uLayout: Unified Room Layout Estimation for Perspective and Panoramic Images (Supplementary Material)

Jonathan Lee[1] Bolivar Solarte[1] Chin-Hsuan Wu[1] Jin-Cheng Jhang[1] Fu-En Wang[1]
Yi-Hsuan Tsai[2] Min Sun[1]
[1]National Tsing Hua University [2]Atmanity Inc.
{jonathanlee19896, enrique.solarte.pardo, wasidennis}@gmail.com
{chinhsuanwu, frank890725, fulton84717}@gapp.nthu.edu.tw sunmin@ee.nthu.edu.tw

1. Outline

In our main paper, we utilize a unified model called uLayout, which is capable of being jointly trained on both perspective and panoramic domain data, enabling it to accurately estimate boundaries within the image compared with two panoramic layout estimation models, LGT-Net [3] and DOP-Net [5] and two perspective layout estimation models, LSUN-ROOM [4] and FUSING [7]. Additionally, our design incorporates optimizations to minimize time, memory usage, and Floating-Point Operations (FLOPs) specifically within the perspective domain. In this supplementary material, we provide more details in the following sections:

- In the ablation study section of our main paper, we mention that the horizon depth becomes infinite when boundaries extend across the middle of the image. In Sec. 2, we will delve into more details regarding the mechanism by calculating the horizon depth and discuss its limitations.
- In Sec. 3, we will provide the experiment with ZInD [2] and LSUN [6] dataset.
- In Sec. 4, we will present the different experiment settings to evaluate the benefit of adding extra training data and a joint training approach.
- In Sec. 5, we will present additional qualitative results showcasing panoramic images sourced from the three datasets we evaluated in the paper, PanoContext [8], Standford 2D-3D [1] and MatterportLayout dataset [9], as well as perspective images captured with varying Field-of-View (FoV) from the LSUN dataset [6].

2. Calculation of Horizon Depth

In this section, we will delve into the intricacies of projecting boundaries onto a bird's-eye view and calculating horizon depth. We operate under the assumption that there exists one ceiling and one floor plane, both perpendicular to the walls.

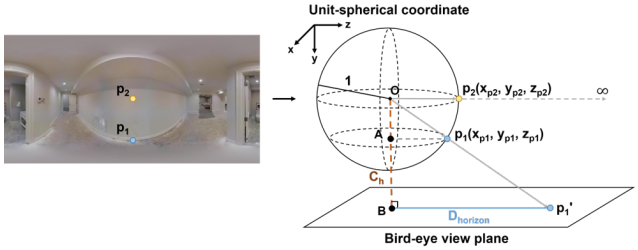


Figure 1. Illustration of calculating Horizon Depth

In Fig. 1, we take p_1 as a point on the floor boundary and p_2 as the horizon point in the middle of the image, providing two examples for clarity. To begin, we project point $p_{1,2}$ from the image space to unit-spherical coordinates, yielding $x_{p_{1,2}}$, $y_{p_{1,2}}$, and $z_{p_{1,2}}$ respectively. Subsequently, by focusing on p_1 and leveraging the similarity between triangles $\triangle Op_1A$ and $\triangle Op_{1'}B$, we can determine the horizon depth D_{horizon} using the relationship expressed in Equation Eq. (1):

$$\frac{D_{\text{horizon}}}{\sqrt{x_p^2 + z_p^2}} = \frac{C_h}{y_p} \tag{1}$$

Here, C_h denotes the camera height, while p signifies any arbitrary point within the image space.

However, in the case of p_2, where y_{p_2} equals zero, the calculation of D_{horizon} becomes infinite according to Equation Eq. (1). Consequently, we are unable to project p_2 onto the bird's eye view plane and compute the Intersection over Union (IoU). In our main paper, to mitigate this issue, we employ vertical shifting to prevent ceiling or floor boundaries from crossing the middle of the image.

3. Experiment Results with ZInD dataset

In this section, we present the experimental results of uLayout on the ZInD dataset [2], trained jointly with the LSUN dataset [6]. We also compare these results against two state-of-the-art panoramic baselines, LGT-Net [3] and DOP-Net [5], as well as two leading perspective-based methods, LSUN-ROOM [4] and FUSING [7].

In Tab. 1, when compared to the two panoramic baselines, uLayout achieves the same 3D IoU as DOP-Net [5], although its 2D IoU is slightly lower. Moreover, our model demonstrates over a 35% improvement in performance compared to the two panoramic baselines in perspective. As for the two perspective baselines, uLayout surpasses them in ceiling 2D IoU, but its floor 2D IoU falls short of FUSING [7]. This performance discrepancy can be attributed to two primary factors. First, the ZInD dataset [2], with over 25,000 panoramic images—nearly ten times the size of the LSUN dataset [6]—causes the model to learn predominantly from the panoramic domain. Second, while the ZInD dataset contains complex room layouts, the images lack furniture, allowing for clearer room structures. In contrast, the LSUN dataset is filled with furniture, which frequently occludes room layouts. This domain gap between the two datasets impacts the model's generalization, particularly in tasks like floor 2D IoU, where occlusions play a significant role.

Table 1. Experiment with ZInD [2] and LSUN [6] datasets.

	ZInD [2]		LSUN [6]	
Method	2DIoU	3DIoU	Ceiling 2DIoU	Floor 2DIoU
LGT-Net [3]	91.77	89.95	45.47	37.70
DOP-Let [5]	**91.94**	90.13	1.49	3.32
LSUN-ROOM [4]	-	-	76.59	73.62
FUSING [7]	-	-	80.68	**80.03**
Ours	91.83	**90.14**	**83.56**	79.72

4. Discussion for Additional data and Joint Training

In this section, we explore four different training settings to demonstrate the benefits of incorporating additional perspective data and using a joint training approach on the MatterportLayout [9] and LSUN [6] datasets. These settings are (a) training solely on MatterportLayout [9] panoramic data, (b) training on both MatterportLayout [9] panoramic data and perspective views extracted from these panoramas, (c) training on LSUN [6] perspective data using a pre-trained model that was initially trained on MatterportLayout [9] panoramic data, and (d) jointly training on both MatterportLayout [9] and LSUN [6] datasets.

In Tab. 2, comparing methods (a), (b), and (c), we find that joint training with perspective data derived from masked panoramic images, without adding new information such as LSUN [6] data, does not improve performance. This indicates that simply combining panoramic and perspective data is not sufficient to boost results. However, the introduction of new training data, as seen in method (d), significantly enhances performance in both domains. Additionally, comparing methods (c) and (d) shows that training panoramic and perspective data separately cannot achieve the same results as joint training. This highlights that both the inclusion of new training data and the joint training approach are essential for improving performance across both domains.

Table 2. Ablation Study with MatterportLayout [9] and LSUN [6] datasets.

		MatterportLayout [9]		LSUN [6]	
	Method	2D IoU	3D IoU	Ceiling 2D IoU	Floor 2D IoU
(a)	only pano	83.08	80.83	3.96	0.57
(b)	pano + pers(masked pano)	82.56	80.20	55.68	61.15
(c)	pretrain pano + pers(LSUN)	76.11	73.55	80.95	76.55
(d)	pano + pers(LSUN)	**84.05**	**81.84**	**83.61**	**80.25**

5. Qualitative Results

In this section, we will present additional qualitative examples of panoramic images compared to two panorama baselines, DOP-Net [5] and LGT-Net [3]. We will also provide 3D visualization examples for panoramic images to illustrate our model's capabilities. Additionally, we will showcase original data with different Field-of-View (FoV) sourced from LSUN dataset [6] as the original image, the original image after preprocessing as input image, our model's prediction, and predictions from two perspective layout models, FUSING [7] and LSUN-ROOM [4].

Qualitative Results on Panoramic Images.

In Figs. 2 to 4, we present additional qualitative results for panoramic images from PanoContext [8], Stanford 2D-3D [1] and MatterportLayout [9]. Panel (a) displays the predictions made by our model and panel (b) shows the predictions made by DOP-Net [5], while panel (c) shows the predictions made by LGT-Net [3]. In Figs. 5 to 7, we provide 3D visualization for panoramic images from three different datasets.

Qualitative Results on Perspective Images.

In Fig. 8, we present additional qualitative results displaying perspective images with various Field-of-View (FoV) sourced from the LSUN dataset [6]. Panel (a) exhibits the original images sourced from the LSUN dataset [6].

To ensure visual clarity, we adjust the original image's height and width to maintain a consistent image height while preserving the aspect ratio of height and width. Panel (b) showcases the input images, which are the original images after preprocessing. Panel (c) displays our model's predictions. Panels (d) and (e) present predictions from two perspective layout model baselines, FUSING [7] and LSUN-ROOM [4], respectively.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 1, 2, 4, 6

[2] Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow indoor dataset: Annotated floor plans with 360º panoramas and 3d room layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2133–2143, June 2021. 1, 2

[3] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *CVPR*, 2022. 1, 2

[4] Hung Jin Lin, Sheng-Wei Huang, Shang-Hong Lai, and Chen-Kuo Chiang. Indoor scene layout estimation from a single image. In *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018. 1, 2, 3

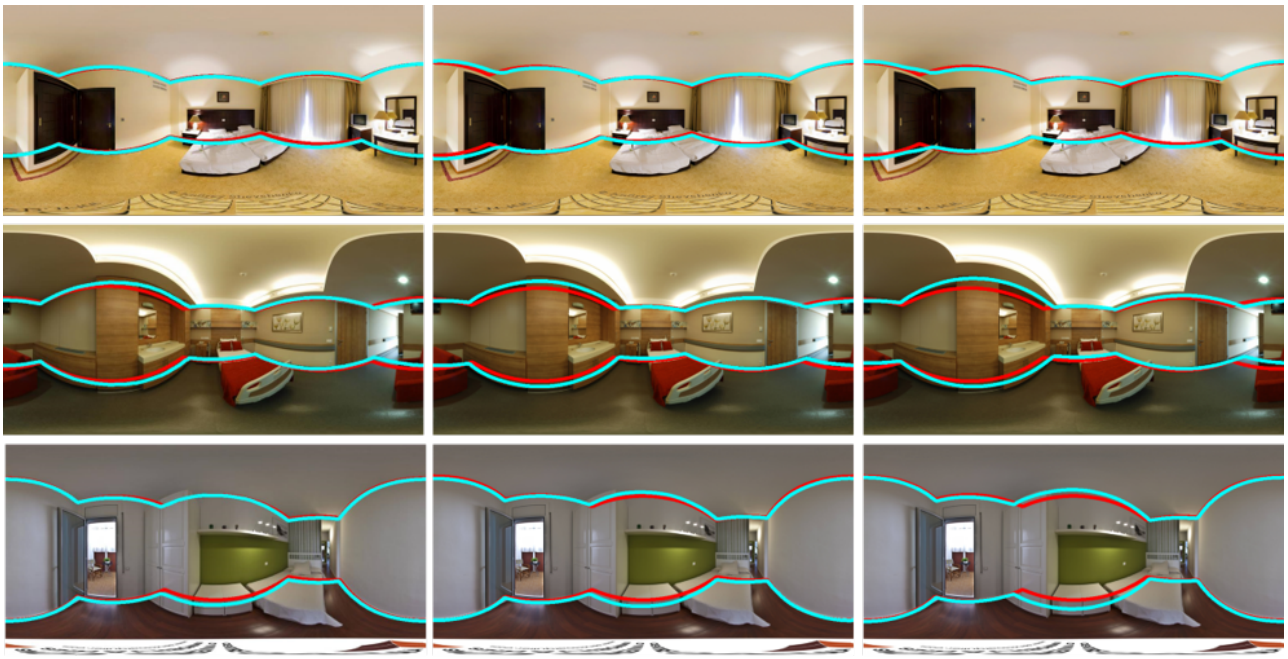
[5] Zhijie Shen, Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Shuai Zheng, and Yao Zhao. Disentangling orthogonal planes for indoor panoramic room layout estimation with cross-scale distortion awareness. In *CVPR*, 2023. 1, 2

[6] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1, 2, 3

[7] Weidong Zhang, Xueke Hu, Ying Liu, and Yuquan Gan. Fusing structural and appearance features for 3d layout estimation. In *2023 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 2195–2200. IEEE, 2023. 1, 2, 3

[8] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. 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 13*, pages 668–686. Springer, 2014. 1, 2, 4, 6

[9] Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. Manhattan room layout reconstruction from a single 360 image: A comparative study of state-of-the-art methods. *International Journal of Computer Vision*, 129(5):1410–1431, 2021. 1, 2, 5, 6



(a) Our result (b) DOP-Net [5] (c) LGT-Net [3]

Figure 2. **Qualitative Results** for Panoramic Images from PanoContext [8] dataset. Red lines denote the ground truth layout. Cyan lines denote the predicted layout.

(a) Our result (b) DOP-Net [5] (c) LGT-Net [3]

Figure 3. **Qualitative Results** for Panoramic Images from Stanford 2D-3D [1] dataset. Red lines denote the ground truth layout. Cyan lines denote the predicted layout.

(a) Our result (b) DOP-Net [5] (c) LGT-Net [3]

Figure 4. **Qualitative Results** for Panoramic Images from MatterportLayout [9] dataset. Red lines denote the ground truth layout. Cyan lines denote the predicted layout.

Figure 5. **3D visualization** for Panoramic Images from PanoContext [8] dataset. Red lines denote the ground truth layout. Cyan lines denote the predicted layout.

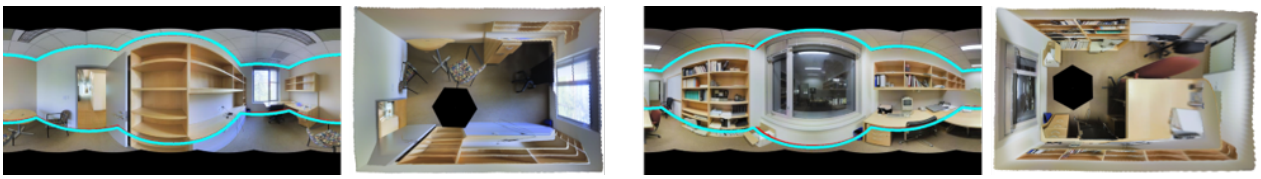


Figure 6. **3D visualization** for Panoramic Images from Stanford 2D-3D [1] dataset. Red lines denote the ground truth layout. Cyan lines denote the predicted layout.

Figure 7. **3D visualization** for Panoramic Images from MatterportLayout [9] dataset. Red lines denote the ground truth layout. Cyan lines denote the predicted layout.

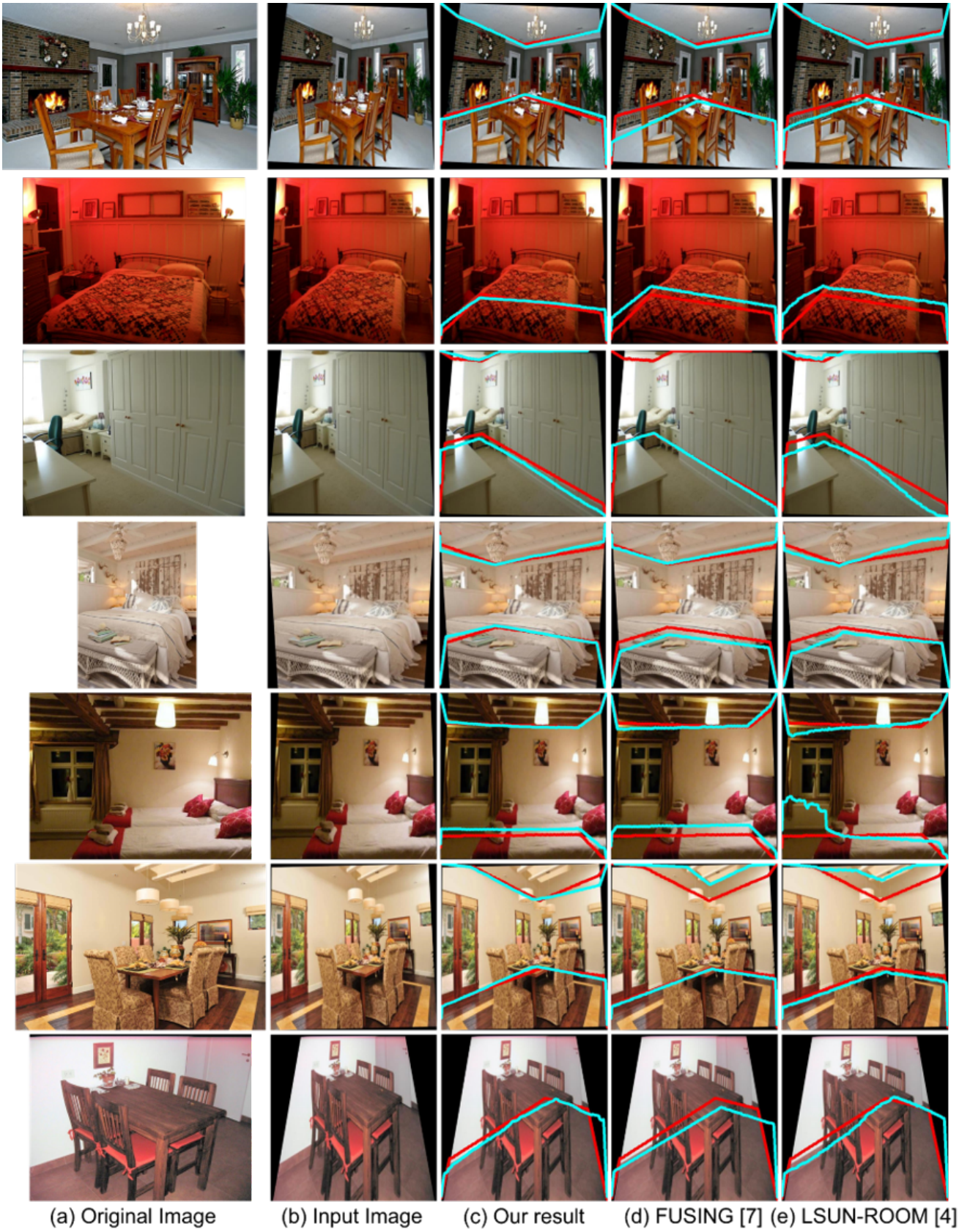


Figure 8. **Qualitative Results** for Perspective Images. Red lines denote the ground truth layout. Cyan lines denote the predicted layout.