360BEV: Panoramic Semantic Mapping for Indoor Bird's-Eye View (Supplementary Materials)

Zhifeng Teng[1,*], Jiaming Zhang[1,*,†], Kailun Yang[2], Kunyu Peng[1],
Hao Shi[3], Simon Reiß[1], Ke Cao[1], Rainer Stiefelhagen[1]
[1]Karlsruhe Institute of Technology, [2]Hunan University, [3]Zhejiang University

1. Data Generation

To perform the data generation, we use an open-source tool[1] to convert the 3D mesh semantic labels in Matterport3D [2] into 194,400 pinhole images with semantic labels. Then, every 18 semantic label pairs are concatenated via a corresponding rotation-translation matrix, yielding 10,800 panoramic semantic ground truth, which is referred to as 360FV-Matterport by us. These panoramic semantic images are originally annotated with 40 object categories. Because many of them are only a small percentage ($\ll$0.1%), we merges some uncommon classes and maintains the 20 most common object categories: `wall`, `floor`, `chair`, `door`, `table`, `picture`, `furniture`, `objects`, `window`, `sofa`, `bed`, `sink`, `stairs`, `ceiling`, `toilet`, `mirror`, `shower`, `bathtub`, `counter`, and `shelving`. For another front-view semantic segmentation dataset, Stanford2D3D [1], we keep the original object classes: `beam`, `board`, `bookcase`, `ceiling`, `chair`, `clutter`, `column`, `door`, `floor`, `sofa`, `table`, `wall`, `window`.

For the presented 360BEV-Stanford dataset, we follow the data split method of Fold-1 of the Stanford2D3D [1] dataset. On the BEV dataset, we use the *area1*, *area2*, *area3*, *area4* and *area6* as the training data for the proposed 360BEV task, and we use the *area5a* and *area5b* as the validation set to evaluate the panoramic semantic mapping performance of models. The results of training and evaluation with the Fold-1 data split is similar the average scores which are calculated by using three-fold cross-validation. Besides, the validation set from Fold-1 is sufficient to evaluate the model performance on panoramic semantic mapping.

For 360BEV-Matterport, we use a different data split compared to Wijmans *et al*. [7]. Instead of using synthetic simulators, all samples on our dataset are converted from the real images and labels of Matterport3D [2] dataset, where there are 86 unique floors on our dataset, including 61 for training, 7 for validations, and 18 for testing.

2. More Quantitative Analysis

2.1. Results on Stanford2D3D

In Table 1, we present the per-class IoU results of front-view semantic segmentation on the Stanford2D3D dataset. The average (Avg.) scores are calculated with three folds [1] of cross validation, where Fold-2 is the most challenging split on the Stanford2D3D dataset. Compared to previous state-of-the-art Trans4PASS [9], our proposed 360Mapper achieves 47.97% mIoU in Fold-2 split. Besides, our 360Mapper model has overall better performance (54.34% in mIoU) in the average result calculated by three folds evaluation, surpassing the previous Trans4PASS model with +2.24% in mIoU. Furthermore, our model achieves the highest scores in 11 of 13 categories, including *board*, *bookcase*, *ceiling*, *chair*, *clutter*, *door*, *floor*, *sofa*, *table*, *wall*, and *window*. Improvements in these categories demonstrate the effectiveness of our 360Mapper model in combating distortions of 360° front-view images by incorporating distortion-aware 360Attention.

2.2. Results on 360FV-Matterport

As shown in Table 2, we present the front-view semantic segmentation results on the `test` set of 360FV-Matterport dataset. We compare our approaches with SegFormer [8], Trans4PASS [9], Trans4PASS+ [10], HoHoNet [6] with RGB and RGB-D, where HoHoNet uses ResNet-101 as backbone and the others use MiT-B2 as backbone. Compared with the well-established existing work SegFormer, our approach obtains a higher mIoU score with 43.16%, having a performance improvement of +0.67% mIoU on the test set. The test set is much more challenging than the validation set of 360FV-Matterport dataset, the results in Table 2 show the superiority of the proposed approach on extracting the underlying cues for the proposed task.

Apart from that, per-class IoU scores on 360FV-Matterport in Table 3. The performance of 360Mapper on both test and validation sets are demonstrated. 360Mapper

[*] Equal contribution.
[†] Corresponding author (e-mail: `jiaming.zhang@kit.edu`).
[1] The matterport utils tool.

Table 1. **Per-class results (360FV)** on the Stanford2D3D dataset. The models are based on the MiT-B2 [8] backbone.

Method	Split	mIoU	beam	board	bookcase	ceiling	chair	clutter	column	door	floor	sofa	table	wall	window
Trans4PASS [9]	Fold-1	53.30	00.40	69.50	62.20	82.80	58.50	34.30	21.90	44.90	91.20	40.80	57.70	74.80	54.20
Trans4PASS [9]	Fold-2	45.70	12.50	46.90	32.60	82.30	64.70	37.50	20.10	42.70	86.60	17.70	45.20	70.30	35.10
Trans4PASS [9]	Fold-3	57.20	21.40	65.40	58.30	80.20	55.80	41.90	28.60	76.30	88.60	45.40	58.80	59.30	63.60
Trans4PASS [9]	Avg.	52.10	**11.40**	60.60	51.10	81.80	59.70	37.90	**23.50**	54.60	88.80	34.60	53.90	68.10	51.00
360Mapper	Fold-1	56.46	00.57	74.61	65.03	83.96	62.41	40.27	18.72	42.22	93.31	53.86	65.90	76.18	58.84
360Mapper	Fold-2	47.97	09.32	41.89	40.45	83.01	62.27	34.92	25.74	57.74	88.02	24.48	42.95	72.19	41.22
360Mapper	Fold-3	58.60	08.05	74.32	61.05	81.05	63.29	44.44	4.64	76.56	90.91	57.28	62.52	64.96	72.77
360Mapper	Avg.	**54.34**	05.98	**63.61**	**55.51**	**82.67**	**62.66**	**39.88**	16.37	**58.84**	**90.75**	**45.21**	**57.12**	**71.11**	**57.61**

Table 2. **Panoramic semantic segmentation (360FV)** on the test set of 360FV-Matterport dataset.

Method	Backbone	Input	mIoU(%)
HoHoNet [6]	ResNet-101	RGB	40.22
HoHoNet [6]	ResNet-101	RGB-D	41.23
Trans4PASS [9]	MiT-B2	RGB	39.70
Trans4PASS+ [10]	MiT-B2	RGB	40.41
SegFormer [8]	MiT-B2	RGB	42.49
Ours	MiT-B2	RGB	**43.16**

delivers 46.35% and 43.16% mIoU performance on validation and test sets of 360FV-Matterport dataset respectively. For per-class IoUs, our model has better performance of challenging class, *e.g.*, *sink* with 25.12% and 28.24% on validation and test sets, surpassing Trans4PASS+ [10] with large margins. It notes that the small objects, *e.g.*, *furniture*, *mirror*, *toilet* on the test set, are still challenging for both methods. Apart from these, our models have better semantic segmentation results on 17 of 20 classes on the 360FV-Matterport dataset.

2.3. Results on 360BEV-Stanford

Per-class IoU scores on 360BEV-Stanford are shown in Table 4. On the 360BEV task, 360Mapper can achieve 45.78% score of mIoU, outperforming the previous Trans4Map [3] method with +9.7%. Specifically, our 360Mapper achieves per-class IoU with 93.33%, 42.52%, 59.14%, 5.06%, 62.66%, 39.75%, 5.48%, 38.74%, 97.76%, 48.92%, 76.76%, 45.86% and 24.89% for *void*, *board*, *bookcase*, *ceiling*, *chair*, *clutter*, *column*, *door*, *floor*, *sofa*, *table*, *wall* and *window*, respectively. Especially, the challenging objects that appear thin lines in bird's-eye views, such as *doors* and *walls*, can be more stably recognized by our method, which improves both IoUs with 10.23%→38.74% and 29.56%→45.86%. The *beam* class is not successfully recognized by both methods, because this BEV mechanism directly ignores objects on the ceiling. Different from the front-view semantic segmentation task, the *void* class is included on the 360BEV task, because this class can be used to indicate the invisible area on the BEV semantic maps, which is important for the downstream task, such as path planing.

2.4. Results on 360BEV-Matterport

The 360BEV results on the test set of 360BEV-Matterport are demonstrated in Table 5. We further compare our approach with three backbones, *e.g.*, MiT-B0, MiT-B2 from SegFormer [8] and MSCA-B from SegNeXt [4] on the test set of the 360BEV-Matterport for the panoramic semantic mapping task. Methods based on intermediate projection show the most promising results compared with those based on early projection and late projection. The result is consistent compared with the ones demonstrated on the validation set of 360BEV-Matterport dataset. 360Mapper still delivers the state-of-the art results for the proposed 360BEV task on the test set, indicating the effectiveness of the proposed architecture. Especially, our 360Mapper with MiT-B2 backbone (38.78%) can surpass Trans4Map with MiT-B2 (31.08%) as well as the one with MiT-B4 (31.79%). Besides, the proposed method based on MSCA-B backbone achieves the best result with 40.27% in mIoU.

Per-class IoU scores on the 360BEV-Matterport dataset are presented in Table 6. The performance of 360Mapper under MiT-B2 from SegFormer [8] and MSCA-B from SegNeXt [4] are included, which achieves promising performance for the 360BEV task. Compared to Trans4Map [3], our 360Mapper with the same MiT-B2 backbone can achieve respective 44.32% and 38.78% in mIoU on the validation set and the test set. The *void* class is also included on the 360BEV-Matterport dataset. Besides, if using a stronger backbone, *e.g.*, MSCA-B [4], our proposed mehods can

Table 3. **Per-class results (360FV)** on the 360FV-Matterport dataset.

Method	Backbone	Data	**mIoU**	wall	floor	chair	door	table	picture	furniture	objects	window	sofa	bed	sink	stairs	ceiling	toilet	mirror	shower	bathtub	counter	shelving
Trans4PASS+ [10]	MiT-B2	val	42.60	63.37	79.11	39.13	40.31	32.76	35.99	30.96	31.52	37.52	44.01	63.17	20.60	41.76	77.55	40.71	24.27	23.73	58.34	34.31	32.90
360Mapper	MiT-B2	val	**46.35**	64.12	83.14	45.75	44.98	37.96	41.08	32.26	35.07	40.61	48.69	69.80	25.12	47.80	80.15	45.96	28.70	22.31	60.05	38.64	34.82
Trans4PASS+ [10]	MiT-B2	test	40.41	64.32	80.12	41.24	41.70	30.86	36.93	35.16	28.27	32.65	33.28	55.98	22.93	37.19	78.36	48.96	17.73	26.51	49.65	28.64	22.82
360Mapper	MiT-B2	test	**43.16**	66.95	82.24	45.12	47.34	32.72	44.35	33.34	29.57	34.59	32.08	62.06	28.24	38.03	81.26	45.47	23.61	29.01	55.44	28.58	23.24

Table 4. **Per-class results (360BEV)** on the 360BEV-Stanford2D3D dataset.

Method	Backbone	**mIoU**	void	beam	board	bookcase	ceiling	chair	clutter	column	door	floor	sofa	table	wall	window
Trans4Map [3]	MiT-B2	36.08	64.17	0.00	28.10	52.96	0.45	52.30	34.71	6.40	10.23	92.18	44.29	68.22	29.56	21.44
360Mapper	MiT-B2	**45.78**	93.33	0.00	42.52	59.14	5.06	62.66	39.75	5.48	38.74	97.76	48.92	76.76	45.86	24.89

Table 5. **Panoramic semantic mapping (360BEV)** on the test set of 360BEV-Matterport dataset.

Method	Backbone	Acc	mRecall	mPrecision	mIoU
(1)Early projection: Proj. → Enc. → Seg.					
SegFormer [8]	MiT-B2	69.72	35.28	40.41	24.04
SegNeXt [4]	MSCA-B	69.99	36.25	41.96	25.22
(2) Late projection: Enc. → Seg. → Proj.					
HoHoNet [6]	ResNet101	62.89	35.18	39.54	22.01
Trans4PASS [9]	MiT-B2	53.50	29.35	33.53	16.53
Trans4PASS+ [10]	MiT-B2	57.24	30.639	34.49	17.72
SegFormer [8]	MiT-B2	62.91	35.35	39.64	22.02
(3) Intermediate projection: Enc. → Proj. → Seg.					
BEVFormer [5]	MiT-B2	72.04	36.69	47.90	27.46
Trans4Map [3]	MiT-B0	71.78	38.27	43.77	26.52
Trans4Map [3]	MiT-B2	72.94	45.45	47.03	31.08
Trans4Map [3]	MiT-B4	73.60	44.33	49.91	31.79
Ours	MiT-B0	76.02	43.11	50.41	31.35 (+4.83)
Ours	MiT-B2	78.04	54.47	54.27	38.78 (+7.70)
Ours	MSCA-B	79.17	55.16	57.27	40.27

achieve higher semantic mapping results on both of validation and test sets of 360BEV-Matterport dataset, which are 46.31% and 40.27% in mIoU, respectively.

3. More Qualitative Analysis

3.1. Analysis on Stanford2D3D

The visualization of front-view semantic segmentation (360FV) on the Stanford2D3D dataset is shown in Fig. 1, where the RGB input, the prediction of the baseline, the prediction of our model and the ground truth are depicted from left to the right. The corresponding color map is showcased at the top of Fig. 1. Compared with the baseline Trans4Pass [9], the panoramic semantic segmentation results of our model have clear boundaries among different objects which is much more similar to the ground truth, *e.g.*, the *door* and the *clutter* of the second sample. Our method also show promising performance on the objects with small spatial size, *e.g.*, *chairs*, compared with the baseline in the last sample, indicating that our 360Attention approach is good at grasping underlying context feature and cues through the deformable sampling locations.

3.2. Analysis on 360FV-Matterport

Fig. 2 is the front-view semantic segmentation visualization of the presented 360FV-Matterport dataset, providing a detailed depiction of the spatial distribution of different semantic classes. Compared with the baseline method Trans4Pass [9], our model produces segmentation results exhibit more precise contours and clearer boundaries between different objects, which closely resemble the ground truth segmentation labels, *e.g.*, the *toilet* and the *door* of the first sample. In the second row, the *door* on the right side is not recognized by the baseline model. In contrast to the baseline method, our model is able to accurately distinguish the *door* class from its surrounding *object* and *wall* classes, despite its small size and low contrast with the surrounding environment. The *table* in the center of the third sample are correctly predicted by our model while it is erroneously segmented by the baseline as *furniture*. This highlights the superior performance of 360Mapper in panoramic semantic segmentation under challenging conditions. In the last two rows, the small *chair* by the wall and the *door* are correctly recognized by our model.

3.3. Analysis on 360BEV-Stanford

We further introduce the qualitative results of 360BEV task on the 360BEV-Standford dataset in Fig. 3. The RGB input, the BEV semantic mapping results of the baseline and 360 Mapper, the BEV semantic mapping ground truth are depicted from left to right, where the color map is shown

Table 6. **Per-class results (360BEV)** on the 360BEV-Matterport dataset.

Method	Backbone	Data	**mIoU**	void	wall	floor	chair	door	table	picture	furniture	objects	window	sofa	bed	sink	stairs	ceiling	toilet	mirror	shower	bathtub	counter	shelving
Trans4Map [3]	MiT-B2	val	36.72	47.87	28.52	82.96	34.44	22.27	39.58	16.28	22.75	26.29	25.08	42.81	62.25	13.95	41.51	37.79	45.82	19.56	48.05	47.71	38.25	27.31
360Mapper	MiT-B2	val	44.32	74.30	31.94	85.85	42.01	26.71	46.40	23.21	25.00	24.87	27.36	51.37	66.59	20.99	47.07	54.97	56.91	29.50	55.70	63.16	45.82	31.04
360Mapper	MSCA-B	val	**46.31**	74.43	35.62	86.17	43.60	28.56	50.61	25.11	25.17	26.26	27.56	53.17	69.36	24.02	50.24	61.26	62.11	31.77	51.60	65.71	47.32	33.06
Trans4Map [3]	MiT-B2	test	31.08	40.51	32.54	80.21	33.23	20.85	37.21	19.01	18.46	23.05	23.56	32.35	52.08	15.34	29.02	18.27	41.90	15.39	25.58	48.19	30.38	15.52
360Mapper	MiT-B2	test	38.78	60.36	36.77	84.34	39.93	24.41	44.58	25.23	21.97	25.20	27.06	36.59	60.84	28.46	35.60	49.69	57.39	19.35	25.84	56.91	37.23	16.60
360Mapper	MSCA-B	test	**40.27**	62.82	40.09	85.22	42.60	25.48	46.00	24.37	25.11	26.08	27.39	39.68	61.45	28.18	36.17	50.88	58.31	19.77	29.85	59.78	35.39	21.14

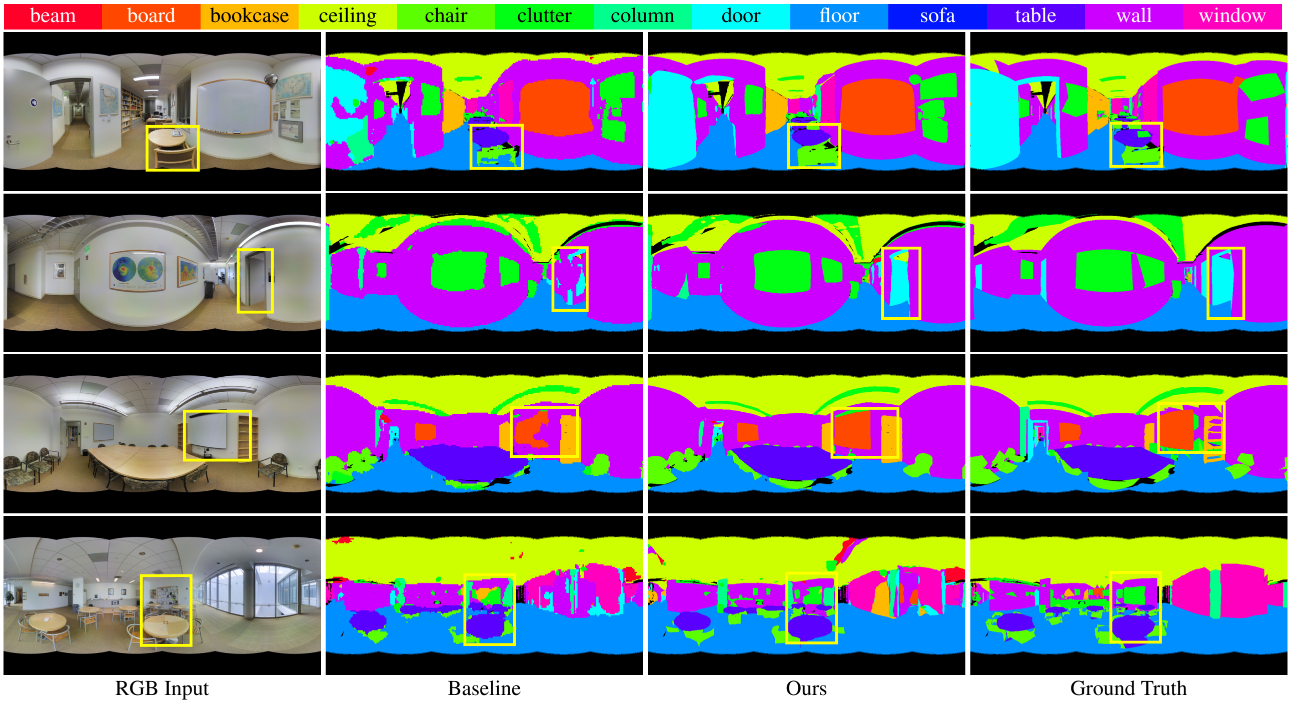

Figure 1. **360FV visualization and qualitative analysis** on the Stanford2D3D dataset.

at the top of Fig. 3. The *chairs* of the first and the second sample are correctly predicted by our method while they are partially or entirely missed by the baseline. Compared with the 360Mapper, the baseline shows more false prediction especially regarding some furniture, *e.g.*, the false predicted *bookcase* at the third sample, which should be predicted as *chairs*. At the last row of Fig. 3, the challenging *door* is not recognized by the baseline model, while our 360Mapper can provide accurate *door* segmentation result, even it is a thin line in the BEV map. Our method shows overall superior performance on the proposed task compared with the baseline in terms of the semantic segmentation performance on small objects, which further illustrates the strength by using 360Attention.

3.4. Analysis on 360BEV-Matterport

Fig. 4 presents qualitative results for the 360BEV task on the 360BEV-Matterport dataset. We observe that our 360Mapper outperforms the baseline method Trans4Map [3] in terms of accurately segmenting small objects. In particular, the baseline method exhibits more false predictions, such as the misclassified *chair* in the first sample and *object* misidentified as *table* in the second sample. Surprisingly, the different steps of *stairs* in the third and the fourth sample are recognized correctly by both methods. However, we find the fifth sample to be particularly challenging, as both the baseline and our 360Mapper recognize the object in the center of the image as a *counter*, which is a *table* as shown in the ground truth. This failure case shows the difficulty of accurately distinguishing between similar object categories from the context of panoramic images to the bird's-eye-view semantic maps.

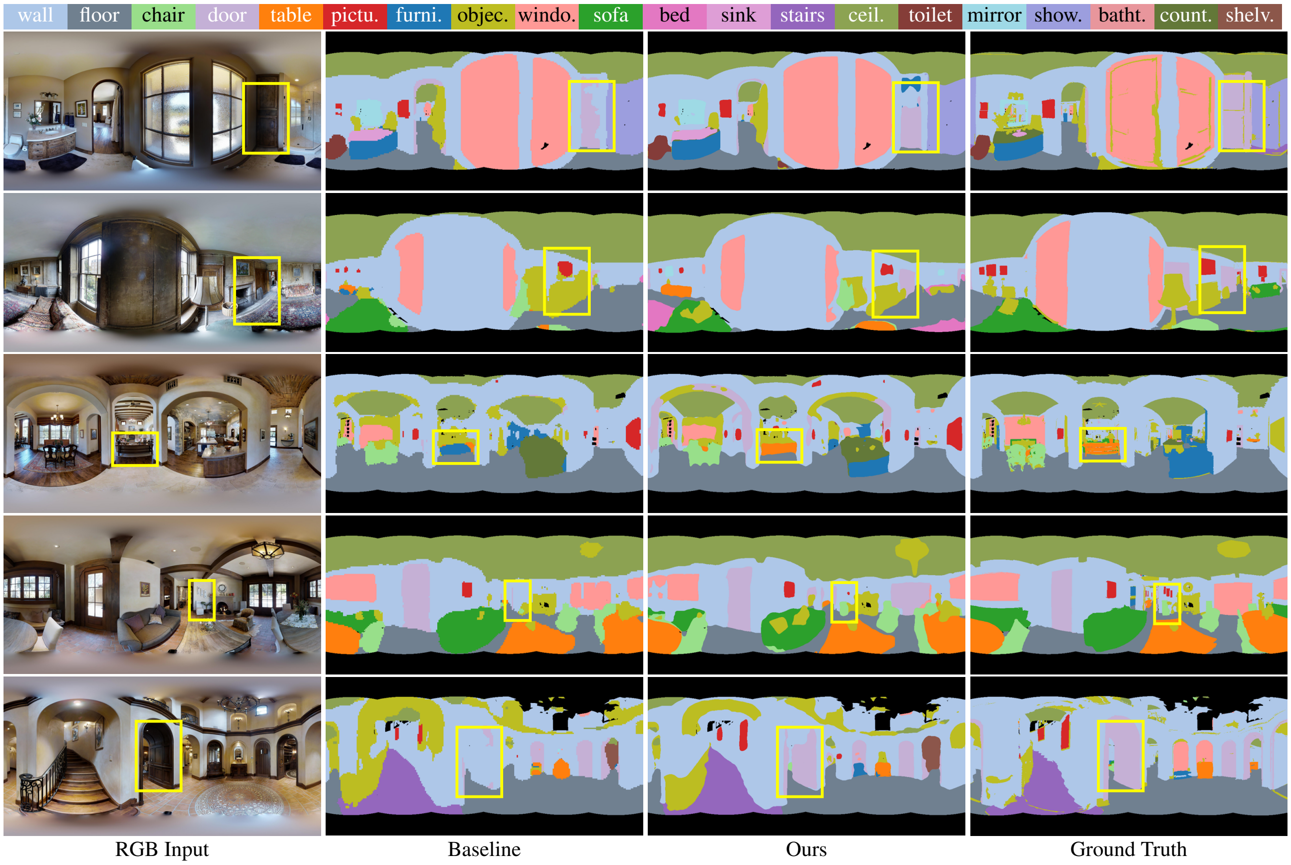

Figure 2. **360FV visualization and qualitative analysis** on the 360FV-Matterport dataset.

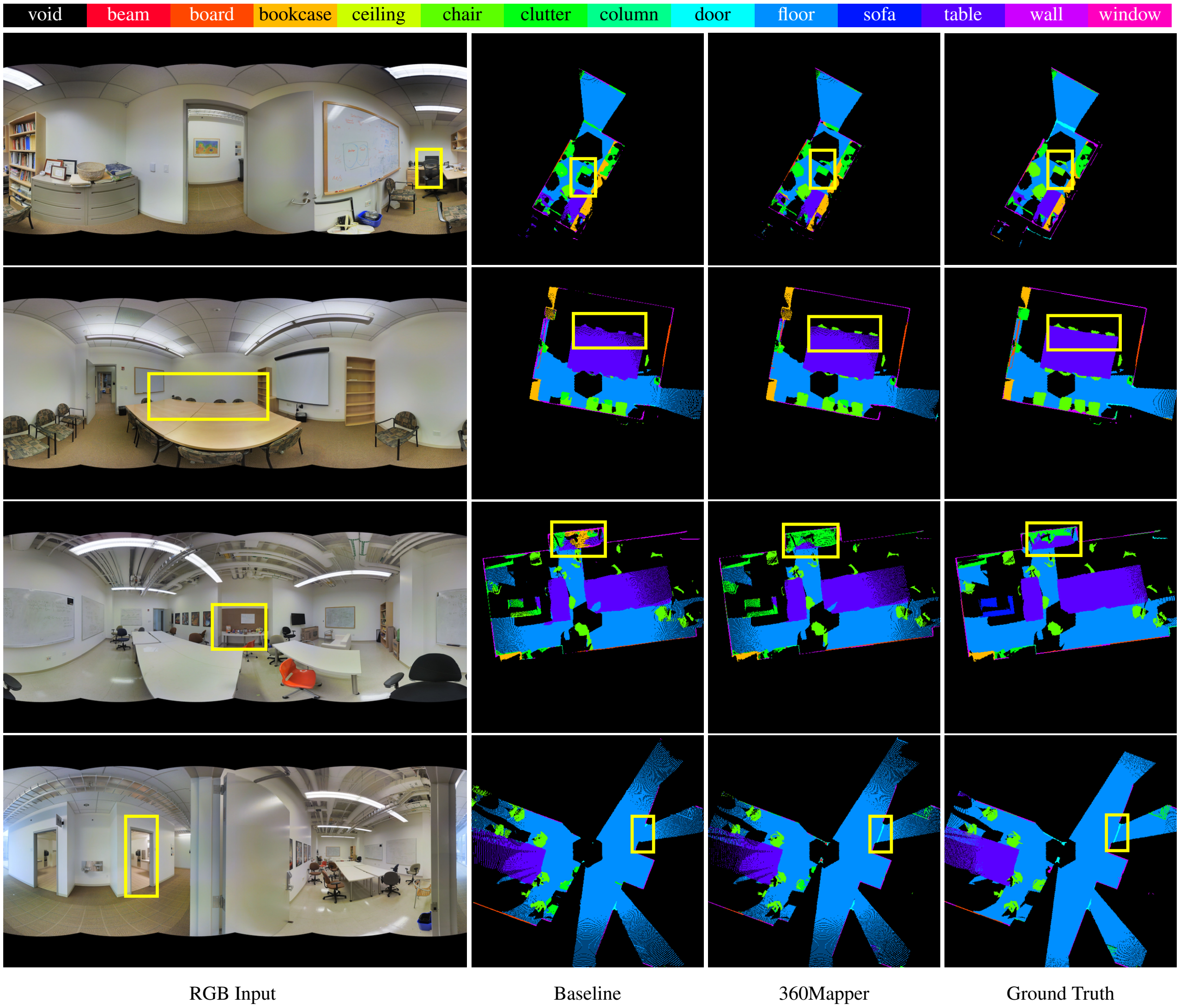

Figure 3. **360BEV visualization and qualitative analysis** on the 360BEV-Stanford dataset. Black regions are the `void` class, indicating the invisible areas in BEV semantic maps. Zoom in for better view.

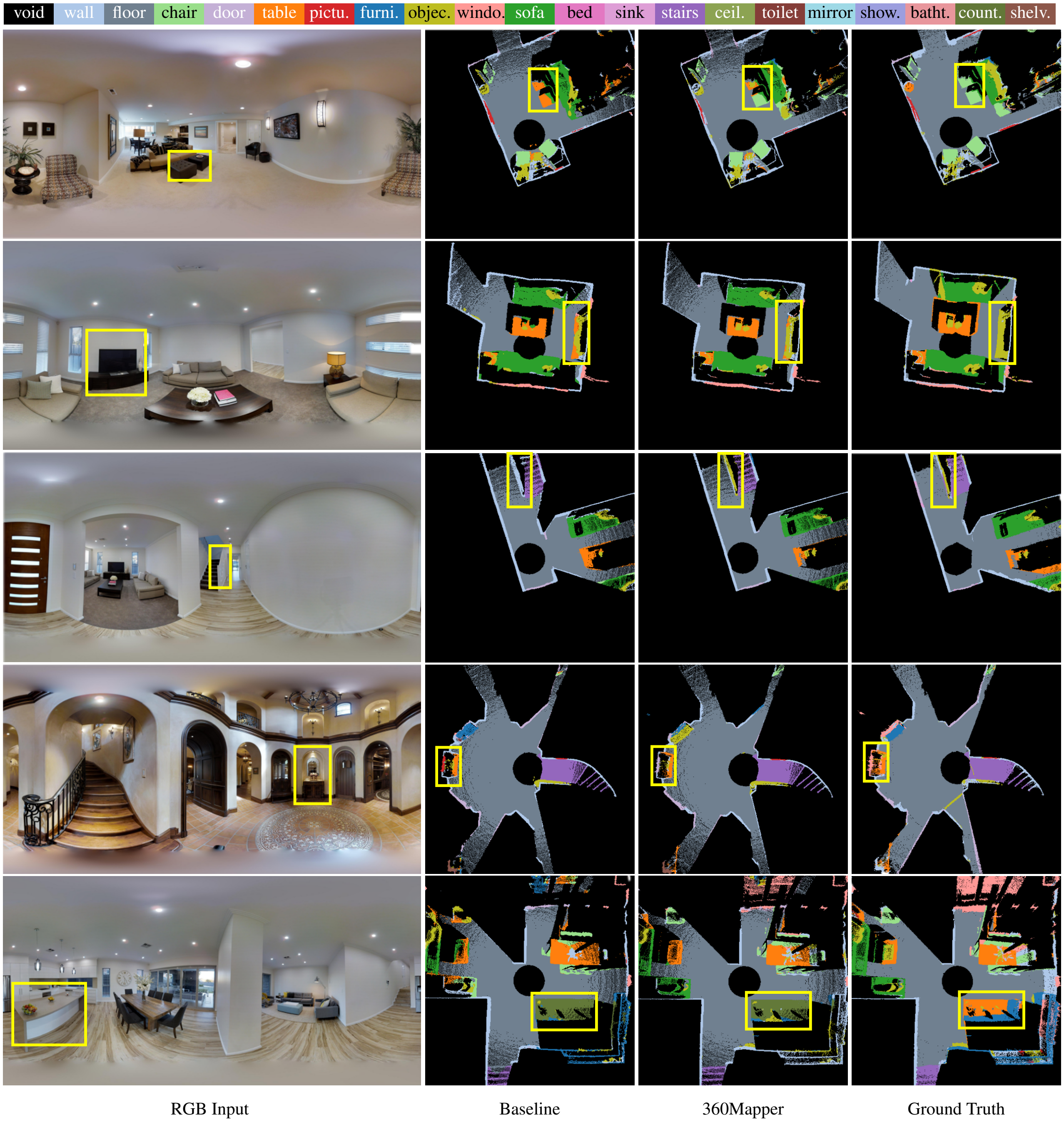

Figure 4. **360BEV visualization and qualitative analysis** on the 360BEV-Matterport dataset. Black regions are the `void` class, indicating the invisible areas in BEV semantic maps. Zoom in for better view.

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] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *3DV*, 2017. 1

[3] Chang Chen, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Trans4Map: Revisiting holistic bird's-eye-view mapping from egocentric images to allocentric semantics with vision transformers. In *WACV*, 2023. 2, 3, 4

[4] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. SegNeXt: Rethinking convolutional attention design for semantic segmentation. In *NeurIPS*, 2022. 2, 3

[5] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 3

[6] Cheng Sun, Min Sun, and Hwann-Tzong Chen. HoHoNet: 360 indoor holistic understanding with latent horizontal features. In *CVPR*, 2021. 1, 2, 3

[7] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. In *CVPR*, 2019. 1

[8] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 1, 2, 3

[9] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *CVPR*, 2022. 1, 2, 3

[10] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *arXiv preprint arXiv:2207.11860*, 2022. 1, 2, 3