PlaneRecNet: Multi-Task Learning with Cross-Task Consistency for Piece-Wise Plane Detection and Reconstruction from a Single RGB Image

Yaxu Xie
yaxu.xie@dkfi.de
Fangwen Shu
fangwen.shu@dkfi.de
Jason Rambach
jason_raphael.rambach@dfki.de
Alain Pagani
alain.pagani@dkfi.de
Didier Stricker
didier.stricker@dkfi.de

German Research Centre for Artificial Intelligence (DFKI)
Kaiserslautern, Germany

Abstract

Piece-wise 3D planar reconstruction provides holistic scene understanding of man-made environments, especially for indoor scenarios. Most recent approaches focused on improving the segmentation and reconstruction results by introducing advanced network architectures but overlooked the dual characteristics of piece-wise planes as objects and geometric models. Different from other existing approaches, we start from enforcing cross-task consistency for our multi-task convolutional neural network, PlaneRecNet, which integrates a single-stage instance segmentation network for piece-wise planar segmentation and a depth decoder to reconstruct the scene from a single RGB image. To achieve this, we introduce several novel loss functions (geometric constraint) that jointly improve the accuracy of piece-wise planar segmentation and depth estimation. Meanwhile, a novel Plane Prior Attention module is used to guide depth estimation with the awareness of plane instances. Exhaustive experiments are conducted in this work to validate the effectiveness and efficiency of our method. github.com/EryiXie/PlaneRecNet

1 Introduction

Man-made environments usually contain rich planar surfaces. In indoor environments, orthogonal or parallel planar regions (walls, floor and ceiling) can form a layout that satisfies the Manhattan World (MW) assumption. Some SLAM systems [2, 6, 14, 15, 30, 33, 46] employ such structural features for fast dense map reconstruction or low drift localization. On the other hand, planar surface representations are also utilized to solve the occlusion-awareness in indoor scene understanding [10]. In outdoor scenarios, an accurate ground

plane estimation [22, 32] is an important prior for object pose estimation and autonomous driving. Overall, detection of 3D planes and scene reconstructing based on piece-wise planes have a very wide range of applications.

With learned prior about the geometry definition of the plane, human vision can easily detect planar regions in real world and have rough estimation of the plane normal no matter if the surface has complex texture or barely any texture, which is a difficult challenge for traditional computer vision methods. Recent deep learning approaches [18, 19, 41, 44] formulate the piece-wise plane estimation and reconstruction as a multi-task learning problem. Neural networks with separate branches provide predictions for segmentation mask, plane parameters and dense depth maps. The final scene reconstruction is then given by assembling the depth estimation of non-planar region and the predicted 3D planes.

However, this assembly may result in inconsistency of the reconstructed scene between planar and non-planar regions whenever plane parameter estimation or the plane segmentation is inaccurate. Furthermore, a loss of local details on the planar region is also inevitable. Addressing this, refinement methods based on multi-view consistency [19, 38] and inter-plane relationship [28] have been proposed to jointly improve the consistency across segmentation, parameter estimation and depth reconstruction. The geometric attributes of planes are also noticed and utilized by [41], so that the network can be trained under indirect supervision without giving piece-wise planar segmentation as ground truth. In addition to the non-optimal assembly as mentioned, we have also observed a common failure mode of existing methods in plane masks segmentation: the predicted mask can either cover multiple planes instance due to the texture similarity of these planes, or cover only a part of a plane instance due to the texture variety of that plane surface.

In this work, we propose PlaneRecNet, a multi-task convolutional neural network with a single-stage instance segmentation branch, a depth decoder with Plane Prior Attention and a shared backbone for piece-wise planar regions estimation and reconstruction. We formulate geometric constraint loss functions specific for 3D planes so that 1) the segmentation network can provide more accurate plane masks using learned occlusion clues, 2) the depth estimation in planar region is regularized with local surface normal. To summarize, our key **contributions** include:

- A multi-task convolutional neural network for piece-wise planar reconstruction with novel Plane Prior Attention module that improves the depth reconstruction with semantic priors.
- Novel depth gradient segmentation and plane surface normal loss functions, which enforce cross-task consistency between depth estimation and instance segmentation.
- Extensive experiments on datasets NYUv2 [24], ScanNet [9] and iBims-1 [13], validating the effectiveness of our method on both plane segmentation and depth estimation.

2 Related Work

Piece-wise Planar Segmentation and Reconstruction: PlaneNet [18] is the first multi-task deep neural network for piece-wise planar reconstruction from a single RGB image. It consists of three prediction branches for plane parameters estimation, plane segmentation and non-planar depth map estimation with a shared encoder. In PlaneRecover [41], Yang and Zhou introduced a novel plane structure-induced loss to train the plane segmentation and the plane parameters estimation together through the supervision on the depth recovered from predicted planar regions and parameters, instead of ground truth mask of plane regions.

From the aspect of segmentation accuracy, improvement was made by Z. Yu et al. [44], who first performed semantic segmentation to distinguish planar/non-planar regions, and then clustered planar pixels into piece-wise instances with their associative embedding vectors. PlaneRCNN [19] proposed a more effective plane segmentation branch built upon Mask R-CNN [9] and jointly refined the segmentation with their novel warping loss function. Qian and Furukawa [28] proposed a post-processing refinement network to optimize the predicted plane parameters and segmentation mask of the existing piece-wise planar reconstruction method by enforcing the inter-plane relationship. Xi and Chen [38] introduced multi-view regularization method to enforce the consistency of plane feature embedding from different views. PlaneSegNet [39] proposed a fast single-stage instance segmentation method for planes, and improved the resolution of the predicted masks. Despite advanced general instance segmentation network architectures improving accuracy, we argue that considering the particularity of piece-wise planar segmentation and reconstruction, the specific problem constraints are not thoroughly exploited in PlaneRecover [41] and Interplane [28]. The specific consistency constraint between the piece-wise planar segmentation and the depth reconstruction can be further explored by fully using the geometry characteristics of planes.

Geometric Constraint Aided Monocular Depth Estimation: Recently, the effectiveness of geometric constraints in monocular depth estimation was exploited to improve the accuracy of reconstructed scene, where the point-wise loss summation over all pixels shows limitations. Novel loss functions based on geometric constraints such as the surface normal and the occlusion boundary were introduced in several approaches [10, 20, 35, 42, 43]. Another open question is the proper method to obtain the reference of the geometry constraint from noisy ground truth depth, since many real datasets are obtained with consumer-level devices. Addressing this, Yin *et al*. [42] proposed Virtual Normal which shows robustness against local noise in experiments, thanks to the long-range triplet points sampling strategy. Long *et al*. [21] introduced an Adaptive Surface Normal Constraint. In their work, a series of random triplets are sampled within a local patch and adaptively weighted with the guidance feature given by another decoder branch in order to provide confident normal on sharp corner and edges. Piece-wise planes are also utilized to formulate geometry constraints, especially for reconstructing indoor scenes. Huynh *et al*. [10] indicated that the depth values of all points belonging to the same plane are linearly dependent, and they introduced a Depth-Attention Volume network to implicitly learn attention map based on co-planar relationships of each pixel. Long *et al*. [20] introduced the Combined Normal Map, which combined the local surface normal of non-planar regions and the mean of the surface normal in a planar region , in order to improve the reconstruction for both local high-curvature and global planar regions. The performance of self-supervised depth estimator in indoor datasets are limited, because large non-texture planar regions are difficult to be learned only with the multi-view photometric consistency. Yu *et al*. [45] utilized piece-wise planar regions annotated using superpixels as priors, to enforce a low plane-fitting error within in each planar region.

Cross-Task Consistency for Multi-Task Learning: As aforementioned, some works opted to train depth estimation networks with novel loss functions rather than introducing one or more separate decoders, while other works [22, 26, 27, 29, 36] focused on explicitly predicting surface normal, depth and other dense maps in separate streams, then jointly regulating the consistency between them and refining the initial scene reconstruction. The consistency across different tasks were not only noticed in 3D reconstruction but also in multi-task learning pairs such as depth estimation + semantic segmentation [31], depth estimation + scenes

parsing [40]. Meanwhile some fundamental questions about multi-task learning in computer vision are also explored. Stanley *et al*. [34] gave an empirical study of influence factors of multi-task learning and a detailed investigation about how tasks influence one another. Zamir *et al*. [47] proposed a general concept to simplify the supervision of the cross-task consistency in multi-task learning, so that the consistency constraint can be learned from data rather than a prior given relationship. So far, most of the works focus on the cross-task consistency between dense maps. Hence, we consider our work as an attempt to tackle the cross-task consistency problem between instance segmentation and dense map prediction.

3 Approach

Given a single color image I as input, we use a multi-branch neural network with a shared backbone to predict piece-wise planar segmentation M_{pred} and global depth estimation D_{pred}. We adapt a light-weight configuration of SOLO V2 [37] for piece-wise planar segmentation, which is a global mask based single-stage instance segmentation method. Segmentation mask is provided through the dynamic convolution operation between predicted mask kernels from prediction head and mask feature from mask head. Mask candidates are then fused into the depth branch as attention through the Plane Prior Attention module. The depth decoder is a lightweight Feature Pyramid Network [16] structure: feature maps from the encoder are upsampled through a series of convolution layers and fused with the corresponding feature maps from the encoder through skip connections.

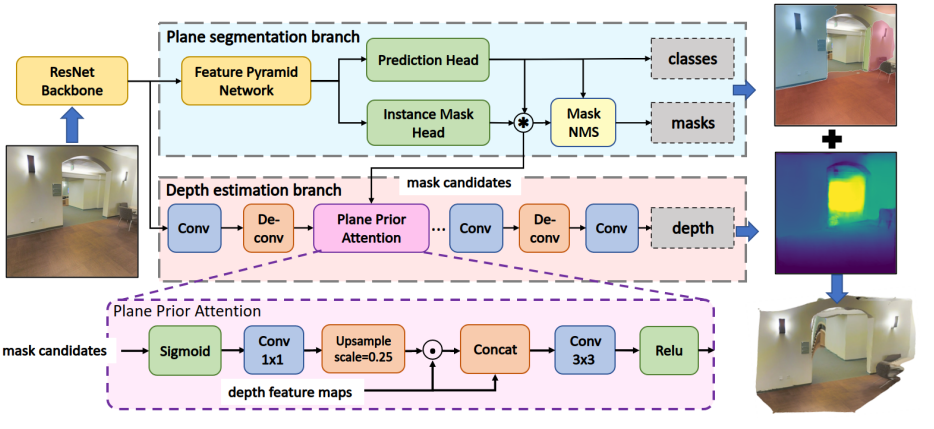


Figure 1: **The PlaneRecNet Architecture:** the network consists of a shared backbone and two task branches for piece-wise plane segmentation and monocular depth estimation. "$\odot$" presents the element-wise multiplication, "$\circledast$" presents the dynamic convolution operation.

Comparing to PlaneRCNN [19], another detection based approach for piece-wise plane reconstruction, our approach is different in following aspects: 1) We utilizes a global-mask based instance segmentation method [37], which provides segmentation masks with higher resolution than local-mask based method [9]. 2) Instead of explicitly predicting plane parameters and complicating the network architecture, we focus on providing a high quality global depth estimation, so that plane parameter can be computed simply with Principal component analysis (PCA) [7] or Random sample consensus (RANSAC) [5] algorithm with the plane

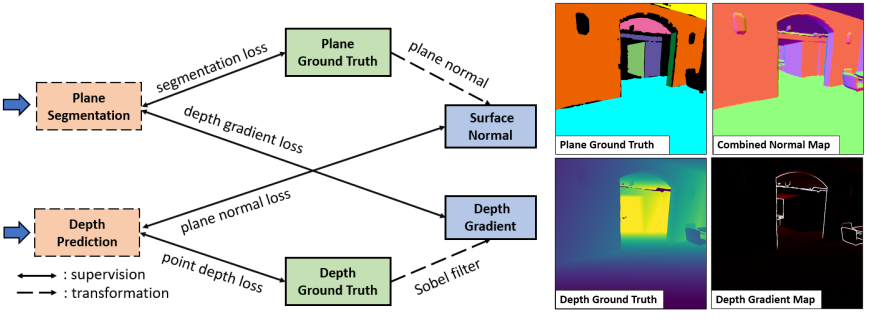
segmentation and the depth prediction. 3) PlaneRCNN introduced an extra segmentation refinement network using global depth, rough segmentation mask and coordinate map of the plane instance as inputs. Our network has no refinement module and it is therefore simpler and faster. The network architecture of our PlaneRecNet is illustrated in Figure 1.

3.1 Plane Prior Attention

Inspired by the Depth Attention Volume [10], we design the Plane Prior Attention (PPA) module, which fuses the plane prior information to the depth decoder, as illustrated in Figure 1. We implement the Plane Prior Attention module by first introducing mask candidates for piece-wise planes into the depth branch. We then reduce the redundant channel numbers of the mask candidates using 1×1 convolution, and resize the mask candidates with interpolation. After that, the mask candidates (as attention weights) are multiplied and concatenated with the depth feature maps. Finally the feature maps are fed into a standard 3×3 convolution block followed with Relu activation.

3.2 Guidance Loss Functions for Cross-Task Consistency

In piece-wise plane reconstruction, the cross-task consistency is naturally linked to the geometry of the planar surface. For depth estimation, we expect the predicted depth map can represent well the surface normal and the planarity of the planar surface of the scene. Moreover, for instance segmentation, a single predicted mask should not cover multiple planar regions, especially when these planar regions are geometrically inconsistent. To enforce such cross-task consistency, we utilize the depth ground truth as a clue of occlusion boundary to constraint the instance segmentation, and the segmentation ground truth as surface normal constraint to regularize the depth estimation, as illustrated in Figure 2(a).



(a) Graph structure of loss functions (b) Ground truth and their conversion

Figure 2: **Graph structure of loss functions for cross-task consistency:** with surface normal and depth gradient as intermediaries, loss functions are built into a graph, so that the plane segmentation is supervised by depth ground truth and the depth prediction is in reverse supervised by plane instance annotation.

Overall Loss: The first three base terms of the overall loss function are: Focal Loss [17] for classification $\mathcal{L}_{Focal}$, Dice Loss [23] for IoU based segmentation supervision $\mathcal{L}_{Dice}$ and RMSE Loss for point-wise depth supervision $\mathcal{L}_{RMSE}$. Except them, we introduce two geometry constraints, the depth gradient segmentation (DGS) loss and plane surface normal

(PSN) loss, to enforce the cross-task consistency:

$$\mathcal{L} = \mathcal{L}_{Focal} + \alpha\mathcal{L}_{Dice} + \beta\mathcal{L}_{RMSE} + \gamma\mathcal{L}_{DGS} + \delta\mathcal{L}_{PSN} \tag{1}$$

where α, β, γ and δ are weighting parameters, which are set to $\alpha = 3$, $\beta = 5$, $\gamma = 1$ and $\delta = 1$ through experiments.

Depth Gradient Segmentation (DGS) Loss: We observe the fact that, when trained only with common segmentation loss functions based on IoU, an instance segmentation network may fail in providing the accurate mask boundary of the plane instance and have the tendency to cover a part of the other nearby plane instance, as shown in Figure 3. From the point of view of mask IoU based loss functions like Dice Loss, Binary Cross-Entropy [3] or Focal Loss, such failure cases only cause a very limited loss overall and are therefore hard to learn. Especially in multiple plane regions with continuous low-texture, the segmentation performance highly relies on edges as clues. Addressing this, we propose to use the occlusion boundary of the depth map as an extra geometry constraint, because the gradient of the depth map within a planar region obviously should not have any significant changes, as illustrated in Figure 2(b). We first computed the depth gradient map G_{gt} of the ground truth depth D_{gt} using Sobel-Filter in $x-$ and $y-$directions, and normalize the gradient map with D_{gt} for the obvious reason that far away pixels have larger gradient than nearby pixels:

$$G_{gt} = \frac{G_x^2 + G_y^2}{D_{gt}^2} \qquad \text{with} \qquad G_x,\ G_y = Sobel_x(D_{gt}),\ Sobel_y(D_{gt}) \tag{2}$$

With the gradient map as an intermediary between the depth ground truth and the segmentation prediction, we built the depth gradient segmentation loss (DGS Loss) simply with the element-wise multiplication of the gradient map G_i of frame i and mask candidates $P_{i,j}$ of this frame, M as the number of mask candidates per frame and N as the batch size:

$$\mathcal{L}_{DGS} = \frac{1}{N}\sum_{i}^{N}\frac{1}{M}\sum_{j}^{M} G_i \odot P_{i,j} \tag{3}$$

Plane Surface Normal (PSN) Loss: Since our network is dedicated to reconstruct planes in the scene and indoor scenes are rich of planar surfaces, it is natural to constrain the depth prediction using the geometric properties of the plane. In addition to the gradient of depth, another link between plane and depth is the surface normal.

In order to minimize the dependence on noisy depth ground truth, we adapt the Combined Normal Map proposed by Long *et al.* [20], which literally combines the normal of planar regions and non-planar regions, as illustrated in Figure 2(b). In our implementation the normal of the planar region is assigned with the normal of the corresponding 3D plane model, rather than the mean value of the surface normal of each planar region as the original design.

To avoid the expensive computation of estimating surface normal with differentiable least square method [26] during training, we utilize the random sampling strategy similar as in [21, 42]. Having pixel $p_i(u_i, v_i)$ of the predicted depth, the 3D location $P_i(x_i, y_i, z_i) \in \mathbb{R}^3$ can be obtained with the camera intrinsic matrix. We then divide the predicted point cloud into N planar groups and one non-planar group using the plane segmentation ground truth: $P_i \in \mathbb{R}^3_j$, $\mathbb{R}^3 = \{\mathbb{R}^3_j | j = 0, ..., N\}$. We randomly sample K point triplets $T_i = \{(P_k^A, P_k^B, P_k^C) | P \in \mathbb{R}^3_j, k =$

$0,...,K-1\}$ within each group, the normal vector of a triplet is defined as:

$$\vec{n_k} = \frac{\overrightarrow{P_k^A P_k^A} \times \overrightarrow{P_k^A P_k^C}}{\left|\overrightarrow{P_k^A P_k^A} \times \overrightarrow{P_k^A P_k^C}\right|} \tag{4}$$

In [42], two restrictions are introduced for each triplet: angle restriction to avoid colinearity and Euclidean distance restriction for sampling more on long-range points and being robust to local noise.

Here we adopt two different restriction strategies for triplets within planar regions and within non-planar region. Since the normal of the triplet from predicted depth T_{pred} in planar regions will be compared with the normal of the fitted 3D plane, we only restrict the triplet not to be colinear, because we expect that the predicted depth in planar regions represents the planarity both locally and globally. Triplets from predicted depth within non-planar regions are compared with their corresponding triplets from noisy ground truth depth, we therefore also restricted the Euclidean distance between triplet points. The planar guidance surface normal loss (PSN Loss) is defined with the cosine similarity function:

$$\mathcal{L}_{PSN} = 1 - N_{pred}^T N_{gt} \tag{5}$$

where N_{pred} is the surface normal sample calculated from the depth prediction and N_{gt} is from the Combined Normal Map with our grouping and restriction strategies.

4 Experiments

In the section, we conduct several experiments to compare our approach against the state-of-the-art plane reconstruction methods and other monocular depth estimation approaches.

4.1 Implementation Details

Our proposed network, PlaneRecNet, is implemented with the Pytorch [25] framework, and trained using a batch size of 8 images with Adam [12] optimizer and a base learning rate of 1×10^{-4}. To stabilize the training in early stage, we set a warm-up phase to linearly increase the base learning rate from 1×10^{-6} to 1×10^{-4} during the first 2,000 iterations. The model is trained for 10 epochs on 10,000 samples (the same amount as the training configuration of PlaneRCNN [19]) from ScanNet with the plane annotation given by [19]. We augment the training data using random photometric distortion, horizontal and vertical flipping and Gaussian noise. All experiments are conducted with the same configuration, as well as the same backbone encoder ResNet101 [8] with deformable convolution [48].

4.2 Datasets and Metrics

To conduct the experiments, we use the labeled subset of **NYUv2** dataset [24] (the plane segmentation ground truth annotated with RANSAC), about 5,000 random sample from **ScanNet** [4], and **iBims** dataset [13] as benchmarks. None of these data samples is seen by our network during training. We utilize Average Precision for mask (AP_m) and bounding box (AP_b) to evaluate the plane segmentation performance. For evaluating the depth estimation, standard metrics: Absolute Relative Error (rel), Log 10 error (log_{10}), Linear Root Mean

Square Error (*RMS*), Accuracy under a threshold ($\sigma_1, \sigma_2, \sigma_3$) are utilized. Details about the new metrics introduced by iBims is explained in [13].

4.3 Ablation Study

In this section, we conduct an ablation study to analyze the details of our approach. To validate the effectiveness of our proposed loss functions and the fusion module, we test our network by adding different components one by one and inference on NYUv2 dataset [24] and evaluate the plane segmentation and depth prediction results. Table 1 shows that all the components have a positive contribution to the final performance.

Methods	Segmentation Metrics						Depth Metrics					
	AP_m	AP_m^{50}	AP_m^{75}	AP_b	AP_b^{50}	AP_b^{75}	$rel \downarrow$	$log_{10} \downarrow$	$RMS \downarrow$	σ_1	σ_2	σ_3
Baseline	26.33	48.47	26.40	26.33	48.14	34.35	0.169	0.070	0.574	0.763	0.944	0.986
Base + DGS	30.05	48.81	**36.60**	34.34	48.02	40.57	0.167	0.070	0.575	0.762	0.945	0.986
Base + DGS + PSN	29.96	48.29	36.16	34.28	48.12	40.99	0.166	0.070	0.575	0.762	0.945	0.986
PlaneRecNet (ours)	**30.10**	**49.06**	36.07	**34.67**	**48.33**	**41.01**	**0.166**	**0.069**	**0.564**	**0.765**	**0.946**	**0.987**

Table 1: Ablation studies on the contributions of proposed geometric constraints and prior module on **NYU V2** dataset [24]. Baseline: SOLO V2 [37] with simple FPN-like depth decoder. PlaneRecNet: Baseline with PPA Module and trained with DGS and PSN losses.

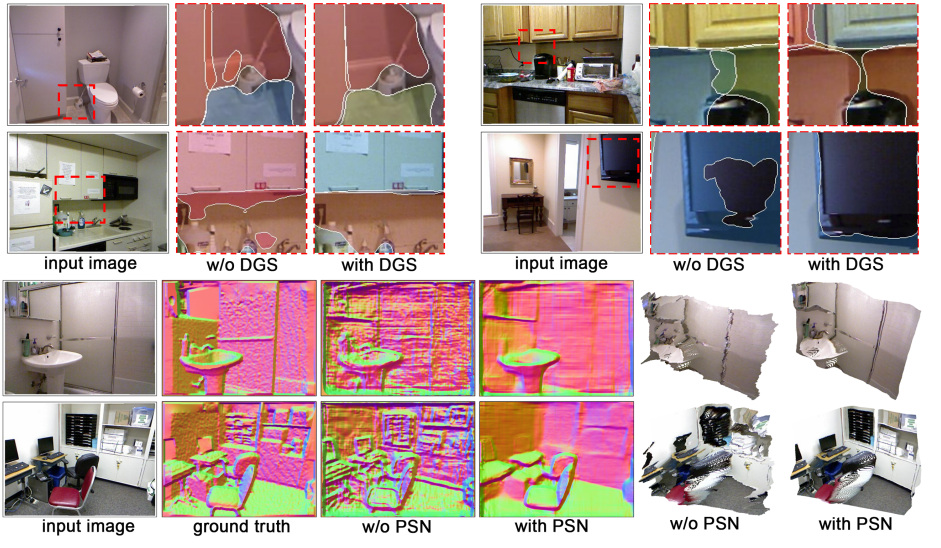


Figure 3: Visualization of the effectiveness of DGS loss and PSN loss. Input images are from NYU V2 dataset.

Effectiveness of DGS: As shown in Figure 3, our PlaneRecNet trained with DGS loss provides more accurate instance boundary, compare to our baseline. With depth gradient as clue, the network learns to concentrate more on occlusion boundary. The "feature leakage" problem is always a weakness of single-stage instance segmentation methods with global mask design [1]. We believe that DGS loss can also be generalized to common instance segmentation task, by using instance boundaries from ground truth instead of depth gradient.

Effectiveness of PSN: With plane surface normal constraint, our depth prediction is improved in providing more precise normal map and point cloud. In Figure 3, we visualize the estimated surface normal using differentiable least square method [26] and the point cloud recovered from the predicted depth. One can see that high order geometry constraint like surface normal improves the scene reconstruction in a significant way.

Effectiveness of Plane Prior Attention: The architecture of our PPA module is very simple. Table 1 shows the improvement by introducing PPA module. The attention weight requires no extra supervision, since the plane segmentation mask is supervised by segmentation loss. The module can be trained along with the whole network, while Depth Attention Volume Network is trained separately in three stages [10].

4.4 Comparison to State-of-the-art

Evaluation on ScanNet: We evaluate our method on about 5,000 random samples from ScanNet [4], and compare the results with other state-of-the-art plane reconstruction methods [19, 24] (trained on the same dataset). Since PlaneRCNN and PlaneAE are trained with SGD optimizer, we also provide the result of our network trained with it. As shown in Table 2, our network outperforms other state-of-the-arts both in segmentation and depth estimation. Some quantitative results are given in Figure 4. Comparing to PlaneRCNN, our method is more unlikely to counter the "uncompleted-segmentation" problem, which we believe is a side effect of the warping loss [19] used in PlaneRCNN. The reason is that, when computing the 3D distance between the warped model and the neighbor view, a under-segmented 3D plane results less warping loss than over-segmented one. Column 1, 4 and 6 of Figure 4 also show that the reconstructed point clouds from PlaneAE and PlaneRCNN have obvious inconsistency and cracks, when the plane parameters of the 3D planes are inaccurate.

Methods	Segmentation Metrics						Depth Metrics					
	AP_m	AP_m^{50}	AP_m^{75}	AP_b	AP_b^{50}	AP_b^{75}	$rel \downarrow$	$log_{10} \downarrow$	$RMS \downarrow$	σ_1	σ_2	σ_3
PlaneAE [24]	15.40	27.62	16.26	19.52	31.77	19.83	0.111	0.049	0.415	0.864	0.967	0.991
PlaneRCNN [19]	32.71	45.18	39.26	35.13	46.27	41.09	0.117	0.048	0.256	0.872	0.973	0.994
Ours (SGD)	33.88	48.95	43.14	34.93	49.00	39.80	0.112	0.049	0.279	0.875	0.979	0.995
Ours (ADAM)	**39.55**	**50.24**	**48.87**	**44.07**	**50.21**	**47.99**	**0.081**	**0.037**	**0.221**	**0.943**	**0.991**	**0.997**

Table 2: Evaluation of segmentation and depth estimation on **ScanNet** dataset [4].

Methods	Standard Metrics						PE (cm)		DBE (px)		DDE (%)		
	rel	log_{10}	*RMS*	$\sigma_1 \uparrow$	$\sigma_2 \uparrow$	$\sigma_3 \uparrow$	ε_{PE}^{plan}	ε_{PE}^{orie}	ε_{DBE}^{acc}	ε_{DBE}^{comp}	$\varepsilon_{DDe}^{0} \uparrow$	ε_{DDe}^{-}	ε_{DDe}^{+}
SharpNet [28]	0.26	0.11	1.07	0.59	0.84	0.94	9.95	25.67	3.52	7.61	84.03	9.48	6.49
DAV-Net [10]	0.24	0.10	1.06	0.59	0.84	0.94	7.21	18.45	3.46	7.43	84.36	**6.84**	6.27
VNL-Net [32]	0.24	0.11	1.06	0.54	0.84	0.94	5.73	16.91	3.65	7.16	82.72	13.91	3.36
Ours	**0.18**	**0.09**	**1.00**	**0.67**	**0.89**	**0.95**	**3.29**	**9.12**	**2.41**	**6.59**	**84.67**	13.91	**1.42**
PlaneNet [18]	0.29	0.17	1.45	0.41	0.70	0.86	7.26	17.24	4.84	8.86	71.24	28.36	0.40
PlaneAE [24] †	0.27	0.15	1.42	0.39	0.74	0.88	4.48	10.65	4.52	8.39	71.67	27.98	**0.35**
PlaneRCNN [19] †	0.20	0.10	1.04	0.66	0.88	**0.96**	4.50	13.38	4.03	8.01	**85.51**	**12.24**	2.25
Ours + PD	**0.18**	**0.09**	**1.02**	**0.67**	**0.89**	0.95	**1.55**	**9.19**	**3.03**	**7.03**	84.54	13.92	1.54

Table 3: Comprehensive depth evaluation on **iBims-1** dataset [13]. † Using author's released models. Ours + PD: Rendering depth map using 3D plane fitted with PCA.

Evaluation on iBims-1: We evaluate our method on iBims-1 dataset [13]. This dataset introduced a set of error metrics to evaluate the planarity of the depth map (ε_{PE}^{plan} and ε_{PE}^{orie}),

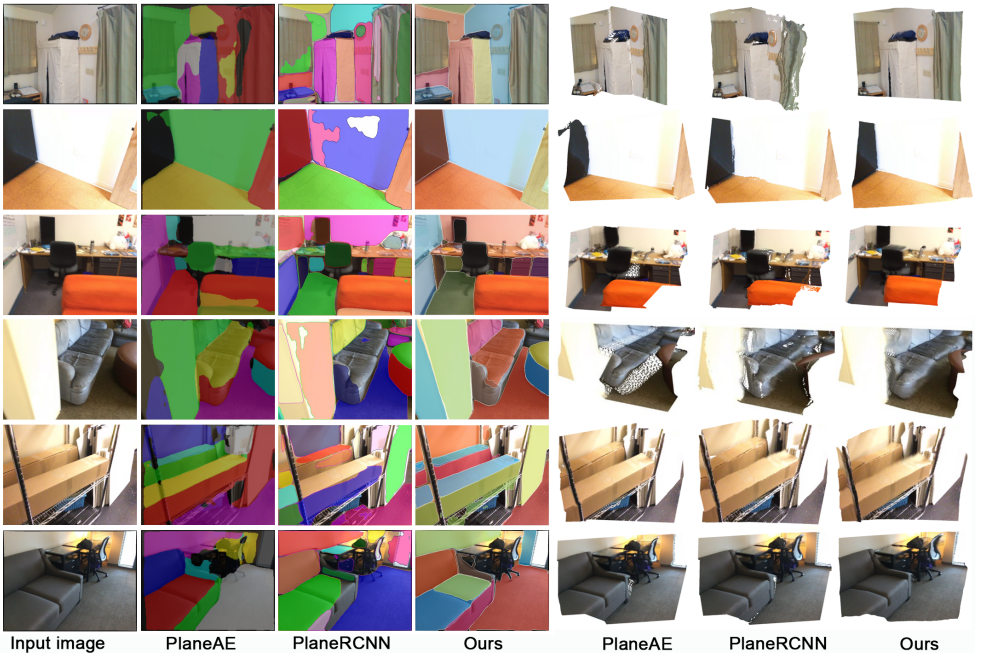


Figure 4: Qualitative comparison of PlaneAE, PlaneRCNN and ours on ScanNet dataset.

the location accuracy of depth boundaries (ε_{DBE}^{acc} and ε_{DBE}^{comp}) and the consistency of depth prediction over the whole image (ε_{DDe}^{-}, ε_{DDe}^{+}, ε_{DDe}^{0}). We provide the quantitative comparison between our method and [10, 18, 19, 29, 42, 44] in Table 3. Our network outperforms other monocular depth estimators without fitting 3D planes to the predicted depth. After fitting 3D planes to predicted depth, our networks shows better performance than PlaneNet and PlaneAE in geometry related metrics in most of the metrics, especially planarity metrics ε_{PE}^{plan} and ε_{PE}^{orie}. The depth estimation accuracy of our network is actually worsen after fitting 3D planes, which confirms our argument in Section 1, that "hard-fitting" with inaccurate segmentation masks leads to less consistent results. Additionally, our network is over **4x faster** than PlaneRCNN (runtime analysis in supplementary materials). This should be attributed to the fact that we utilize the light-weight single-stage instance segmentation method [37], which is much faster than two-stage method like Mask R-CNN [9] used in PlaneRCNN. Moreover, the refinement sub-network of PlaneRCNN is also very time-consuming.

5 Conclusion and Future Work

This paper proposes PlaneRecNet, a multi-task convolutional neural network for piece-wise planar reconstruction from a single RGB image. We introduce depth gradient segmentation loss function and planar surface normal loss function to enforce the cross-task consistency between piece-wise plane segmentation and monocular depth estimation. An interesting future direction is to introduce a self-supervised occlusion boundary prediction branch which may further improve our network's ability of geometry understanding.

6 Acknowledgments

The research leading to these results has been partially funded by the German BMBF project MOVEON (Funding reference number 01IS20077) and by the German BMBF project RACKET (01IW20009).

References

[1] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9157–9166, 2019.

[2] Alejo Concha, Muhammad Wajahat Hussain, Luis Montano, and Javier Civera. Manhattan and piecewise-planar constraints for dense monocular mapping. In *Robotics: Science and systems*, 2014.

[3] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.

[5] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. ISSN 0001-0782. doi: 10.1145/358669.358692. URL https://doi.org/10.1145/358669.358692.

[6] Alex Flint, David Murray, and Ian Reid. Manhattan scene understanding using monocular, stereo, and 3d features. In *2011 International Conference on Computer Vision*, pages 2228–2235. IEEE, 2011.

[7] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[10] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä. Guiding monocular depth estimation using depth-attention volume. In *European Conference on Computer Vision*, pages 581–597. Springer, 2020.

[11] Ziyu Jiang, Buyu Liu, Samuel Schulter, Zhangyang Wang, and Manmohan Chandraker. Peek-a-boo: Occlusion reasoning in indoor scenes with plane representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–121, 2020.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[14] Yanyan Li, Nikolas Brasch, Yida Wang, Nassir Navab, and Federico Tombari. Structure-slam: Low-drift monocular slam in indoor environments. *IEEE Robotics and Automation Letters*, 5(4):6583–6590, 2020.

[15] Yanyan Li, Raza Yunus, Nikolas Brasch, Nassir Navab, and Federico Tombari. Rgb-d slam with structural regularities. *arXiv preprint arXiv:2010.07997*, 2020.

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[18] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2579–2588, 2018.

[19] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4450–4459, 2019.

[20] Xiaoxiao Long, Lingjie Liu, Christian Theobalt, and Wenping Wang. Occlusion-aware depth estimation with adaptive normal constraints. *arXiv e-prints*, pages arXiv–2004, 2020.

[21] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, and Wenping Wang. Adaptive surface normal constraint for depth estimation. *arXiv preprint arXiv:2103.15483*, 2021.

[22] Yunze Man, Xinshuo Weng, Xi Li, and Kris Kitani. Groundnet: Monocular ground plane normal estimation with geometric consistency. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2170–2178, 2019.

[23] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[24] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[26] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018.

[27] Xiaojuan Qi, Zhengzhe Liu, Renjie Liao, Philip HS Torr, Raquel Urtasun, and Jiaya Jia. Geonet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[28] Yiming Qian and Yasutaka Furukawa. Learning pairwise inter-plane relations for piecewise planar reconstruction. In *European Conference on Computer Vision*, pages 330–345. Springer, 2020.

[29] Michael Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2019.

[30] Jason Rambach, Paul Lesur, Alain Pagani, and Didier Stricker. Slamcraft: Dense planar rgb monocular slam. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–6. IEEE, 2019.

[31] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. In *Asian Conference on Computer Vision*, pages 298–313. Springer, 2018.

[32] Akshay Rangesh and Mohan Manubhai Trivedi. Ground plane polling for 6dof pose estimation of objects on the road. *IEEE Transactions on Intelligent Vehicles*, 5(3): 449–460, 2020.

[33] Fangwen Shu, Yaxu Xie, Jason Rambach, Alain Pagani, and Didier Stricker. Visual slam with graph-cut optimized multi-plane reconstruction. *arXiv preprint arXiv:2108.04281*, 2021.

[34] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pages 9120–9132. PMLR, 2020.

[35] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5038–5047, 2017.

[36] Peng Wang, Xiaohui Shen, Bryan Russell, Scott Cohen, Brian Price, and Alan L Yuille. Surge: Surface regularized geometry estimation from a single image. *Advances in Neural Information Processing Systems*, 29:172–180, 2016.

[37] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[38] Weijie Xi and Xuejin Chen. Reconstructing piecewise planar scenes with multi-view regularization. *Computational Visual Media*, 5(4):337–345, 2019.

[39] Yaxu Xie, Jason Rambach, Fangwen Shu, and Didier Stricker. Planesegnet: Fast and robust plane estimation using a single-stage instance segmentation cnn. *arXiv preprint arXiv:2103.15428*, 2021.

[40] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018.

[41] Fengting Yang and Zihan Zhou. Recovering 3d planes from a single image via convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[42] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[43] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. *arXiv preprint arXiv:2012.09365*, 2020.

[44] Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, and Shenghua Gao. Single-image piece-wise planar 3d reconstruction via associative embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1037, 2019.

[45] Zehao Yu, Lei Jin, and Shenghua Gao. P^2net: Patch-match and plane-regularization for unsupervised indoor depth estimation. In *ECCV*, 2020.

[46] Raza Yunus, Yanyan Li, and Federico Tombari. Manhattanslam: Robust planar tracking and mapping leveraging mixture of manhattan frames. *arXiv preprint arXiv:2103.15068*, 2021.

[47] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2020.

[48] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.