PlaneNet: Piece-wise Planar Reconstruction from a Single RGB Image

Chen Liu[1] Jimei Yang[2] Duygu Ceylan[2] Ersin Yumer[3] Yasutaka Furukawa[4]

[1]Washington University in St. Louis [2]Adobe Research [3]Argo AI [4]Simon Fraser University

chenliu@wustl.edu {jimyang,ceylan}@adobe.com meyumer@gmail.com furukawa@sfu.ca

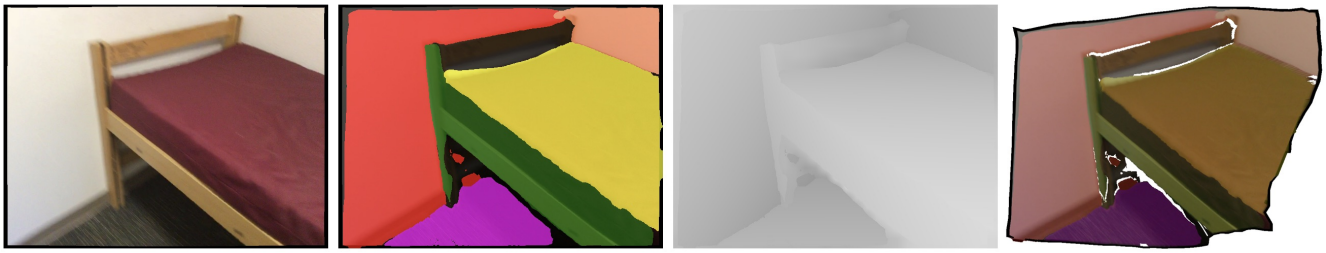

Figure 1: This paper proposes a deep neural architecture for piece-wise planar depthmap reconstruction from a single RGB image. From left to right, an input image, a piece-wise planar segmentation, a reconstructed depthmap, and a texture-mapped 3D model.

Abstract

This paper proposes a deep neural network (DNN) for piece-wise planar depthmap reconstruction from a single RGB image. While DNNs have brought remarkable progress to single-image depth prediction, piece-wise planar depthmap reconstruction requires a structured geometry representation, and has been a difficult task to master even for DNNs. The proposed end-to-end DNN learns to directly infer a set of plane parameters and corresponding plane segmentation masks from a single RGB image. We have generated more than 50,000 piece-wise planar depthmaps for training and testing from ScanNet, a large-scale RGBD video database. Our qualitative and quantitative evaluations demonstrate that the proposed approach outperforms baseline methods in terms of both plane segmentation and depth estimation accuracy. To the best of our knowledge, this paper presents the first end-to-end neural architecture for piece-wise planar reconstruction from a single RGB image. Code and data are available at https://github.com/art-programmer/PlaneNet.

1. Introduction

Human vision has a remarkable perceptual capability in understanding high-level scene structures. Observing a typical indoor scene (e.g., Fig. 1), we can instantly parse a room into a few number of dominant planes (e.g., a floor, walls, and a ceiling), perceive major surfaces for a furniture, or recognize a horizontal surface at the table-top. Piece-wise planar geometry understanding would be a key for many applications in emerging domains such as robotics or augmented reality (AR). For instance, a robot needs to identify the extent of a floor to plan a movement, or a table-top segmentation to place objects. In AR applications, planar surface detection is becoming a fundamental building block for placing virtual objects on a desk [17], replacing floor textures, or hanging artworks on walls for interior remodeling. A fundamental problem in Computer Vision is to develop a computational algorithm that masters similar perceptual capability to enable such applications.

With the surge of deep neural networks, single image depthmap inference [8, 7, 20, 35, 36] and room layout estimation [21] have been active areas of research. However, to our surprise, little attention has been given to the study of *piece-wise planar depthmap reconstruction*, mimicking this remarkable human perception in a general form. The main challenge is that the piece-wise planar depthmap requires structured geometry representation (i.e., a set of plane parameters and their segmentation masks). In particular, we do not know the number of planes to be inferred, and the order of planes to be regressed in the output feature vector, making the task challenging even for deep neural networks.

This paper proposes a novel deep neural architecture "PlaneNet" that learns to directly produce a set of plane

*Work done during Chen Liu's internship at Adobe Research.

parameters and probabilistic plane segmentation masks from a single RGB image. Following a recent work on point-set generation [9], we define a loss function that is agnostic to the order of planes. We further control the number of planes by allowing probabilistic plane segmentation masks to be all 0 [33]. The network also predicts a depthmap at non-planar surfaces, whose loss is defined through the probabilistic segmentation masks to allow back-propagation. We have generated more than 50,000 piece-wise planar depthmaps from ScanNet [6] as ground-truth by fitting planes to 3D points and projecting them to images. Qualitative and quantitative evaluations show that our algorithm produces significantly better plane segmentation results than the current state-of-the-art. Furthermore, our depth prediction accuracy is on-par or even superior to the existing single image depth inference techniques that are specifically trained for this task.

2. Related work

Multi-view piece-wise planar reconstruction. Piece-wise planar depthmap reconstruction was once an active research topic in multi-view 3D reconstruction [12, 31, 13, 40]. The task is to infer a set of plane parameters and assign a plane-ID to each pixel. Most existing methods first reconstruct precise 3D points, perform plane-fitting to generate plane hypotheses, then solve a global inference problem to reconstruct a piece-wise planar depthmap. Our approach learns to directly infer plane parameters and plane segmentations from a single RGB image.

Learning based depth reconstruction. Saxena *et al*. [28] pioneered a learning based approach for depthmap inference from a single image. With the surge of deep neural networks, numerous CNN based approaches have been proposed [8, 23, 27]. However, most techniques simply produce an array of depth values (i.e., depthmap) without plane detection or segmentation. More recently, Wang *et al*. [35] enforce planarity in depth (and surface normal) predictions by inferring pixels on planar surfaces. This is the closest work to ours. However, they only produce a binary segmentation mask (i.e., if a pixel is on a planar surface or not) without plane parameters or instance-level plane segmentation.

Layout estimation. Room layout estimation also aims at predicting dominant planes in a scene (e.g., walls, floor, and ceiling). Most traditional approaches [16, 22, 14, 29, 10, 34] rely on image processing heuristics to estimate vanishing points of a scene, and aggregate low-level features by a global optimization procedure. Besides low-level features, high-level information has been utilized, such as human poses [4, 10] or semantics [4, 2]. Attempts have been made to go beyond room structure, and predict object geometry [14, 34, 2, 42]. However, the reliance on hand-crafted features makes those methods less robust, and the Manhattan World assumption limits their operating ranges. Recently, Lee *et al*. [21] proposed an end-to-end deep neural network, RoomNet, which simultaneously classifies a room layout type and predicts corner locations. However, their framework is not applicable to general piece-wise planar scenes.

Line analysis. Single image 3D reconstruction of line drawings date back to the 60s. The earliest attempt is probably the Robert's system [26], which inspired many follow-up works [32, 37]. In real images, extraction of line drawings is challenging. Statistical analysis of line directions, junctions, or image segments have been used to enable 3D reconstruction for architectural scenes [25] or indoor panoramas [38]. Attributed grammar was used to parse an image into a hierarchical graph for 3D reconstruction [24]. However, these approaches require hand-crafted features, grammar specification, or algorithmic rules. Our approach is purely data-driven harnessing the power of deep neural networks.

3. PlaneNet

We build our network upon Dilated Residual Networks (DRNs) [39, 5] (See Fig. 2), which is a flexible framework for both global tasks (e.g., image classification) and pixel-wise prediction tasks (e.g., semantic segmentation). Given the high-resolution final feature maps from DRN, we compose three output branches for the three prediction tasks.

Plane parameters: For each scene, we predict a fixed number (K) of planar surfaces $\mathcal{S} = \{S_1, \cdots S_K\}$. Each surface S_i is specified by the three plane parameters P_i (i.e., encoding a normal and an offset). We use D_i to denote a depth image, which can be inferred from the parameters P_i *.

Non-planar depthmap: We model non-planar structures and infer its geometry as a standard depthmap. With abuse of notation, we treat it as the $(K+1)^{th}$ surface and denote the depthmap as D_{K+1}. This does not explain planar surfaces.

Segmentation masks: The last output is the probabilistic segmentation masks for the K planes $(M_1, \cdots M_K)$ and the non-planar depthmap (M_{K+1}).

To summarize, the network predicts 1) plane parameters $(P_1, \cdots, P_K)$, 2) a non-planar depthmap (D_{K+1}), and 3) probabilistic segmentation masks $(M_1, \cdots, M_{K+1})$. We now explain more details and the loss function for each task.

3.1. Plane parameter branch

The plane parameter branch starts with a global average-pooling to reduce the feature map size to 1x1 [39], followed by a fully connected layer to produce $K \times 3$ plane parameters. We do not know the number of planes as well as their order in this prediction task. By following prior works [9, 33], we predict a constant number (K) of planes, then allow some

*The depth value calculation requires camera intrinsic parameters, which can be estimated via vanishing point analysis, for example. In our experiments, intrinsics are given for each image through the database information.

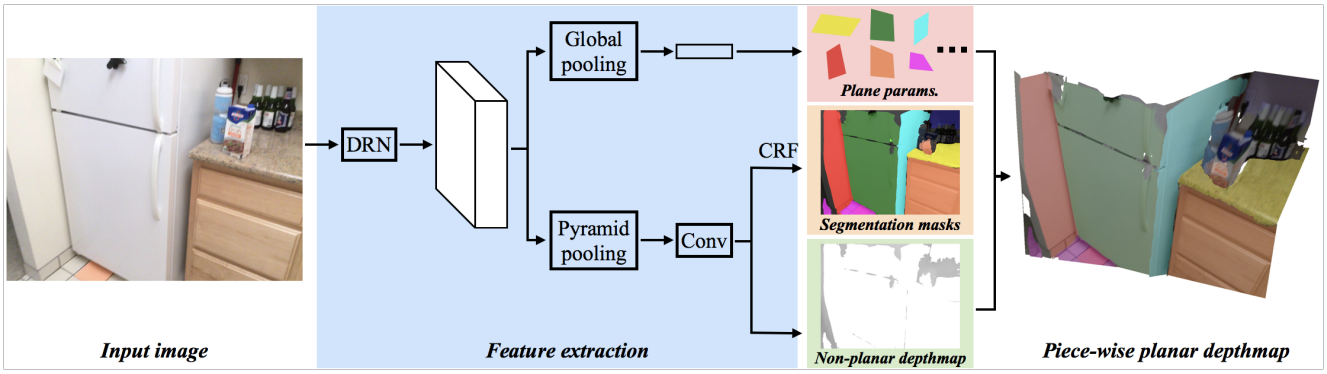

Figure 2: PlaneNet predicts plane parameters, their probabilistic segmentation masks, and a non-planar depthmap from a single RGB image.

predictions to be invalid by letting the corresponding probabilistic segmentation masks to be 0. Our ground-truth generation process (See Sect. 4) produces at most 10 planes for most examples, thus we set $K = 10$ in our experiments. We define an order-agnostic loss function based on the Chamfer distance metric for the regressed plane parameters:

$$\mathcal{L}^P = \sum_{i=1}^{K^*} \min_{j \in [1,K]} \|P_i^* - P_j\|_2^2 . \quad (1)$$

The parameterization P_i is given by the 3D coordinate of the point that is closest to the camera center on the plane. P_i^* is the ground-truth. K^* is the number of ground-truth planes.

3.2. Plane segmentation branch

The branch starts with a pyramid pooling module [43], followed by a convolutional layer to produce $K + 1$ channel likelihood maps for planar and non-planar surfaces. We append a dense conditional random field (DCRF) module based on the fast inference algorithm proposed by Krahenbuhl and Koltun [19], and jointly train the DCRF module with the precedent layers as in Zheng *et al.* [44]. We set the number of meanfield iterations to 5 during training and to 10 during testing. Bandwidths of bilateral filters are fixed for simplicity. We use a standard softmax cross entropy loss to supervise the segmentation training:

$$\mathcal{L}^M = \sum_{i=1}^{K+1} \sum_{p \in I} (\mathbf{1}(M^{*(p)} = i) \log(1 - M_i^{(p)})) \quad (2)$$

The internal summation is over the image pixels (I), where $M_i^{(p)}$ denotes the probability of pixel p belonging to the i^{th} plane. $M^{*(p)}$ is the ground-truth plane-id for the pixel.

3.3. Non-planar depth branch

The branch shares the same pyramid pooling module, followed by a convolution layer to produce a 1-channel depthmap. Instead of defining a loss specifically for non-planar regions, we found that exploiting the entire ground-truth depthmap makes the overall training more effective. Specifically, we define the loss as the sum of squared depth differences between the ground-truth and either a predicted plane or a non-planar depthmap, weighted by probabilities:

$$\mathcal{L}^D = \sum_{i=1}^{K+1} \sum_{p \in I} (M_i^{(p)} (D_i^{(p)} - D^{*(p)})^2) \quad (3)$$

$D_i^{(p)}$ denotes the depth value at pixel p, while $D^{*(p)}$ is the ground truth depth value.

4. Datasets and implemenation details

We have generated 51,000 ground-truth piece-wise planar depthmaps (50,000 training and 1,000 testing) from ScanNet [6], a large-scale indoor RGB-D video database. A depthmap in a single RGB-D frame contains holes and the quality deteriorates at far distances. Our approach for ground-truth generation is to directly fit planes to a consolidated mesh and project them back to individual frames, while also exploiting the associated semantic annotations.

Specifically, for each sub mesh-models of the same semantic label, we treat mesh-vertices as points and repeat extracting planes by RANSAC with replacement. The inlier distance threshold is $5cm$, and the process continues until 90% of the points are covered. We merge two (not necessarily adjacent) planes that span different semantic labels if the plane normal difference is below $20°$, and if the larger plane fits the smaller one with the mean distance error below $5cm$. We project each triangle to individual frames if the three vertices are fitted by the same plane. After projecting all the triangles, we keep only the planes whose projected area is larger than 1% of an image. We discard entire frames if the ratio of pixels covered by the planes is below 50%. For training samples, we randomly choose 90% of the scenes from ScanNet, subsample every 10 frames, compute piecewise planar depthmaps with the above procedure, then use

the final random sampling to produce 50,000 examples. The same procedure generates 1,000 testing examples from the remaining 10% of the scenes.

We have implemented PlaneNet using TensorFlow [1] based on DeepLab [5]. Our system is a 101-layer ResNet [15] with Dilated Convolution, while we have followed a prior work and modified the first few layers to deal with the degridding issue [39]. The final feature map of the DRN contains 2096 channels. We use the Adam optimizer [18] with the initial learning rate set to 0.0003. The input image, the output plane segmentation masks, and the non-planar depthmap have a resolution of 256x192. We train our network for 50 epochs on the 50,000 training samples.

5. Experimental results

Figure 3 shows our reconstruction results for a variety of scenes. Our end-to-end learning framework has successfully recovered piece-wise planar and semantically meaningful structures, such as floors, walls, table-tops, or a computer screen, from a single RGB image. We include many more examples in the supplementary material. We now provide quantitative evaluations on the plane segmentation accuracy and the depth reconstruction accuracy against the competing baselines, followed by more analyses of our results.

5.1. Plane segmentation accuracy

Piece-wise planar reconstruction from a single RGB image is a challenging problem. While existing approaches have produced encouraging results [11, 24, 25], they are based on hand-crafted features and algorithmic designs, and may not match against big-data and deep neural network (DNN) based systems. Much better baselines would then be piece-wise planar depthmap reconstruction techniques from 3D points [12, 31, 13, 40], where input 3D points are either given by the ground-truth depthmaps or inferred by a state-of-the-art DNN-based system [20].

In particular, to infer depthmaps, we have used a variant of PlaneNet which only has the pixel-wise depthmap branch, while following Eigen *et al*. [7] to change the loss. Table 1 shows that this network, PlaneNet (Depth rep.), outperforms the current top-performers on the NYU benchmark [30].

For piece-wise planar depthmap reconstruction, we have used the following three baselines from the literature.

- "NYU-Toolbox" is a plane extraction algorithm from the official NYU toolbox [30] that extracts plane hypotheses using RANSAC, and optimizes the plane segmentation via a Markov Random Field (MRF) optimization.
- Manhattan World Stereo (MWS) [12] is very similar to NYU-Toolbox except that MWS employs the Manhattan World assumption in extracting planes and exploits vanishing lines in the pairwise terms to improve results.
- Piecewise Planar Stereo (PPS) [31] relaxes the Manhattan World assumption of MWS, and uses vanishing lines to generate better plane proposals. Please see the supplementary document for more algorithmic details on the baselines.

Figure 4 shows the evaluation results on two recall metrics. The first metric is the percentage of correctly predicted ground-truth planes. We consider a ground-truth plane being correctly predicted, if one of the inferred planes has 1) more than 0.5 Intersection over Union (IOU) score and 2) the mean depth difference over the overlapping region is less than a threshold. We vary this threshold from 0 to 0.6m with an increment of 0.05m to plot graphs. The second recall metric is simply the percentage of pixels that are in such overlapping regions where planes are correctly predicted. The figure shows that PlaneNet is significantly better than all the competing methods when inferred depthmaps are used. PlaneNet is even better than some competing methods that use ground-truth depthmaps. This demonstrates the effectiveness of our approach, learning to infer piece-wise planar structures from many examples.

Figure 5 shows qualitative comparisons against existing methods with inferred depthmaps. PlaneNet produces significantly better plane segmentation results, while existing methods often generate many redundant planes where depthmaps are noisy, and fail to capture precise boundaries where the intensity edges are weak.

5.2. Depth reconstruction accuracy

While the capability to infer a plane segmentation mask and precise plane parameters is the key contribution of the work, it is also interesting to compare against depth prediction methods. This is to ensure that our structured depth prediction does not compromise per-pixel depth prediction accuracy. PlaneNet makes (K+1) depth value predictions at each pixel. We pick the depth value with the maximum probability in the segmentation mask to define our depthmap.

Depth accuracies are evaluated on the NYUv2 dataset at 1) planar regions, 2) boundary regions, and 3) the entire image, against three competing baselines. [†] Eigen-VGG [7] is a convolutional architecture to predict both depths and surface normals. SURGE [35] is a more recent depth inference network that optimizes planarity. FCRN is the current state-of-the-art single-image depth inference network [20] [‡].

Depthmaps in NYUv2 are very noisy and ground-truth plane extraction does not work well. Thus, we fine-tune our network using only the depth loss (3). Note that the key factor in this training is that the network is trained to generate a depthmap through our piece-wise planar depthmap represen-

[†] Following existing works, we choose NYUv2 to evaluate depth accuracy and consider only the valid 561x427 area as the entire image evaluation.

[‡] The numbers are different from the numbers reported in [20] since [20] evaluate on the original resolution, 640x480, and their numbers are influenced by the issue reported at https://github.com/iro-cp/FCRN-DepthPrediction/issues/42.

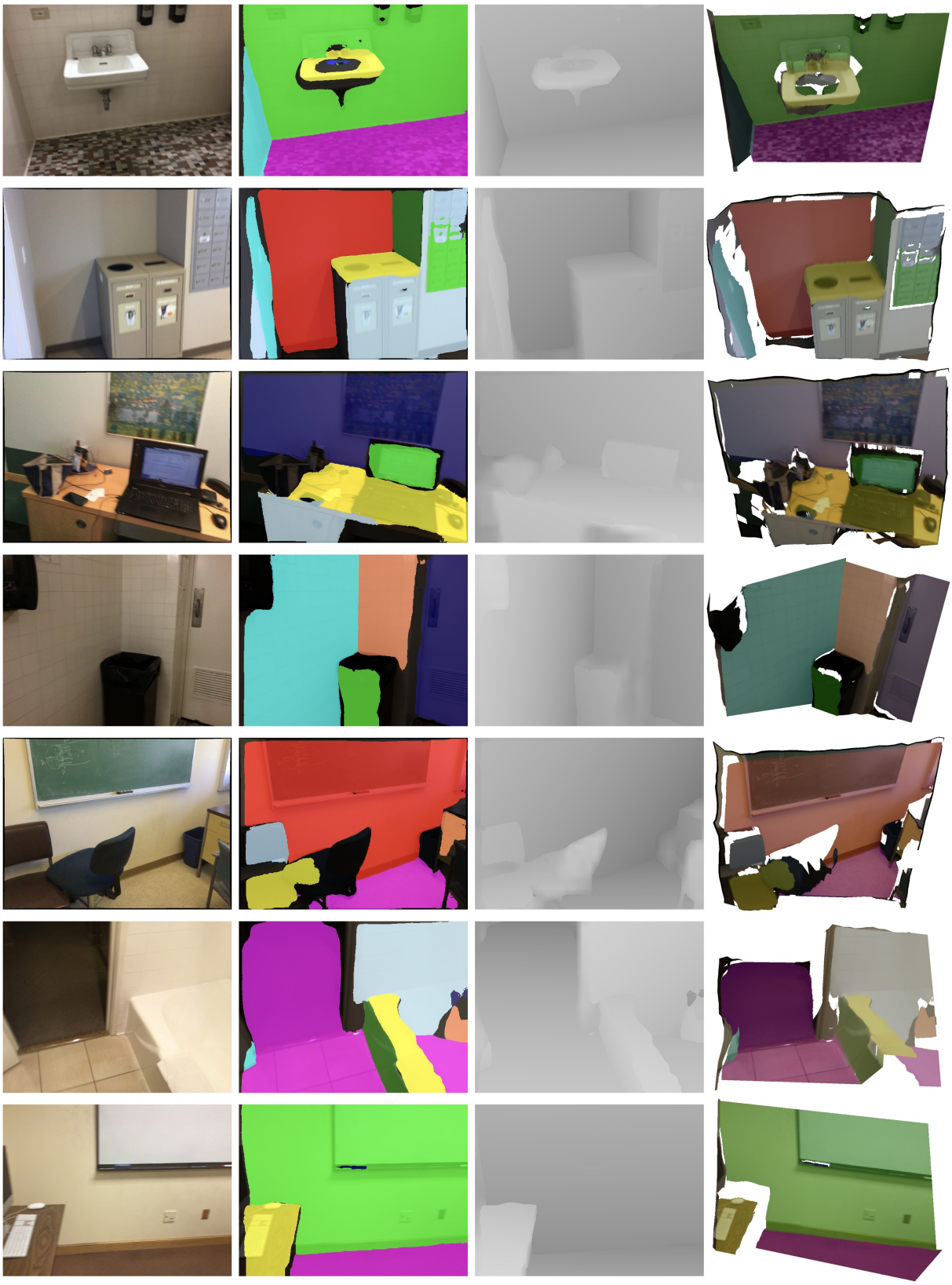

Figure 3: Piece-wise planar depthmap reconstruction results by PlaneNet. From left to right: input image, plane segmentation, depthmap reconstruction, and 3D rendering of our depthmap. In the plane segmentation results, the black color shows non-planar surface regions.

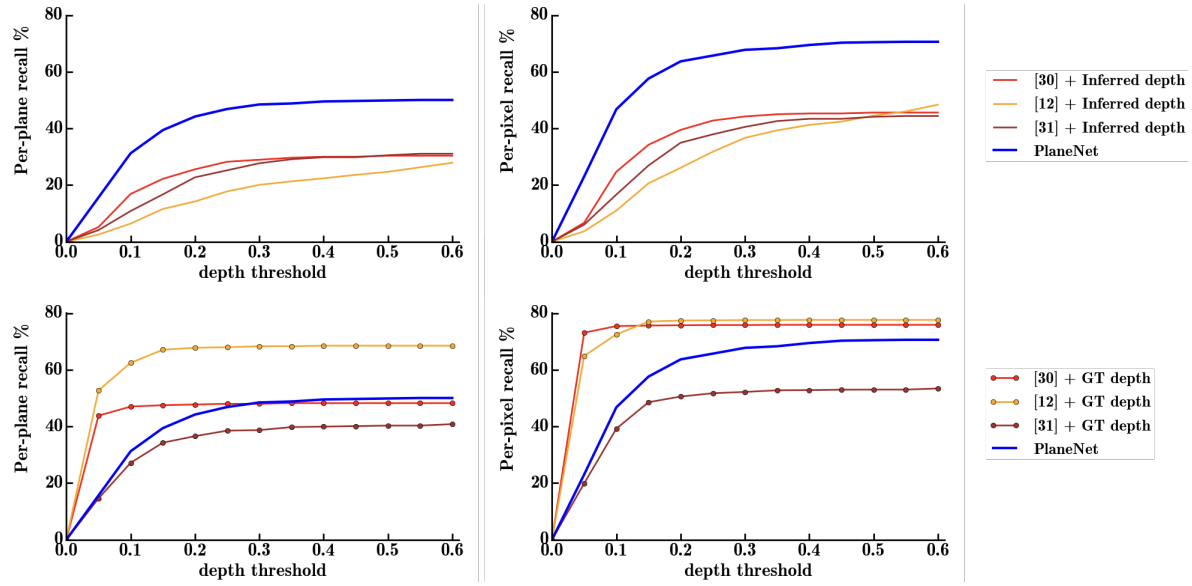

Figure 4: Plane segmentation accuracy against competing baselines that use 3D points as input [30, 12, 31]. Either ground-truth depthmaps or inferred depthmaps (by a DNN-based system) are used as their inputs. PlaneNet outperforms all the other methods that use inferred depthmaps. Surprisingly, PlaneNet is even better than many other methods that use ground-truth depthmaps.

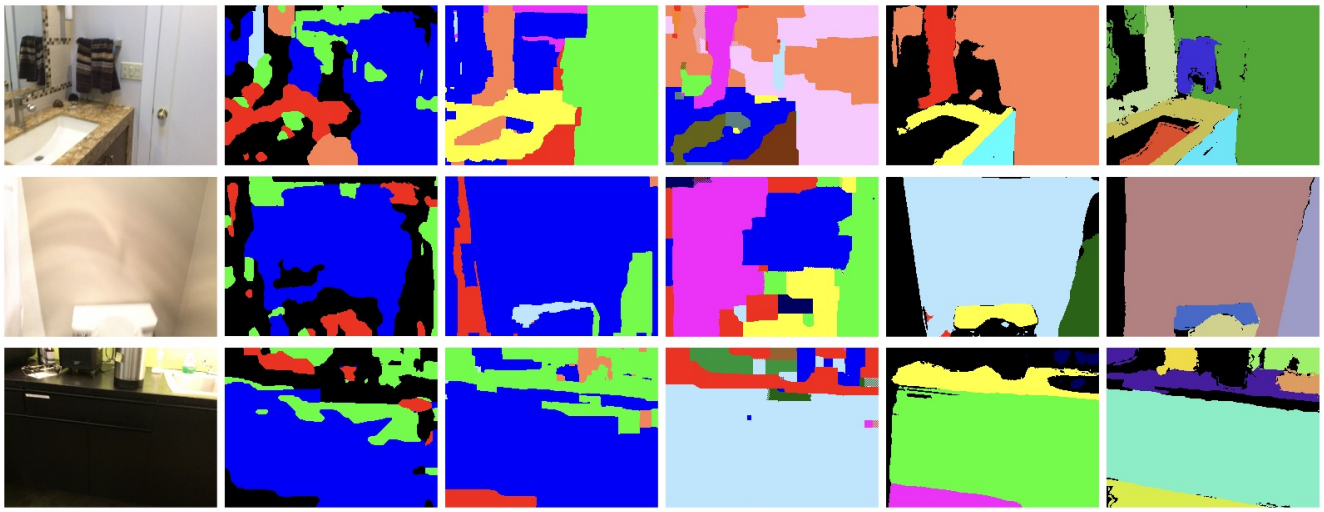

Figure 5: Qualitative comparisons between PlaneNet and existing methods that use inferred depthmaps as the inputs. From left to right: an input image, plane segmentation results for [30], [12], [31], and PlaneNet, respectively, and the ground-truth.

tation. To further verify the effects of this representation, we have also fine-tuned our network in the standard per-pixel depthmap representation by disabling the plane parameter and the plane segmentation branches. In this version, denoted as "PlaneNet (Depth rep.)", the entire depthmap is predicted in the $(K+1)^{th}$ depthmap (D_{K+1}).

Table 1 shows the depth prediction accuracy on various metrics introduced in the prior work [8]. The left five metrics provide different error statistics such as relative difference (Rel) or rooted-mean-square-error (RMSE) on the average per-pixel depth errors. The right three metrics provide the ratio of pixels, for which the relative difference between the predicted and the ground-truth depths is below a threshold. The table demonstrates that PlaneNet outperforms the state-of-the-art of single-image depth inference techniques. As observed in prior works [35, 3], the planarity constraint makes differences in the depth prediction task, and the improvements are more significant when our piece-wise planar representation is enforced by our network.

Table 1: Depth accuracy comparisons over the NYUv2 dataset.

Method	Lower the better (LTB) Rel	Rel(sqr)	log_{10}	$RMSE_{iin}$	$RMSE_{log}$	Higher the better (HTB) 1.25	1.25^2	1.25^3
Evaluation over planar regions								
Eigen-VGG [7]	0.143	0.088	0.061	0.490	0.193	80.1	96.4	**99.3**
SURGE [35]	0.142	0.087	0.061	0.487	0.192	80.2	96.6	**99.3**
FCRN [20]	0.140	0.087	0.065	0.460	0.183	79.2	95.6	99.0
PlaneNet (Depth rep.)	0.130	0.080	**0.054**	0.399	0.156	**84.4**	96.7	99.2
PlaneNet	**0.129**	**0.079**	**0.054**	**0.397**	**0.155**	84.2	**96.8**	99.2
Evaluation over edge areas								
Eigen-VGG [7]	0.165	0.137	0.073	0.727	0.228	72.9	74.3	98.7
SURGE [35]	0.162	0.133	0.071	0.697	0.221	74.7	94.7	98.7
FCRN [20]	0.154	0.111	0.073	0.548	0.208	74.7	94.1	98.5
PlaneNet (Depth rep.)	**0.145**	**0.099**	**0.061**	0.480	0.179	**80.9**	95.9	99.0
PlaneNet	**0.145**	**0.099**	**0.061**	**0.479**	**0.178**	80.7	**96.1**	**99.1**
Evaluation over the entire image								
Eigen-VGG [7]	0.158	0.121	0.067	0.639	0.215	77.1	95.0	98.8
SURGE [35]	0.156	0.118	0.067	0.643	0.214	76.8	95.1	**98.9**
FCRN [20]	0.152	0.119	0.072	0.581	0.207	75.6	93.9	98.4
PlaneNet (Depth rep.)	0.143	**0.107**	**0.060**	0.518	0.180	**81.3**	95.5	98.7
PlaneNet	**0.142**	**0.107**	**0.060**	**0.514**	**0.179**	81.2	**95.7**	**98.9**

5.3. Plane ordering consistency

The ordering ambiguity is a challenge for piece-wise depthmap inference. We found that PlaneNet automatically learns a consistent ordering without supervision, for example, the floor is always regressed as the second plane. In Fig. 3, colors in the plane segmentation results are defined by the order of the planes in the network output. Although the ordering loses consistency for small objects or extreme camera angles, major common surfaces such as the floor and walls have a consistent ordering in most cases.

We have exploited this property and implemented a simple room layout estimation algorithm. More specifically, we look at reconstruction examples and manually select the entries of planes that correspond to the ceiling, the floor, and the left/middle/right walls. For each possible room layout configuration [21], (e.g., a configuration with a floor, a left wall, and a middle wall visible), we construct a 3D concave hull based on the plane parameters and project it back to the image to generate a room-layout. We measure the score of the configuration by the number of pixels, where the constructed room layout and the inferred plane segmentation (determined by the winner-takes-all) agree. We pick the constructed room layout with the best score as our prediction. Figure 6 shows that our algorithm is able to generate reasonable room layout estimations even when the scene is cluttered and contain many occluding objects. Table 2 shows the quantitative evaluations on the NYUv2 303 dataset [41], where our method is comparable to existing techniques which are designed specifically for this task. [§]

Table 2: Room layout estimations. Quantitative evaluations against the top-performers over the NYUv2 303 dataset.

	Input	Layout error
Schwing *et al.* [29]	RGB	13.66%
Zhang *et al.* [41]	RGB	13.94%
Zhang *et al.* [41]	RGB+D	8.04%
RoomNet [21]	RGB	12.96%
PlaneNet	RGB	12.64%

5.4. Failure modes

While achieving promising results on most images, PlaneNet has some failure modes as shown in Fig. 7. In the first example, PlaneNet generates two nearly co-planar vertical surfaces in the low-light region below the sink. In the second example, it cannot distinguish a white object on the floor from a white wall. In the third example, it misses a column structure on a wall due to the presence of object

[§]RoomNet paper [21] does not provide code or evaluation numbers for the NYUv2 benchmark. We have implemented their system using Torch7 and trained on LSUN dataset as described in their paper.

Figure 6: Room layout estimations. We have exploited the ordering consistency in the predicted planes to infer room layouts.

Figure 7: Typical failure modes occur in the absence of enough image texture cues or at the presence of small objects and clutter.

clutter. While the capability to infer precise plane parameters is already super-human, there is a lot of room for improvement on the planar segmentation, especially in the absence of texture information or at the presence of clutter.

6. Applications

Structured geometry reconstruction is important for many application in Augmented Reality. We demonstrate two image editing applications enabled by our piece-wise planar representation: texture insertion and replacement (see Fig. 8).

Figure 8: Texture editing applications. From top to bottom, an input image, a plane segmentation result, and an edited image.

We first extract Manhattan directions by using the predicted plane normals through a standard voting scheme [12]. Given a piece-wise planar region, we define an axis of its UV coordinate by the Manhattan direction that is the most parallel to the plane, while the other axis is simply the cross product of the first axis and the plane normal. Given a UV coordinate, we insert a new texture by alpha-blending or completely replace a texture with a new one. Please see the supplementary material and the video for more AR application examples.

7. Conclusion and future work

This paper proposes PlaneNet, the first deep neural architecture for piece-wise planar depthmap reconstruction from a single RGB image. PlaneNet learns to directly infer a set of plane parameters and their probabilistic segmentation masks. The proposed approach significantly outperforms competing baselines in the plane segmentation task. It also advances the state-of-the-art in the single image depth prediction task. An interesting future direction is to go beyond the depthmap framework and tackle structured geometry prediction problems in a full 3D space.

8. Acknowledgement

This research is partially supported by National Science Foundation under grant IIS 1540012 and IIS 1618685, Google Faculty Research Award, and Adobe gift fund. We thank Nvidia for a generous GPU donation.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 4

[2] S. Y. Bao, A. Furlan, L. Fei-Fei, and S. Savarese. Understanding the 3d layout of a cluttered room from multiple images. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 690–697. IEEE, 2014. 2

[3] A. Chakrabarti, J. Shao, and G. Shakhnarovich. Depth from a single image by harmonizing overcomplete local network predictions. In *Advances in Neural Information Processing Systems*, pages 2658–2666, 2016. 6

[4] Y.-W. Chao, W. Choi, C. Pantofaru, and S. Savarese. Layout estimation of highly cluttered indoor scenes using geometric and semantic cues. In *International Conference on Image Analysis and Processing*, pages 489–499. Springer, 2013. 2

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 2, 4

[6] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *arXiv preprint arXiv:1702.04405*, 2017. 2, 3

[7] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015. 1, 4, 7

[8] 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. 1, 2, 6

[9] H. Fan, H. Su, and L. Guibas. A point set generation network for 3d object reconstruction from a single image. *arXiv preprint arXiv:1612.00603*, 2016. 2

[10] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single view geometry. *International journal of computer vision*, 110(3):259–274, 2014. 2

[11] D. F. Fouhey, A. Gupta, and M. Hebert. Unfolding an indoor origami world. In *European Conference on Computer Vision*, pages 687–702. Springer, 2014. 4

[12] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Manhattan-world stereo. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1422–1429. IEEE, 2009. 2, 4, 6, 8

[13] D. Gallup, J.-M. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1418–1425. IEEE, 2010. 2, 4

[14] A. Gupta, M. Hebert, T. Kanade, and D. M. Blei. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *Advances in neural information processing systems*, pages 1288–1296, 2010. 2

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[16] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *Computer vision, 2009 IEEE 12th international conference on*, pages 1849–1856. IEEE, 2009. 2

[17] K. Karsch, K. Sunkavalli, S. Hadap, N. Carr, H. Jin, R. Fonte, M. Sittig, and D. Forsyth. Automatic scene inference for 3d object compositing. *ACM Trans. Graph.*, 33(3):32:1–32:15, June 2014. 1

[18] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[19] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. 3

[20] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016. 1, 4, 7

[21] C.-Y. Lee, V. Badrinarayanan, T. Malisiewicz, and A. Rabinovich. Roomnet: End-to-end room layout estimation. *arXiv preprint arXiv:1703.06241*, 2017. 1, 2, 7

[22] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2136–2143. IEEE, 2009. 2

[23] 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. 2

[24] X. Liu, Y. Zhao, and S.-C. Zhu. Single-view 3d scene parsing by attributed grammar. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 684–691, 2014. 2, 4

[25] S. Ramalingam and M. Brand. Lifting 3d manhattan lines from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 497–504, 2013. 2, 4

[26] L. G. Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963. 2

[27] A. Roy and S. Todorovic. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5506–5514, 2016. 2

[28] 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. 2

[29] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Efficient structured prediction for 3d indoor scene understanding. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2815–2822. IEEE, 2012. 2, 7

[30] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. *Computer Vision–ECCV 2012*, pages 746–760, 2012. 4, 6

[31] S. Sinha, D. Steedly, and R. Szeliski. Piecewise planar stereo for image-based rendering. 2009. 2, 4, 6

[32] K. Sugihara. *Machine interpretation of line drawings*, volume 1. MIT press Cambridge, 1986. 2

[33] S. Tulsiani, H. Su, L. J. Guibas, A. A. Efros, and J. Malik. Learning shape abstractions by assembling volumetric primitives. *arXiv preprint arXiv:1612.00404*, 2016. 2

[34] H. Wang, S. Gould, and D. Roller. Discriminative learning with latent variables for cluttered indoor scene understanding. *Communications of the ACM*, 56(4):92–99, 2013. 2

[35] P. Wang, X. Shen, B. Russell, S. Cohen, B. Price, and A. L. Yuille. Surge: Surface regularized geometry estimation from a single image. In *Advances in Neural Information Processing Systems*, pages 172–180, 2016. 1, 2, 4, 6, 7

[36] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of CVPR*, 2017. 1

[37] T. Xue, J. Liu, and X. Tang. Example-based 3d object reconstruction from line drawings. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 302–309. IEEE, 2012. 2

[38] H. Yang and H. Zhang. Efficient 3d room shape recovery from a single panorama. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5422–5430, 2016. 2

[39] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. *arXiv preprint arXiv:1705.09914*, 2017. 2, 4

[40] L. Zebedin, J. Bauer, K. Karner, and H. Bischof. Fusion of feature-and area-based information for urban buildings modeling from aerial imagery. *Computer Vision–ECCV 2008*, pages 873–886, 2008. 2, 4

[41] J. Zhang, C. Kan, A. G. Schwing, and R. Urtasun. Estimating the 3d layout of indoor scenes and its clutter from depth sensors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1273–1280, 2013. 7

[42] Y. Zhang, S. Song, P. Tan, and J. Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *European Conference on Computer Vision*, pages 668–686. Springer, 2014. 2

[43] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. *arXiv preprint arXiv:1612.01105*, 2016. 3

[44] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015. 3