SG-BEV: Satellite-Guided BEV Fusion for Cross-View Semantic Segmentation

Supplementary Material

In this supplementary material, we provide additional details about our method in Section 1, a detailed introduction to the datasets in Section 2, specifics of experimental settings and additional experimental results in Section 3.

1. Additional details of our proposed method

1.1. Visualization of point clouds reprojection using our SGR module

In this section, we provide additional visualizations of point clouds offset obtained before and after using Satellite-Guided Reprojection (SGR) module (introduced in Section 3 of the main paper). The original point clouds obtained through depth information projection often appear as shown on the left side of Figure 1, where the point clouds are noticeably concentrated along the edges of building exteriors, leaving larger interior areas of the buildings devoid of point clouds distribution. After passing through the SGR module, the reprojected point clouds gradually shift from the exterior walls of the buildings to their interiors, as depicted on the right side of Figure 1.

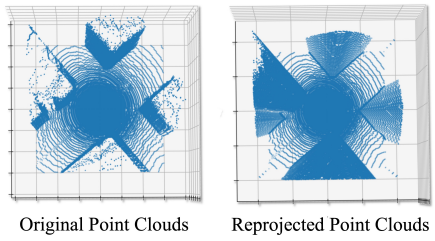

Figure 1. Distribution of point clouds obtained before and after using SGR module.

1.2. Visualization comparison of different cross-view projection algorithms

To intuitively demonstrate the feature mapping results of various methods, RGB values are utilized to substitute the features requiring mapping, facilitating a visual comparison. As clearly shown in Figure 2, the geometric projection methods Spherical Transform (ST) [6] and Geometric Projection (GP) [5] can only map features of low building facades, such as railings and low walls, and these features are severely distorted. The original BEV method leads to a concentration of features at the wall locations, resulting in sparse interior building features. Our approach employs the SGR module to efficiently map facade features onto the BEV plane, ensuring maximal transfer while maintaining visual continuity.

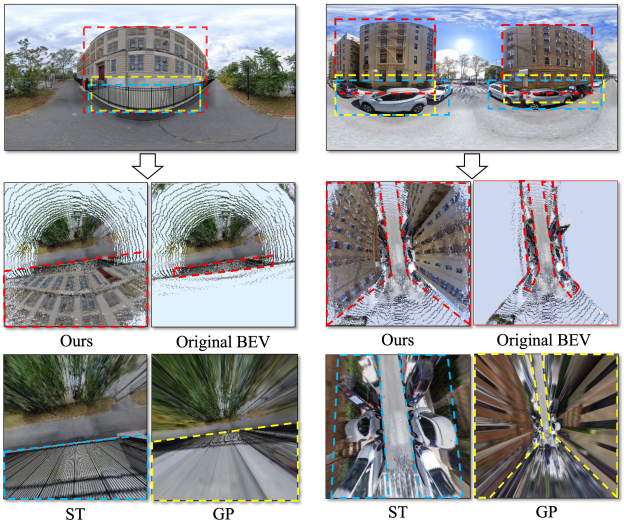

Figure 2. Intuitive visualization comparisons of different cross-view projection algorithms.

1.3. More details about the structure of cross-view feature fusion module

The input satellite features and Bird's Eye View (BEV) features are both unified within the same top-down perspective feature space. As shown in Figure 3, we initially employ an align module to shift the BEV features for alignment with the satellite features. Subsequently, a dynamic fusion module is employed to optimize the feature fusion process.

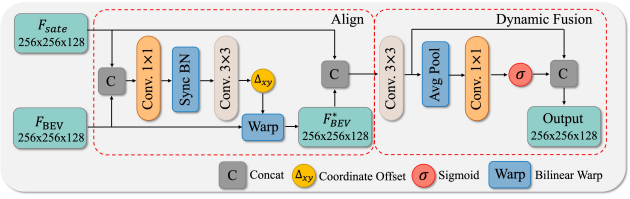

Figure 3. The structure of our cross-view feature fusion module.

2. Additional details of the datasets

2.1. Details of panorama depth maps

For monocular depth estimation of street-view panorama images, we can employ established depth estimation algorithms like ZoeDepth [1], or use Google Street View Download 360[1] to download corresponding depth maps, as shown in Figure 4. Although both methods yield favorable results, the depth maps provided by Google Street View Download 360 are more accurate for depth estimation in building areas. Hence, we use it as the data source of depth estimation information in our method.

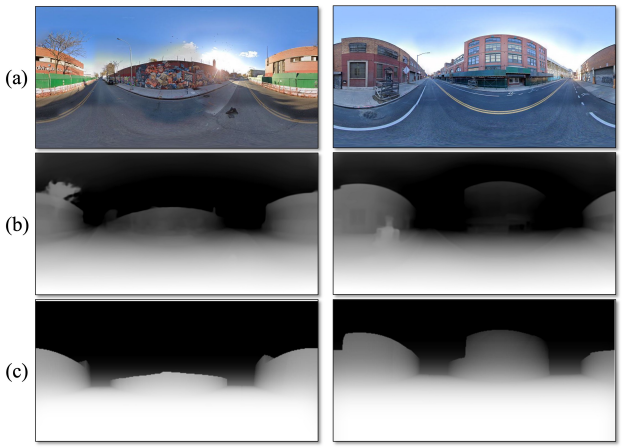

Figure 4. Depth estimation result based on street-view image. (a) Street-view image from OmniCity dataset. (b) Depth estimation results from ZoeDepth [1]. (c) Depth maps provided by Google Street View Download 360.

2.2. Details of each dataset

For the public Vigor [7] dataset, we supplemented it with land use information provided by DataSF[2] in the San Francisco area. Additionally, to facilitate subsequent BEV tasks, we augmented each street-view image in both the Vigor and OmniCity [3] datasets with depth maps using Google Street View Download 360[1]. For our self-collected Brooklyn and Boston datasets, we used property data provided by PLUTO[3] and Boston Maps[4], as well as the OpenStreetMap (OSM) building outlines to obtain attribute data of individual buildings following the approach used in OmniCity dataset. Compared with OmniCity dataset, Brooklyn dataset covers the entire Brooklyn and Manhattan areas, with the step distance of street view images increased to 97.5 meters to reduce the overlap of satellite images. The Boston dataset covers the urban area of Boston with the same step distance of street view images as Brooklyn dataset. For each dataset, we strictly divide the training and test samples by regional partition (train: test = 4: 1). Table 1 shows the number of training and test samples of the four datasets, along with the corresponding categories for each dataset shown in Table 5.

[1] https://svd360.istreetview.com/

[2] https://data.sfgov.org/Housing-and-Buildings/Land-Use/us3s-fp9q

[3] https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page

[4] https://data.boston.gov/dataset/boston-buildings-with-roof-breaks

Table 1. Number of training and test samples in each dataset.

Dataset	OmniCity [3]	Vigor [7]	Brooklyn	Boston
Train	14,400	11,960	7,600	7,036
Test	3,600	2,990	1,900	1,759

2.3. The impact of different data partition methods

As mentioned above, the training and test sets for all datasets were collected through regional partitioning. To further explore the impact of different data partition methods, we compare the experimental results of the OmniCity and Brooklyn datasets using random partitioning and regional partitioning. As illustrated in Table 2, the performance metrics obtained through the random partitioning approach are remarkably higher than those achieved with regional partitioning, a trend particularly evident in the densely sampled OmniCity dataset. This overestimation is primarily attributed to the nature of the cross-view datasets that are densely sampled and randomly partitioned, where a single satellite image often covers multiple street view images. Random partitioning of training and test sets in cross-view image pairs might lead to significant overlaps among satellite images, compromising the dataset's independence and resulting in inflated performance metrics.

Table 2. Quantitative analysis of random and regional partitioning in OmniCity and Brooklyn datasets, which indicates the overestimation of performance metrics using random partitioning.

Partition	Method	Dataset			
		OmniCity		Brooklyn	
		Land use	Floor	Land use	Floor
Random	SegNext [2]	77.31	77.92	45.00	44.61
	BEVFormer [4]	79.15	78.86	48.89	50.32
	Ours	80.13	81.88	52.50	56.09
Regional	SegNext [2]	31.25	31.19	35.77	33.16
	BEVFormer [4]	31.95	32.18	41.89	43.32
	Ours	37.95	40.02	47.20	48.81

3. Additional details of experimental results

3.1. Additional experimental analysis on hyperparameter settings

In our Satellite-Guided Reprojection (SGR) method, the calculation of offset Δ involved two critical hyperparameters: d_0 and α, as shown in Eq. 1.

$$\Delta = \log(1 + d - d_0) \times \alpha \tag{1}$$

In our study, the parameter d_0 is introduced to mitigate the feature offset from areas such as roads to building regions. Typically, d_0 represents the depth of the ground within a certain range from the camera. In practical computation, if d is less than d_0, the offset will not occur. Considering that the width of urban roads typically ranges between 10 to 14 meters[5], together with the additional sidewalks of about 2 meters wide on each side, we set d_0 to approximately 10 meters. This distance represents the average distance from a vehicle to either edge of the road. With access to more specific road width information, the parameter d_0 can be further optimized to achieve enhanced performance.

Another critical hyperparameter in our study was α, which determined the amplitude of variation in Δ at different depths under the same building. Initially, the entire area is divided into 3×3 blocks, and then the proportion of building footprint pixels in each block is calculated to obtain the pixel ratio parameter ρ. Below is our formula for calculating α based on ρ. The hyperparameter t in the formula is adjustable. We tested three different values of t: 10, 20, and 30, with a higher value of t indicating a larger amplitude of change. As shown in Table 3, our experiments on the OmniCity and Brooklyn datasets with these t values yielded results, demonstrating that $t = 20$ achieved the best performance. Furthermore, as illustrated in Figure 5, $t = 20$ is the most balanced choice in terms of visual effects. Therefore, $t = 20$ was selected as the parameter for our experiments.

$$\alpha = \begin{cases} 0 & \text{if } \rho \leq 0.1, \\ 5 + t \times \rho & \text{if } \rho > 0.1. \end{cases} \tag{2}$$

3.2. Additional experimental analysis on satellite image size

We also provide experimental analysis to validate the model performance using satellite images of different sizes, including (1) 128×128 pixels, (2) 256×256 pixels, and (3) 512×512 pixels.

From Table 4, it is evident that the model performs worst when using the image size of 512×512 pixels. This is because with the increasing of image size, the satellite image may cover multiple blocks and streets, making it challenging for a single street-view panorama to provide sufficient effective information. As observed in Figure 6, beyond a certain range, the model's segmentation performance begins to decline sharply. The model achieves better performance using image sizes of 128×128 and 256×256 pixels. However, too small image size may limit the model's perceptual field, which is detrimental to downstream fine-grained segmentation tasks and leads to inefficiencies in large-scale applications. Considering the above factors, we select 256×256 pixels as the satellite image size used in our experiments.

[5] https://safety.fhwa.dot.gov/geometric/pubs/mitigationstrategies/chapter3/3_lanewidth.cfm

Table 3. Quantitative analysis of different α, in terms of mIoU (%), which indicates that $\alpha = 20$ yields the best performance.

Parameter	Dataset			
	OmniCity		Brooklyn	
	Land use	Floor	Land use	Floor
$t = 10$	36.18	39.45	46.99	47.74
$t = 20$	37.54	**40.64**	**47.19**	**49.51**
$t = 30$	**37.68**	40.41	46.15	48.42

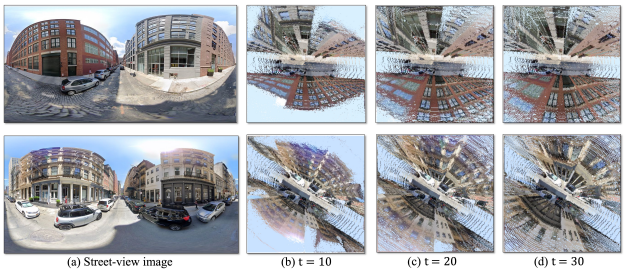

Figure 5. Visualization of the reprojection results using different α values.

Table 4. Quantitative analysis of different sizes of satellite images, in terms of mIoU (%), which illustrates that 256×256 pixels yield the best performance.

Satellite Image Size	Method	Dataset			
		OmniCity		Brooklyn	
		Land use	Floor	Land use	Floor
128×128	SegNext [2]	28.65	24.17	32.44	30.07
	SG-BEV	36.82	38.23	46.82	**49.89**
256×256	SegNext [2]	31.38	25.27	36.85	34.55
	SG-BEV	**37.54**	**40.64**	**47.19**	49.51
512×512	SegNext [2]	29.88	26.03	37.31	30.77
	SG-BEV	32.55	31.98	43.04	37.35

3.3. More quantitative experimental results

The performance of our method in fine-grained attribute segmentation on four different datasets is demonstrated in

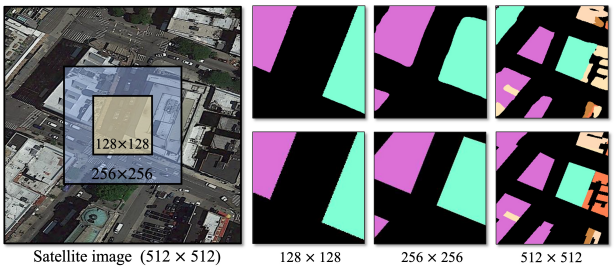

Figure 6. Visualization of the semantic segmentation results by applying different sizes of satellite images. The first column shows the satellite image. The rest columns represent the semantic segmentation results obtained from satellite images of different sizes (first row) alongside their corresponding ground truths (second row).

Tables 6, 7, 8, and 9. We compare our approach with the state-of-the-art satellite-based method (SegNext [2]) and cross-view method (BEVFormer [4]). Our method significantly enhances the performance in almost all building categories across all datasets, demonstrating its robustness across a wide range of urban architectural styles and task attributes.

3.4. More qualitative experimental results

We provide additional visualizations for various datasets. Figure 7 illustrates the comparison of our SG-BEV method with different satellite-based method. Unlike other methods that roughly identify building outlines without discerning fine-grained attributes, our method is capable of differentiating buildings with distinct attributes. In Figure 8, SG-BEV is compared with various cross-view methods, demonstrating more comprehensive feature mapping within the same building, leading to more consistent internal attributes and superior performance. Moreover, Figure 9 shows several typical failure cases, such as occlusions by trees and vehicles, or the shooting locations too far from the buildings.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2

[2] 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. *Advances in Neural Information Processing Systems*, 35:1140–1156, 2022. 2, 3, 4, 5

[3] Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Song Xia, and Dahua Lin. Omnicity: Omnipotent city understanding with multi-level and multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17397–17407, 2023. 2, 5

[4] 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 *European conference on computer vision*, pages 1–18. Springer, 2022. 2, 4, 5

[5] Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. Boosting 3-dof ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21516–21526, 2023. 1

[6] Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[7] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 2, 5

Table 5. Label Categories for Each Dataset. OmniCity [3] dataset contains detailed land use information. Brooklyn, Boston, and Vigor's land use information comes from PLUTO[3], Boston Maps[4], and DataSF[2], respectively.

Category	OmniCity [3]/ Brooklyn		Boston	Vigor [7]
	Land Use	Floor	Land Use	Land Use
1	Background (BG)	Background (BG)	Background (BG)	Background (BG)
2	1/2 Family Buildings (A1/C1)	Level 1 (B1/D1)	Industrial Manufacturing (E1)	Residential (F1)
3	Walk-Up Buildings (A2/C2)	Level 2 (B2/D2)	Commercial (E2)	Mixed Use (F2)
4	Elevator Buildings (A3/C3)	Level 3 (B3/D3)	High Residential (E3)	Industrial (F3)
5	Mixed Residential/ Commercial (A4/C4)	Level 4 (B4/D4)	Low Medium Residential (E4)	Cultural/ Institutional/ Educational (F4)
6	Office Buildings (A5/C5)	Level 5 (B5/D5)	Low Residential (E5)	Others (F5)
7	Industrial/ Transportation/ Utility (A6/C6)	Level 6 (B6/D6)	Public (E6)	-
8	Others (A7/C7)	Level 7 and Above (B7/D7)	-	-

Table 6. Fine-grained attribute segmentation results of different methods on the OmniCity dataset. Our method demonstrates an improvement in the mIoU by 0.95% - 10.37% and 0.1% - 11.53% for the land use attribute, respectively, and 0.57% - 22.39% and 0.75% - 12.43% for the floor level attribute, respectively, compared with the state-of-the-art.

Method	mIoU (%) of each category															
	Land use								Floor							
	BG	A1	A2	A3	A4	A5	A6	A7	BG	B1	B2	B3	B4	B5	B6	B7
SegNext [2]	86.31	13.16	29.53	19.58	26.70	34.23	21.75	19.73	83.17	7.04	5.69	7.33	15.17	16.59	15.41	51.75
BEVFormer [4]	87.16	15.37	30.47	18.42	28.48	33.38	23.55	20.55	82.99	14.82	17.15	17.63	22.44	20.48	26.45	52.56
Ours	**87.26**	**22.75**	**33.10**	**29.95**	**31.67**	**38.91**	**29.75**	**26.91**	**83.74**	**19.93**	**27.56**	**29.72**	**34.87**	**30.29**	**36.94**	**62.06**
Δ_1	+0.95	+9.59	+3.57	+10.37	+4.97	+4.68	+8.00	+7.18	+0.57	+12.89	+21.87	+22.39	+19.70	+13.70	+21.53	+10.31
Δ_2	+0.10	+7.38	+2.63	+11.53	+3.19	+5.53	+6.20	+6.36	+0.75	+5.11	+10.41	+12.09	+12.43	+9.81	+10.49	+9.50

Δ_1: The improvement compared with SegNext. Δ_2: The improvement compared with BEVFormer.

Table 7. Fine-grained results of different models on the Brooklyn dataset. Except for BG in cross-view method, our method improves the mIoU by 2.27% - 17.34% and 3.92% - 6.18% for land use attribute, respectively, 0.64% - 29.07% and 0.22% - 21.24% for floor level attribute, respectively, compared with current state-of-the-art.

Method	mIoU (%) of each category															
	Land use								Floor							
	BG	C1	C2	C3	C4	C5	C6	C7	BG	D1	D2	D3	D4	D5	D6	D7
SegNext [2]	83.62	35.32	35.22	35.17	30.41	22.09	39.96	12.97	84.73	31.92	34.92	34.51	29.47	5.29	31.25	27.28
BEVFormer [4]	**86.28**	48.74	42.76	39.78	31.68	25.39	48.68	20.37	84.67	38.93	46.73	45.00	44.23	21.48	40.59	35.11
Ours	85.89	**52.66**	**47.38**	**45.96**	**36.14**	**29.43**	**53.09**	**27.05**	**85.37**	**42.65**	**48.47**	**45.74**	**44.45**	**27.07**	**46.01**	**56.35**
Δ_1	+2.27	+17.34	+12.16	+10.79	+5.73	+7.34	+13.13	+14.08	+0.64	+10.73	+13.55	+11.23	+14.98	+21.78	+14.76	+29.07
Δ_2	-0.39	+3.92	+4.62	+6.18	+4.46	+4.04	+4.41	+6.68	+0.70	+3.72	+1.74	+0.74	+0.22	+5.59	+5.42	+21.24

Table 8. Fine-grained results of different models on the Boston dataset. Except for BG and E1 in cross-view method, our method improves the mIoU by 1.19% - 13.59% and 3.08% - 4.24% for land use attribute, respectively, compared with current state-of-the-art.

Method	mIoU (%) of each category: Land use						
	BG	E1	E2	E3	E4	E5	E6
SegNext [2]	86.99	13.59	35.90	34.02	30.08	19.93	7.38
BEVFormer [4]	**88.24**	**15.97**	39.85	37.05	37.75	29.28	11.96
Ours	88.18	15.90	**44.08**	**40.40**	**40.83**	**33.52**	**15.16**
Δ_1	+1.19	+2.31	+8.18	+6.38	+10.75	+13.59	+7.78
Δ_2	-0.06	-0.07	+4.23	+3.35	+3.08	+4.24	+3.20

Table 9. Fine-grained results of different methods on the Vigor dataset. Our method improves the mIoU by 1.30% - 11.27% and 0.24% - 9.57% for land use attribute, respectively, compared with current state-of-the-art.

Method	mIoU (%) of each category: Land use					
	BG	F1	F2	F3	F4	F5
SegNext [2]	81.03	61.07	19.49	21.25	15.04	11.63
BEVFormer [4]	82.09	63.51	25.74	23.86	16.55	12.25
Ours	**82.33**	**67.65**	**30.76**	**28.76**	**26.12**	**14.54**
Δ_1	+1.30	+6.58	+11.27	+7.51	+11.08	+2.91
Δ_2	+0.24	+4.14	+5.02	+4.90	+9.57	+2.29

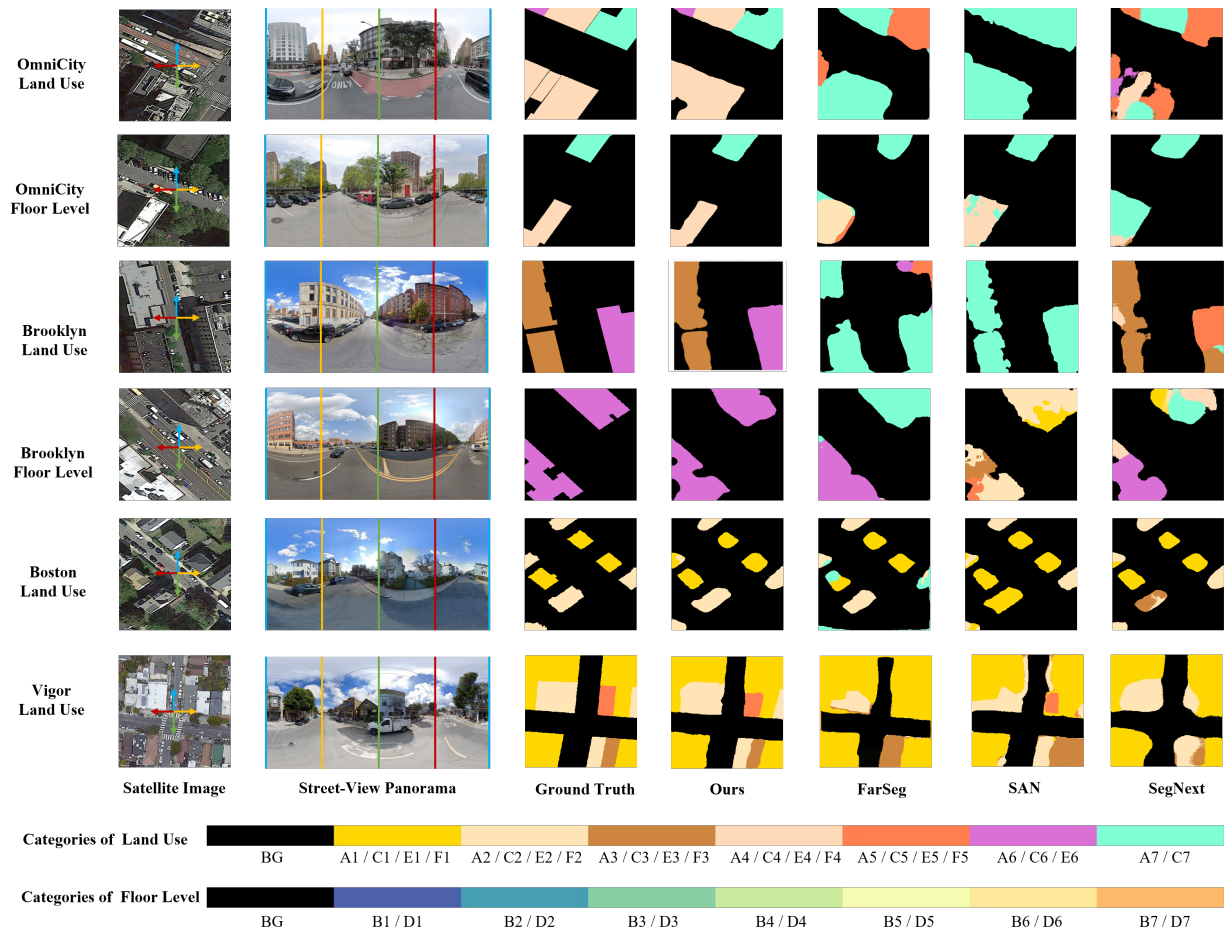

Figure 7. **Comparisons of SG-BEV (Ours) and Satellite-Based Methods for Fine-Grained Segmentation.** The first two rows show results of OmniCity on land use and floor level segmentation tasks. The third and forth rows present land use and floor level segmentation results of Brooklyn. The fifth and sixth rows show the land use segmentation results of Boston and Vigor. The street-view panoramas, from left to right, correspond to a 360-degree clockwise rotation starting from the north direction in the satellite imagery.

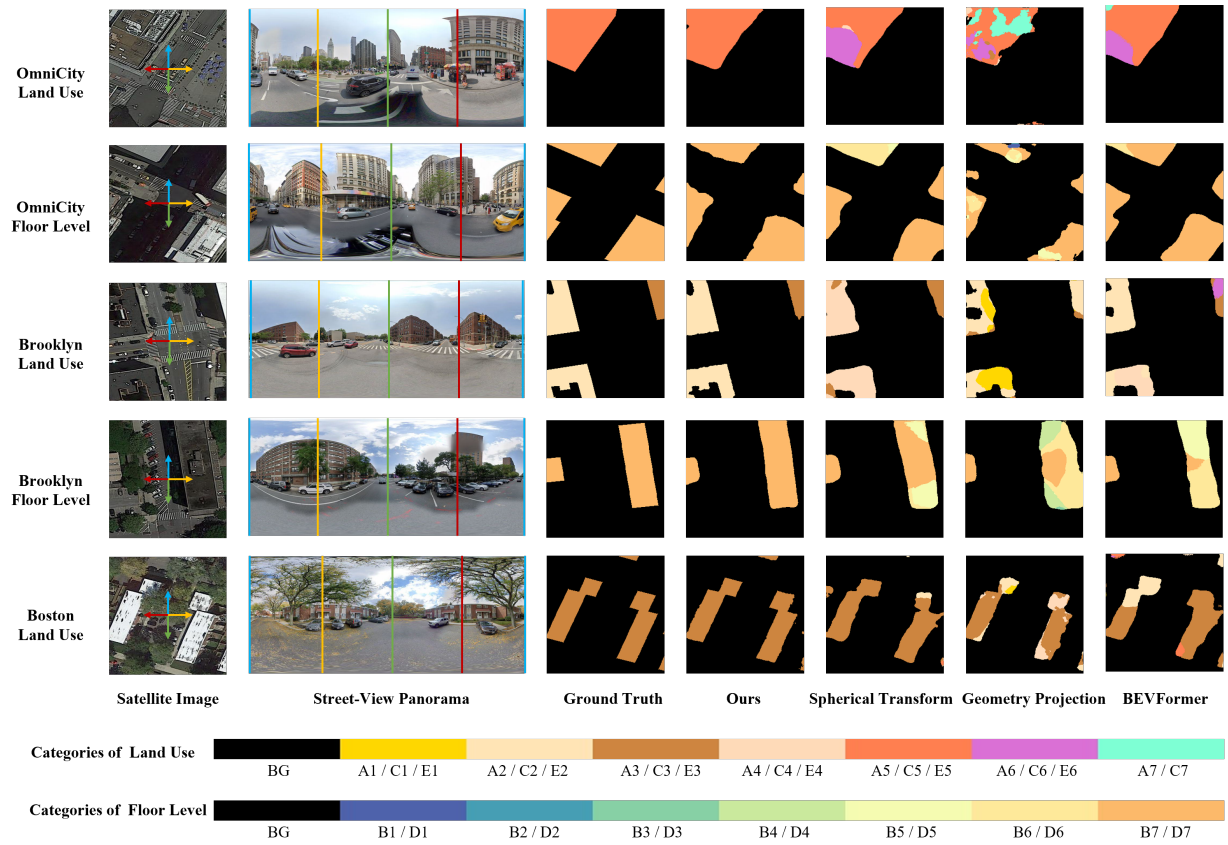

Figure 8. **Comparisons of SG-BEV (Ours) and Cross-View Methods for Fine-Grained Segmentation.** The first two rows show results of OmniCity on land use and floor level segmentation tasks. The third and forth rows present land use and floor level segmentation results of Brooklyn. The fifth row shows the land use segmentation results of Boston. Results for the Vigor dataset are not included, as the offset problem in this dataset makes the Spherical Transform and Geometric Projection methods inapplicable. The street-view panoramas, from left to right, correspond to a 360-degree clockwise rotation starting from the north direction in the satellite imagery.

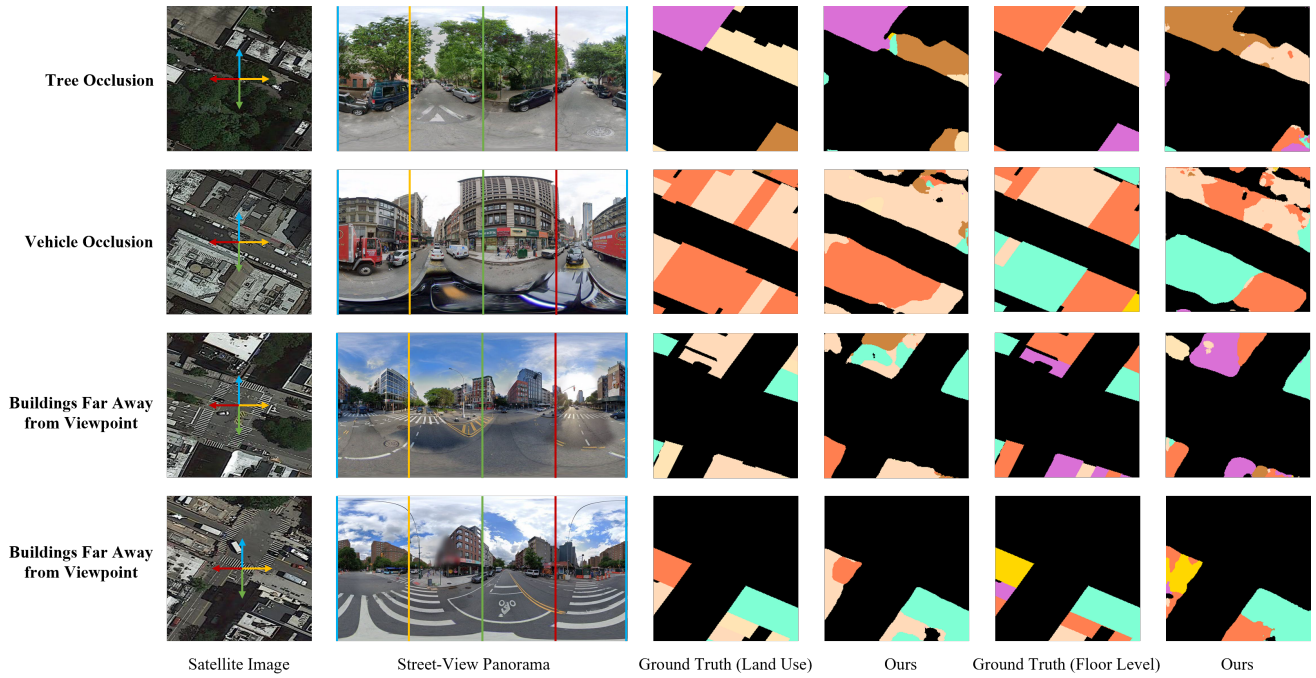

Figure 9. **Limitations of the SG-BEV (Ours) Method.** The first row depicts common errors in land use and floor level segmentation tasks, primarily due to occlusion by trees. The second row presents inaccuracies in land use and floor level predictions, resulting from occlusion caused by large vehicles. The third and fourth rows display unclear segmentation results, arising from the considerable distance of buildings from the viewpoint.