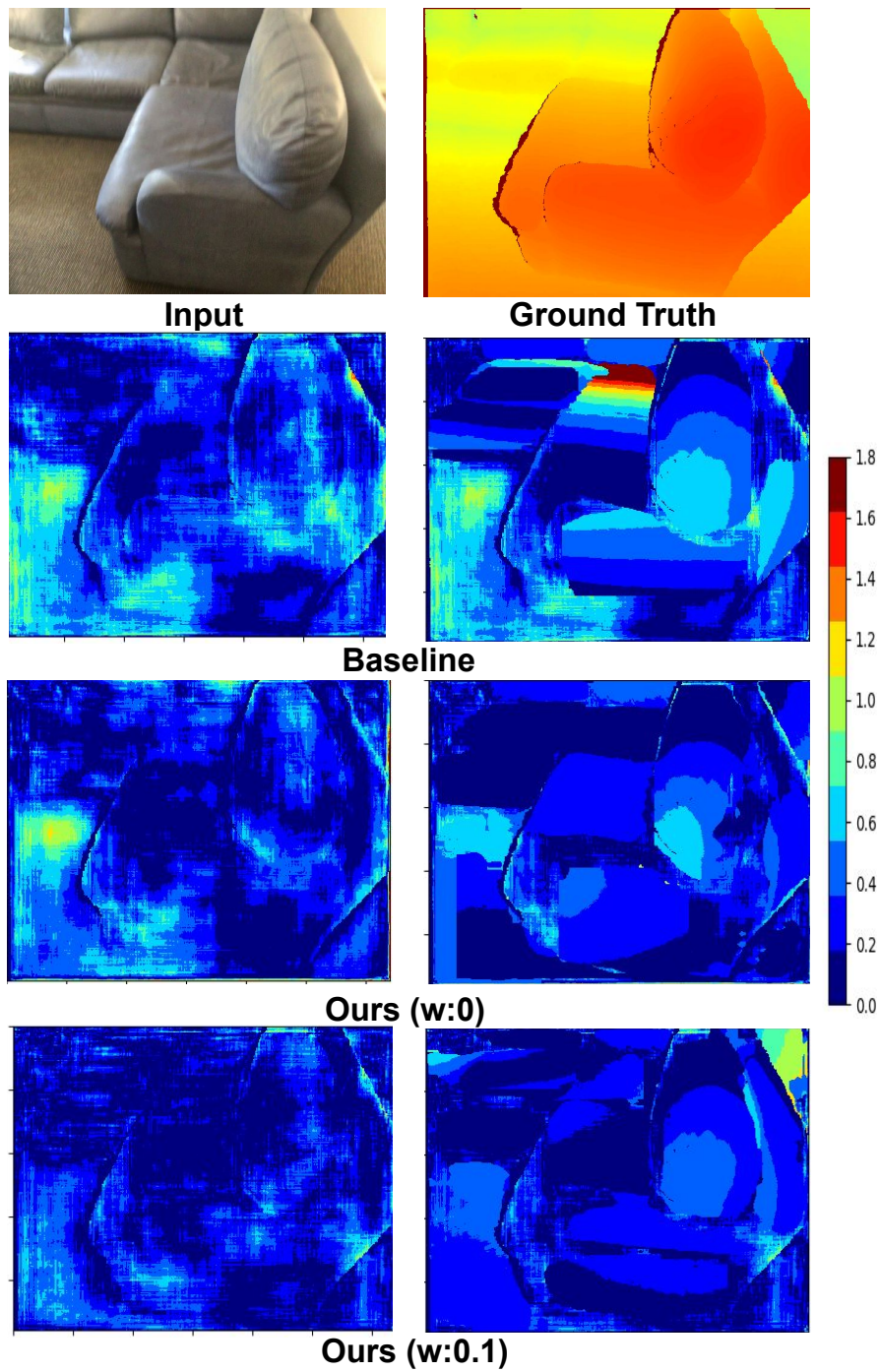


Figure 5.12: Error visualization in comparison with baseline with global depth on left and piecewise planar depth on right side

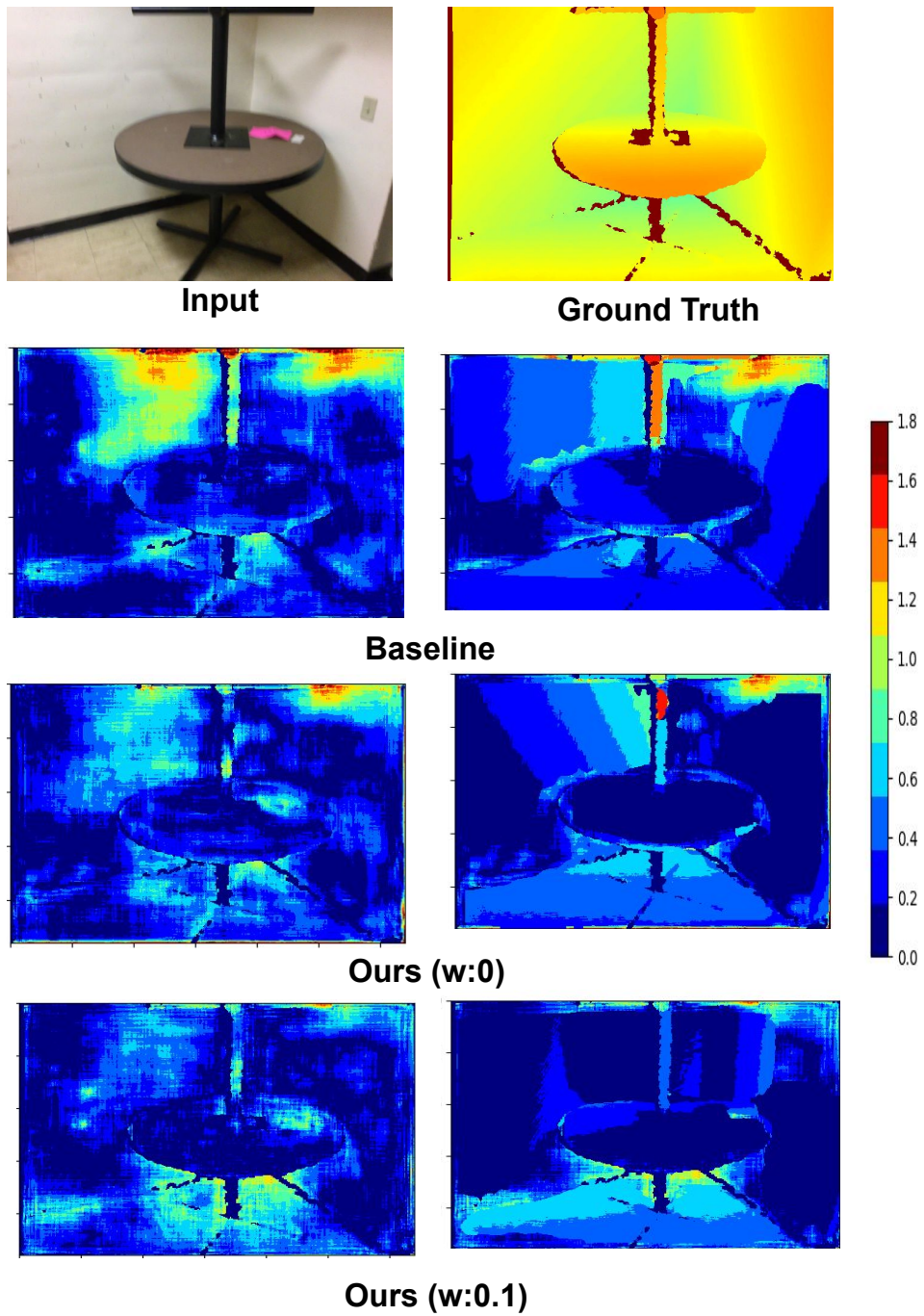


Figure 5.13: Error visualization in comparison with baseline with global depth on left and piecewise planar depth on right side

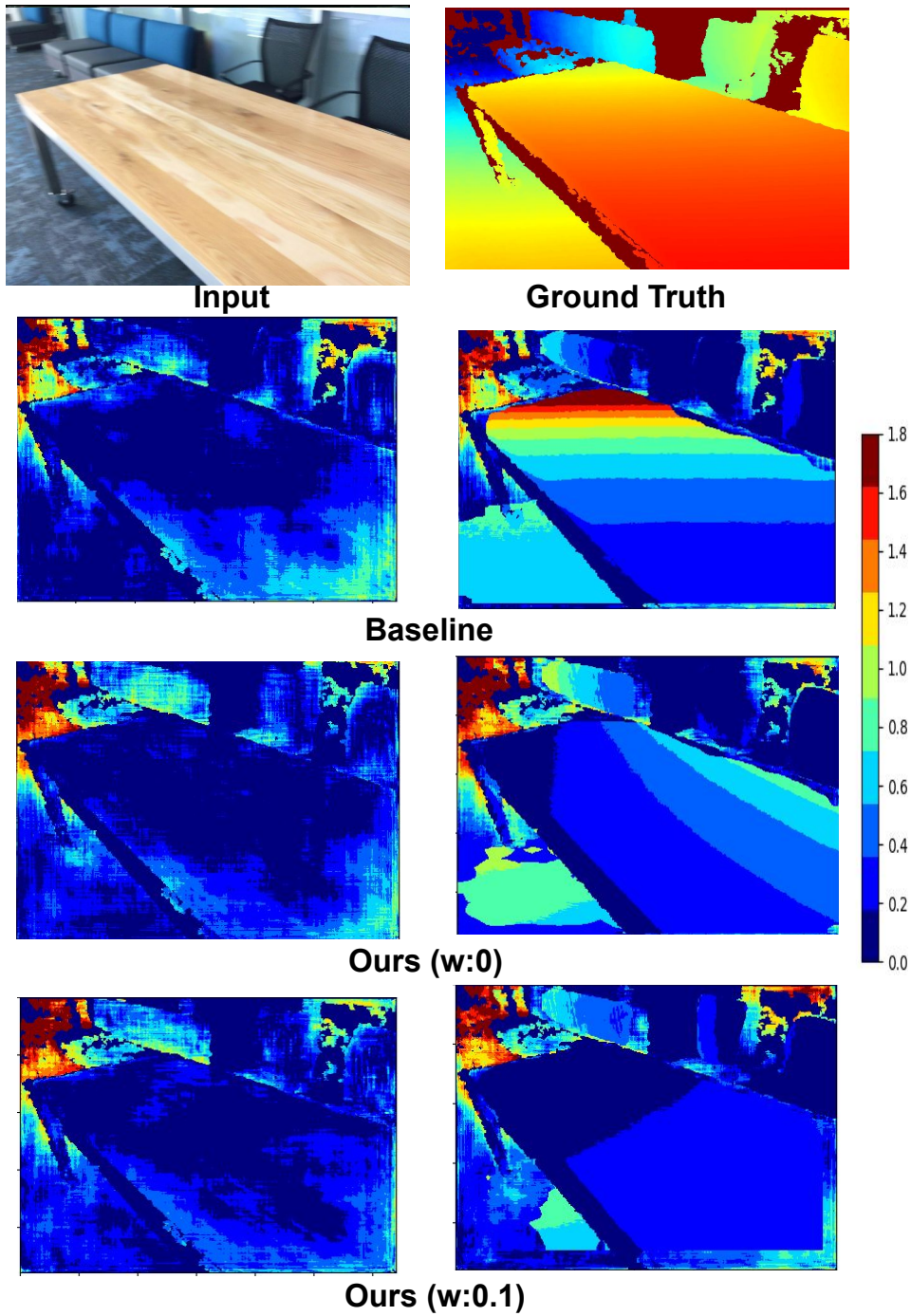


Figure 5.14: Error visualization in comparison with baseline with global depth on left and piecewise planar depth on right side

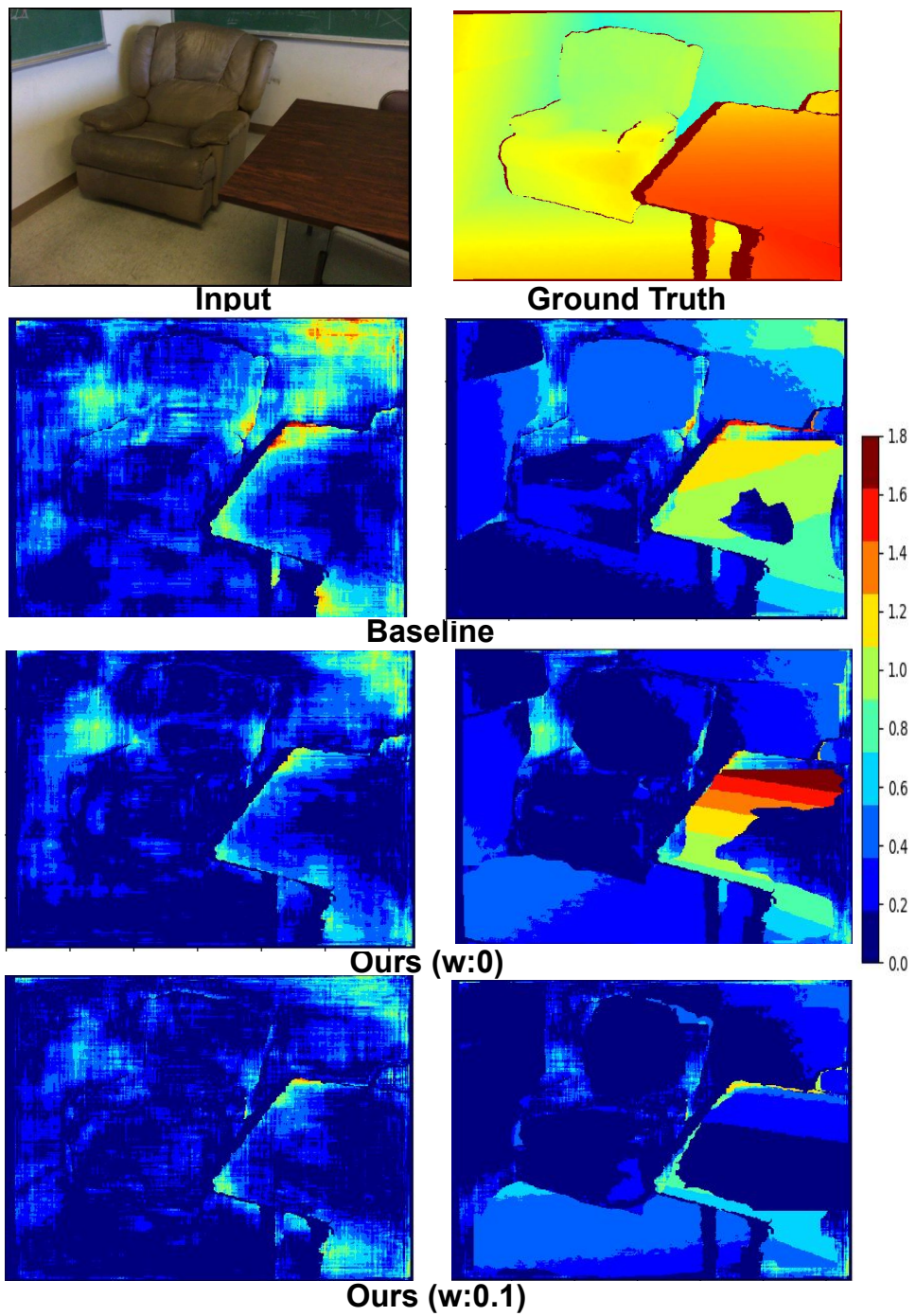


Figure 5.15: Error visualization in comparison with baseline with global depth on left and piecewise planar depth on right side

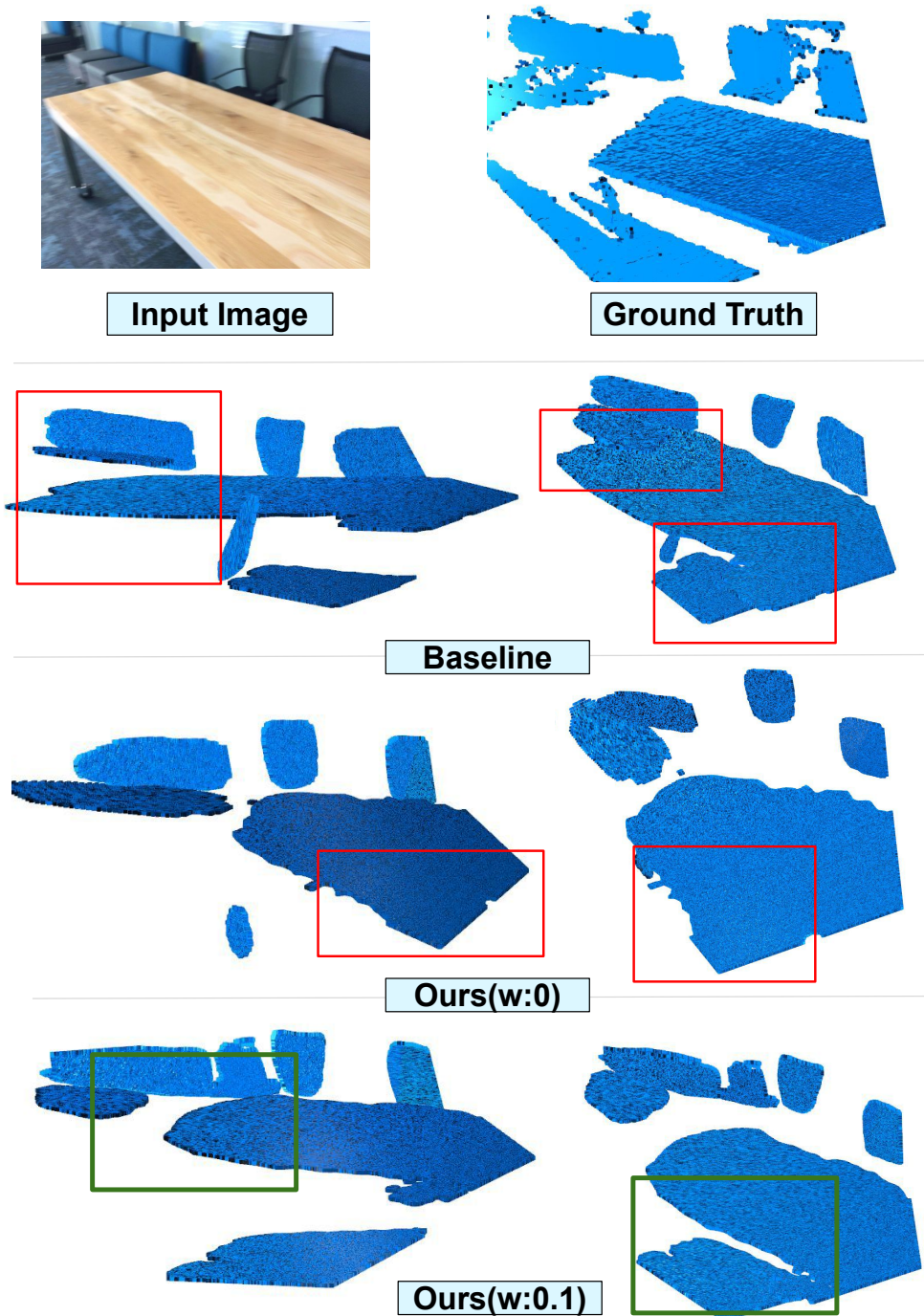


Figure 5.16: Comparison of piecewise planar models of baseline and ours with different weights with side view on the left hand side and front view on the right side. It can be observed from the red squares that there is improvement in orientation and extent of the table surface with our both terms. With only pixel level loss, the baseline creates an inaccurate reconstruction of table surface.

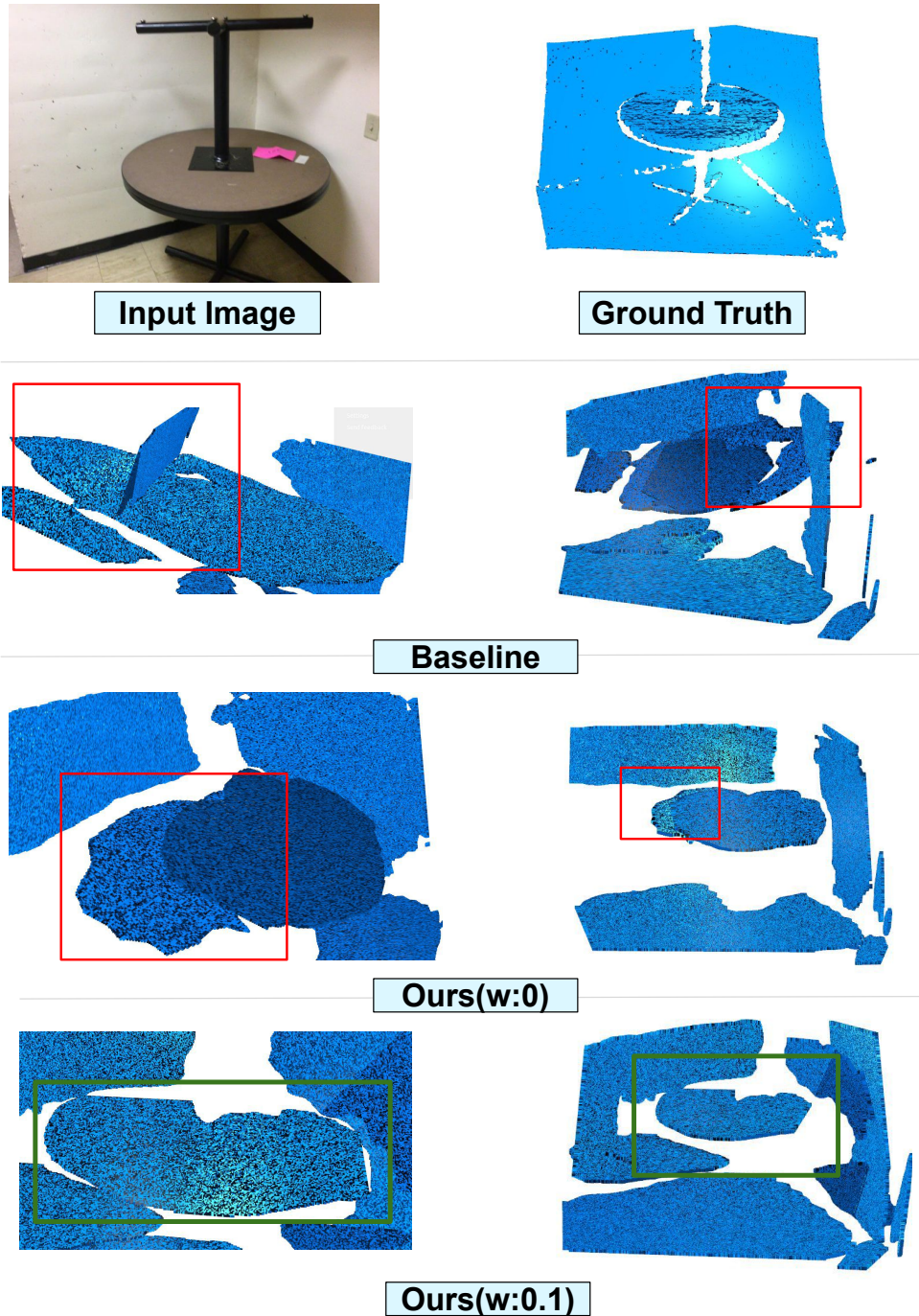


Figure 5.17: Comparison of piecewise planar models of baseline and ours with different weights with side view on the left hand side and front view on the right side. It can be observed from the red markings that the curvature of the round table is better represented with both terms of our loss function. With only first term, there is improvement in the round curvature with another false surface is reconstructed nearby while in baseline the whole table is broken and going inside the wall.

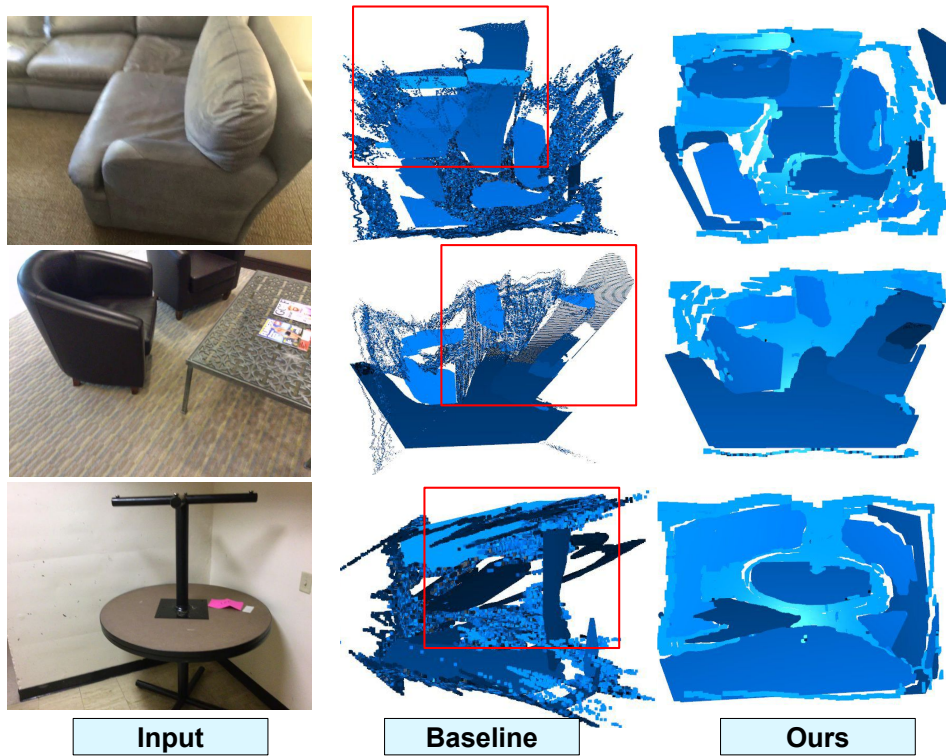


Figure 5.18: Comparison of point clouds from baseline and ours with both terms after using consistency term during 3D Reconstruction. It can be observed that there is improvement in scene understanding from non-planar regions in our case with respect to the baseline, apart from the improvement in planar structure. There is inconsistency in non-planar regions in the original point cloud which is partially resolved using consistency term.

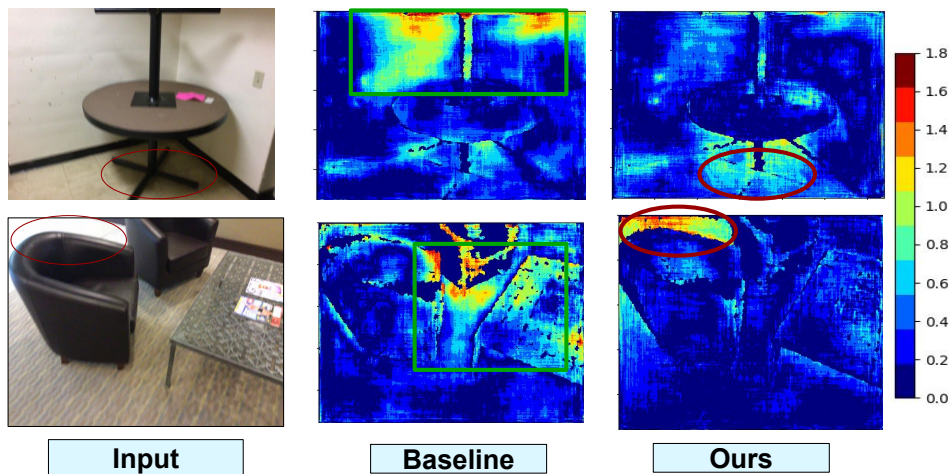


Figure 5.19: Depiction of limitations from our loss function by comparing the error in baseline and ours. The top input image. From Left to Right: input, baseline and ours

5.3 LIMITATIONS

- When there is a region in the image where it is difficult to distinguish the object boundaries, then there is a high error occurring within that area. In [Figure 5.19](#), two examples have been provided to showcase the limitation of our loss function. In the top row, it can be seen that the error on the walls, table and cylinder stand decreases as represented by green box on baseline. However, the error in the lower part of the legs of table increases where the legs are under the shadow of table. This may be attributed to the effect of our second term which influences the nearby depth values to be closer to each other based on the color consistency. In another case (bottom row), the error on the edges of sofa and table surface decreases, however, there is high error in the immediate background of the first sofa. These cases give cue for further investigation on using superpixels in loss functions.
- We do see improvement with our optimization approach based on superpixels and explore the effect of individual terms on the 3D reconstruction, however, within the scope of this project, the exact reason why there is improvement is not explored. To do this research, several propositions will need to be tested to pinpoint the reason. One speculation is that there is tackling of noise present in ground truth data collected by laser scanners. There are several types of noise in an acquired 3D dataset such as range error, instrument error and error due to surface properties. These become points with holes or missing information, and are not used in training. To investigate this argument, one needs to do experiments with indoor synthetic dataset such as one provided by [\[McCormac et al., 2017\]](#) and choose a particular noise model to insert noise in the perfect dataset. Creating a noise model will itself require further research as traditional approach of Gaussian noise will not work in this case. One will have to develop a statistical model such as one proposed in [\[Sun et al., 2008\]](#). Moreover, it can be further beneficial to use a refined real world dataset, instead of synthetic dataset as there is generally inconsistency between results when using the synthetic versus real world dataset [\[Sun et al., 2008\]](#). Such completed dataset can be generated by using deep learning techniques such as one proposed in [\[Zhang and Funkhouser, 2018\]](#).
- The non-planar region in the scene is biased towards the camera coordinate system. From the image perspective, the 3D reconstruction is refined but from another view, the quality of the point cloud is not good. This can be observed in [Figure 5.8](#)

and [Figure 5.18](#). One reason is the small size of training dataset used to learn the features. If more images are used and training is done on a large dataset, the noise is expected to decrease and quality of 3D reconstruction may be improved. If warped loss module proposed in [\[Liu et al., 2019\]](#) is used, then, further improvement is expected.

- In our research, we choose a particular neural network architecture based on the availability of time and the reproducibility of the code for training and testing. Using other neural networks with similar loss function could be beneficial for further research and testing the effect of the extent of receptive field on 3D reconstruction.
- The training time taken by the loss with both terms is three times than the baseline. This can be reduced by saving the ground truth metadata, neighbors and superpixel segmentation for the loss function in the training dataset. This would require higher storage size but will save computing time during training. The algorithm time can be reduced by using CUDA during segmentation as well as for finding neighbors if there is compatibility of the programming environment of the deep learning model with available modules. This was not possible in current scenario.
- Due to the restriction of time and computing power available for the chosen framework, a small training and testing dataset is used. This can be increased if more power is available. It will give further insights on the parameters of loss function and validation on the generalization of neural network under different conditions. Also, the optimum values of hyper-parameters is dependent on the experimentation by a particular individual, making it prone to bias.
- The terms of the loss function use a particular superpixel segmentation algorithm and histogram comparison method due to less computational effort and relatively better outcome. However, the limitations of these techniques are also inherited in the process such as problems in accurate boundary regions. A better segmentation and a color comparison technique can be used at the cost of execution time. This will bring out better insights on the process.
- There is an inherent bias in this kind of research, where visual analysis and tuning of parameters is required. Also, there is a constant struggle between qualitative versus quantitative analysis, and human vision versus computer vision. A third party survey of the 3D reconstruction results can be conducted to tackle these issues, if time and logistics are available.

6 | CONCLUSIONS

6.1 DISCUSSION

Our research objective was to investigate the effect of the proposed optimization approach on the 3D reconstruction process and answer the following research question :

“Can optimization based on the spatial and color compatibility of pixels within image, help in the improvement of 3D reconstruction from a single image?”

Based on the quantitative and qualitative analysis, we observe that the proposed approach helps in improving the 3D reconstruction from a single image in indoor environment. This improvement is attributed to both terms in the loss function. While the depth consistency term based on connected superpixel neighborhood has effect on non-planar and planar regions, the first term is more effective on the curved regions. In planar surfaces, the superpixel loss has similar effect in general except improvement in the boundaries. From examples of visual comparison of estimated depths and reconstructed models, we find that an improvement in global depth estimation leads to an improvement in piecewise planar depth reconstruction. In planar reconstruction, the extent and curvature of the surfaces is better and the orientation and object pose is maintained with respect to the scene, while in the point cloud, there is a better understanding of the scene due to reorientation of normals of non planar regions.

The overall improvement in depth estimation is higher in curved objects than the planar objects. The two terms of our loss function have their individual effect on the depth estimation. The first term provides a good context for the loss at global level for supervision against ground truth while the second term acts at local level to improve the predicted depth with respect to its own local neighborhood. The second term has a very high influence on depth estimation and is sensitive to the geometry of objects, thus, a low weight is required to balance the effect. The first term influences the curved surfaces more than the planar surfaces while providing the smoothness at surface level while the second term handles the error better at the edges and boundaries. Whenever there is change in the depth values, the model

performs better than the baseline method. Although there is improvement in the boundary regions of planar surfaces, the effect is less over large surfaces due to less changes in nearby depth values.

A better global depth estimation leads to a better piecewise planar depth estimation and 3D reconstruction. While the plane segmentation highly influences the piecewise planar reconstruction, the non planar objects are more influenced by the global depth. The final 3D reconstruction is affected by both plane instances and global depth map. Thus, an improvement or reduction in quality in any step will result in proportional change in the accuracy of 3D reconstruction.

Since the loss function uses neighborhood regions to provide geometric awareness, the error is higher when the boundary between objects are not distinguishable based on color consistency. The research conducted has some limitations dependent on the size and distribution of the datasets used for experimentation. The 3D reconstruction quality has potential for improvement to provide better representation of surfaces when compared to the ground truth acquired data. This can be improved by further training on a large dataset subject to the availability of computing power and time. For more generalization and validation of results, training and testing with different datasets and other neural networks can be helpful to further research into the reasons of improvement due to proposed loss function.

Contribution

- We propose a new loss function in this research for geometry aware depth optimization in the process of 3D reconstruction from a single image. This function provides depth consistency over the scene based on color similarity and spatial compatibility, reducing the error based on the local and sub-local neighborhood. Similarly, we provide an orientation consistency term during 3D reconstruction for further refinement. This provides a new direction of research in an unexplored area, wherein, superpixels are directly used for designing a learning algorithm for the neural networks in the context of 3D reconstruction. The simplicity and ease of use of the loss function can be helpful for other researchers to further experiment with other neural network architectures and loss functions. We will also provide the code¹ for open access and use, to encourage further research on this topic.
- We provide insights into the state of the art method for 3D Reconstruction using single images by showcasing a full pipeline

¹ <https://github.com/cgarg-tud/GeomAwareLoss>

of the process. It is important to understand how each step in the whole process affects the geometric accuracy of the final reconstruction. We show that global depth estimation is crucial for 3D reconstruction, especially for non-planar objects while the plane segmentation helps in extracting the planar surfaces of the objects.

6.2 FUTURE WORK

- One avenue for the future work is using a normal consistency loss in which depth information is replaced by normal orientation to provide geometric awareness during supervision.
- Since the warped loss module proposed in [Liu et al., 2019] is influenced by the estimated depth, the effect of loss function can further be tested by training on this module
- Another area of future work could be using the reconstructed model for indoor navigation and localisation using images
- One direction could also be comparing the point clouds reconstructed from the neural networks to the traditional techniques for various types of applications in 3D simulation and virtual reality environment
- Testing the methodology against a synthetic dataset and investigating the role of different types of noises present in various sensing technologies use to gather real world 3D information.
- Another avenue of research is either comparing the current available techniques of single image based 3D reconstruction methods for various applications related to Geomatics. These can be as follows:
 - Using 3D output of a scene from multiple views for full reconstruction
 - Direct analysis on the 3D output using semantics labels and room layout for post-processing to get a mesh or voxel level representation
 - Using a signature of scene from 3D output for indoor localisation and navigation
 - Exploring the benefit of single image 3D reconstruction in culture and heritage field to obtain 3D output using historic images or paintings

BIBLIOGRAPHY

- Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2010). Slic superpixels. Technical report.
- Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282.
- Arbelaez, P., Maire, M., Fowlkes, C., and Malik, J. (2010). Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916.
- Chen, C., Gong, W., Chen, Y., and Li, W. (2019). Object detection in remote sensing images based on a scene-contextual feature pyramid network. *remote sensing*, 11(3):339.
- Choi, S., Zhou, Q., and Koltun, V. (2015). Robust reconstruction of indoor scenes. *IEEE Conference on Computer Vision and Pattern Recognition*. 10.1109/CVPR.2015.7299195.
- Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes.
- Delaunay, B. N. (1934). Sur la sphère vide. *Izvestia Akademia Nauk SSSR, Otdelenie Matematicheskii i Estestvoennyka Nauk*, 7:793–800.
- Donaubauer, A., Kohoutek, T. K., and Mautz, R. (2010). *CityGML als Grundlage für die Indoor Positionierung mittels Range Imaging*. abc-Verl., Heidelberg.
- Durrant-Whyte, H. and Bailey, T. (2006). Simultaneous localization and mapping: part i. *IEEE Robotics Automation Magazine*, 13(2):99–110.
- Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374.
- Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R. (2009). Manhattan-world stereo. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1422–1429. IEEE.

- Furukawa, Y. and Ponce, J. (2009). Accurate, dense, and robust multi-view stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376.
- Galliani, S., Lasinger, K., and Schindler, K. (2015). Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881.
- Gallup, D., Frahm, J.-M., and Pollefeys, M. (2010). Piecewise planar and non-planar stereo for urban scene reconstruction. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1418–1425. IEEE.
- Gkioxari, G., Malik, J., and Johnson, J. (2019). Mesh r-cnn.
- Goldlücke, B., Aubry, M., Kolev, K., and Cremers, D. (2014). A super-resolution framework for high-accuracy multiview reconstruction. *International journal of computer vision*, 106(2):172–191.
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. *IEEE International Conference on Computer Vision (ICCV)*.
- Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.
- Kang, Z., Yang, J., Yang, Z., and Cheng, S. (2020). A review of techniques for 3d reconstruction of indoor environments. *ISPRS International Journal of Geo-Information*, 9(5):330.
- Kim, I. (2017). Deep object detectors. techreport, Slideshare.
- Li, B., Shen, C., Dai, Y., Van Den Hengel, A., and He, M. (2015). Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127.
- Liu, C., Kim, K., Gu, J., Furukawa, Y., and Kautz, J. (2019). Planercnn: 3d plane detection and reconstruction from a single image. *IEEE In Computer Vision and Pattern Recognition (CVPR)*. <https://arxiv.org/pdf/1812.04072.pdf>.
- Liu, C., Yang, J., Ceylan, D., Yumer, E., and Furukawa, Y. (2018). Planenet: Piece-wise planar reconstruction from a single rgb image. In *Computer Vision and Pattern Recognition (CVPR)*. <https://arxiv.org/abs/1804.06278.pdf>.

- Liu, F., Shen, C., and Lin, G. (2015). Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.
- Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., and Brox, T. (2015). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. <https://arxiv.org/pdf/1512.02134.pdf>.
- McCormac, J., Handa, A., Leutenegger, S., and Davison, A. J. (2017). Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2678–2687.
- Meilă, M. (2005). Comparing clusterings: an axiomatic view.
- Mousavian, A., Pirsiavash, H., and Kosecka, J. (2016). Joint semantic segmentation and depth estimation with deep convolutional networks.
- Nathan Silberman, Derek Hoiem, P. K. and Fergus, R. (2012). Indoor segmentation and support inference from rgb-d images. In *ECCV*.
- Newell, A., Huang, Z., and Deng, J. (2016). Associative embedding: End-to-end learning for joint detection and grouping.
- Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., and Zhang, J. J. (2020). Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image.
- Saxena, A., Chung, S. H., and Ng, A. Y. (2006). Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168.
- Silberman, N. and Fergus, R. (2011). Indoor scene segmentation using a structured light sensor. In *Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition*.
- Sinha, S., Steedly, D., and Szeliski, R. (2009). Piecewise planar stereo for image-based rendering.
- Smith, E. J., Fujimoto, S., Romero, A., and Meger, D. (2019). Geometrics: Exploiting geometric structure for graph-encoded objects.

- Sun, X., Rosin, P. L., Martin, R. R., and Langbein, F. C. (2008). Noise in 3d laser range scanner data. In *2008 IEEE International Conference on Shape Modeling and Applications*, pages 37–45. IEEE.
- Ullman, S. (1979). The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426.
- Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., and Jiang, Y.-G. (2018). Pixel2mesh: Generating 3d mesh models from single rgb images.
- Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., and Yuille, A. L. (2015). Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2809.
- Yang, F. and Zhou, Z. (2018). Recovering 3d planes from a single image via convolutional neural networks. In *Computer Vision – ECCV 2018*, pages 87–103. Springer International Publishing.
- Yu, F., Koltun, V., and Funkhouser, T. (2017). Dilated residual networks. <https://arxiv.org/pdf/1705.09914.pdf>.
- Yu, Z., Zheng, J., Lian, D., Zhou, Z., and Gao, S. (2019). Single-image piece-wise planar 3d reconstruction via associative embedding. *arXiv preprint arXiv:1902.09777*.
- Zhang, Y. and Funkhouser, T. (2018). Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 175–185.
- Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., and Torralba, A. (2018). Semantic understanding of scenes through the ade20k dataset. <https://arxiv.org/pdf/1608.05442.pdf>.
- Zlatanova, S. and Isikdag, U. (2015). *3D Indoor Models and Their Applications*, pages 1–12. Springer International Publishing, Cham.
- Zlatanova, S. and Isikdag, U. (2017). *3d indoor models and their applications*. Springer.

A

REPRODUCIBILITY SELF-ASSESSMENT

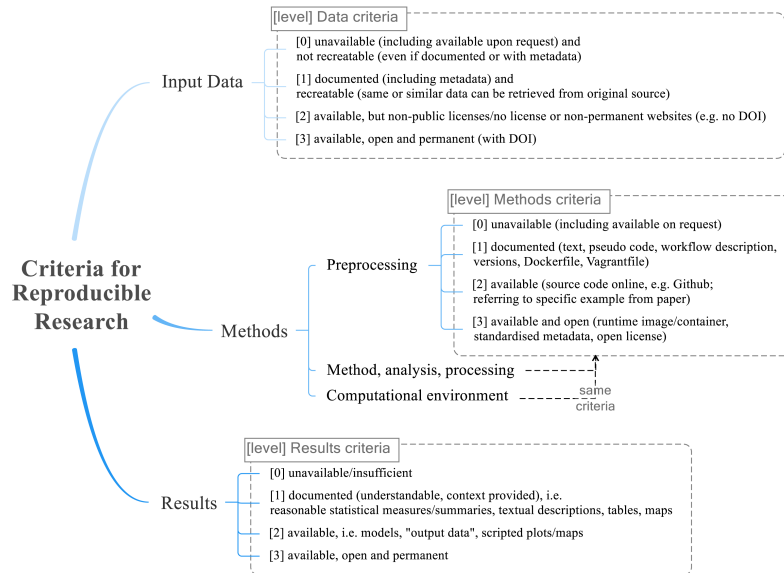


Figure A.1: Reproducibility criteria to be assessed.

Based on the criteria mentioned in [Figure A.1](#), we do the self-assessment for the research. For *input data* and *pre-processing*, we use publicly available datasets for our experiments [[Dai et al., 2017](#)] and [[Nathan Silberman and Fergus, 2012](#)]. Anyone can get the access to the download link by signing the terms of use provided by the corresponding authors. Thus, both are marked as 3. For *methods*, we provide the code through Github. Apart from the code provided by [[Liu et al., 2019](#)] for model setup, training and testing, we introduce new functions in the code for loss calculations and analysis of error and provide it in the Github. Based on this, we mark it as 3. For *computational environment*, we use openly available modules such as Pytorch, python and CUDA. Although since it may not be possible for all users to get access to hardware such as high power computing cluster for conducting experiments, we mark it as 2.5. For *results*, both qualitative and quantitative, we document everything in the thesis, these can be generated using code as well. The model used for experiment, a checkpoint is provided in Github to generate 3D models. However, to generate one's results from scratch, the user needs to set up specific hardware and software for running experiments and based on the chosen hardware, software, hyper parameters and type of dataset,

the results might vary depending on the experiment setup. Overall, the results are marked as 2.

B | DATA OF QUANTITATIVE EVALUATION

Scannet Curved Dataset : Mean values of 3 times run experiment						
Models	Piecewise Planar Depth			Global Depth		
	relative error (P)	rmse (P)	Accuracy (1.25 thresh)	relative error(G)	rmse (G)	Accuracy (1.25 thresh)
b	0.175	0.382	0.763	0.185	0.385	0.73
ours(w:0)	0.157	0.36	0.774	0.161	0.36	0.771
ours(w:0.1)	0.144	0.341	0.806	0.147	0.338	0.798

Scannet Curved Dataset : Standard deviation of 3 times run experiment						
Models	Piecewise Planar Depth			Global Depth		
	relative error (P)	rmse (P)	Accuracy (1.25 thresh)	relative error(G)	rmse (G)	Accuracy (1.25 thresh)
b	0.0085	0.0092	0.0111	0.0067	0.009	0.035
ours(w:0)	0.0035	0.004	0.0125	0.005	0.0068	0.0032
ours(w:0.1)	0.004	0.0078	0.0055	0.0053	0.0075	0.0067

Scannet Planar Dataset : Mean values of 3 times run experiment						
Models	Piecewise Planar Depth			Global Depth		
	relative error (P)	rmse (P)	Accuracy (1.25 thresh)	relative error(G)	rmse (G)	Accuracy (1.25 thresh)
b	0.205	0.441	0.625	0.202	0.445	0.657
ours_0	0.194	0.452	0.656	0.196	0.446	0.658
ours_0.1	0.171	0.428	0.726	0.183	0.438	0.672

Scannet Planar Dataset : Standard deviation of 3 times run experiment						
Models	Piecewise Planar Depth			Global Depth		
	relative error (P)	rmse (P)	Accuracy (1.25 thresh)	relative error(G)	rmse (G)	Accuracy (1.25 thresh)
b	0.0215	0.0187	0.0488	0.0155	0.0068	0.0181
ours(w:0)	0.0147	0.0275	0.0465	0.0064	0.008	0.021
ours(w:0.1)	0.0035	0.017	0.0144	0.009	0.0155	0.0204

Figure B.1: Quantitative Evaluation of reconstructed depth maps on Scannet Dataset

Scannet Curved Dataset : Mean values of 3 times run experiment						
	Plane Segmentation			Plane Reconstruction		
	Lower is better		Higher is better	Higher is better (Average Precision : AP at a particular depth error)		
Models	Random Index	VOI↓	Segmentation Cover	AP : 0.4m	AP : 0.6m	AP : 0.9m
b	0.813	2.253	0.547	0.298	0.468	0.484
ours(w:0)	0.81	2.254	0.546	0.369	0.471	0.478
ours(w:0.1)	0.814	2.21	0.553	0.4	0.484	0.487

Scannet Curved Dataset : Standard Deviation values of 3 times run experiment						
	Plane Segmentation			Plane Reconstruction		
	Random Index	VOI↓	Segmentation Cover	AP : 0.4m	AP : 0.6m	AP : 0.9m
b	0.0038	0.038	0.0076	0.046	0.0176	0.0032
ours(w:0)	0.004	0.0099	0.0045	0.0201	0.0273	0.0255
ours(w:0.1)	0.005	0.0674	0.0132	0.0144	0.0095	0.0102

Scannet Planar Dataset : Mean values of 3 times run experiment						
	Plane Segmentation			Plane Reconstruction		
	Lower is better		Higher is better	Higher is better (Average Precision : AP at a particular depth error)		
Models	Random Index	VOI↓	Segmentation Cover	AP : 0.4m	AP : 0.6m	AP : 0.9m
b	0.0215	0.0187	0.0488	0.0155	0.0068	0.0181
ours(w:0)	0.0147	0.0275	0.0465	0.0064	0.008	0.021
ours(w:0.1)	0.0035	0.017	0.0144	0.009	0.0155	0.0204

Scannet Planar Dataset : Standard Deviation values of 3 times run experiment						
	Plane Segmentation			Plane Reconstruction		
	Random Index	VOI↓	Segmentation Cover	AP : 0.4m	AP : 0.6m	AP : 0.9m
b	0.0095	0.0223	0.0095	0.0257	0.0061	0.0067
ours(w:0)	0.0243	0.1667	0.03	0.0338	0.0219	0.0162
ours(w:0.1)	0.0045	0.0098	0.002	0.0166	0.0115	0.0121

Figure B.2: Quantitative Evaluation of planar segmentation and reconstruction on Scannet Dataset

NYU Dataset : Mean values of 3 times run experiment										
	<i>Piecewise Planar Depth</i>					<i>Global Depth</i>				
	Lower is better				Higher is better	Lower is better				Higher is better
Models	<i>relative error</i>	<i>rel_sqrt</i>	<i>rmse</i>	<i>rmse_log</i>	<i>Acc (1.25)</i>	<i>relative error</i>	<i>rel_sqrt</i>	<i>rmse</i>	<i>rmse_log</i>	<i>Acc (1.25)</i>
b	0.33	0.406	0.949	0.349	0.497	0.334	0.416	0.961	0.349	0.491
ours_0	0.307	0.363	0.935	0.36	0.491	0.312	0.373	0.949	0.369	0.485
ours_0.1	0.291	0.335	0.944	0.342	0.473	0.295	0.344	0.943	0.341	0.469

NYU Dataset : Standard Deviation values of 3 times run experiment										
	<i>Piecewise Planar Depth</i>					<i>Global Depth</i>				
Models	<i>relative error</i>	<i>rel_sqrt</i>	<i>rmse</i>	<i>rmse_log</i>	<i>Acc (1.25)</i>	<i>relative error</i>	<i>rel_sqrt</i>	<i>rmse</i>	<i>rmse_log</i>	<i>Acc (1.25)</i>
b	0.018	0.035	0.0087	0.0078	0.0036	0.0175	0.0345	0.0089	0.0038	0.0035
ours_0	0.0101	0.0172	0.0017	0.0188	0.0062	0.0095	0.0187	0.0006	0.021	0.0053
ours_0.1	0.0017	0.004	0.0074	0.005	0.0056	0.0015	0.0046	0.0095	0.004	0.0072

Figure B.3: Quantitative Evaluation of reconstructed depth maps on NYU Dataset

COLOPHON

This document was typeset using \LaTeX . The document layout was generated using the `arsclassica` package by Lorenzo Pantieri, which is an adaption of the original `classiethesis` package from André Miede.

