Supplementary Materials for Structured3D: A Large Photo-realistic Dataset for Structured 3D Modeling

Jia Zheng[1,2*†], Junfei Zhang[1*], Jing Li[2], Rui Tang[1],
Shenghua Gao[2,3], and Zihan Zhou[4]

[1] KooLab, Kujiale.com
[2] ShanghaiTech University
[3] Shanghai Engineering Research Center of Intelligent Vision and Imaging
[4] The Pennsylvania State University
https://structured3d-dataset.org

1 Additional Details about the Dataset

Dataset Statistics. Fig. 1 reports the statistics for the number of rooms per scene, room types, and instances with 3D bounding box annotations for the top 20 categories. We also report the proportion of large (area $\geqslant 96^2$), medium ($32^2 \leqslant$ area $< 96^2$) and small objects (area $< 32^2$) in each category. The area is measured as the number of pixels in the segmentation mask.

More Annotations. Based on the "primitive+relationship" representation, we can generate various types of structures, such as wireframe, planes, and floorplan. Some examples are shown in Fig. 4.

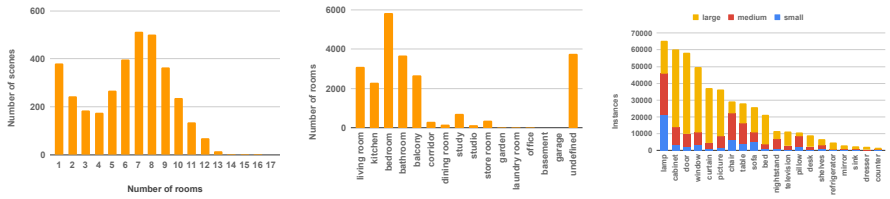
Fig. 5 and Fig. 6 show more layout annotations of panoramic and perspective images in our dataset, respectively. The first two rows show cuboid layouts. The layouts become more complicated from top to bottom in Fig. 5. Note that all ground truth 3D annotations are automatically extracted from the original house design files.

Our dataset also provides 3D bounding box annotations in all scenes. Some examples are shown in Fig. 7.

2 Impact of Rendering Quality

To illustrate the importance of photo-realistic rendering, we have conducted additional experiments to investigate how rendering quality affects room layout estimation accuracy. In this experiment, we render low-quality images by (i) disabling global illumination, (ii) using only ambient light, and (iii) for BSDF, only keeping the diffuse reflection. We show the qualitative comparison between high-quality and low-quality rendering methods in Fig. 2. Table 1 shows the

*: Equal contribution.
†: The work was partially done when Jia Zheng interned at KooLab, Kujiale.com.



(a) (b) (c)

Fig. 1. Statistics of semantic annotation in our dataset. (a) The number of scenes w.r.t. different room numbers. (b) The number of rooms w.r.t. different room types. (c) The number of instances w.r.t. the top 20 object categories.

Table 1. Quantitative evaluation of the room layout estimation with different rendering qualities.

Methods	Rendering quality	Config.	PanoContext			2D-3D-S		
			3D IoU(%)↑	CE(%)↓	PE(%)↓	3D IoU(%)↑	CE(%)↓	PE(%)↓
LayoutNet	low	s	67.92	1.64	5.54	**58.15**	**1.73**	**6.10**
		s → r	84.59	0.67	2.06	83.46	0.75	2.46
	high	s	**75.64**	**1.31**	**4.10**	57.18	2.28	7.55
		s → r	**84.77**	**0.63**	**1.89**	**84.04**	**0.66**	**2.08**
HorizonNet	low	s	70.35	1.89	5.22	64.75	1.22	4.50
		s → r	83.00	0.75	2.09	85.21	0.82	2.16
	high	s	**75.89**	**1.13**	**3.15**	**67.66**	**1.18**	**3.94**
		s → r	**85.27**	**0.66**	**1.86**	**86.01**	**0.61**	**1.84**

quantitative results. As expected, using low-quality images generally leads to degraded performance.

3 Experiments on Room Layout Estimation

3.1 Implementation Details

LayoutNet. We use LayoutNet v2 [11,12] with ResNet-34 as the backbone. Instead of following the step-by-step training procedure in [11,12], we directly train the whole network jointly with a large batch size, which also leads to a comparable result. We use Adam optimizer [3] with learning rate 10^{-3}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We use a mini-batch size of 16. For "s + r", each batch contains 8 images from the real dataset and 4 from the synthetic dataset. We train the whole network on the synthetic dataset for 14k iterations (20 epochs of the synthetic dataset) and fine-tune the network on the real dataset for 5k iterations (100 epochs of the real dataset).

HorizonNet. Following the [6], we use Adam optimizer [3] with learning rate 3×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We use a mini-batch size of 16. For "s + r", each batch contains 16 images from the real dataset and 8 from the synthetic dataset. We train the whole network on the synthetic dataset for 14k iterations (20 epochs of the synthetic dataset), and fine-tune the network on the real dataset for 15k iterations (300 epochs of the real dataset).

Fig. 2. Qualitative comparison of different rendering qualities. **Odd rows**: high-quality images. **Even rows**: low-quality images.

Domain adaptation. We use a PatchGAN [2] for discriminator and LSGAN [4] as the adversarial learning objective function. We use Adam optimizer [3] with learning rate 1×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Each batch contains 4 images from the real dataset and 4 from the synthetic dataset. We train the whole network on the synthetic dataset for 10k iterations. In the training, the weights for depth estimation network and the discriminator network are set as 0.1, 0.001.

We train all our model on 4 NVIDIA GeForce GTX TITAN X GPUs with 12GB GPU memory.

3.2 Qualitative Results

We show qualitative results of LayoutNet [11,12] on PanoContext [9] and 2D-3D-S [1] in Fig. 8 and Fig. 9, respectively. As one can see, by first training on our synthetic data and then fine-tuning on real images, the network generates more accurate estimations than only using synthetic data or real images for training (*e.g.*, see the last column of Fig. 8 and the last column of Fig. 9).

We show qualitative results of HorizonNet [6] on the two datasets in Fig. 10 and Fig. 11, respectively. Similar to the case of LayoutNet, the network is able to generate better layout estimations when trained on both synthetic data and real images.

Finally, we show visual results for the domain adaptation experiment in Fig. 12 and Fig. 13.

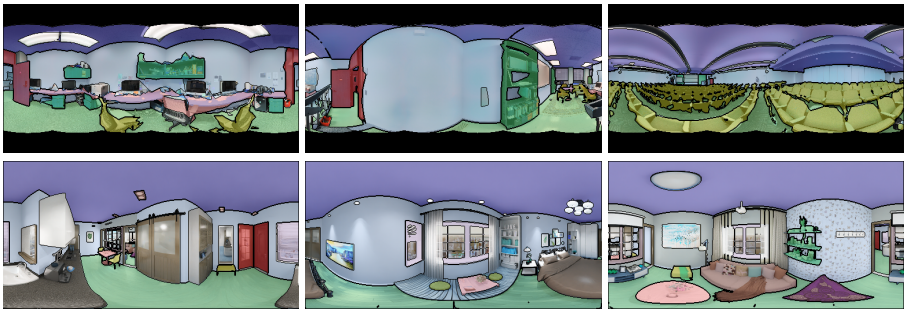


Fig. 3. Comparison of semantic annotation between 2D-3D-3S and Structured3D datasets. **First row**: 2D-3D-S dataset. **Second row**: Structured3D dataset. Different colors indicate different semantic categories.

4 Experiments on Semantic Segmentation

4.1 Experiment Setup

Dataset. In this experiment, we use 2D-3D-S [1] as the real dataset. We split the images into 955 for training, 84 for validation, and 373 for testing. Then, we select a subset of panoramas in our Structured3D Dataset with the original lighting and full configuration. Each panorama corresponds to a different room in our dataset. We divide our synthetic dataset at the scene level into train/val/test which contains 3000/250/250 scenes and 18362/1776/1697 images.
Evaluation metrics. We adopt three standard metrics: i) Mean IoU: intersection over union between the predicted and ground truth pixels, average over all semantic classes; ii) Pixel Accuracy: the proportion of the correctly predicted pixels; iii) Boundary Accuracy [5]: F-measure along the boundary of the predicted and ground truth pixels. We do not use boundary accuracy metric on the real dataset, since the annotation in the 2D-3D-S is not well aligned with the semantic boundary, as shown in Fig. 3.
Methods for comparison. In this experiment, we compare PSPNet [10] with dilated ResNet-50 as the backbone, UPerNet [8] with ResNet-50 as the backbone, and HRNet [7].
Implementation details. We use SGD optimizer with initial learning rate 2×10^{-2} with polynomial decay policy, momentum 0.9, and weight decay 10^{-4}. We set the mini-batch size to 8. For "s + r", each batch contains 4 images from the real dataset and 4 from the synthetic dataset. We train the whole network on the synthetic dataset for 10k iterations and fine-tune the network on the real dataset for 10k iterations. During training and testing, we resize images to 512×1024.

4.2 Experiment Results

Evaluation on Structured3D dataset. The results on the test set of the Structured3D are shown in Table 2. Since the synthetic dataset contains accurate

Table 2. Quantitative evaluation of the semantic segmentation on Structured3D dataset. The best and the second best results are boldfaced and underlined, respectively.

Methods	Mean IoU (%) ↑	Pixel Accuracy (%) ↑	Boundary Accuracy (%) ↑
PSPNet [10]	30.10	77.08	71.29
UPerNet [8]	<u>32.64</u>	<u>82.30</u>	<u>75.89</u>
HRNet [7]	**37.77**	**83.94**	**80.84**

Table 3. Quantitative evaluation of the semantic segmentation under different training schemes. The best and the second best results are boldfaced and underlined, respectively.

Methods	Config.	Mean IoU (%) ↑	Pixel Accuracy (%) ↑
PSPNet [10]	s	26.13	66.26
	r	<u>47.08</u>	80.02
	s → r	46.06	<u>80.85</u>
	s + r	**49.71**	**82.05**
UPerNet [8]	s	28.75	65.49
	r	45.09	<u>81.07</u>
	s → r	<u>46.14</u>	80.87
	s + r	**49.60**	**82.62**
HRNet [7]	s	37.92	73.38
	r	48.92	<u>82.29</u>
	s → r	<u>49.15</u>	81.37
	s + r	**52.00**	**84.12**

pixel-wise annotations, we further evaluate the boundary accuracy. Qualitative results are shown in Fig. 14.

We also report the IoU per category in Table 4. As one can see, the per-class score is strongly correlated with the number and size of instances in each class.

Augmenting real datasets. Since the label set of the 2D-3D-S dataset is not consistent with our Structured3D dataset, we select nine overlapping categories for evaluation: wall, floor, chair, sofa, door, window, bookcase, ceiling, and table. We follow the four training strategies in the "augmenting real datasets" experiment: "s", "r", "s → *r*" and "s + r". The results are shown in Table 3. As expected, augmenting real datasets with our synthetic data boosts the performance of all methods. We also show qualitative results of PSPNet, UPerNet and HRNet on 2D-3D-S dataset in Fig. 15, Fig. 16 and Fig. 17, respectively.

References

1. Armeni, I., Sax, A., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. CoRR **abs/1702.01105** (2017) 3, 4, 12, 13, 14, 15, 16
2. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134 (2017) 3
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) 2, 3
4. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV. pp. 2794–2802 (2017) 3

Table 4. Quantitative evaluation of the semantic segmentation on Structured3D dataset.

Category	PSPNet [10]	UPerNet [8]	HRNet [7]
wall	72.60	80.56	**84.36**
floor	81.27	81.77	**88.53**
cabinet	35.39	47.60	**52.78**
bed	48.20	**54.31**	25.99
chair	11.95	15.11	**15.30**
sofa	28.06	**41.62**	39.09
table	25.27	15.52	**29.52**
door	55.82	64.70	**69.03**
window	52.80	59.38	**62.15**
picture	31.56	42.13	**47.72**
counter	0.01	0.00	**0.36**
desk	14.75	4.67	**24.21**
shelves	**3.42**	0.62	2.56
curtain	56.79	51.68	**68.40**
dresser	0.00	**1.26**	1.04
pillow	6.82	**8.80**	8.55
mirror	6.40	12.10	**12.74**
ceiling	84.94	90.52	**92.55**
refrigerator	13.62	9.43	**37.10**
television	45.16	56.37	**63.30**
box	0.00	0.00	0.00
nightstand	26.03	35.22	**38.70**
toilet	34.16	21.60	**42.66**
sink	1.47	0.87	**4.36**
lamp	33.91	46.06	**47.26**
bathtub	32.59	18.53	**39.31**
otherstructure	31.53	39.97	**47.37**
otherfurniture	12.87	17.36	**19.49**
otherprop	25.57	28.81	**30.95**
mean	30.10	32.64	37.77

5. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR. pp. 724–732 (2016) 4
6. Sun, C., Hsiao, C.W., Sun, M., Chen, H.T.: Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In: CVPR. pp. 1047–1056 (2019) 2, 3, 13
7. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE TPAMI (2020) 4, 5, 6, 15, 16
8. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV. pp. 432–448 (2018) 4, 5, 6, 15, 16
9. Zhang, Y., Song, S., Tan, P., Xiao, J.: Panocontext: A whole-room 3d context model for panoramic scene understanding. In: ECCV. pp. 668–686 (2014) 3, 12, 13, 14
10. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017) 4, 5, 6, 15
11. Zou, C., Colburn, A., Shan, Q., Hoiem, D.: Layoutnet: Reconstructing the 3d room layout from a single RGB image. In: CVPR. pp. 2051–2059 (2018) 2, 3, 12

12. Zou, C., Su, J., Peng, C., Colburn, A., Shan, Q., Wonka, P., Chu, H., Hoiem, D.: 3d manhattan room layout reconstruction from a single 360 image. CoRR **abs/1910.04099** (2019) 2, 3, 12

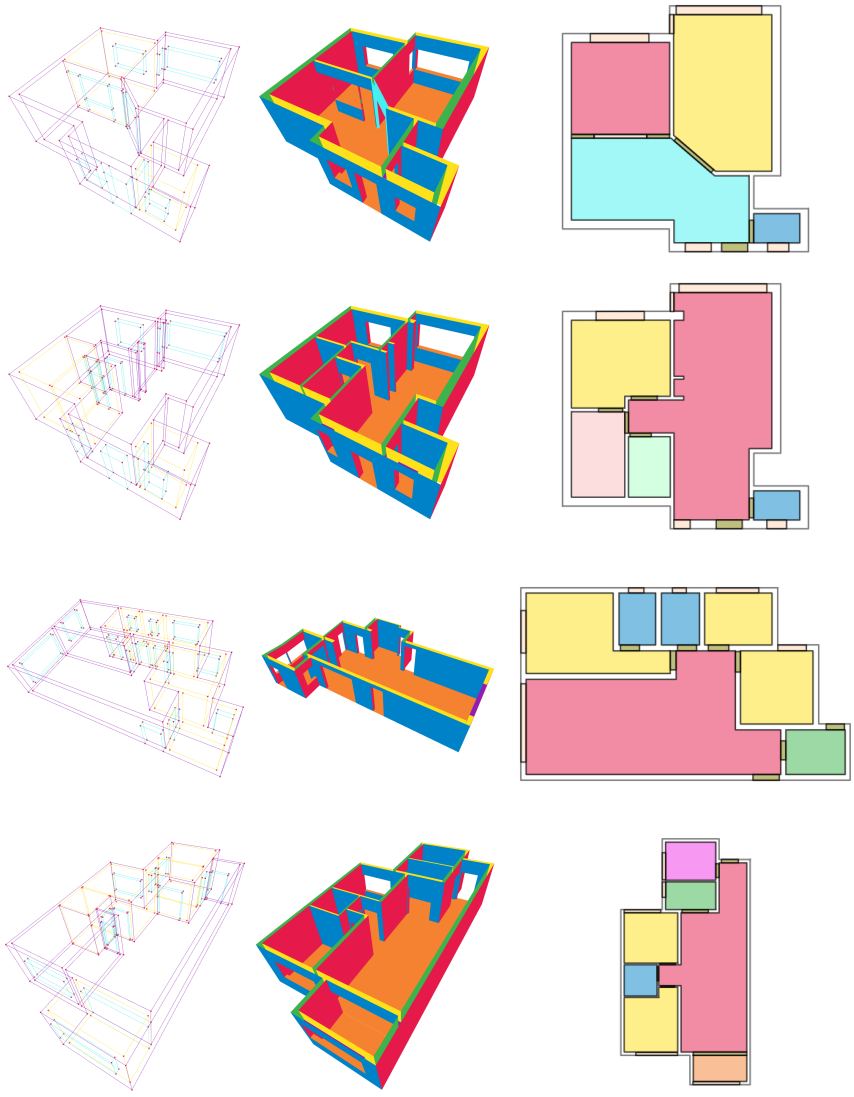


Fig. 4. Example 3D ground truth structures in the Structured3D dataset. In the wireframe, the yellow wireframe denotes the cuboid-shaped room and the blue one denotes the hole (such as window and door), respectively. The planes are colored by the normal. Difference colors denote different room types, windows or doors in the floorplan.

Fig. 5. Example ground truth room layouts for panoramic images in the Structured3D dataset. **From top to bottom**: Simple to complicated cases. The ground truth layouts are drawn as blue lines.

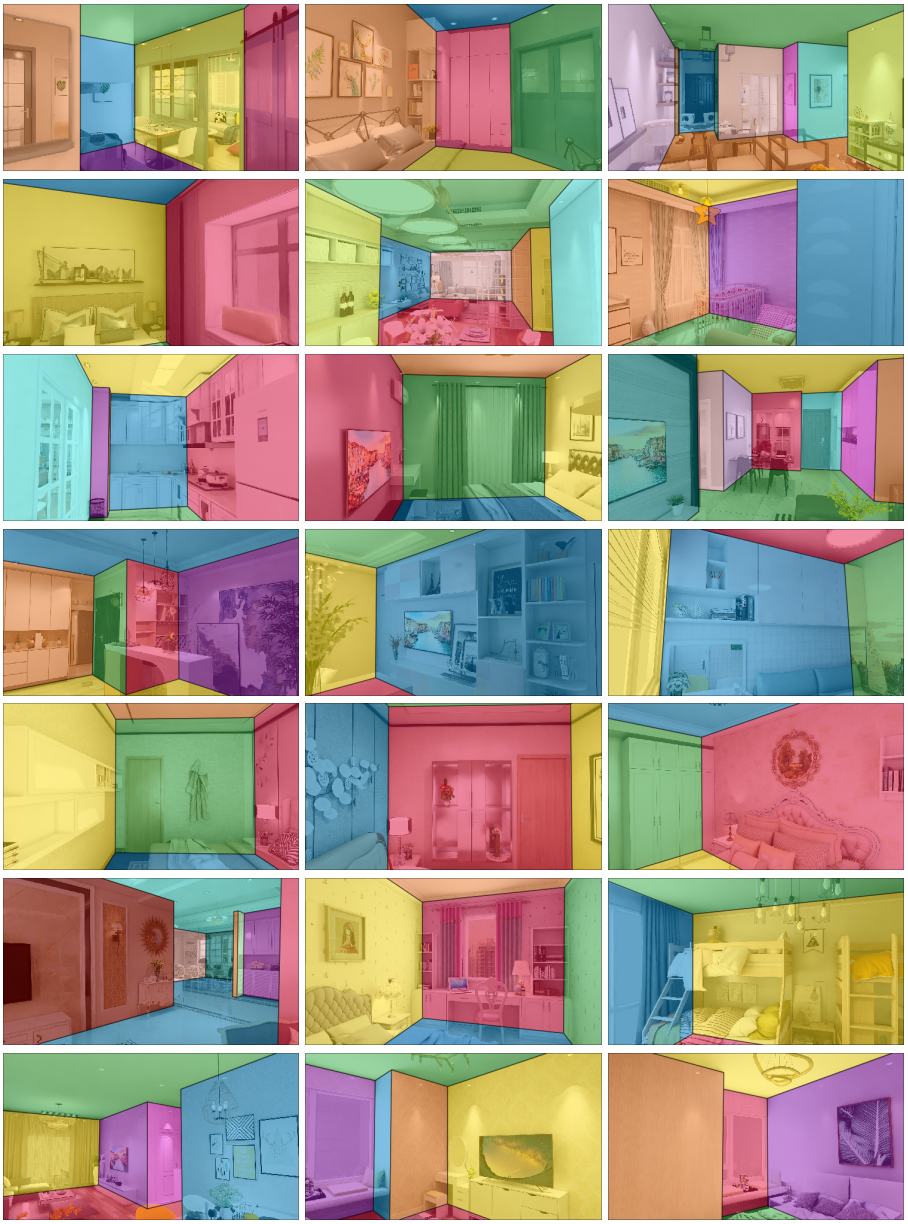


Fig. 6. Example ground truth room layouts for perspective images in the Structured3D dataset. Different colors denote different planes.

Fig. 7. Example ground truth 3D bounding boxes in the Structured3D dataset.

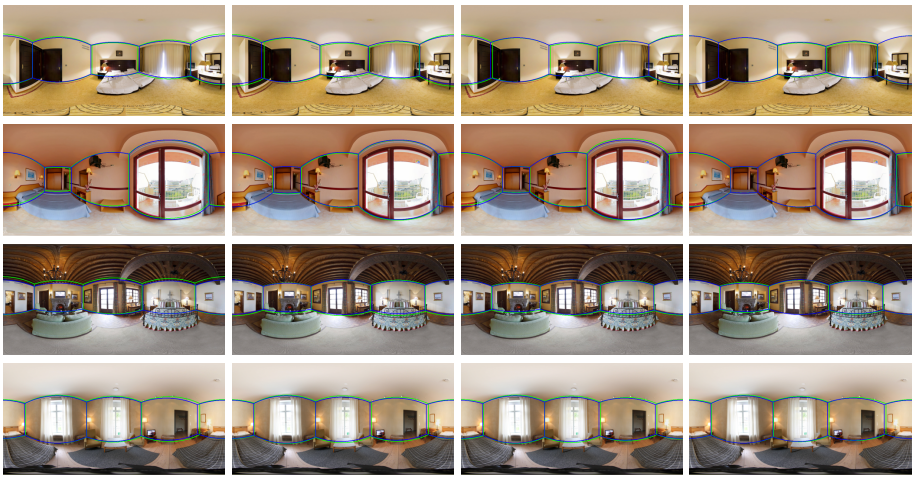


Fig. 8. Qualitative results of LayoutNet [11,12] on the PanoContext dataset [9]. The blue lines are ground truth layout and the green lines are predictions.

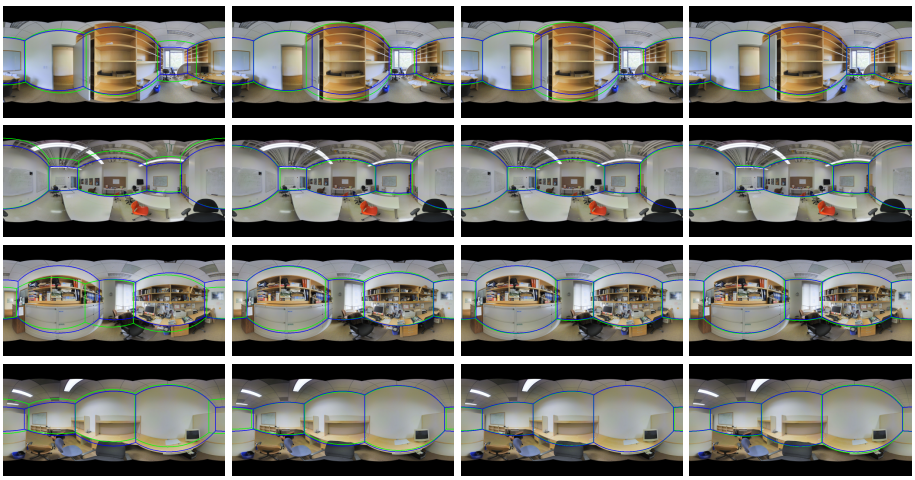


Fig. 9. Qualitative results of LayoutNet [11,12] on the 2D-3D-S dataset [1]. The blue lines are ground truth layout and the green lines are predictions.

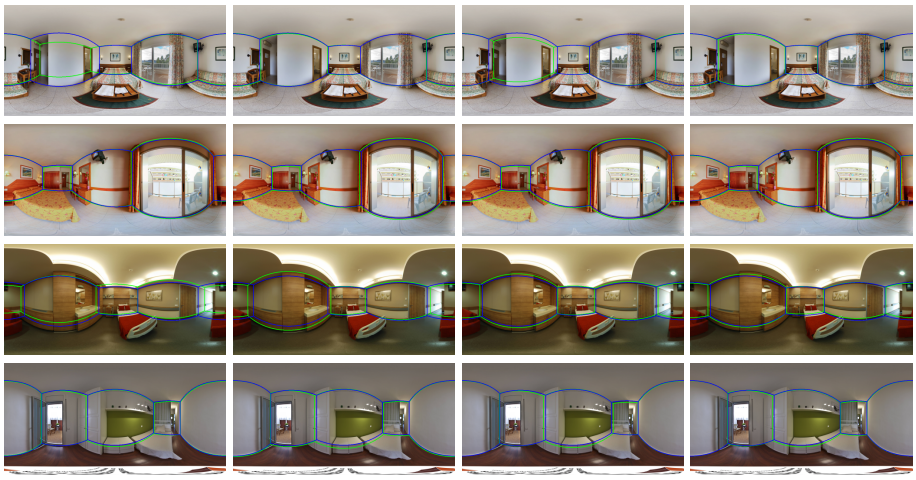


Fig. 10. Qualitative results of HorizonNet [6] on the PanoContext dataset [9]. The blue lines are ground truth layout and the green lines are predictions.

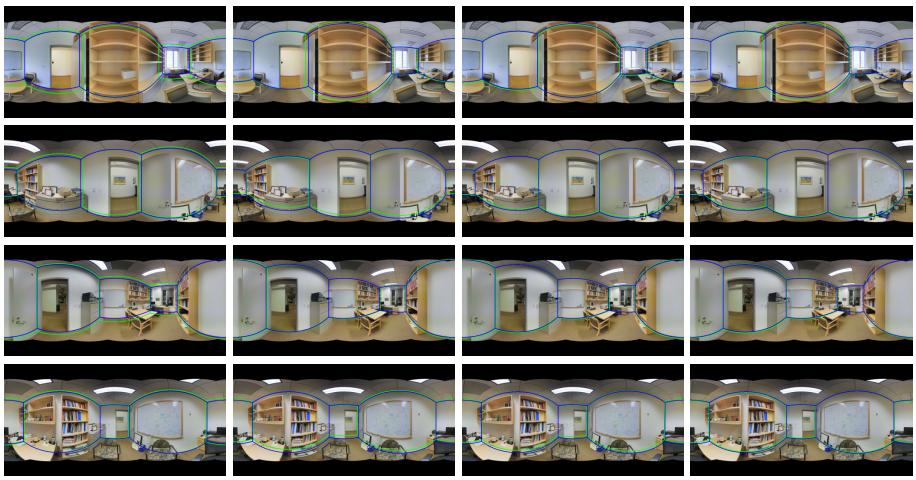


Fig. 11. Qualitative results of HorizonNet [6] on the 2D-3D-S dataset [1]. The blue lines are ground truth layout and the green lines are predictions.

Fig. 12. Domain adaptation results on the PanoContext dataset [9]. NA: non-adaptive baseline. DA: align layout estimation output. DA + Depth: align both layout estimation and depth outputs. Real: train in the target domain. The blue lines are ground truth layout and the green lines are predictions.

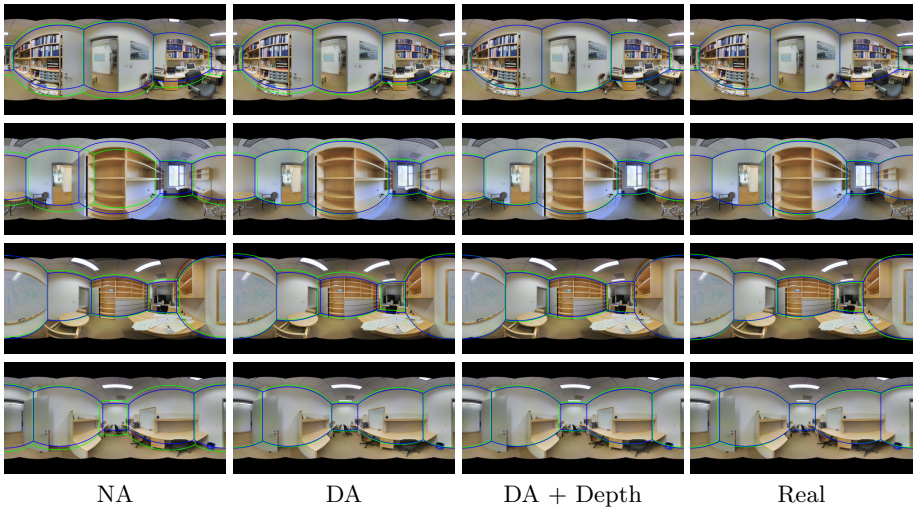


Fig. 13. Domain adaptation results on the 2D-3D-S dataset [1]. NA: non-adaptive baseline. DA: align layout estimation output. DA + Depth: align both layout estimation and depth outputs. Real: train in the target domain. The blue lines are ground truth layout and the green lines are predictions.

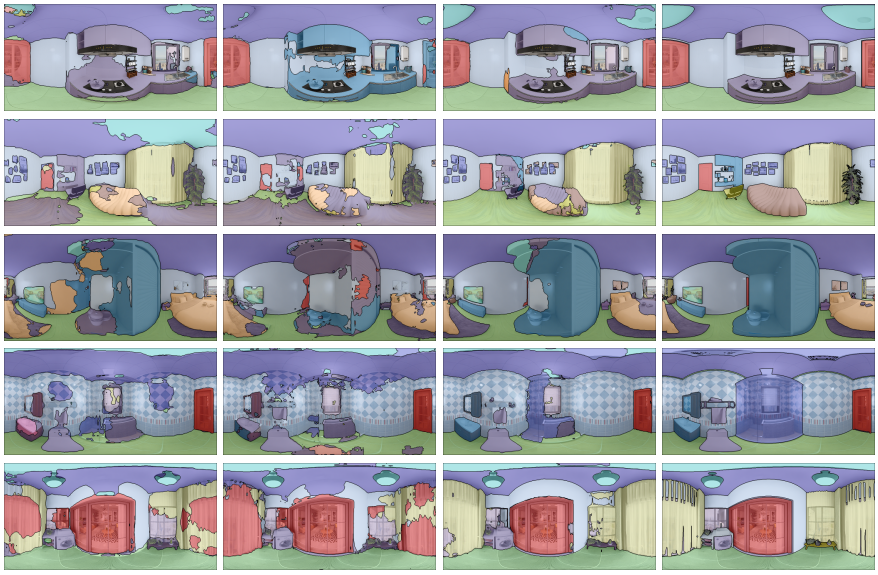


Fig. 14. Semantic segmentation results of PSPNet [10], UPerNet [8] and HRNet [7] on our Structured3D dataset. Difference colors denote different semantic categories.

Fig. 15. Semantic segmentation results of PSPNet [10] on the 2D-3D-S dataset [1]. Difference colors denote different semantic categories.

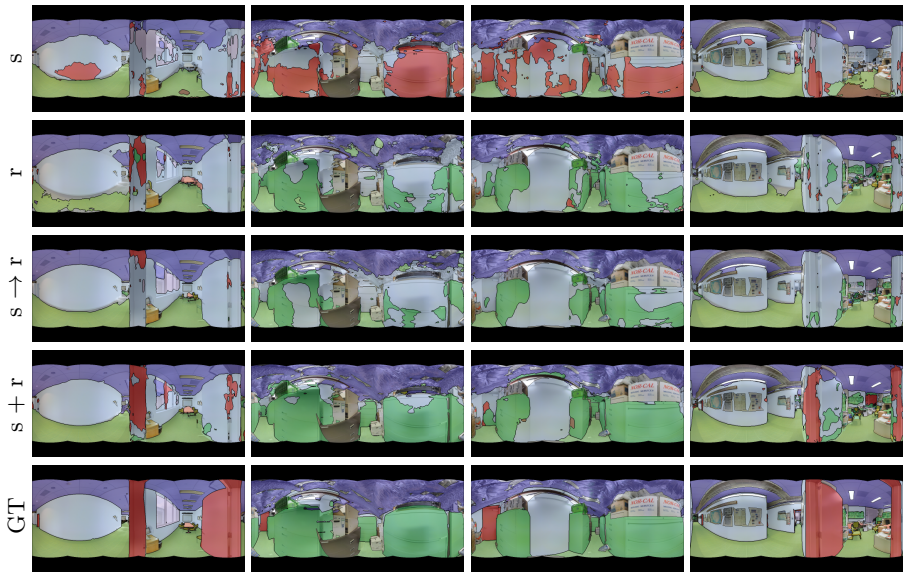


Fig. 16. Semantic segmentation results of UPerNet [8] on the 2D-3D-S dataset [1]. Difference colors denote different semantic categories.

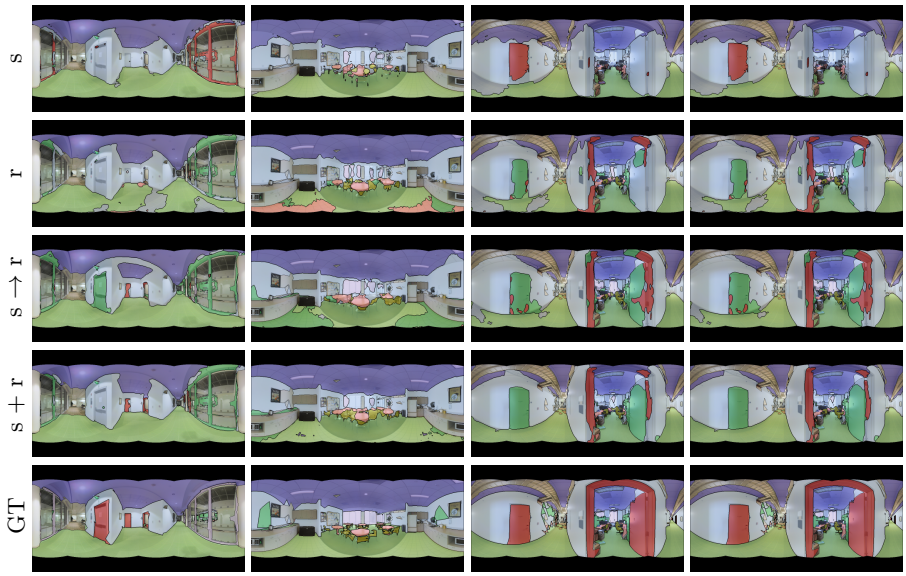


Fig. 17. Semantic segmentation results of HRNet [7] on the 2D-3D-S dataset [1]. Difference colors denote different semantic categories.