Learning Pairwise Inter-Plane Relations for Piecewise Planar Reconstruction

Yiming Qian and Yasutaka Furukawa

Simon Fraser University, Canada
{yimingq,furukawa}@sfu.ca

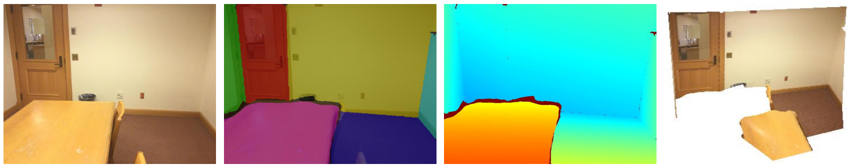

Fig. 1. This paper takes a piecewise planar reconstruction and improves its plane parameters and segmentation masks by inferring and utilizing inter-plane relationships. From left to right, input image, segmented plane instances, recovered depthmap, reconstructed 3D planar model

Abstract. This paper proposes a novel single-image piecewise planar reconstruction technique that infers and enforces inter-plane relationships. Our approach takes a planar reconstruction result from an existing system, then utilizes convolutional neural network (CNN) to (1) classify if two planes are orthogonal or parallel; and 2) infer if two planes are touching and, if so, where in the image. We formulate an optimization problem to refine plane parameters and employ a message passing neural network to refine plane segmentation masks by enforcing the inter-plane relations. Our qualitative and quantitative evaluations demonstrate the effectiveness of the proposed approach in terms of plane parameters and segmentation accuracy.

Keywords: piecewise planar; reconstruction; deep learning; single-view

1 Introduction

Inter-plane relationships convey rich geometric information for underlying scene structure. Man-made environments are full of parallelism and orthogonality, whose information would constrain surface orientations. Planes meet along a line, where knowing the presence and location of such contact lines would further refine plane parameters and produce precise plane segmentation.

With the emergence of deep learning, state-of-the-art piecewise planar reconstruction methods are capable of finding plane instances and estimating their parameters even from a single image [12, 10, 24]. However, these approaches reconstruct plane instances independently, and reconstructions suffer from inter-plane inconsistencies. For example, depth values are inconsistent at plane boundaries, leading to clear visual artifacts in the 3D models. Plane segmentation is also erroneous at its boundaries with many holes in-between, yielding gaps and cracks in the 3D models, which could cause unpleasant user experiences in AR applications (*e.g.*, a virtual ball gets stuck or goes through cracks into walls).

This paper proposes a novel single-image piecewise planar reconstruction technique that takes and improves existing piecewise planar reconstruction by detecting and enforcing inter-plane relationships. More concretely, given a piecewise planar reconstruction (*i.e.*, a set of plane parameters and segmentation masks), convolutional neural networks (CNNs) first infer two types of inter-plane relationships: 1) If two planes are orthogonal, parallel, or neither; and 2) If two planes are in contact and, if so, where in the image.

With the relationships, we formulate an optimization problem to refine plane parameters so that 1) plane normals agree with the inferred orthogonality or parallelism; and 2) plane intersections project onto the estimated contact lines. Lastly, we employ message passing neural networks to refine plane segmentation while ensuring that the plane segmentation and parameters become consistent.

We have utilized ScanNet [3] to generate ground-truth inter-plane relationships and introduced three new inter-plane consistency metrics. We have built the proposed algorithm in combination with two state-of-the-art piecewise planar reconstruction methods (PlaneRCNN [10] and the work by Yu *et al.* [24]). Our qualitative and quantitative evaluations demonstrate that the proposed approach consistently improves the accuracy of the plane parameters and segmentation. Code and data are available at https://github.com/yi-ming-qian/interplane.

2 Related Work

We first review piecewise planar reconstruction literature, and then study other techniques relevant to our paper.

Piecewise planar reconstruction: Traditional approaches for piecewise planar reconstruction require multiple views or depth information [5, 6, 14, 19, 20, 25]. They generate plane proposals from 3D points by heuristics (*e.g.*, RANSAC based plane fitting), then assign a proposal to each pixel via a global inference (*e.g.*, Markov Random Field). Deng *et al.* [4] proposed a learning-based approach to recover planar regions, while still requiring depth information as input. Recently, Chen *et al.* [12] revisited the piecewise planar depthmap reconstruction problem from a single image with an end-to-end learning framework (PlaneNet). PlaneRecover [23] later proposed an unsupervised learning approach. Both PlaneNet and PlaneRecover require the maximum number of planes in an image as a prior (*i.e.*, 10 in PlaneNet and 5 in PlaneRecover). To handle arbitrary number of planes, PlaneRCNN employs a detection architecture from the recognition community to handle arbitrary number of planes [10]. Yu *et al.* employs an associative embedding technique instead [24]. These methods produce impressive reconstructions, but plane segmentation masks are almost always imprecise at their boundaries. For example, a plane boundary should often be an exact line shared by another plane, which is rarely the case in these methods. Furthermore, plane depth values are not consistent at their contacts.

Room layout estimation: Under special structural assumptions, piecewise planar reconstruction with exact segmentation boundary has been possible. Room layout estimation is one such example, where the methods seek to find the boundary lines between the horizontal floor and vertical walls. Estimation of plane geometry/parameters is automatic from the segmentation thanks to the structural assumption [7, 13, 21, 27].

Segmentation with piecewise linear boundary: While not necessarily a reconstruction task, segmentation with compact linear boundary is a closely related work. KIPPI is a polygonal image segmentation technique, which detects and extends line segments to form polygonal shapes [1]. Planar graph reconstruction is a similar task, effective for floorplan reconstruction [2], floorplan vectorization [11], or outdoor architectural parsing [26]. The key difference in our problem is that we solve reconstruction and segmentation, where plane parameters and segmentation boundaries are tightly coupled. In fact, the earlier work by Kushal and Seitz [9] exploits this relationship to reconstruct a piecewise smooth 3D model. Their method extracts boundary first, which is often challenging and requires manual work, then performs piecewise smooth surface reconstruction. Our work solves reconstruction and segmentation simultaneously.

3D primitive-based reconstruction: 3D primitive based reconstruction produces piecewise planar/smooth models with clean boundary lines. Constructive solid geometry with 3D solid primitives was used for large-scale building reconstruction with an assumption of a block world [22]. More recently, a data-driven approach was proposed for CSG model reconstruction [17]. They produce high-quality 3D models but were demonstrated mostly on synthetic objects. This work tackles complex cluttered indoor scenes.

Plane identification: Given a pair of images, a CNN was trained to identify the same plane in the image pair, which was used for the loop-closing in the SLAM application [18]. This work exploits much richer class of pairwise plane relationships for a single image planar reconstruction.

3 Algorithm

Our system takes a piecewise planar reconstruction as input, and refines its plane parameters and segmentation masks by exploiting inter-plane relationships. In practice, we have used two state-of-the-art methods for generating our inputs: PlaneRCNN by Liu *et al.* [10] and the work by Yu *et al.*, which we refer to as PlaneAE for convenience [24].

Our process consists of three steps (See Fig. 2). First, we use CNNs to infer two inter-plane relationships for every pair of plane instances. Second, we

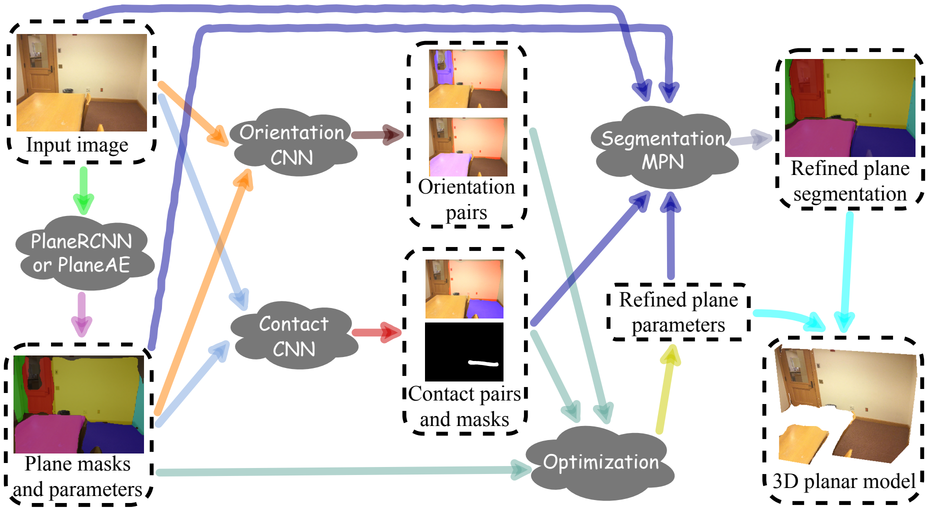

Fig. 2. System overview. Given a single RGB input, we use PlaneRCNN or PlaneAE to produce initial plane segmentation masks and parameters. We then use Orientation-CNN and Contact-CNN for pairwise relationship reasoning. Next, we solve an optimization problem to refine the plane parameters, followed by a message passing neural network (Segmentation-MPN) to refine plane segmentation

formulate an optimization problem to refine plane parameters by enforcing the inter-plane relationships. Third, we jointly refine the plane segmentation masks by a message passing neural network to be consistent with the refined plane parameters and the inter-plane relationships. We now explain the details.

3.1 Inter-plane relationships learning

We consider two types of pairwise inter-plane relationships.
- **Orientation**: Are two planes parallel or orthogonal?
- **Contact**: Are two planes in contact? Where is the contact line in an image?

Manhattan structure is prevalent for man-made environments and the orientation relationship would effectively constrain surface normals for many pairs of planes. Plane-contacts are straight lines, whose information would constrain surface normal/offset parameters and provide precise segmentation boundaries. We employ standard CNNs to infer both relationships.

Orientation-CNN: A pair of planes is classified into three orientation types: *parallel*, *orthogonal*, or *neither*. The input is an 8-channel tensor: the RGB image (3 channels), the binary segmentation masks of the two planes (2 channels), the depthmaps of the two planes (2 channels), and the dot-product between the two plane normals (1 channel). The depthmap has depth values at every pixel, not just over the plane masks. The dot-product is a scalar and we copy the same value at every pixel location to form an image. ResNet-50 is adopted as the architecture, where we change the first layer to accommodate the input 8-channel tensor and change the last layer to output 3 values that corresponds to the three orientation types. Orientation-CNN is trained with a cross-entropy loss against the ground-truth (See Sec. 4 for the GT preparation).

Contact-CNN: The network classifies if two planes are in contact (binary classification) and estimates the contact line as a pixel-wise segmentation mask (binary segmentation). The input is the same 8-channel tensor as in Orientation-CNN. We again take ResNet-50 as the backbone and replaced the first layer. For binary contact classification, we change the last layer to output a 2D vector. For binary contact segmentation, we attach a branch of bilinear upsampling and convolution layers to the last residual block of ResNet-50, which is adopted from the binary segmentation branch of [24]. The output size of the segmentation mask is the same as the input image. The network is trained with a cross-entropy loss against the ground-truth, where the loss for the segmentation mask is averaged over all the pixels. The classification and segmentation losses are enforced with equal weights. For a non-contact plane pair, an empty mask is the supervision.

3.2 Plane parameter refinement

We solve the following optimization problem to refine the plane parameters. Our variables are the plane normal $\mathbf{N}_i$ and the offset d_i for each plane: $\mathbf{N}_i \cdot \mathbf{X} = d_i$. $\mathbf{X}$ is a 3D point coordinate.

$$\min_{\{\mathbf{N}_i, d_i\}} E_{unit} + E_{input} + E_{parallel} + E_{ortho} + E_{contact}, \tag{1}$$

$$E_{unit} = 10 \sum_{i \in P} (\mathbf{N}_i \cdot \mathbf{N}_i - 1)^2, \tag{2}$$

$$E_{input} = 10 \sum_{i \in P} w_i (\mathbf{N}_i \cdot \overline{\mathbf{N}}_i)^2 + \sum_{i \in P} \sum_{p \in M_i} w_i (\mathbf{N}_i \cdot \overline{\mathbf{X}}_i^p - d_i)^2 / |M_i|, \tag{3}$$

$$E_{parallel} = \sum_{(i,j) \in P_{pa}} w_i w_j (\mathbf{N}_i \cdot \mathbf{N}_j - 1)^2, \tag{4}$$

$$E_{ortho} = \sum_{(i,j) \in P_{or}} w_i w_j (\mathbf{N}_i \cdot \mathbf{N}_j)^2, \tag{5}$$

$$E_{contact} = \sum_{(i,j) \in P_{co}} \sum_{p \in M_{i,j}} w_i w_j (\mathbf{D}_i^p - \mathbf{D}_j^p)^2 / |M_{i,j}|. \tag{6}$$

• E_{unit} enforces N_i to have a unit norm. P denotes the set of planes.

• E_{input} keeps the solution close to the original and breaks the scale/rotational ambiguities inherent in the other terms. The first term is on the plane normal. $\overline{\mathbf{N}}_i$ denotes the initial plane normal. w_i is a rescaled segmentation area: $\sum_{i \in P} w_i = 1$. The second term measures the deviation from the initial depthmap. $\overline{\mathbf{X}}_i^p$ denotes the 3D coordinate of a pixel p based on the initial plane parameters. The plane equation residual is summed over the mask M_i of the i_{th} plane.

• $E_{parallel}$ and E_{ortho} enforces the parallel and orthogonal relationships, respectively. P_{pa} and P_{or} denotes the pairs of planes with the inferred parallel and orthogonal relationships, respectively.
• $E_{contact}$ measures the consistency of depth values along the plane contacts between pairs of planes P_{co} with the inferred contact relationships. For every pixel p in the plane contact area $M_{i,j}$ estimated by the Contact-CNN, we compute the depth values $(\mathbf{D}_i^p, \mathbf{D}_j^p)$ based on their plane parameters. The term evaluates the average depth value discrepancy.

The energy terms are well normalized by the rescaled segmentation areas (w_i). We rescale E_{unit} and the first term of E_{input} by a factor of 10 and keep the balancing weights fixed throughout the experiments. E_{unit} has a large weight because its role is close to a hard constraint (ensuring that the plane normal is a unit vector). E_{input} has a large weight because the normal estimation is usually more accurate than the offset estimation in the initial input from PlaneRCNN and PlaneAE. We use an off-the-shelf BFGS optimization library in SciPy to solve the problem [15].

3.3 Plane segmentation refinement (Segmentation-MPN)

PlaneRCNN [10] jointly refines plane segmentation by "ConvAccu Module", which is a special case of more general convolutional message passing neural architecture (Conv-MPN) [26]. PlaneAE [24] estimates segmentation masks jointly by associative embedding. However, there are two major issues in their segmentation results. First, they ignore inter-plane contact relationships: Plane parameters would not be consistent with the boundaries, and the 3D model would look broken (gaps and discontinuities). Second, they tend to under-segment (especially PlaneRCNN) because the ground-truth is often under-segmented, too. [1]

We follow PlaneRCNN and utilize Conv-MPN for joint segmentation refinement. We add the binary split mask as the 5_{th} channel to the input so that the network knows when the segmentation boundary becomes consistent with the plane parameters. A plane may have multiple contacts, and the union of all the split-masks is formed for each plane. We make the following modifications to address the above two issues (See Fig. 3).

Resolving parameter inconsistencies: Our idea is simple. For each pair of planes with the inferred contact mask, we compute the exact plane boundary inside the mask as a line from the refined plane parameters. We split the mask into two regions along the line and send them as images to Conv-MPN. The split mask serves as the input as well as the loss so that Conv-MPN will learn to satisfy the contact consistency. More precisely, we compute the 3D intersection line from the refined plane parameters and project the line into the image, where the intrinsic parameters are given for each image in the database. After splitting the mask into two regions along the line at the pixel-level, we can determine

[1] Ground-truth segmentation comes from plane-fitting to 3D points [10]. For being conservative, they focus on high confidence areas with high point densities only, dropping the plane boundaries.

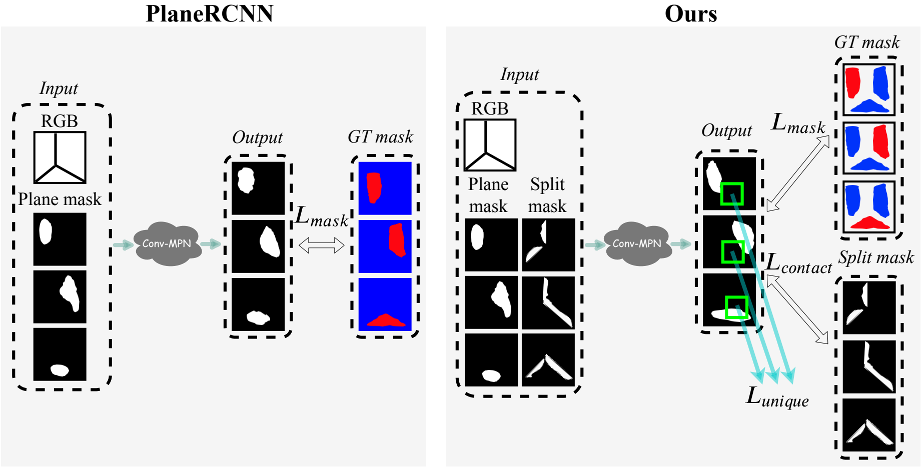

Fig. 3. Contact-aware joint segmentation refinement. We follow PlaneRCNN [10] and utilize Conv-MPN [26] architecture. (Left) In the original formulation in PlaneRCNN, the mask loss (L_{mask}) tries to make the output equal to the under-segmented "ground-truth". (Right) In the new formulation, we change the definition of negative samples (blue pixels) in the mask loss (L_{mask}) to prevent under-segmentation. Furthermore, we add a contact loss $L_{contact}$ with the split mask so that the plane segmentation boundary becomes consistent with the plane parameters. The same split mask is also added to the input of the network. Lastly, we add L_{unique} to prevent over-segmentation and ensure that a pixel (green boxes) belongs to at most one plane

easily which split mask should belong to which plane by comparing the current plane segmentation and the line. We simply add a cross entropy loss $L_{contact}$ for pixels inside the split contact mask.

Resolving under-segmentation: In the original PlaneRCNN formulation [10], the cross entropy loss was defined with the under-segmented "ground-truth" plane mask. The red and blue pixels in Fig. 3 illustrate the positive and negative pixels for the loss (L_{mask}). We modify the definition of negative samples in this mask loss (L_{mask}) to be the union of the other under-segmented "ground-truth" regions instead, allowing Conv-MPN to grow beyond the under-segmented "ground-truth". In order to prevent over-segmentation this time, we introduce a new loss (L_{unique}) which prevents a pixel from belonging to multiple planes. To be precise, the loss is defined at each pixel as

$$L_{unique} = -\log(2 - \max(1, \alpha)). \tag{7}$$

α is the sum of the top 2 mask probabilities at a pixel. The sum of the three terms $L_{mask}, L_{contact}$, and L_{unique} becomes the loss without rescaling.

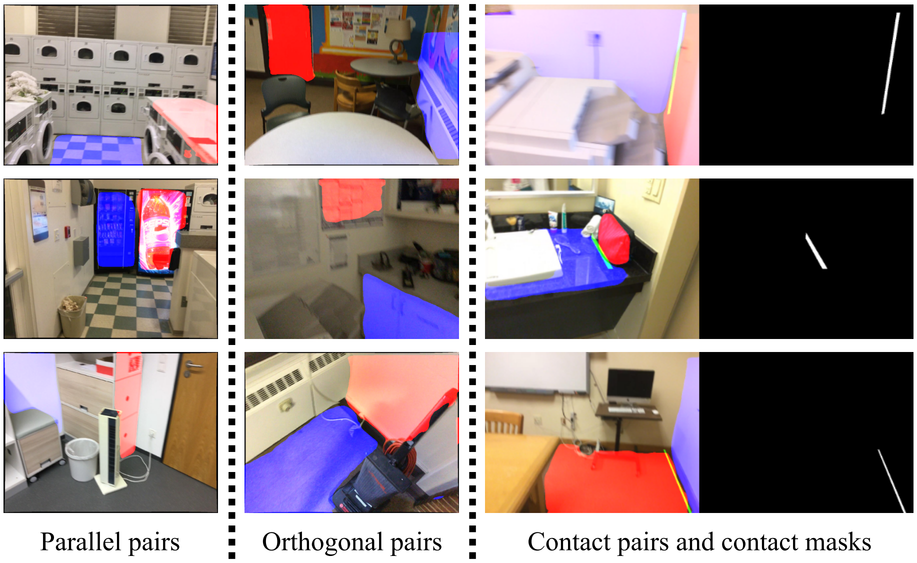

Fig. 4. Inter-plane relationships dataset. Here we show three example pairs for each relationship type, where the PlaneRCNN plane masks are colored by red and blue. For contact pairs, we also show the contact lines as overlays (green) in the third column and as the binary masks in the last column. Note that our relationship annotations are automatically generated from the ScanNet database [3]

4 Dataset and Metrics

Inter-plane relationship learning is a new task, where this paper generates the ground-truth and introduces new evaluation metrics (See Fig. 4).

4.1 Dataset

We borrow the piecewise planar reconstruction dataset by Liu *et al.* [12, 10], which was originally constructed from ScanNet [3]. We follow the same process in splitting the dataset into training and testing sets. Note that the camera intrinsic parameters are associated with each image. We generate inter-plane relationship labels (parallel, orthogonal, and contact) as follows.

First, we associate each PlaneRCNN plane segment to a corresponding GT segment with the largest overlap. To detect parallel and orthogonal plane pairs, we check if the angle between plane normals are either 0 or 90 degrees with a tolerance of 10 degrees. To detect the contact relationship, we compute the 3D intersection line using the GT plane parameters, and project onto the image. After rastering the line into a set of pixels, we filter out pixels if the distance to the closest pixel in the two plane masks is more than 20 pixels. Two planes are declared to be in contact if more than 5% pixels survive the filtering. We apply 5×5 dilation (OpenCV implementation, 5 times) to the remaining pixels and apply 5×5 Gaussian blur (OpenCV implementation) to obtain the contact mask. The process produced roughly $2,472,000$ plane pairs. 16%, 47%, and 12% of the pairs are labeled as parallel, perpendicular, and in contact, respectively.

4.2 Metrics

Piecewise planar reconstruction with inconsistent plane parameters and segmentation information leads to 3D models with large visual artifacts. Standard reconstruction metrics are angular errors in the plane normals, errors in the plane offset parameters, or depth errors inside the plane mask. Similarly, standard segmentation metrics are variation of information (VoI), rand index (RI), segmentation covering (SC), and intersection over union (IoU) [24, 10]. However, these metrics are evaluated per-plane and do not reflect the visual quality of 3D models well, where inter-plane inconsistencies become more noticeable. This paper introduces three new metrics.

Relative orientation error (ROE): For each plane pair, we compute the angle between their normals using the GT plane parameters and the reconstructed parameters. The discrepancy of the two angles averaged over all the plane pairs is the metric.

Contact consistency error (CCE): Depth values of the two planes must be the same along the contact line. CCE measures the average depth value discrepancy along the ground-truth contact line. Note a difference from the contact energy term $E_{contact}$ Eq.(6) from the optimization, which measures the discrepancy along the contact line predicted by Contact-CNN. This metric measures the discrepancy along the ground-truth line.

Segmentation metric over contact mask: This is a simple modification to the standard segmentation metrics. We simply compute the standard metrics (VoI, RI, SC, and IoU) inside the contact line mask instead of an entire image. Segmentation masks must be accurate at its boundaries to achieve high scores.

5 Experiments

We have implemented our approach in Python using the PyTorch library. We train our networks with the Adam optimizer [8] and set the learning rate to 10^{-4}. The batch size of 24 is used for Orientation-CNN and Contact-CNN, and the batch size of 1 is used for segmentation-MPN. The training of each network takes about 10 to 15 hours on an NVIDIA GTX 1080 Ti GPU with 11GB of RAM. The average run-times of our algorithm (pairwise relationship prediction, BFGS optimization, and segmentation refinement) are shown in Table 1 when testing on an image with the resolution of 224×224.

5.1 Planar reconstruction

Table 2 provides the quantitative evaluation of the geometrical reconstruction accuracy. The table reports the three standard reconstruction metrics (mean

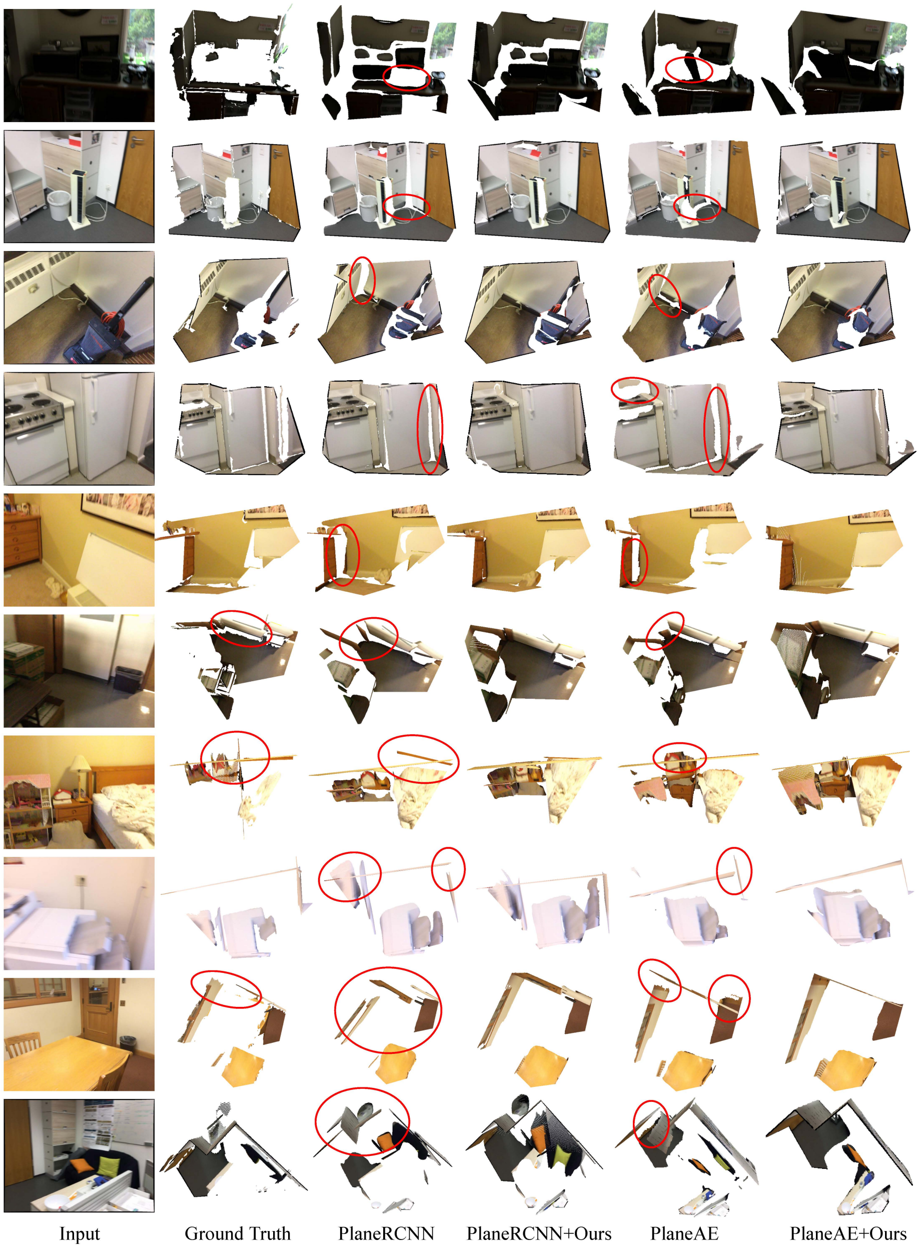

Fig. 5. Visual comparison of 3D planar reconstruction results. From left to right, it shows the input RGB image, the ground truth, the PlaneRCNN results [10], our results with PlaneRCNN, the PlaneAE results [24], our results with PlaneAE. The proposed approach fixes incomplete and inconsistent reconstructions at various places, highlighted in red ovals. Also refer to the supplementary video for the best assessment

Table 1. Average running time in seconds of our method

Input method	Relationship prediction	BFGS optimization	Segmentation-MPN	Total
PlaneRCNN	0.35	1.36	1.24	2.95
PlaneAE	0.15	0.84	0.95	1.94

Table 2. Quantitative evaluation of reconstruction accuracy. The unit of the offset error, the depth error and CCE is in centimeters. The unit of normal error and ROE is in degrees. The color cyan denotes the best result in each triplet

	Normal Error	Offset Error	Depth Error	ROE	CCE
PlaneRCNN	12.37	20.12	21.97	12.12	12.92
+Ours (w/o contact)	11.38	19.81	21.98	10.09	14.82
+Ours (all)	11.11	20.09	21.93	10.06	9.25
PlaneAE	9.77	15.53	17.60	11.28	13.05
+Ours (w/o contact)	9.38	15.55	17.40	10.71	13.24
+Ours (all)	9.68	15.85	17.36	10.69	11.59

angular error of plane normals, mean absoluate error of plane offset parameters, and mean depth error inside the ground-truth segmentation mask) and the new ROE and CCE metrics. We tested the proposed system with PlaneRCNN or PlaneAE while running their released official code. As an ablation study, we also run our optimization process while removing the contact term ($E_{contact}$).

As shown in Table 2, our approach consistently improves normal-error, depth-error, ROE, and CCE metrics, in particular, the last two inter-plane consistency metrics. The offset error rather increased, because it conflicts with the depth error which our optimization minimizes. The offset-error does not take into account the surface region on the plane, and we believe that the depth-error is more informative. The use of the plane-contact constraints have dramatic effects on the CCE metric as expected. ROE and CCE reflect the visual quality of the reconstructed 3D models more accurately as shown in Figure 5. The models are rendered from viewpoints close to the original in the top half, where the proposed method consistently improves segmentation at plane boundaries, often completely closing the gaps in-between. Models are rendered from lateral viewpoints in the bottom half. It is clear that planes meet exactly at their contacts with our approach, while 3D models by PlaneRCNN or PlaneAE often suffer from severe artifacts due to plane gaps and intersections. Also refer to the supplementary video for the best assessment of the visual quality.

5.2 Plane instance segmentation

"Ground-truth" plane segmentation in the PlaneRCNN dataset have large errors. We have randomly chosen 50 testing images and manually annotated ground-truth plane segmentation by the LabelMe tool [16] (See the second column of Fig. 6). Following [24, 10], we employ four segmentation metrics mentioned in Sec. 4.2 (VoI, RI, SC, and IoU). To further evaluate the segmentation accuracy

Table 3. Quantitative evaluation of plane segmentation. The smaller the better for VoI, while the larger the better for the other metrics. The color cyan denotes the best result in each pair

	Evaluation on the entire image				Evaluation on contact line only			
Method	VoI↓	RI	SC	IoU%	VoI↓	RI	SC	IoU%
PlaneRCNN	0.967	0.910	0.788	78.98	2.301	0.744	0.456	39.90
PlaneRCNN+Ours	0.822	0.936	0.830	80.90	2.146	0.778	0.508	48.96
PlaneAE	1.183	0.881	0.735	69.86	2.253	0.735	0.466	38.90
PlaneAE+Ours	1.002	0.882	0.753	69.26	2.101	0.733	0.486	41.72

along the contact plane boundary, we have annotated the plane contact lines with LabelMe and computed the same metrics only inside the contact lines (See Sec. 4.2). Table 3 shows that our approach consistently improves segmentation accuracy over both PlaneRCNN and PlaneAE, especially along the contact lines. Fig. 6 qualitatively demonstrates that our approach produces more complete segmentation, especially at plane boundaries. Furthermore, our plane segmentation boundaries are exact straight lines that are consistent with the plane parameters, while the boundaries are usually curved in the raw PlaneRCNN and PlaneAE results. It is also noteworthy that we faithfully recover the T-junctions in an indoor scene as shown in both Fig. 5 and Fig. 6.

5.3 Pairwise relationship inference

Table 4 evaluates the inter-plane relationship classification (parallel, orthogonal, and contact) and the contact mask estimation by our two CNN modules (Orientation-CNN and Contact-CNN). We compare against three baseline methods that use PlaneRCNN for reconstruction and simple heuristics to infer the relationships between planes.

- *PlaneRCNN-Angle* is a baseline for the orientation classification utilizing PlaneRCNN reconstruction. It simply takes the plane surface normals from PlaneRCNN, calculates the angle differences for pairs of planes, then classifies the relationship (parallel, orthogonal, or neither) based on the angular difference with a tolerance of 10 degrees.
- *PlaneRCNN-Contact1* is a baseline for the contact inference utilizing PlaneRCNN plane masks. We perform the dilation operation 5 times (by OpenCV implementation) to expand each PlaneRCNN plane mask. A pair of planes is deemed to be in contact if the intersection of their expanded masks have more than 10 pixels. The intersection region is reported as the contact mask.
- *PlaneRCNN-Contact2* is a baseline for the contact inference utilizing PlaneRCNN plane masks as well as parameters. We follow the steps of generating ground-truth contact information in Sec. 4, while replacing the GT plane parameters by the PlaneRCNN parameters. This baseline takes into account both 2D segmentation masks and 3D plane depths in judging the contact relationship.

Table 4 demonstrates that the proposed approach performs the best in all the metrics. While all the baselines perform reasonably well by utilizing the

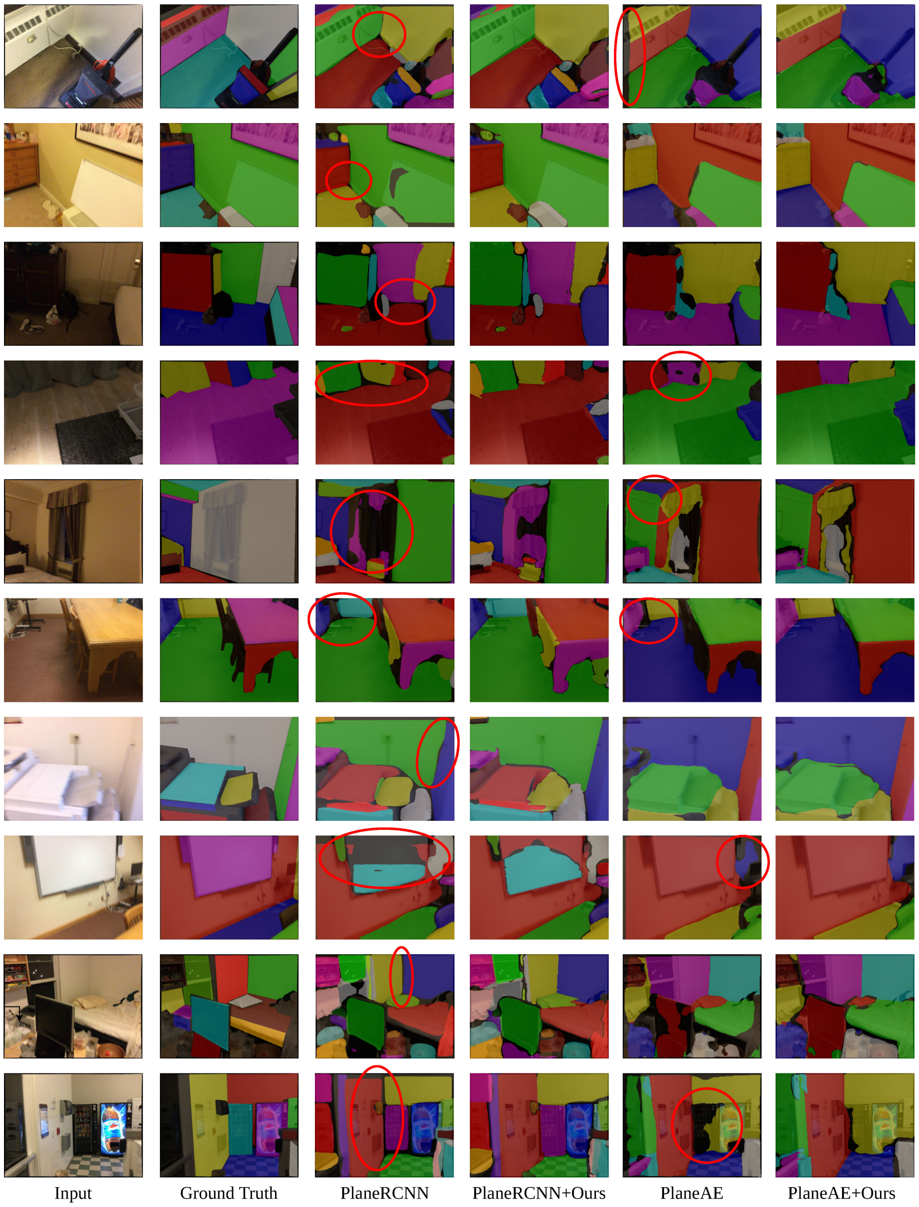

Fig. 6. Visual comparison of planar segmentation results. From left to right, it shows the input RGB image, our ground-truth manual annotation by LabelMe [16], the PlaneRCNN results [10], our results with PlaneRCNN, the PlaneAE results [24], and our results with PlaneAE. Improvements by the proposed approach are noticeable at various places, as highlighted in the red ovals

Table 4. Inter-plane relationship evaluation on the two CNN modules (Orientation-CNN and Contact-CNN). F1-scores are used for the parallel, orthogonal, and contact relationship classifier. IoU metric is used for the contact mask prediction. We compare against a few baseline methods (PlaneRCNN-Angle, PlaneRCNN-Contact1, and PlaneRCNN-Contact2). We also conduct an ablation study where we control the amount of input information. The color cyan and orange denote the best and the second best results

Method	F1-score (parallel)	F1-score (orthogonal)	F1-score (contact)	IoU% (contact mask)
PlaneRCNN-Angle	0.51	0.68	—	—
PlaneRCNN-Contact1	—	—	0.64	35.75
PlaneRCNN-Contact2			0.69	21.60
Ours (mask)	0.37	0.51	0.69	42.84
Ours (mask+RGB)	0.45	0.58	0.69	41.40
Ours (mask+RGB+depth)	0.59	0.74	0.72	42.64
Ours (all)	0.60	0.76	0.75	45.43

PlaneRCNN reconstruction results, our simple CNN solutions (Orientation-CNN and Contact-CNN) infer inter-plane relationships the best. The ablation study (the last 4 rows) shows that the CNN modules consistently improves numbers as more input information is given.

6 Conclusion

This paper proposed a novel single-image piecewise planar reconstruction technique that infers and enforces inter-plane relationships. Our approach utilizes CNNs to infer the relationships, refines the plane parameters by optimization, and employs a message passing neural network for jointly refining the plane segmentation, while enforcing the inter-plane consistency constraints. We have generated ground-truth inter-plane relationship labels and introduced three new metrics in assessing reconstruction and segmentation. Qualitative and quantitative evaluations demonstrate the effectiveness of the proposed method.

Acknowledgments: This research is partially supported by NSERC Discovery Grants, NSERC Discovery Grants Accelerator Supplements, and DND/NSERC Discovery Grant Supplement.

References

1. Bauchet, J.P., Lafarge, F.: Kippi: kinetic polygonal partitioning of images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3146–3154 (2018)
2. Chen, J., Liu, C., Wu, J., Furukawa, Y.: Floor-sp: Inverse cad for floorplans by sequential room-wise shortest path. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2661–2670 (2019)

3. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5828–5839 (2017)
4. Deng, Z., Todorovic, S., Latecki, L.J.: Unsupervised object region proposals for rgb-d indoor scenes. Computer Vision and Image Understanding **154**, 127–136 (2017)
5. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Manhattan-world stereo. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1422–1429. IEEE (2009)
6. Gallup, D., Frahm, J.M., Pollefeys, M.: Piecewise planar and non-planar stereo for urban scene reconstruction. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1418–1425. IEEE (2010)
7. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: 2009 IEEE 12th international conference on computer vision. pp. 1849–1856. IEEE (2009)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Kushal, A., Seitz, S.M.: Single view reconstruction of piecewise swept surfaces. In: 2013 International Conference on 3D Vision-3DV 2013. pp. 239–246. IEEE (2013)
10. Liu, C., Kim, K., Gu, J., Furukawa, Y., Kautz, J.: Planercnn: 3d plane detection and reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4450–4459 (2019)
11. Liu, C., Wu, J., Kohli, P., Furukawa, Y.: Raster-to-vector: Revisiting floorplan transformation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2195–2203 (2017)
12. Liu, C., Yang, J., Ceylan, D., Yumer, E., Furukawa, Y.: Planenet: Piece-wise planar reconstruction from a single rgb image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2579–2588 (2018)
13. Liu, C., Schwing, A.G., Kundu, K., Urtasun, R., Fidler, S.: Rent3d: Floor-plan priors for monocular layout estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3413–3421 (2015)
14. Monszpart, A., Mellado, N., Brostow, G.J., Mitra, N.J.: Rapter: rebuilding man-made scenes with regular arrangements of planes. ACM Trans. Graph. **34**(4), 103–1 (2015)
15. Nocedal, J., Wright, S.: Numerical optimization. Springer Science & Business Media (2006)
16. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. International journal of computer vision **77**(1-3), 157–173 (2008)
17. Sharma, G., Goyal, R., Liu, D., Kalogerakis, E., Maji, S.: Csgnet: Neural shape parser for constructive solid geometry. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5515–5523 (2018)
18. Shi, Y., Xu, K., Niessner, M., Rusinkiewicz, S., Funkhouser, T.: Planematch: Patch coplanarity prediction for robust rgb-d reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 750–766 (2018)
19. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)
20. Sinha, S., Steedly, D., Szeliski, R.: Piecewise planar stereo for image-based rendering (2009)

21. Sun, C., Hsiao, C.W., Sun, M., Chen, H.T.: Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1047–1056 (2019)
22. Xiao, J., Furukawa, Y.: Reconstructing the world's museums. International journal of computer vision **110**(3), 243–258 (2014)
23. Yang, F., Zhou, Z.: Recovering 3d planes from a single image via convolutional neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 85–100 (2018)
24. Yu, Z., Zheng, J., Lian, D., Zhou, Z., Gao, S.: Single-image piece-wise planar 3d reconstruction via associative embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1029–1037 (2019)
25. Zebedin, L., Bauer, J., Karner, K., Bischof, H.: Fusion of feature-and area-based information for urban buildings modeling from aerial imagery. In: European conference on computer vision. pp. 873–886. Springer (2008)
26. Zhang, F., Nauata, N., Furukawa, Y.: Conv-mpn: Convolutional message passing neural network for structured outdoor architecture reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
27. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. arXiv preprint arXiv:1908.00222 (2019)