YawP^3: Yaw-invariant Parametrization and Panoramic Planar-reconstruction

Anonymous ECCV submission

Paper ID 5951

Abstract. This paper presents the first method for indoor planar reconstruction from a panoramic image. We leverage three unique characteristics in indoor scenes: 1) most planar surfaces are either horizontal or vertical, 2) the 2D orientation of vertical planes co-vary with the change of the yaw angle of the camera, and 3) most vertical planes share common 2D orientations. To inherently ingrain these priors, our model first segments horizontal/vertical planes (HV-planes) for separate treatments. Most importantly, we propose a novel yaw-invariant parameterization for vertical planes to solve the yaw ambiguity problem of 360° and effectively cluster vertical planar segments with a shared 2D orientation. Finally, we fuse the predicted geometry information and plane instance segmentation into a piece-wise planar reconstruction. We evaluate our method on our newly extracted panoramic piece-wise HV-planar dataset derived from three large-scale RGB-D panorama datasets. For benchmark baselines, we train two state-of-the-art planar models on our dataset with modifications to help them adapt to our 360° HV-planar dataset. Our method significantly outperforms all baselines and achieves superior visually pleasing reconstruction of indoor scenes.

Keywords: Planar reconstruction, 3D reconstruction, panoramic image, 360° dataset

1 Introduction

Reconstructing planar surfaces from single-view images have many applications such as interior modeling, virtual and augmented reality, robot navigation, scene understanding, and 3D reconstruction. Generally, an indoor scene can be approximated with a small set of dominant planes, which makes planar reconstruction suitable for 3D indoor modeling.

Recent works on planar reconstruction [12–14] built upon recent state-of-the-art instance segmentation methods achieve promising results. However, these works are mostly trained on the planar datasets derived from ScanNet [5] and NYUv2 [16] with a small field-of-view (FoV). Data of this kind require multiple images to reconstruct entire scenes, which would cost more computational time and resources. As 360° devices get popularized, the amount of 360° data has significantly increased. With 360° images as input, 3D reconstruction of an entire scene can be done with only one snapshot on the input data. Considering the

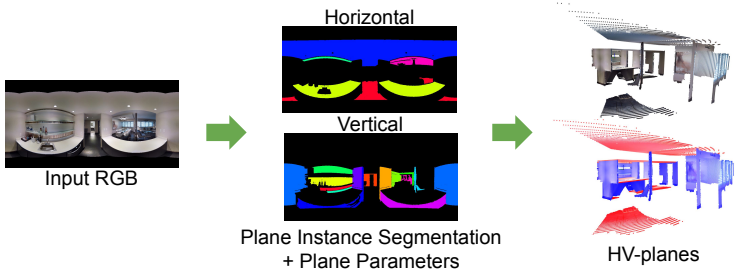


Fig. 1: Planar surfaces of man-made structures are mostly horizontal or vertical with respect to the gravity direction (HV-planes). Given an RGB panorama (left), we propose to model a 3D scene as HV-Planes (middle). The red and blue points represent pixels belonging to H-planes and V-planes respectively (right).

benefits of 360° data in planar reconstruction and the gap of existing literature in this research field, we believe the task of planar reconstruction from a panoramic image is worthy of investigation.

In this paper, we propose YawP^3, the first deep neural network addressing 360° planar reconstruction. We manage to simplify the tasks of planar instance segmentation based on surface orientation grouping, which provides geometric awareness to the succeeding module of instance segmentation. Moreover, we propose a yaw-invariant parameter in V-plane prediction to solve the yaw ambiguity problem in omnidirectional images. We show that the yaw-invariant parameterization brings about significant improvements in 3D evaluation metrics that consider geometric accuracy. To evaluate our proposed method, we present a 360° planar dataset as well as baseline models adapted from state-of-the-art models of perspective views. We further simplify this task by presenting novel HV-plane annotations as data ground truth to reduce the degree-of-freedom in planar parameters. Fig. 1 illustrates the key concepts of the design. Our model splits the prediction into H/V branches based on the new representation and combines the yaw-invariant representation into V-branch since H-plane only consists of one degree of freedom, which is perpendicular to yaw-direction. The plane geometry information is fused with planar instance segmentation for final reconstruction. Our model achieves state-of-the-art performance on the newly presented benchmark.

The contributions of this work are manifold. In terms of **technical contribution**, the proposed YawP^3 framework consists of *i*) a novel yaw-invariant vertical plane parameter representation that addresses the 360° yaw ambiguity and also boosts other existing methods; *ii*) a new geometry-aware mechanism of plane-based segmentation which exploits the geometry information to improve planar instance segmentation. In terms of **system contribution**, we construct a new 360° piece-wise planar benchmark, which focuses on horizontal and vertical planes, where the proposed YawP^3 outperforms the two adapted state-of-the-art planar reconstruction methods that are originally designed for perspective images.

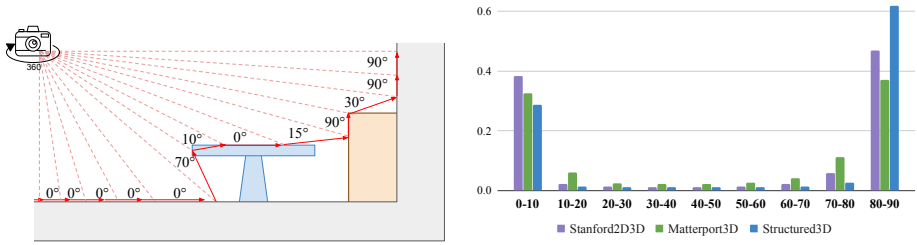
2 Related Work

Reconstructing piece-wise 3D planar surfaces from an image could be decomposed into two problems: segmenting plane instances and estimating plane geometric parameters. To solve these problems, PlaneNet [13] reconstructs a fixed number of planes by estimating plane parameters and plane segmentation masks both in an instance-wise manner by jointly training CNN and DCRF as [27]. Recent state-of-the-art approaches relax the constraint on the number of planes by exploiting popular frameworks in instance segmentation. PlaneRCNN [12] modifies the two-stage architecture Mask R-CNN [9] with object category classification replaced by plane geometry prediction, followed by a network to refine the segmentation masks. PlanarReconstruction [24] predicts per-pixel plane parameters and adopts the method similar to [3, 7, 10, 24], which trains a network to map each pixel to embedding space and then clusters the embedded pixels to generate instances. In the field of piece-wise planar reconstruction, geometric clues such as plane orientation could be useful for segmenting plane instances. However, existing works utilize limited plane geometry information in their instance segmentation processes. For instance, PlanarReconstruction [24] integrates the segmented plane instance masks into geometry prediction with the instance-level loss for plane parameters, but the segmentation process is still independent to the estimated geometry. PlaneRCNN [12] uses the initial estimation of segmentation maps and depth as input to train a refinement network for segmentation. In contrast, our method explicitly involves the estimated plane orientations in our plane instance segmentation process.

Previous works for planar reconstruction focus on perspective images, while this paper targets 360° images. In 360° domain, estimating per-pixel surface normals from an equirectangular image is challenging for the CNN layers as the yaw-rotation of the 360° camera is ambiguous. In other words, 360° camera yaw-rotation affects plane parameters, but the counterpart left-right circular shifting on the equirectangular image is less aware by the CNN layers due to its translation invariant property. Although surface normals are fundamental to many 360° applications aside from planar reconstruction, none of the existing works which estimate surface normals from an equirectangular image [22, 17, 20, 6] discuss the 360° camera yaw ambiguous problem. To this end, we propose a method that targets 360° images, with a yaw-invariant parameterization of vertical plane orientation for our applications.

3 PanoHV Dataset

In this section, we first introduce some large-scale panoramic public datasets that are used to construct our dataset. We then show the statistical analysis on these datasets to support the gravity aligned assumption and the validity of scene approximation with HV-planes, followed by the description of the pre-processing mechanism. Finally, we briefly describe the ground truth extraction algorithm. Please refer to supplementary material for the full description of our



(a) (b)

Fig. 2: The histograms of angles yielded by adjacent points on a vertical 'slice' of a point cloud. We divide the angles from 0° to 90° into nine bins. An angle is calculated between the horizon and the vector connecting two 3D points projected from the corresponding top-bottom adjacent pair of pixels in the image. This local analysis provides a strong cue to support the gravity aligned assumption and will be used in benchmark construction. Please refer to Sec. 5 for more details.(a) An illustration of a vertical 'slice' taken from a point cloud. (b) Statistics of angles quantized as histograms

data preparation algorithm. All panoramic images in this work are represented in the equirectangular format if not specified otherwise.

3.1 360° Dataset Sources

Inspired by [12, 13], we construct our dataset from three public 360° RGB-D datasets, including Structure3D [26] in synthetic environments, and Matterport3D [4] and Stanford2D3D [2] in real-world scenes. These three datasets consist of large-scale aggregation of panoramic RGB images and depth maps. Our annotations of planar masks and parameters are derived from the ground-truth depths.

3.2 Gravity Aligned and HV-plane Assumption

Similar to the *Manhattan* assumption made by most of the previous methods (*e.g.* [15, 18, 28]), we assume the gravity aligned property for simplifying the task of *indoor piece-wise planar reconstruction from a single panorama*. More specifically, we aim to approximate the indoor scene by horizontal planes (***H-planes*** with the normal parallel to the gravity direction) and vertical planes (***V-planes*** with the normal perpendicular to the gravity direction). Although such an assumption complies with our intuition, we further perform two quantitative analyses on the aforementioned three large-scale indoor panorama datasets. These datasets include various indoor scenes (*e.g.*, classroom, office, living room, kitchen) to justify the validity and applicability of the assumption.

1) For every pair of adjacent pixels in the y-direction of an image, we calculate the angle between the horizon and the vector connecting the two 3D points

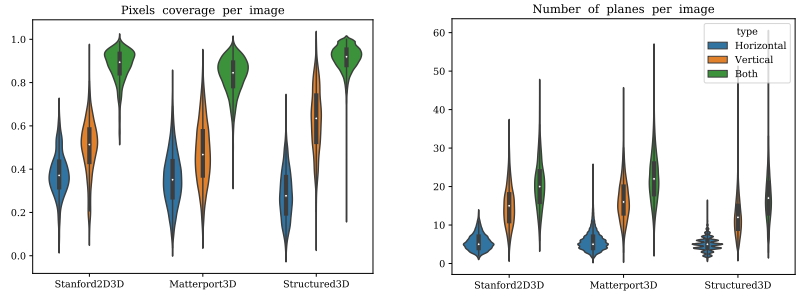


Fig. 3: The violin plot illustrates the statistics of the constructed HV-plane benchmark. We show all the combinations between the three panoramic datasets (X-axis) and the type of planes (blue bars for H-planes, orange bars for V-planes, and green bars for both). The left canvas shows the per-image pixel coverage of HV-planes. The Y-axis is the percentage of covered pixels with respect to valid pixels (having depth). The right canvas shows the number of HV-planes per image

that correspond to those two neighboring pixels. The histogram of angles computed from all images on the three indoor panoramic datasets is shown in Fig. 2. Imagine that we take a vertical 'slice' of the 3D point cloud and move vertically between the adjacent points along with the slice—Most of the moving directions will be either horizontal (0°-10°) or vertical (80°-90°). Such a cue suggests the gravity aligned nature of indoor scene structures.

2) We approximate the scenes with our H-planes and V-planes extraction algorithms, which will be described later. The results of statistics over indoor panoramic datasets are shown in Fig. 3. We can see that, in general, more than 80% of the pixels in an image can be covered by roughly 20 HV-planes. This result suggests that our HV-plane approximation derived from gravity aligned planes are suitable for modeling the gist of an indoor scene.

3.3 Panoramic Image Alignment

We assume all images are aligned with the gravity direction. In case that g-sensor and tripod are not equipped with the 360° camera, and the image is not aligned, we can use the voting-based algorithm mentioned in [8, 23, 25, 29] for panoramic image alignment and vanishing point (VP) detection. Panoramas generally provide enough context to extract the gravity direction, and we will not lose any pixel or introduce any padding after image alignment.

3.4 Ground-truth HV-planes Preparation

Following the gravity aligned and HV-plane assumptions, we present the first 360° planar dataset with training, validation, and test sets. The annotations include HV-plane masks and plane parameters, and all images and annotations are

in the same resolution of 512×1024. We use the same local analysis as in Fig. 2 to classify each pixel into H-pixel, V-pixel, or other. RANSAC is then performed on H-pixels and V-pixels to extract instance mask and plane geometry for H-planes and V-planes, respectively. The statistical information of the constructed dataset is depicted in Fig. 3. Please refer to supplementary material for detailed ground truth extracting algorithm.

4 Approach

Our task is to reconstruct HV-planes from a single 360° RGB image. We first segment an image into H-planar, V-planar, and non-planar regions (Sec. 4.1). Based on the HV-planar segmentation and 360° image geometry, pixel-level depth map, and V-planar orientation with the novel yaw-invariant representation are transformed into planes geometry (Sec. 4.2). A geometry-aware plane instance segmentation is proposed to exploit the aforementioned estimation to first group pixels with similar planar orientations, and then identify individual plane instances within each group by pixel embedding (Sec. 4.3). An overview of our approach is depicted in Fig. 4. We use a shared backbone with multiple heads for different prediction targets.

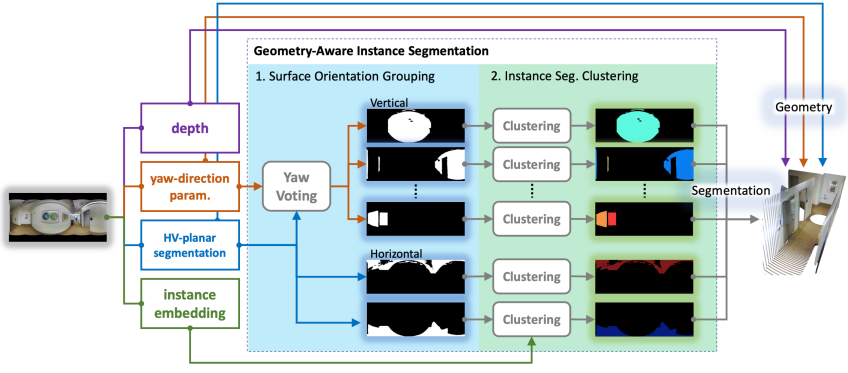


Fig. 4: The pipeline of the proposed YawP^3 framework. The deep model predicts pixel-level geometry information, HV-planar segmentation, and instance embedding. In the proposed geometry-aware plane instance segmentation process, we first group pixels with a shared orientation and then identify individual planes from each group via instance embedding

4.1 HV-Planar Segmentation

To distinguish the H-planar and V-planar regions, we predict each pixel as 'H-planar,' 'V-planar,' or 'Non-planar' with two binary classifiers. Both classifiers

are trained to be activated on pixels of H-planes and V-planes respectively, with Binary Cross Entropy loss. For inference, a pixel is recognized as 'Non-planar' when the corresponding probabilities from both classifiers are below a certain threshold (we set to 0.5 in this work); otherwise, it will be classified as 'H-planar' or 'V-planar' according to which classifier gives a higher probability. Based on the segmented H-planar pixels and V-planar pixels, we exploit their prior to give different treatments in the following plane reconstruction process.

4.2 Planar Geometry Estimation

We train the model to estimate two types of pixel-level geometry information—depth d and V-planar surface normal in radian θ. Two heads are attached to the final layer of the backbone for per-pixel d, θ estimation. In below, a planar geometry is denoted as $\vec{n} = \begin{bmatrix} x & y & z \end{bmatrix}$, which is the unit surface normal multiplied by the plane offset; image coordinate is denoted as $u \in [-\pi, \pi]$ and $v \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Note that the geometry re-projection described below is pixel-level where we use subscript to denote a pixel.

H-planar geometry. The unit surface normal of an H-plane is either $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$ or $\begin{bmatrix} 0 & 0 & -1 \end{bmatrix}$ (corresponding to horizontal plane above or below the camera respectively) and can be determined accordingly as the pixel is located at the upper half or the bottom half of an equirectangular image. The H-plane offset of a pixel with index i can be derived from d_i by

$$z_i = d_i \cdot \sin(v_i).$$

The training loss for H-planes is $L_H = \frac{1}{|I_H|} \sum_{i \in I_H} |z_i - z_i^*|$ where I_H is the set of all H-planar indices and z_i^* is the ground truth H-plane offset of the pixel. In testing phase, the offset of a detected H-plane instance mask M is determined by the median of $\{z_i \mid i \in I_M\}$ where I_M is all the indices of the instance mask.

V-planar geometry. We derive the V-planar unit surface normal from the estimated θ as $\begin{bmatrix} \cos\theta_i & \sin\theta_i & 0 \end{bmatrix}$. Following the manner of H-planar geometry, the V-planes offset is derived by

$$o_i = d_i \cdot \cos(v_i) \cdot \begin{bmatrix} \cos\theta_i & \sin\theta_i \end{bmatrix} \cdot \begin{bmatrix} \cos(u_i) & \sin(u_i) \end{bmatrix}^T.$$

Thus, $\vec{n}_i = o_i \cdot \begin{bmatrix} \cos\theta_i & \sin\theta_i & 0 \end{bmatrix}$ is the geometry representation of the V-plane. The V-planar loss is $L_V = \frac{1}{|I_V|} \sum_{i \in I_V} \|\vec{n}_i - \vec{n}_i^*\|$, where $\vec{n}_i^*$ is the ground truth V-planar normal vector. We also train θ directly using cosine similarity loss $L_\theta = 1 - \begin{bmatrix} \cos\theta_i & \sin\theta_i \end{bmatrix} \cdot \begin{bmatrix} \cos\theta_i^* & \sin\theta_i^* \end{bmatrix}^T$ where θ_i^* is derived from $\vec{n}_i^*$. Similar to the testing phase of H-plane geometry, we take the dimensional wise median of $\vec{n}$ in a V-plane instance region.

Yaw-invariant parameterization for V-planes. The accuracy of the estimated V-planar orientation is critical in our framework. It affects the geometry quality and also relates to the proposed geometry-aware plane instance segmentation process (Sec. 4.3). However, V-planar orientations estimation in a 360° image

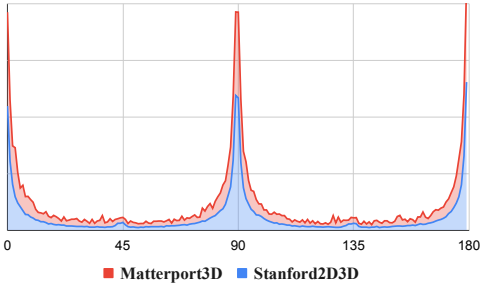
is a challenging task for CNN layers. The reason is that the V-planar surface orientation co-varies with the 360° camera yaw-rotation (which corresponds to left-right circular shifting on the equirectangular image), but the CNN layers are less aware of it due to the translation invariant property. Adding u-coordinate to the input image as an additional channel could be a workaround. To achieve better performance, we propose re-parameterizing θ_i into residual form with respect to the pixel yaw viewing angle u_i such that it is invariant to the 360° camera yaw-rotation. More specifically, $\theta'_i = \theta^*_i - u_i = \arctan2\,(y^*_i, x^*_i) - u_i$ is the re-parameterized yaw-invariant V-planar orientation representation. The proposed representation enables the model to infer the V-planar orientation without the knowledge about the 360° camera yaw-rotation.

4.3 Geometry-Aware Plane Instance Segmentation

Motivated by the strong prior of indoor scenes where most plane instances share a small number of distinct orientations, we propose to integrate plane surface information into the main process of plane instance segmentation. Our geometry-aware plane instance segmentation consists of two stages: *i)* aggregate pixels by surface orientation grouping; *ii)* in each group, a pixel embedding clustering algorithm is conducted to identify individual plane instance.

Stage 1: Surface orientation grouping. The analysis on our dataset (see Fig. 5) reveals that most plane instances in an indoor scene share similar planar orientations with other planes. Hence, the per-pixel V-planar surface orientations are distributed primarily around a small number of angles. Utilizing such regularity, we divide pixels of an image into groups of similar plane normals, which preliminarily separates plane instances of dissimilar orientation. For H-planes, the two surface orientation groups (*i.e.* $\vec{n} = (0, 0, \pm 1)$) can be easily determined by the HV-planar segmentation mask (Sec. 4.1). Specifically, pixels classified as H-planar are assigned to two groups according as they are located above or below the equator (corresponding to H-planes above or below the camera center). For V-planes, we apply a voting process to detect the prominent V-planes surface orientation peaks. Namely, we quantize the estimated V-planar orientations of pixels classified as V-planar into circular bins following by a peak-finding algorithm (see supplementary material for detail). V-planar pixels are assigned to the nearest peak to form surface orientation groups.

Stage 2: Pixel embedding clustering. To further identify the plane instances in each surface orientation group, we exploit pixel embedding, which is widely used for instance segmentation [3, 7, 10] and also found to be effective for plane instances segmentation [24]. We follow the loss in [24] and train the model to pull the pixels of a plane instance toward their centroid in the embedding space and push the centroids of different planes away from each other. In testing phase, We apply the mean shift clustering to each orientation group separately.



(a)

(b)

Fig. 5: (a) Statistic of angles between all pairs of planes in an image. The results are averaged across the entire Stanford2D3d and subsampled Matterport3D. (b) One result of surface orientation grouping. Different colors denote different groups. The outer circle represents the distribution of yaw-rotations, where each point on the circle is a vote by a pixel

5 Experiments

5.1 Baselines Construction

We manage to adapt two competitive planar reconstruction approaches – PlanarReconstruct [24][1] and PlaneRCNN [12][2] – to 360° H-FoV panoramic images based on their official implementation. The following is the description of the general changes for both baselines and the respective adaptations.

Common adaptation. The x-axis of a perspective image is linear, but in our 360° H-FoV panorama case, the x-axis is circular. We exploit the left-right circular padding [19] for all CNN layers with 360° H-FoV features as input. The u-coordinate is concatenated as one of the input channels to alleviate the ambiguity of yaw-rotation on 360° images. All methods, including ours, use Resnet101-FPN [11] as the backbone for a fair comparison.

PlaneRCNN [12]. We carefully update RoIAlign and NMS in Mask R-CNN to deal with the left-right circular coordinate system. To refine a b-box from an anchor box, the left-right movement can be ambiguous, so we train the model by choosing the direction with the minimum shifting. We have normal clusters for both vertical and horizontal planes. The resolution of the segmentation refinement module is scaled according to the resolution of the original input.

PlanarReconstruct [24]. We duplicate all heads for H-planes and V-planes. We enforce the plane parameters $\vec{\mathbf{n}}_x = 0, \vec{\mathbf{n}}_y = 0$ for H-planes, and $\vec{\mathbf{n}}_z = 0$ for V-planes by ignoring the prediction of the corresponding output channels.

[1] https://github.com/svip-lab/PlanarReconstruct
[2] https://github.com/NVlabs/planercnn

Hyperparameters (*i.e.*, number of sampled points and number of iterations) of the Efficient Mean Shift are scaled according to the change of image resolution.

5.2 Data Split

We follow the official setting to split the scenes into training, validation, and testing sets, but remove data with too many missing pixels in depth value while extracting the ground-truth HV-planes (refer to Sec. 3 for ground truth extraction detail). Finally, Stanford2D3D [2] contains 1,040 images for training and 372 images for validation; Matterport3D [4] contains 7,275, 1,189, and 1,005 for training/validation/test respectively; Structured3D [26] contains 18,332 for training, 1,771 for validation, and 1,691 for testing.

5.3 Evaluation Metrics

Following previous works [12, 21, 24], we evaluate the performance of plane instances segmentation with some common clustering metrics [1]: Adjusted Rand Index (ARI↑), Variation of Information (VI↓), and Segmentation Covering (SC↑) which only consider the segmentation result on the 2D image. To evaluate 3D reconstruction quality, plane and pixel recall are used under different geometric thresholds and a segmentation criteria. We report the results by averaging the plane recall and pixel recall under depth threshold of 5cm, 10cm, 20cm, 30cm and 60cm.

5.4 Results

Quantitative evaluation. In Table 1, we report comparisons between our approach and two competitive baselines on three presented 360° datasets. The performance of Stanford2D3D [2] is shown on the validation set as it lacks test set. Our method achieves state-of-the-art performance consistently on all metrics. For the metrics that only relate to segmentation quality, our method shows more improvement on Matterport3D, which contains more complex scenes captured in luxury houses comparing to the performance gain on Stanford2D3D with mainly official buildings and the synthetic Structured3D dataset. For metrics considering 3D reconstruction quality, our approach outperforms both baselines by a large margin. The results in Table 1 clearly demonstrate the effectiveness of our proposed method.

Qualitative results. In Fig. 7, we show the 3D models reconstructed by PlaneRCNN [12], PlanarReconstruct [24], and the proposed YawP^3. Owing to our yaw-invariant representation, the geometric quality of YawP^3 is generally better than the others. In addition, the planes reconstructed by ours are aligned better with each other. In contrast, using a detection-based method to predict segmentation mask and plane normal separately for each instance, PlaneRCNN has larger gaps between plane instances, which is especially obvious for the large planes in the first, second, and sixth examples. In general, we achieve higher quality in both geometry and plane instance segmentation.

Table 1: The performance of our method and the two baselines reported on the three datasets

Method	Segmentation Quality ARI↑	VI↓	SC↑	Per-pixel recall ↑	Per-plane recall ↑
Matterport3D [4] test set					
P.RCNN [12]	0.574	2.022	0.632	0.473	0.336
P.Recon. [24]	0.673	1.944	0.640	0.498	0.381
Ours	**0.686**	**1.894**	**0.660**	**0.544**	**0.410**
Stanford2D3D [2] validation set					
P.RCNN [12]	0.682	1.677	0.703	0.452	0.297
P.Recon. [24]	0.765	1.536	0.733	0.520	0.341
Ours	**0.768**	**1.514**	**0.742**	**0.627**	**0.430**
Structured3D [26] test set					
P.RCNN [12]	0.726	1.393	0.743	0.654	0.522
P.Recon. [24]	0.821	1.175	0.785	0.728	0.591
Ours	**0.824**	**1.150**	**0.794**	**0.794**	**0.657**

5.5 Ablation Study

In this section, several ablation studies are shown to further prove the effectiveness of the proposed method and identify the source of improvement.

Can the proposed 360° yaw-invariant plane parameters benefit other baselines? As mentioned in Sec. 4.2, one challenge of 360° planar orientation estimation is that the non-horizontal normals are related to camera yaw rotation, but the CNN layers are less aware of the counterpart 2D left-right circular shifting on the equirectangular image. In addition to the input u-coordinate workaround, we propose a 360° camera yaw-invariant representation for V-plane normals, which ensures the independence between the ground truth normal and camera yaw rotation (see Sec. 4.2 for detail). The proposed representation is applied to other baselines, and the results are reported in Table 2. We observe consistent improvements on all metrics by applying the proposal on PlaneRCNN [12]. PlanarReconstruct [24] also achieves better 3D reconstruction quality with a slightly degraded 2D segmentation. Please note that even by applying the proposed yaw-invariant to other baselines, our method still achieves the best performance, which suggests that the proposed plane representation is not the only source for our superior results.

Can the proposed planar orientation parameterization along resist the yaw ambiguity of 360° camera? To show the yaw ambiguity problem of 360° images and how the proposed yaw-invariant representation resists to it, we conduct an ablation study on our method with all the combinations of input u-coordinate and the proposed representation. In Table 3, we show the per-pixel

Table 2: Ablation study of the proposed yaw-invariant parameterization on the two baselines

Method	yaw-invariant	Segmentation Quality ARI↑	VI↓	SC↑	Per-pixel recall ↑	Per-plane recall ↑
Matterport3D [4] validation set						
P.RCNN		0.531	2.118	0.613	0.471	0.313
	✓	**0.571**	**2.053**	**0.632**	**0.495**	**0.326**
P.Recon		**0.634**	**2.102**	**0.615**	0.484	0.362
	✓	0.625	2.171	0.602	**0.495**	**0.369**
ours	✓	0.654	2.044	0.640	0.545	0.398
Stanford2D3D [2] validation set						
P.RCNN		0.682	1.677	0.703	0.452	0.297
	✓	**0.683**	**1.665**	**0.705**	**0.481**	**0.313**
P.Recon		**0.766**	**1.536**	**0.733**	0.520	0.341
	✓	0.763	1.537	0.728	**0.555**	**0.368**
ours	✓	0.768	1.514	0.742	0.627	0.430

normal error in degree over the V-planar region of Stanford2D3D [2] validation set. As mentioned in Sec. 4.2, V-plane normals depend on 360° camera yaw-rotation, but CNN layers are less aware of it. Consequently, the model with neither input u-coordinate nor the proposed yaw-invariant representation leads to inferior V-planar geometry quality. On the other hand, our yaw-invariant reparameterization, which is designed to be invariant to the 360° camera yaw-rotation, can achieve superior orientation quality.

Table 3: Ablation study of the V-plane normal error under different settings. A row with the first column checked indicates that the model has an extra input of image u-coordinate. A checkmark on the second column means the proposed yaw-invariant representation is used. Other than the two factors, all models share the same setting and are trained on Stanford2D3D [2] training set. The results are evaluated on Stanford2D3D validation set

input u-coord.	**yaw-invariant**	V-plane orientation error (deg°)↓
		7.29
✓		5.97
	✓	**5.18**
✓	✓	5.21

The effectiveness of the surface orientation grouping stage. To examine the impact of the surface orientation grouping stage, we modify PlanarReconstruction [24], the baseline which only adopts pixel embedding, by integrating our proposed yaw-invariant representation into their method. In Table 2, the comparison between the results of ours and the modified PlanarReconstruction [24] suggests that that the geometry-aware plane instance segmentation apart from yaw-invariant representation also improves the performance. We show some representative qualitative comparisons in Fig 6.

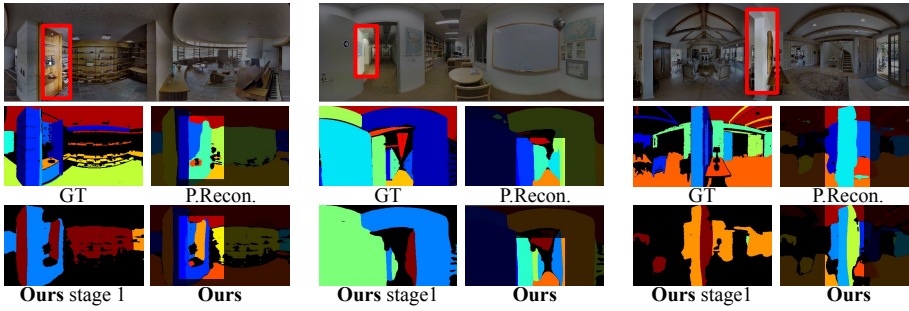


Fig. 6: Three representative examples to demonstrate the effectiveness of the proposed geometry-aware plane instance segmentation. In comparison with the baseline considering only pixel embedding, which fails in the highlighted regions, our geometry-aware plane instance segmentation can separate the undetected plane from its neighboring planes by the first stage orientation voting. Please see Sec. 4.3 for the intuition and the technical detail

6 Conclusion and Future Work

This paper presents YawP^3, a novel deep neural network targeting indoor 3D planar reconstruction from a single panoramic image with HV-planes. YawP^3 applies surface orientation grouping to panoramic images, which improves the instance segmentation module through geometric awareness. The proposed yaw-invariant V-plane parameter of YawP^3 is capable of both solving surface normal ambiguity of panoramic images and boosting the performance of reconstruction. We created the first 360° planar reconstruction benchmark with datasets, and baseline models for experimentation and comparison. The presented planar datasets are derived from large-scale RGB-D datasets and the baseline models are modified from representative perspective view planar models to adapt to 360° data. Our model outperforms the two baselines by a large margin and achieves state-of-the-art on this newly presented benchmark. Two future directions that are worth investigating are *i*) finding better approaches for reconstructing HV-planes and *ii*) completing occluded and non-estimated regions in HV-planes.

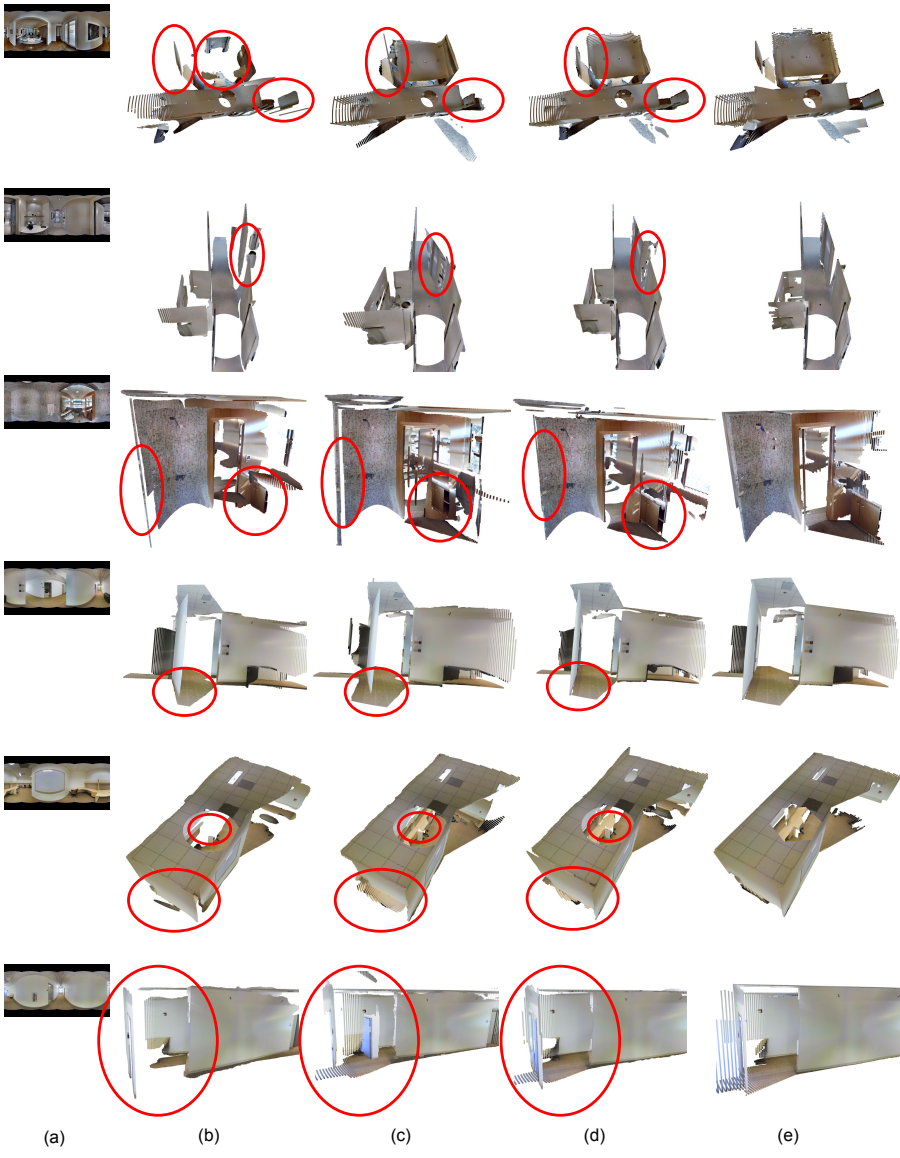


Fig. 7: We show some qualitative comparisons of the 3D reconstruction results among the three methods. We highlight the differences in red circles for easier to distinguish the reconstruction quality. (a) Input RGB, (b) PlaneRCNN [12], (c) PlanarReconstruction [24], (d) ours, and (e) ground truth

References

1. Arbelaez, P., Maire, M., Fowlkes, C.C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 898–916 (2011)
2. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. CoRR **abs/1702.01105** (2017)
3. Brabandere, B.D., Neven, D., Gool, L.V.: Semantic instance segmentation with a discriminative loss function. CoRR **abs/1708.02551** (2017)
4. Chang, A.X., Dai, A., Funkhouser, T.A., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: learning from RGB-D data in indoor environments. In: 2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017. pp. 667–676 (2017)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T.A., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 2432–2443 (2017)
6. Eder, M., Moulon, P., Guan, L.: Pano popups: Indoor 3d reconstruction with a plane-aware network. In: 2019 International Conference on 3D Vision, 3DV 2019, Québec City, QC, Canada, September 16-19, 2019. pp. 76–84 (2019)
7. Fathi, A., Wojna, Z., Rathod, V., Wang, P., Song, H.O., Guadarrama, S., Murphy, K.P.: Semantic instance segmentation via deep metric learning. CoRR **abs/1703.10277** (2017)
8. Fernandez-Labrador, C., Pérez-Yus, A., López-Nicolás, G., Guerrero, J.J.: Layouts from panoramic images with geometry and deep learning. IEEE Robotics and Automation Letters **3**(4), 3153–3160 (2018)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2980–2988 (2017)
10. Kong, S., Fowlkes, C.C.: Recurrent pixel embedding for instance grouping. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 9018–9028 (2018)
11. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 936–944 (2017)
12. Liu, C., Kim, K., Gu, J., Furukawa, Y., Kautz, J.: PlaneRCNN: 3d plane detection and reconstruction from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 4450–4459 (2019)
13. Liu, C., Yang, J., Ceylan, D., Yumer, E., Furukawa, Y.: PlaneNet: piece-wise planar reconstruction from a single RGB image. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 2579–2588 (2018)
14. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. pp. 2277–2287 (2017)
15. Purkait, P., Zach, C., Leonardis, A.: Rolling shutter correction in manhattan world. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 882–890 (2017)

16. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V. pp. 746–760 (2012)
17. Song, S., Zeng, A., Chang, A.X., Savva, M., Savarese, S., Funkhouser, T.A.: Im2pano3d: Extrapolating 360° structure and semantics beyond the field of view. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 3847–3856 (2018)
18. Sun, C., Hsiao, C., Sun, M., Chen, H.: HorizonNet: learning room layout with 1d representation and pano stretch data augmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 1047–1056 (2019)
19. Wang, T., Huang, H., Lin, J., Hu, C., Zeng, K., Sun, M.: Omnidirectional CNN for visual place recognition and navigation. In: 2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018. pp. 2341–2348 (2018)
20. Xu, J., Stenger, B., Kerola, T., Tung, T.: Pano2cad: Room layout from a single panorama image. In: 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017. pp. 354–362 (2017)
21. Yang, F., Zhou, Z.: Recovering 3d planes from a single image via convolutional neural networks. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X. pp. 87–103 (2018)
22. Yang, H., Zhang, H.: Efficient 3d room shape recovery from a single panorama. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 5422–5430 (2016)
23. Yang, Y., Jin, S., Liu, R., Kang, S.B., Yu, J.: Automatic 3d indoor scene modeling from single panorama. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 3926–3934 (2018)
24. Yu, Z., Zheng, J., Lian, D., Zhou, Z., Gao, S.: Single-image piece-wise planar 3d reconstruction via associative embedding. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 1029–1037 (2019)
25. Zhang, Y., Song, S., Tan, P., Xiao, J.: PanoContext: A whole-room 3d context model for panoramic scene understanding. In: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI. pp. 668–686 (2014)
26. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3D: A large photo-realistic dataset for structured 3d modeling. CoRR **abs/1908.00222** (2019)
27. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S.: Conditional random fields as recurrent neural networks. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 1529–1537. IEEE Computer Society (2015)
28. Zhou, Y., Qi, H., Zhai, Y., Sun, Q., Chen, Z., Wei, L., Ma, Y.: Learning to reconstruct 3d manhattan wireframes from a single image (2019)
29. Zou, C., Colburn, A., Shan, Q., Hoiem, D.: LayoutNet: reconstructing the 3d room layout from a single RGB image. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 2051–2059 (2018)