SEMANTIC 3D INDOOR SCENE MODELLING FROM A SINGLE IMAGE

CHIRAG GARG
4817818
c.garg@student.tudelft.nl

1st Supervisor: dr. Liangliang Nan
2nd Supervisor: dr. Jan Van Gemert
Date P2: 2020-01-15

CONTENTS

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 are 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 semantic 3D information has many applications. For example, it can be used for by 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. This also has 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

Figure 1.1: Applications of semantic 3D information using images : (a) Recognizing indoor space furniture can help in asset management[Donaubauer et al., 2010],(b) Division of floor into small recognizable spaces for indoor navigation[Zlatanova and Isikdag, 2017] (c) An application to create 3D model of a room[Canvas, 2016]

To get 3D information such as depth or planar surfaces from indoor space, 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 easier than sensor based approach. In this, multiple images having a minimum overlap to reconstruct geometric primitives such as line segments, vanishing points, planes and local features such as corners, blobs, which are grouped together into planar or surface patches[Gallup et al., 2010]. However, these bottom up

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 extraction algorithms for feature reconstruction, and 3) the complexity of placement of various objects challenges the Manhattan world norms. The top-down approach using neural networks,[Liu et al., 2015] tries to tackle these challenges by looking at an image from a holistic perceptive. Keeping this in mind, using a single image to extract 3D information can make the data acquisition process easier.

With the evolution of deep learning techniques, the Convolution Neural networks have been utilized to infer information such as semantic labels, depth maps, surface normals from a single image[Eigen et al., 2014]. 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]. In literature, many models focus on designing neural networks to predict depth and semantic information independently. Recently, new models have come which performs 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]. PlaneRCNN [Liu et al., 2019] achieved a breakthrough in reconstructing 3D model outperforming other models. It uses MaskRCNN,[He et al., 2017] as the backbone network and make improvements for extracting planar surfaces and depthmap from single image. Through analysis of existing methods, it is expected that using an energy function to enforce semantic and color consistency, spatially with depth consistency can influence the reconstruction process. The aim of this research project is to generate a depthmap or point cloud along with semantic labels of objects from a single RGB image by jointly optimizing depth estimation and semantic segmentation. The main objectives are to investigate the current state of the art methods, establish a benchmark, utilize the new energy functions to design a new model and generate a 3D semantic point cloud from single image

2
Related Work

2.1. Semantic Segmentation

Image segmentation is a process of identifying certain certain objects or parts of objects in an image. This technique helps in eliminating the need of considering every pixel as unit of observation and provides granular perspective of details in image. Initial techniques for such detections used certain templates such as bounding boxes,[Chen et al., 2016], complex 3d representations[Wu et al., 2016] and shape compositions. These methods generally have coarse level features and do not handle the complexity of modelling chaotic indoor scenes. Before the advent of CNN, Conditional Random Field(CRF) was used to segment image in a hierarchical manner in [L. Ladicky and Torr., 2009] and as fully connected in [Krahenbuhl and Koltun, 2012] . On a basic level, CRF is sort of a probabilistic network to determine labels of objects and segmentation by incorporating certain relationships within pixels of image. For example, nearby pixels are likely to have similar labels, same depth pixels probably form a plane surface. [Chen et al., 2014], combined the deep CNN with fully connected CRF, thus combining the positives of both methods. The model is trained independently in the proposed method however there were improvements in training schemes. In [Zheng et al., 2015], CRFs are utilized with RNNs to exploit global level information and local information through differential operations for end to end training. In [Zhao et al., 2016], PSPNet model is proposed which aggregates features at various scales using a Pyramid Pooling Module (PPM) to derive the per-pixel prediction. In [Xiao et al., 2018], UPerNet model is proposed which combines PPM with a Feature Pyramid Network (FPN),[Lin et al., 2016], providing less training time and power consumption than PSPNet. Recently,[Sun et al., 2019] proposed HRNet, a model that retains high resolution feature representation by incrementally increasing the convolution networks at different scales rather than in a single stage.

2.2. Instance Segmentation

Semantic segmentation only provides pixels belonging to a category in image but does not distinguish between several objects of same class in image, hence it is not instance aware. This has been depicted in figure 2.1. In semantic segmentation all chairs belong to blue pixels in an image but in instance segmentation, these are further classified into 9 separate chairs. The most common benchmark for instance level segmentation is MaskRCNN proposed in [He et al., 2017], which is built upon Faster-RCNN[Ren et al., 2015]. The MaskRCNN in first stage creates proposals, regions of interest with maximum likelihood of containing an object, then in second stage,these proposals are classified into a class and bounding box and a binary mask is generated for each classified proposal. In another approach, [Newell et al., 2016],the concept of associative embedding is used wherein,the pixels belonging to same instance level have similar properties. By

using a discriminative loss introduced in [Brabandere et al., 2017], the model learns instance embedding to form proposals using a mean shift clustering algorithm.

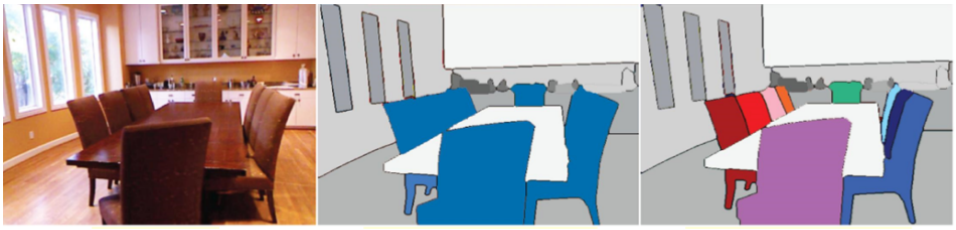

Figure 2.1: Image segmentation : This image from a blog shows difference between semantic and instancec level segmentation by labelling chairs[Tsang, 2018]

2.3. 3D RECONSTRUCTION FROM IMAGES

Traditional methods for 3d reconstruction from images use multiple views of the scene and sometimes, the depth information as well.([Sinha et al., 2009],[Furukawa et al., 2009]). These methods generally first reconstruct the points in 3D, then, planes are fitted through the points by inferring corresponding plane parameters and assigning a plane ID to each point[Gallup et al., 2010]. A piece-wise planar depthmap can be produced by solving global-inference problem. These methods are often time-consuming and statistical optimization is required to achieve accurate results. This also restricts their utility in real-time applications. One of the early pioneers in this field, [Saxena et al., 2006] infers depth from outdoor scenes using Markov Random field (MRF) to incorporate both global and local features of an image to refine depth prediction. With the advent of deep neural networks, many CNN based techniques have been produced to infer depthmaps 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. But these methods do not provide planar segmentation or parameters which can help in inferring topological relationships among various elements in the scene.

Recently, a novel model, "Planenet", was proposed by [Liu et al., 2018] to reconstruct a "piece-wise depthmap", given a single RGB-image using end-to-end deep neural network built upon dilated Residual Networks(DRNs) proposed in [Yu et al., 2017]. 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, plane segmentation masks and a global depthmap.[Liu et al., 2018]. In another approach, "PlaneRecover" a fully convolution network (FCN) based on Disp-Net,[Mayer et al., 2015] simultaneously predicts plane segmentation map and plane parameters, taking advantage of ground truth semantic labels, depthmap 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 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 loosing 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., 2018a], 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]. This approach uses an proposal-free approach. 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]. It contains three networks: firstly, a plane detection network based on MaskRCNN,[He et al., 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 2.2, clipped from [Liu et al., 2019].

Figure 2.2: From left to right: input image, Planenet [Liu et al., 2018], PlaneRecover [Yang and Zhou, 2018], and PlaneRCNN[Liu et al., 2019].

3
RESEARCH QUESTIONS

3.1. OBJECTIVES

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

"To what extent, a 3D model of an indoor scene can be generated from a single image ?"

To support our main research objective, some sub questions have also been formulated as following :

- Is it possible, and to what extent, 3D information(e.g., plane/depth) of an indoor scene can be extracted from a single image?
- Can joint depth estimation and semantic segmentation improve the accuracy of the 3D model?
- How accurate is the output of the current techniques compared to traditional and benchmark algorithms?

3.2. SCOPE OF THE THESIS

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

- Only indoor scenes will be used for research Hence, no outdoor scenes or buildings will be considered.
- Only single image will be used as data input, thus, no multiple images are utilized.
- The output will be a depth image or a point cloud with semantic information but no solid model or surface representation will be produced. Reconstructing a mesh or surface representation from point cloud or depth-map is beyond the scope of this research.

4

METHODOLOGY

To model a 3D semantic scene from a single RGB image, a convolution neural network will be developed in such a way that the model jointly optimizes the depth estimation and semantic segmentation. To conduct our research, we will perform the following tasks.

4.1. BACKBONE NEURAL NETWORK

The main model will be developed using the principles of transfer learning, wherein, a pre-trained model is used to access its feature maps and customized to design a new model for performing another task. Based on the literature review, for depth estimation, PlaneRCNN,[Liu et al., 2019] depicted in figure A.1, and PlanarReconstruction,[Yu et al., 2019], depicted in figure A.2, are suitable candidates for our investigation and will be analyzed empirically to test the quality of output. They provide piece-wise depthmaps from single image and use instance level information for planar segmentation. Another advantages of using these methods include no restriction on the number of planes considered in a scene and the combination of local and global features for inference. A pre-trained model can be used to produce piece-wise planar depthmap from single RGB image after creating a virtual environment and installing dependencies on a Linux machine. Using known camera intrinsic parameters a point cloud will be generated from the depthmap. For a point P_c in the camera coordinate system, let $(\mathrm{x_c}, \mathrm{y_c}, \mathrm{z_c}, 1)^{\mathrm{t}}$ be the homogeneous coordinates for a pixel $\mathrm{P_I}$ in image with $(\mathrm{i}, \mathrm{j}, 1)^{\mathrm{t}}$ as homogeneous coordinates. The pinhole camera equation gives us:

$$\mathrm{P}_I = \pi\left(\mathrm{P}_c\right) = \left(\frac{f_x x_c + c_x}{z\left(\mathrm{P}_I\right)}, \frac{f_y y_c + c_x}{z\left(\mathrm{P}_I\right)}, 1\right)^t, \tag{4.1}$$

where f_x, f_y are focal lengths in x and y direction; c_x, c_y are respective principal point offsets; $z\left(\mathrm{p}_I\right)$ 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 :

$$\mathrm{P}_c = \pi^{-1}\left(\mathrm{P}_I, z\left(\mathrm{P}_I\right)\right) = z\left(\mathrm{P}_I\right)\left(\frac{i - c_x}{f_x}, \frac{j - c_y}{f_y}, 1\right)^t \tag{4.2}$$

Once the point cloud is obtained, semantic segmentation is performed using the benchmark methods discussed in section 2.1. The models proposed in [Zhao et al., 2016], [Xiao et al., 2018] and [Sun et al., 2019] will be analyzed using pre-trained models provided by [Zhou et al., 2017] and [Zhou et al., 2018a] in an encoder-decoder architecture. These models will be used to generate pixel wise semantic labels which are then further pro-

jected onto the point cloud generated by using equation 4.2. The final backbone network will be chosen based on experimenting the existing methods.

4.2. JOINT OPTIMIZATION

In order to design a model that can jointly optimize the depth and semantic information, the neural network needs to enforce a constraint such that in an image, the pixels in a neighborhood with similar colors and semantic labels should have similar depth values. Thus, sudden depth changes should be penalised in a neighborhood of similar colors. We can formulate this as an energy function which has to be minimized by the neural network. If depth changes too much, the patch should be split. This problem belongs to the category of Conditional Random Field(CRF) problem[Liu et al., 2015]. Thus, incorporating CRF energy function through convolution layers in a model can possibly improve the results of depth estimation and corresponding 3d model. Broadly, there are 3 components of CRF: a) a unary network to define self-energy of a pixel b) a pairwise network to define mutual energy between a pair of pixels using their shared properties which in our case is color and semantic labels c) a loss layer to penalize the energy function.

To implement CRF for depth estimation alone, in [Liu et al., 2015], a fully convolution network is used, while in [Xue et al., 2019] a recurrent neural network is used . In [Mousavian et al., 2016], a multi-scale CRF neural network was used to further optimize semantic using depth information in constraints. We will instead optimize depth map using semantic information for pairwise network. Adopting the concept of super pixels representing depth of a pixel at its centroid from previous work, [Saxena et al., 2006],[Liu et al., 2015], for an Image **I**, let **N** be the number of super-pixels and $X = \{x_1, x_2, \ldots, x_N\}$ be the continuous label vector of N superpixels of depthmap, where $x_i \in \{1, \ldots, D\}$, given D as the number of depth labels. The probability of assigning a depth label can be defined using Gibbs distribution as :

$$\Pr(\mathbf{x}|\mathbf{I}) = \frac{1}{\mathrm{Z}(\mathbf{I})} \exp(-E(\mathbf{x}, \mathbf{I})), \tag{4.3}$$

where **E(x)** is the energy function describing the cost of label assignment, and **Z** is the normalization or partition factor defined as :

$$\mathrm{Z(I)} = \int_{\mathrm{x}} \exp\{-E(\mathrm{x}, \mathrm{I})\}\mathrm{dx}, \tag{4.4}$$

. For simplicity, I term will be dropped for now. The features representative of each pixel i can be depicted as $f_i = \{p_i, I_i, d_i, S_i\}$ where p_i is the spatial location, I_i is the RGB value, d_i is the depth value and S_i is the semantic label. The energy function for the fully connected CRF becomes:

$$E(x, f) = \sum_i \psi_u(x_i) + \sum_{i,j} \psi_p\left(x_i, f_i, x_j, f_j\right), \tag{4.5}$$

where unary potential $\sum_i \psi_u(x_i)$ come from the depth decoder of the PlaneRCNN network and the second term depicting pair-wise potentials have the form,

$$\psi_p\left(x_i, f_i, x_j, f_j\right) = \mu\left(x_i, x_j\right) k\left(f_i, f_j\right), \tag{4.6}$$

where where $\mu\left(x_i, x_j\right)$ depicts the compatibility between depth label assignments of pixel i and j. Gaussian kernel $k\left(f_i, f_j\right)$ calibrates the evidence that should be disseminated between x_i and x_j based on the spatial distance, RGB distance, depth distance and semantic affinity between pairs of pixels. $k\left(f_i, f_j\right)$ consists of four different weights w^i for $i \in 1,2,3,4$ and hyper-parameters $\theta(.)$ that control the tolerance with respect to difference in semantic pixel labels, depth pixel values, RGB pixel values and spatial location of pairs of pixels. $k\left(f_i, f_j\right)$is computed using the following equation:

$$\begin{aligned} k\left(f_i, f_j\right) = w^{(1)} \exp\left(\frac{\left|p_i - p_j\right|^2}{2\theta_\alpha^2} + \frac{\left|I_i - I_j\right|^2}{2\theta_\beta^2}\right) \\ + w^{(2)} \exp\left(\frac{\left|p_i - p_j\right|^2}{2\theta_\gamma^2} + \frac{\left|d_i - d_j\right|^2}{2\theta_\zeta^2}\right) \\ + w^{(3)} \exp\left(\frac{\left|p_i - p_j\right|^2}{2\theta_\kappa^2} + \frac{\left|S_i - S_j\right|^2}{2\theta_\lambda^2}\right) \\ + w^{(4)} \exp\left(\frac{\left|p_i - p_j\right|^2}{2\theta_\tau^2}\right) \end{aligned} \tag{4.7}$$

The inference will be done using the mean-field approximation used in previous works, [Zheng et al., 2015]. The unary potentials, compatibility parameters and pairwise weights are learnt during the training stage by back propagation. The derivatives are back-propagated to further refine the feature representation. It is important to note here that there are various approaches to apply constraints. The semantic label can be replaced with instance mask as well. To enforce constraint, for pixels belonging to same semantic label, the depth is regularized, while when the labels are different, surface normal can be used to guide the topological relationship as used in [Ji et al., 2016]. Another approach can be considered from [Xu et al., 2018], where the a structured attention guided model is used to correlate latent feature maps at different scales to guide the CRF network weights.

Using pre-trained models and adding CRF module to jointly optimize depth and semantic labels will be an iterative task. Fine tuning the whole pipeline will be bottom up approach.For CRF, we will start from incorporating color consistency, then semantic and surface normal information will be used.

4.3. EVALUATION

4.3.1. DEPTH ESTIMATION

Following the previous works, the new model will be evaluated by using metrics adopted in [Eigen et al., 2014] and [Wang et al., 2015]. If d_i^{pr} represents the predicted depth and d_i^{gt} represents the ground truth depth of a pixel i and N is the number of pixels in images to be tested, then the following errors will be calculated using their respective equations:

- mean relative error :

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\left|d_i^{pr}-d_i^{gt}\right|}{d_i^{gt}} \tag{4.8}$$

- Root mean square error(rmse) :

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(d_i^{pr}-d_i^{gt}\right)^2} \tag{4.9}$$

- mean log 10 error :

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\log_{10}\left(d_i^{pr}\right)-\log_{10}\left(d_i^{gt}\right)\right\| \tag{4.10}$$

- 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 4.11 :

$$\max\left(\frac{d_i^{gt}}{d_i^{pr}},\frac{d_i^{pr}}{d_i^{gt}}\right)=\delta<th\left(th\in\left[1.25,1.25^2,1.25^3\right]\right) \tag{4.11}$$

4.3.2. SEMANTIC SEGMENTATION

The semantic segmentation will be measured by adopting the four metrics used in previous works, [Long et al., 2015] as follows :

- **Pixel accuracy** : It represents the percentage of pixels which are classified correctly
- **Mean accuracy** : It represents the percentage of pixels which are classified correctly averaged over all classes.
- **Mean IoU** : It represents the the intersection-over-union between the ground truth and predicted pixels averaged over all classes.
- **Weighted IoU** : It represents the the IoU weighted by the ratio of all pixels of every class.

5
SCHEDULE

The following plan has been set up to complete tasks for fulfilling research objectives :

Start	End	Activity
1 sept	31 sept	Exploring graduation topics
		P1 - Progress review Graduation Plan
1 oct	31 dec	Literature study
7 oct	15 dec	Research existing methods for 3d reconstruction using single image
10 oct	30 dec	Study existing CNN models and reproduce results
15 oct	15 jan	Study semantic segmentation from single image and reproduce results
10 nov	15 jan	Generate 3d semantic point clouds using single image and analyze
		P2 - Formal assessment Graduation Plan
10 dec	1 feb	Define the new energy terms
25 jan	15 mar	Implement the energy terms and the deep learning network
1 feb	15 mar	Quantitative and Qualitative evaluation
		P3 - Colloquium midterm
20 mar	12 may	Write final implementation
1 may	12 may	Thesis writing
		P4 - Formal process assessment
12 may	15 jun	Finalize thesis
12 jun	22 jun	Prepare final presentation
		P5 - Public presentation and final assessment

The tentative graduation calendar is shown below. The exact dates of the presentations will be determined during the year.

Event	Date
P1	11 Nov
P2	15 Jan
P3	12-20 Mar
P4	15-30 May
P5	10-30 June

Weekly meetings will be held with the daily supervisor dr. Liangliang Nan. Additional guidance and feedback will be provided by another supervisor, dr. Jan Van Gemert. The co-reader is yet to be decided.

6
EXPERIMENTS

6.1. TOOLS

In order to conduct experiments following hardware and softwares will be used. For using deep learning techniques, Ubuntu 18.04 with graphics card, NVIDIA QUADRO P1000 having 4GB GDDR5 on-board memory is used. For training and testing, graphics card provided by HPC cluster, TU Delft will be used. For each experiment, a certain virtual environment is required with some dependencies such as skit learn, cffi, numpy, opencv,python, scikit-image, torch, tqdm. Thus, either conda or venv is used to do this. For normal operations, python is used in Spyder while Open3d is used for visualization and redering mages and point clouds.

6.2. DATA

In order to conduct our research, we will use established benchmark datasets which provide RGB-D ground truth with rich annotations at indoor level and toolbox to do preprocessing. Also both real and synthetic datasets will be used for training and testing. Considering this, following datasets will be used:

- **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.
- **Sun RGB-D** : Provided by [Song et al., 2015], the dataset contains 10335 indoor images with dense annotations in 2D and 3D for both objects and indoor scenes.
- **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 level semantic category labels with 1205 scans for training and other 312 scans for testing.
- **Matterport3D** :In [Chang et al., 2017], it provides 194400 RGBD images for indoor environments of 90 buildings annotated for 2D and 3D semantic segmentation, camera poses and surface reconstructions

6.3. EXPERIMENTS WITH EXISTING METHODS

6.3.1. SEMANTIC SEGMENTATION

For producing semantic segmentation, models provided in [Zhou et al., 2017] and [Zhou et al., 2018a] are used. They provide normalized pre-trained models, whose training is benchmarked on a server with 8 NVIDIA Pascal Titan Xp GPUs (12 GB GPU memory). After experimentation with several models and visually comparing them, a model is selected to project semantic labels onto the point cloud generated in first step. A visual comparison for some of the models using two nearby images is shown in figure 6.1.

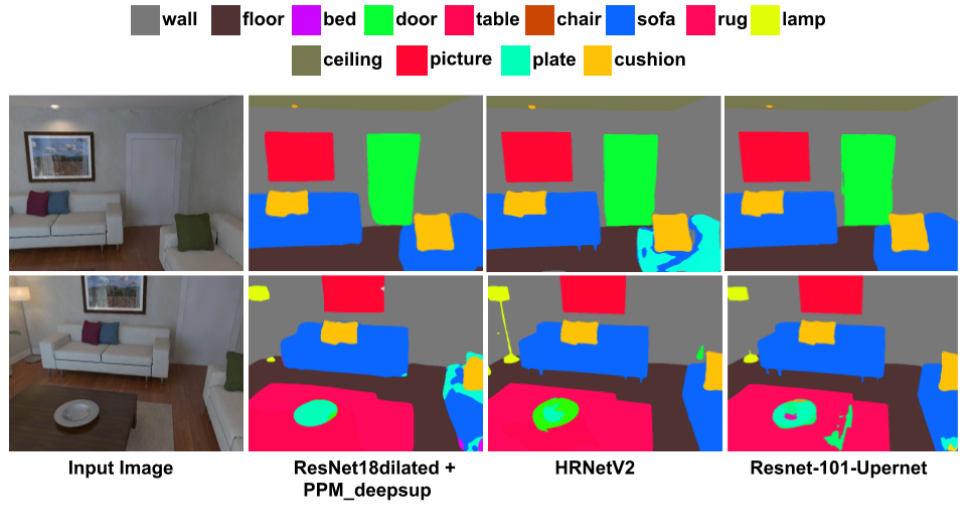

Figure 6.1: From left to right: input image, segmentation from resnet18dilated-ppm-deepsup model, hrnetv2 model and resnet-101-upernet model

6.3.2. 3D RECONSTRUCTION

To create a 3D model from image, several models were considered in the beginning. Implementing Planenet,[Liu et al., 2018] was difficult due to problems in compatibility of software and hardware. It was possible to reproduce results from Planerecover, but it was tested for outdoor images hence, is not shown here. Finally, we choose two models for testing. Firstly, PlaneRCNN, [Liu et al., 2019] is used to load a pre-trained model on NVIDIA TitanX GPU for 10 epochs with 100,000 randomly sampled images from training scenes in ScanNet. A single image is processed using parameters stated in [Liu et al., 2019] to get depth map, plane segmentation and masks. After creating an ad-hock function to generate point clouds using the estimated piece-wise depth-map. Three separate ply files: mesh reproduced from model, a point cloud with xyz coordinates and another colourised point cloud. The results of the same are shown in top row in figure 6.2 The second model implemented is from [Yu et al., 2019]. A pre-trained model for 50 epochs on one NVIDIA TITAN XP GPU device is used for testing images. The results of both PlaneRCNN and PlanarReconstruction can be seen in figure 6.2. From visual analysis, it can

be observed that the point cloud of PlaneRCNN provides more information about local objects and surfaces. PlanarReconstruction generalizes the scene into larger planes and thus the depth information does not preserve the topological relationships.

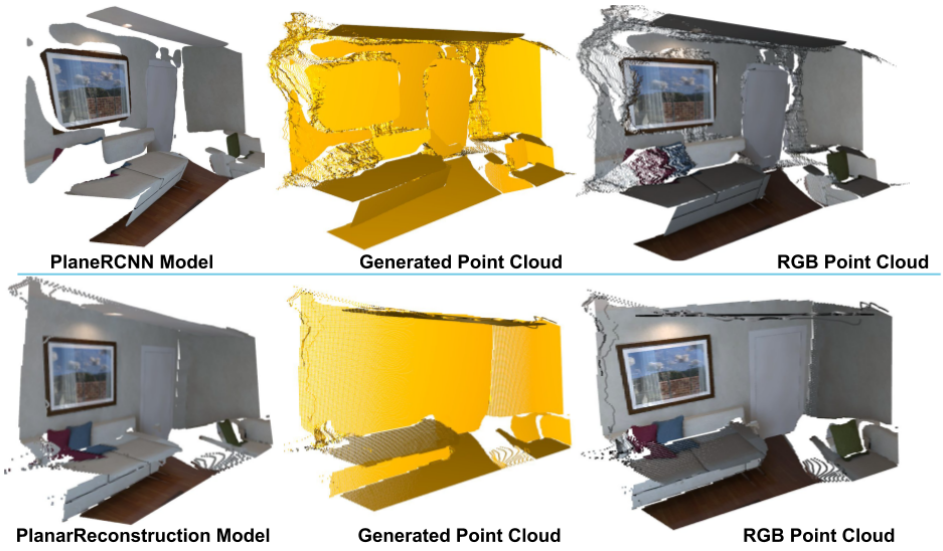

Figure 6.2: From left to right: 3D mesh reproduced from model, generated point cloud using predicted depth, and colored point cloud using original image rgb values

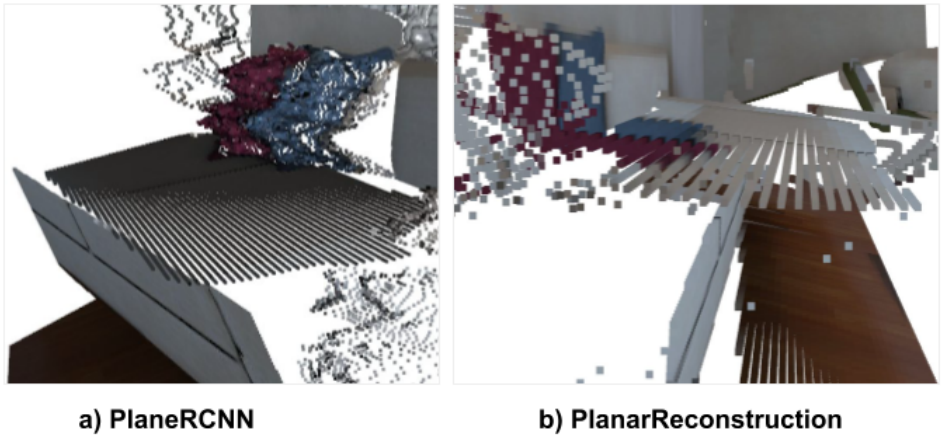

Figure 6.3: From left to right: point cloud generated from PlaneRCNN and PlanarReconstruction respectively.

6.3.3. 3D SEMANTIC POINT CLOUD GENERATION AND RENDERING

To create a semantic point cloud which can represent indoor scene, the output from step 1 and 2 are combined. For first image, resnet-101-upernet model is chosen for semantics while for second, hrnetv2 model is chosen. The point clouds generated from models of PlaneRCNN and PlanarReconstruction are labelled with semantic labels using equation 4.2. Open3d, [Zhou et al., 2018b] is then used to load all point clouds and rendering for visual analysis.

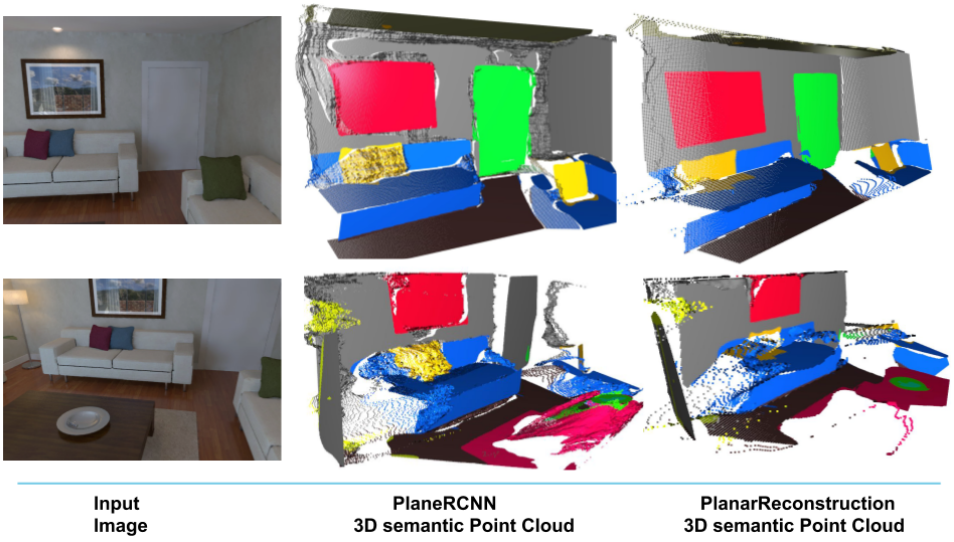

Figure 6.4: From left to right: input image, 3D semantic point cloud from PlaneRCNN model, 3D semantic point cloud from PlanarReconstruction model

6.4. WAY FORWARD

From visual analysis, with the observations in figures 6.2 and 6.3 that the first approach by PlaneRCNN performs qualitatively better over second approach. It is globally and locally, better representation of indoor scene. In figure 6.3, and 6.5, both approaches does 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 semantic segmentation provides a good perspective on the quality of depth-map. It can be observed in figure 6.3 that PlaneRCNN provides better representation of points to semantic labels than second approach. After analysing the existing methods, it is expected that jointly optimizing depth estimation and semantic segmentation can potentially help the process of extracting 3D information using single image but it needs it to be implemented and tested.

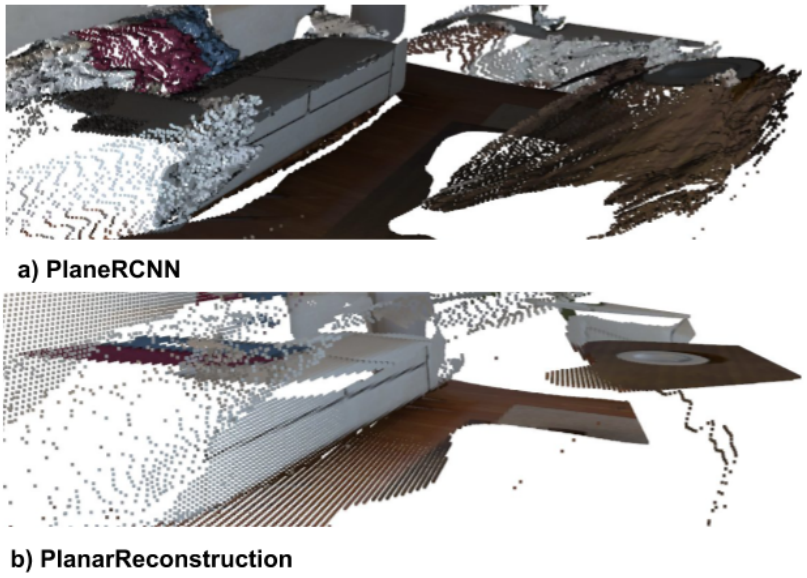

Figure 6.5: Side view of colored point cloud generated from second input image zoomed in on couch, pillows and table generated using a)PlaneRCNN b) PlanarReconstruction

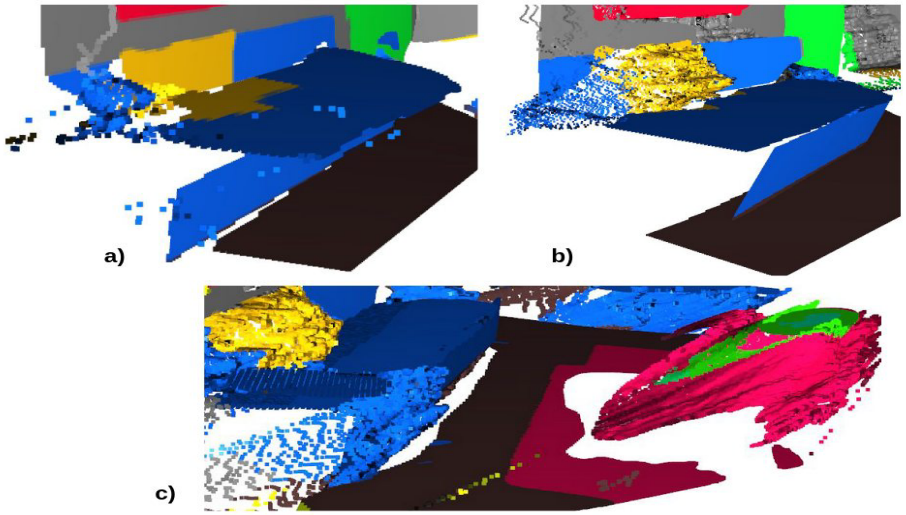

Figure 6.6: Side view of 3D semantic point cloud zoomed in on couch and pillows generated using a) PlaneR-CNN, b)PlanarReconstruction,c) another view of 3D model focusing on table

The current approaches often fails in critical boundary conditions resulting in misplaced planes or inconsistency in depth values. For example, in figure 6.1 a table and rug in the image are together predicted as table while in the 3D model shown in figure 6.5, major

points of table are predicted between ground and table and only a small portion is table is appeared at a height in PlaneRCNN while a good part of table appears as planar surface in PlanarReconstruction. Similarly for pillows in 6.6, there is no depth consistency maintained for single object and when the boundary is changing. Hence, adding a new constraint to enforce the relationship between depth and semantic information has a potential to improve 3D reconstruction. Taking cue from this evidence, in next phase, the energy terms defined in methodology will be implemented in an iterative manner. Then the new model and the baseline methods will be evaluated qualitatively and quantitatively.

REFERENCES

B. D. Brabandere, D. Neven, and L. V. Gool. Semantic instance segmentation with a discriminative loss function. 2017. https://arxiv.org/pdf/1708.02551.pdf.

Canvas. Canvas: Create a 3d model of your home in minutes, 2016. https://www.youtube.com/watch?v=XA7FMoNAK9M.

A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. 2017.

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. 2014. https://arxiv.org/pdf/1412.7062.pdf.

X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals using stereo imagery for accurate object class detection. 2016.

S. Choi, Q. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 10.1109/CVPR.2015.7299195.

A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. 2017.

A. Donaubauer, T. K. Kohoutek, and R. Mautz. *CityGML als Grundlage für die Indoor Positionierung mittels Range Imaging*. abc-Verl., Heidelberg, 2010. ISBN 978-3-938833-42-1.

D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Manhattan-world stereo. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1422–1429. IEEE, 2009.

D. Gallup, J.-M. Frahm, and M. Pollefeys. 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, 2010.

K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *IEEE International Conference on Computer Vision (ICCV)*, 2017.

R. Ji, L. Cao, and Y. Wang. Joint depth and semantic inference from a single image via elastic conditional random field. *Pattern Recognition*, 59:268–281, 2016.

P. Krahenbuhl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in Neural Information Processing Systems 24 (2011) 109-117*, pages 108–119, 2012. https://arxiv.org/pdf/1210.5644.pdf.

P. K. L. Ladicky, C. Russell and P. H. Torr. Associative hierarchical crfs for object class image segmentation. *In IEEE ICCV*, 2009. http://www.robots.ox.ac.uk/~lubor/iccv09.pdf.

B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He. 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, 2015.

T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. 2016.

C. Liu, J. Yang, D. Ceylan, E. Yumer, and Y. Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. *In Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1804.06278.pdf.

C. Liu, K. Kim, J. Gu, Y. Furukawa, and J. Kautz. Planercnn: 3d plane detection and reconstruction from a single image. *IEEE In Computer Vision and Pattern Recognition (CVPR)*, 2019. https://arxiv.org/pdf/1812.04072.pdf.

F. Liu, C. Shen, and G. Lin. 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, 2015.

J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. 2015. https://arxiv.org/pdf/1512.02134.pdf,.

A. Mousavian, H. Pirsiavash, and J. Kosecka. Joint semantic segmentation and depth estimation with deep convolutional networks. 2016.

P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

A. Newell, Z. Huang, and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. 2016.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 2015.

A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006.

N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition*, 2011.

S. Sinha, D. Steedly, and R. Szeliski. Piecewise planar stereo for image-based rendering. 2009.

S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-dscene understanding benchmark suite. *in Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

S.-H. Tsang. Review: Deepmask, an instance segment proposal method driven by convolution neural networks. 2018. https://towardsdatascience.com/review-deepmask-instance-segmentation-30327a072339.

P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. 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, 2015.

J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. 2016. doi: 10.1007/978-3-319-46466-4_22.

T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. 2018.

Y. Xue, J. Chen, W. Wan, Y. Huang, C. Yu, T. Li, and J. Bao. Mvscrf: Learning multi-view stereo with conditional random fields. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4312–4321, 2019.

F. Yang and Z. Zhou. Recovering 3d planes from a single image via convolutional neural networks. In *Computer Vision – ECCV 2018*, pages 87–103. Springer International Publishing, 2018. ISBN 978-3-030-01249-6.

F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. 2017. https://arxiv.org/pdf/1705.09914.pdf.

Z. Yu, J. Zheng, D. Lian, Z. Zhou, and S. Gao. Single-image piece-wise planar 3d reconstruction via associative embedding. *arXiv preprint arXiv:1902.09777*, 2019.

H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. 2016.

S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. 2015. doi: 10.1109/ICCV.2015.179. https://arxiv.org/pdf/1502.03240.pdf.

B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. 2018a. https://arxiv.org/pdf/1608.05442.pdf.

Q. Zhou, J. Park, and V. Koltun. Open3d: A modern library for 3d data processing. *arXiv:1801.09847*, 2018b.

S. Zlatanova and U. Isikdag. *3d indoor models and their applications*. Springer, 2017.

A
APPENDIX

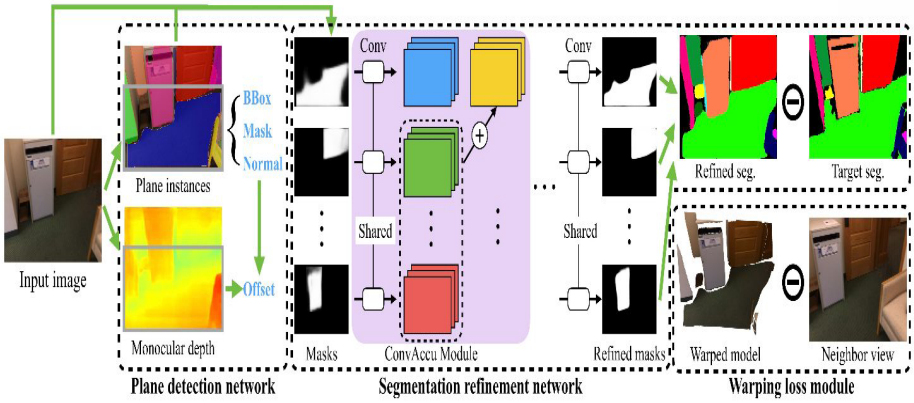

Figure A.1: overview of methodology by planercnn,[Liu et al., 2019] used in chapter 4 and 6

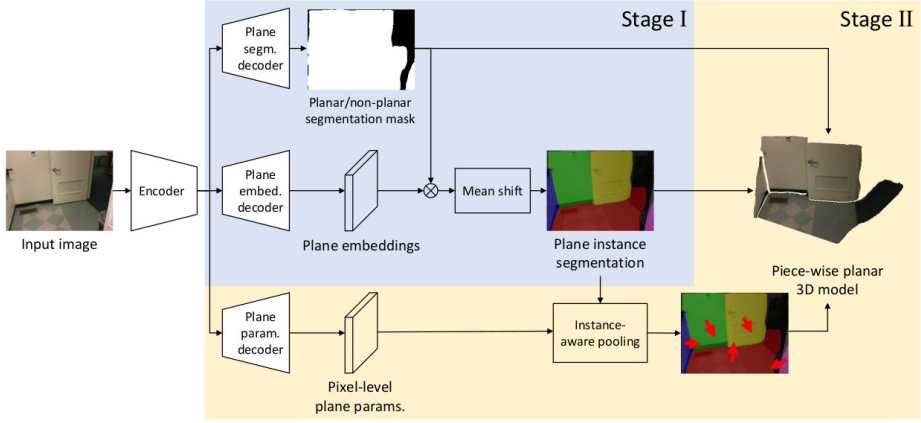

Figure A.2: overview of methodology by planarreconstruction, [Yu et al., 2019] used in chapter 4 and 6