# Self-training Room Layout Estimation via Geometry-aware Ray-casting (Supplementary Material)

Bolivar Solarte[1,2], Chin-Hsuan Wu[1], Jin-Cheng Jhang[1], Jonathan Lee[1], Yi-Hsuan Tsai[3], and Min Sun[1]

[1] National Tsing Hua University, Taiwan
[2] Industrial Technology Research Institute ITRI, Taiwan
[3] Google

In this document, we provide additional information that complements our main manuscript, serving as supplementary material.

## 1 Projection Function $\pi(\cdot)$

The projection function $\pi(\cdot)$, as presented in Eq. 2 in our main manuscript, stands as the camera-model projection function that projects the outputs' model into Euclidean space. This projection can defined as follows:

$$\mathbf{y}_i = \pi(f_\Theta(I_i), \mathbf{T}_i), \tag{1}$$

Where, $f_\Theta(I_i)$ is the estimation of a model with $\Theta$ parameters, $I_i$ is the input panoramic image, and $\mathbf{T}_i$ is the corresponding camera pose. Note that $\pi(\cdot)$ function is defined differently for each layout model since each layout model has different geometry representations.

In the case of HorizonNet [5], the projection function presented in Eq. (1) can be described as follows:

$$f_\Theta(I_j) = \{(\phi_i, \theta_i)\}_{i=1:W}, \tag{2}$$

$$\mathcal{B} = \{\mathbf{b_i}\} = \left\{ \begin{bmatrix} h_j \cot\phi_i \sin\theta_i \\ h_j \\ h_j \cot\phi_i \cos\theta_i \end{bmatrix} \right\}_{i=1:W}, \tag{3}$$

$$\mathbf{y}_j = \text{concat}(\{\mathbf{R}_j\mathbf{b}_i + \mathbf{t}_j\}_{i=1:W}), \quad \mathbf{y}_j \in \mathbb{R}^{3\times W} \tag{4}$$

where $(\phi_i, \theta_i)$ represents the layout estimation parametrized in spherical coordinates, $\mathcal{B}$ stands as a set of Euclidean points $\mathbf{b}_i$ computed from each pair $(\phi_i, \theta_i)$ with $h_j$ as the camera height to the floor, and $\mathbf{R}_j \in SO(3)$ with $\mathbf{t}_j \in \mathbb{R}^3$ as he rigid transformation in world coordinates.

In the case of LGTNet, the layout prediction as a set of horizon-depth values can be presented as follows:

$$f_\Theta(I_j) = \{d_i\}_{i=1:W}, \tag{5}$$

$$\mathcal{R} = \{\mathbf{r}_i\}_{i=1:W} \ , \quad \mathbf{r}_i \in \mathbb{R}^3 \ , \quad |\mathbf{r}_i| = 1, \tag{6}$$

$$\mathbf{y}_j = \text{concat}(\{d_i \mathbf{R}_j \mathbf{r_i} + \mathbf{t}_j\}_{i=1:W}) + [0, h_j, 0]^\top, \tag{7}$$

where $\{d_i\}$ is a set of scalar values that represent the distance from the camera view to the layout geometry, $\mathcal{R}$ is a set of ray vectors $\mathbf{r_i}$ in BEV, $h_j$ is the camera height to the floor, and $\mathbf{R}_j \in SO(3)$ with $\mathbf{t}_j \in \mathbb{R}^3$ represent the rigid transformation in world coordinates.

## 2 Impact of the Layout Model Backbone

In this section, we present evidence that our proposed self-training framework can effectively leverage a better backbone model unlike the baseline 360-MLC [4].

To be specific, when adopting a stronger model for room layout estimation, e.g., LGTNet [2], the pseudo-label errors computed by 360-MLC produce less desirable estimations than the ones using HorizonNet [5], particularly for occluded geometries. This evidence can be verified in Tab. 2 and Tab. 3 in our main manuscript, where 360-MLC shows a drastic drop in performance in the occlusion subset when using the LGTNet backbone, e.g., from 79.19% 2D IoU using HorizonNet (Tab. 2, row 2, column 6) to 71.29% in the case of using LGTNet (Tab. 3, row 2, column 5).

To further corroborate this evidence, in Fig. 1, we present qualitative results of 360-MLC using both proposed backbone models, HorizonNet and LGTNet. Notably, we observe that 360-MLC drastically decreases performance when using the LGTNet model. In contrast, our proposed self-training framework effectively leverage the stronger LGTNet model showing substantial difference to the baseline and corroborating our finding in Tab. 2 and Tab. 3 in our main manuscript. Based on these results, we assert that our proposed formulation can generalize well and achieve consistently better results when using a stronger base model in compassion to the 360-MLC [4] approach.

## 3 Cross-domain Evaluation

**Table 1:** Cross-domain Evaluation Experiment.

| | Testing set | | Occlusion Subset | |
|---|---|---|---|---|
| Method | 2D IoU ↑ | 3D IoU ↑ | 2D IoU ↑ | 3D IoU ↑ |
| pre-trained on ZIND | 67.66 | 63.95 | 70.75 | 69.38 |
| Ours | **73.55** | **70.01** | **78.45** | **77.20** |

In Tab. 1, we evaluate our solution under a cross-domain setting by using a pre-trained HorizonNet [5] on the ZInD dataset [1] to create pseudo-labels and

then self-train on HM3D-MVL. We find that our proposed strategy still improves over the pre-trained model and reduces domain gaps despite the severe difference in the data domain.

## 4 Evaluations on Occluded Geometries

To evaluate our claim that our solution particularly addresses occluded geometries, we compute 2D-IoU only for occluded regions per image view. Results show that our solution significantly outperforms the baseline 360-MLC [4] across all datasets. These results are presented in Tab. 2.

**Table 2:** Evaluation on Occluded Geometries.

| | Only on occluded regions - 2D IoU (%) | | |
|---|---|---|---|
| Method | HM3D-MVL | MP3D-FPE | ZInD |
| 360-MLC [17] | 76.03 | 75.97 | 74.88 |
| Ours | **84.31** | **83.18** | **79.63** |

## 5 Extended Ablation Study

We extend the ablation study presented in our manuscript by evaluating several pseudo-labels using different $\delta_r$ values in the HM3D-MVL testing split. Results show that our ray-casting is robust to a large $\delta_r$ range using the median function as a sampler. We argue that sampling along several positions and directions across the scene is the key component to handling the influence of the occluded casting points. Results are presented in Tab. 3.

**Table 3:** Ablation Study over $\delta_r$ hyper-parameter.

| | 2D IoU (%) | | | |
|---|---|---|---|---|
| Sampler | $\delta_r = 10$ | $\delta_r = 20$ | $\delta_r = 30$ | $\delta_r = 50$ |
| Mean | 77.39 | 78.39 | 78.11 | 78.07 |
| Median | **81.04** | **82.76** | **82.16** | **82.16** |

Additionally, to complement the ablation study presented in Table 4 of our manuscript, we present results using 100% of the data with two backbone models. This ablation validates the effectiveness of the proposed components. Results are presented in Tab. 4.

**Table 4:** Extended Ablation Study.

| Method | Testing set 2D IoU ↑ | Testing set 3D IoU ↑ | Occlusion Subset 2D IoU ↑ | Occlusion Subset 3D IoU ↑ |
|---|---|---|---|---|
| HorizonNet [5] - 100% data | | | | |
| pre-trained | 76.71 | 71.79 | 78.74 | 75.72 |
| Pseudo-labels | 82.72 | 77.80 | 81.78 | 79.84 |
| $\omega = \sigma^{-1}$ | 81.98 | 77.50 | 81.67 | 79.95 |
| $\omega = Eq.(11)$ | **82.99** | **78.95** | **83.01** | **81.38** |
| LGTNet [2] - 100% data | | | | |
| pre-trained | 78.90 | 74.04 | 80.22 | 78.10 |
| Pseudo-labels | 84.11 | 79.65 | 81.96 | 80.43 |
| $\omega = \sigma^{-1}$ | 83.10 | 78.43 | 82.01 | 81.87 |
| $\omega = Eq.(11)$ | **86.49** | **81.90** | **83.75** | **82.06** |

## 6 Layout Results on Panoramic View

To further corroborate the effectiveness of our proposed self-training framework, we present in Fig. 2 and Fig. 3 qualitative results on MP3D-FPE [3] and HM3D-MVL datasets, respectively. These results complement the results in Fig. 7 presented in our main manuscript. Moreover, we show evaluations using both backbone models, i.e., HorizonNet [5] in panel (a), and LGTNet [2] in panel (b).

Based on the results of these experiments, we find evidence that the proposed self-training framework with ray-casting pseudo labels significantly outperforms 360-MLC in both datasets using both backbone models.

## 7 Pseudo-labels on BEV

In this experiment, we aim to validate our claim that the proposed multi-cycle ray-casting pseudo-labeling is capable of handling complex geometry scenes. For this purpose, we significantly extend the qualitative results presented in Fig. 5 in our main manuscript by projecting pseudo-label geometries in BEV of several complex scenes.

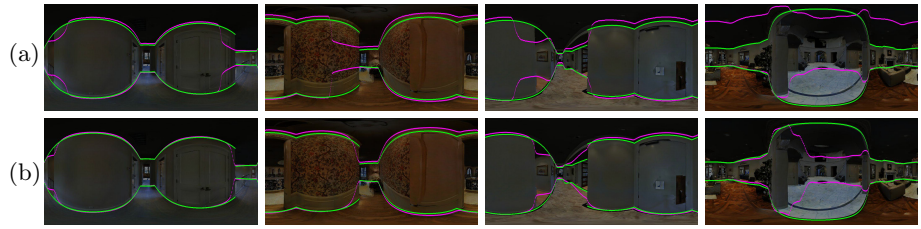
The results are shown in Fig. 4, where the first column displays layout estimates from a pre-trained HorizonNet [5] model, the second column shows all pseudo-labels computed by 360-MLC [4], the third column depicts all pseudo-labels of our proposed solution using the HorizonNet backbone, the fourth column shows ours pseudo-labels of using the LGTNet backbone, and the last column serves as a reference with the point cloud of the scene.

Based on these results, we can assert that our proposed ray-casting pseudo-labeling significantly outputs a less noisy room geometry compared with 360-MLC, showing consistency along all pseudo-labels in the scene. Furthermore,

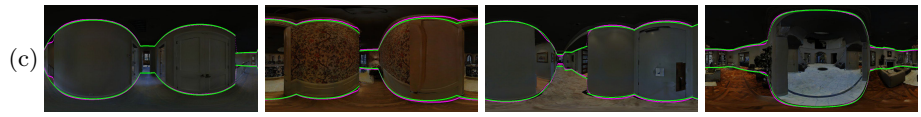
our proposed solution demonstrates proficiency in handling complex geometries, showing the versatility of our contribution.

## References

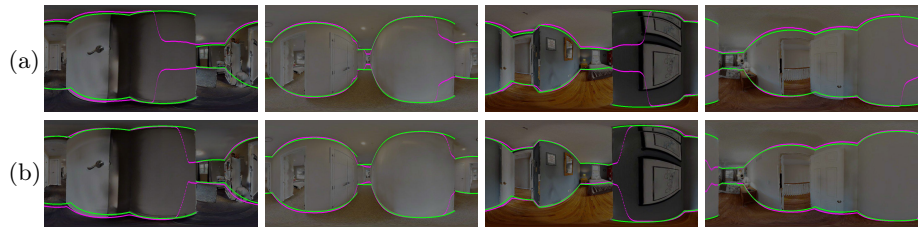

1. Cruz, S., Hutchcroft, W., Li, Y., Khosravan, N., Boyadzhiev, I., Kang, S.B.: Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3d room layouts. In: CVPR (2021) 2
2. Jiang, Z., Xiang, Z., Xu, J., Zhao, M.: Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In: CVPR (2022) 2, 4, 6, 7, 8, 9
3. Solarte, B., Liu, Y.C., Wu, C.H., Tsai, Y.H., Sun, M.: 360-dfpe: Leveraging monocular 360-layouts for direct floor plan estimation. IEEE Robotics and Automation Letters **7**(3), 6503–6510 (2022) 4, 6
4. Solarte, B., Wu, C.H., Liu, Y.C., Tsai, Y.H., Sun, M.: 360-mlc: Multi-view layout consistency for self-training and hyper-parameter tuning. In: NeurIPS (2022) 2, 3, 4, 6, 7, 8, 9
5. Sun, C., Hsiao, C., Sun, M., Chen, H.: Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In: CVPR (2019) 1, 2, 4, 6, 7, 8, 9




360-MLC [4] with different backbone models in the MP3D-FPE [3] dataset.




Ours using LGTNet [2] backbone model in the MP3D-FPE [3] dataset.

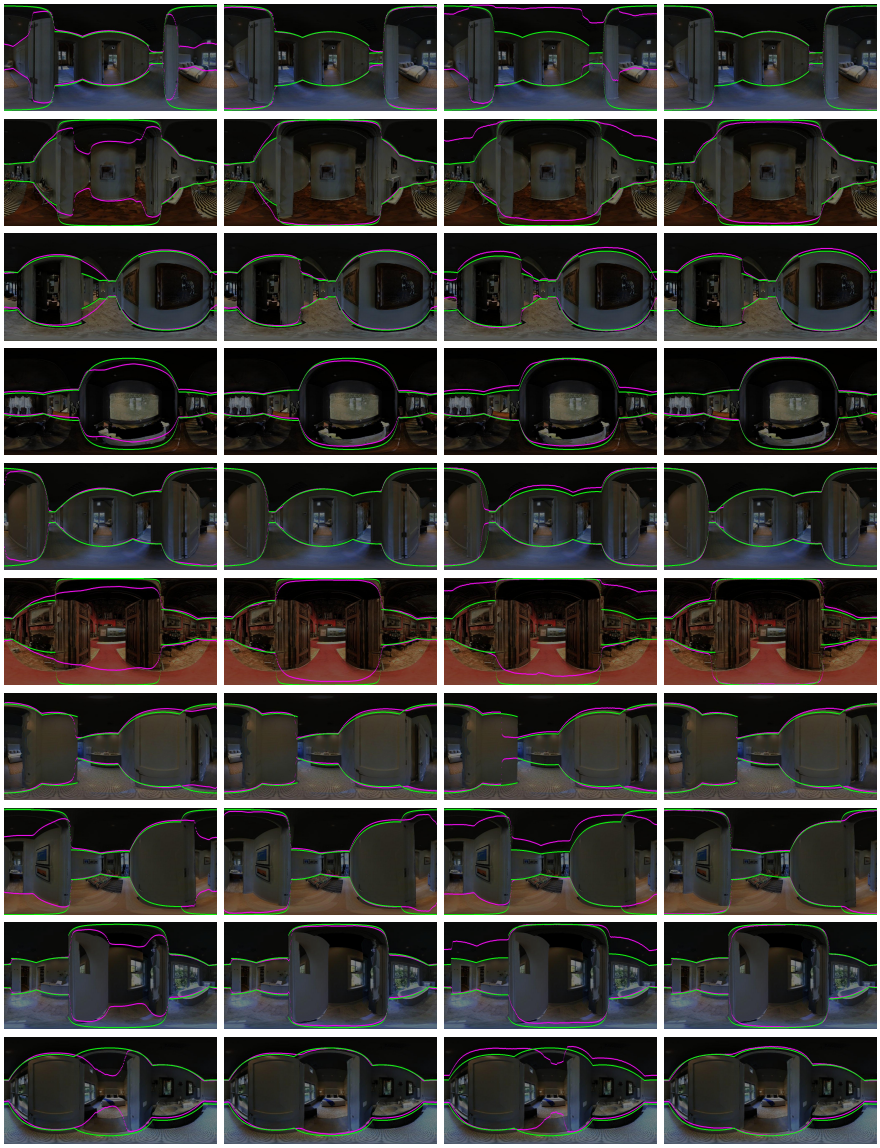



360-MLC [4] with different backbone models in our HM3D-MVL dataset.


Ours using LGTNet [2] backbone model in our HM3D-MVL dataset.

**Fig. 1: Impact of a strong backbone model.** We present qualitative results of self-training using HorizonNet [5] and LGTNet [2] backbone models. Estimations are presented in magenta, while the ground truth is shown in green. In panels (a) and (b), we present the baseline 360-MLC [4] using LGTNet [2] and HorizonNet [5] models, respectively. In panel (c), we show the results of our proposed self-training using LGTNet [2]. Note that a strong backbone model like LGTNet hurts 360-MLC performance significantly. Meanwhile, our proposed self-training framework is capable of exploiting the benefits of a robust backbone model.

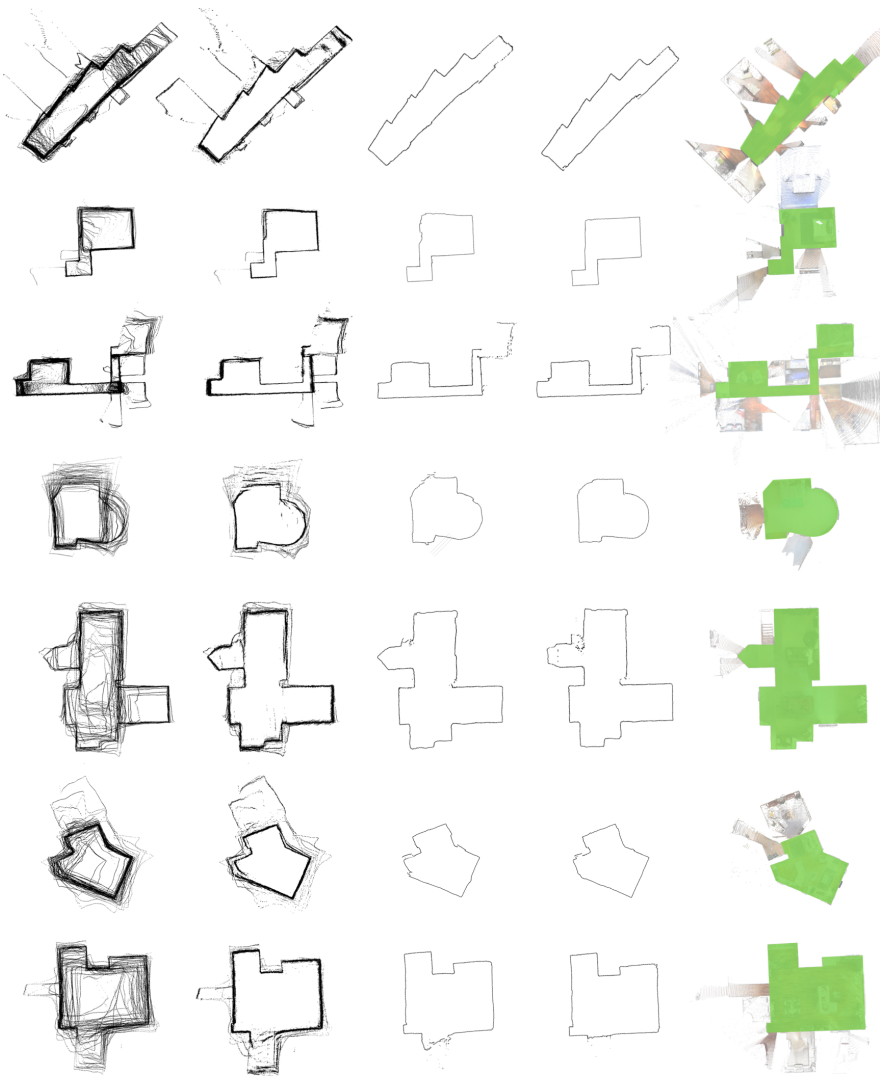
(a) with HorizonNet [5] backbone (b) with LGTNet [2] backbone



360-MLC [4] Ours 360-MLC Ours

**Fig. 2: Qualitative results on the MP3D-FPE dataset.** We present results comparing 360-MLC [4] and our self-training formulation using different backbone models. Panel (a) is with HorizonNet [5], and panel (b) is with the LGTNet [2] backbone. The greed line represents the ground truth label, while the magenta line is the estimated layout after self-training.


**Fig. 3: Qualitative results on the HM3D-MVL dataset.** We present results comparing 360-MLC and our self-training framework using different backbone models. Panel (a) shows results using HorizonNet. Panel (b) shows results with the LGTNet backbone. The greed line represents the ground truth label, while the magenta line is the estimated layout after self-training.



HorizonNet [5] 360-MLC [4] Ours (a) Ours (b) Point cloud

**Fig. 4: Qualitative comparisons of estimated pseudo-labels.** We present a qualitative visualization of layout geometries projected in BEV. In the first column, we show all layouts estimated from a pre-trained HorizonNet [5] model. In the second column, all pseudo-labels from the baseline 360-MLC [4] are visualized. In the third and fourth columns, we present a visualization of our ray-casting pseudo-labels, using HorizonNet [5] and LGTNet [2] backbones in (a) and (b), respectively. In the last column, we present the point cloud of the scene, highlighting the room scene in green for reference purposes. Note that our pseudo-labels do not only present less noisy geometries than 360-MLC but also is capable of defining complex and not trivial room geometries.