REL-SF4PASS: Panoramic Semantic Segmentation with REL Depth Representation and Spherical Fusion

Supplementary Material

A. Details for Constructing 3D Point Cloud

In this section, we show the way that HHA and REL convert depth information to 3D point cloud, respectively, and whether camera intrinsics are used. For convenience, This part shows the coordinate calculation before gravity correction, and the gravity correction is the same for HHA / REL and is not dependent on camera intrinsics. Suppose that a pixel is (u, v) of an image with width W and height H in 3D Cartesian coordinate. Using HHA representation, we calculate:

$$p_x = \frac{(u - u_0) \cdot d}{f_u}, \tag{S-1}$$

$$p_y = d, \tag{S-2}$$

$$p_z = \frac{(v - v_0) \cdot d}{f_v}, \tag{S-3}$$

where camera optical center is (u_0, v_0), and focal length $f = (f_u, f_v)$.

Using REL representation, we calculate:

$$p_x = r \cos\phi \sin\theta, \tag{S-4}$$

$$p_y = r \cos\phi \cos\theta, \tag{S-5}$$

$$p_z = r \sin\phi, \tag{S-6}$$

where $r = d$, $\theta = \frac{u}{W} \cdot 360° - 180°$, and $\phi = 90° - \frac{v}{h} \cdot 180°$.

Using REL representation in cylindrical coordinate, we calculate:

$$p_\rho = r \cos\phi, \tag{S-7}$$

$$p_\theta = \theta, \tag{S-8}$$

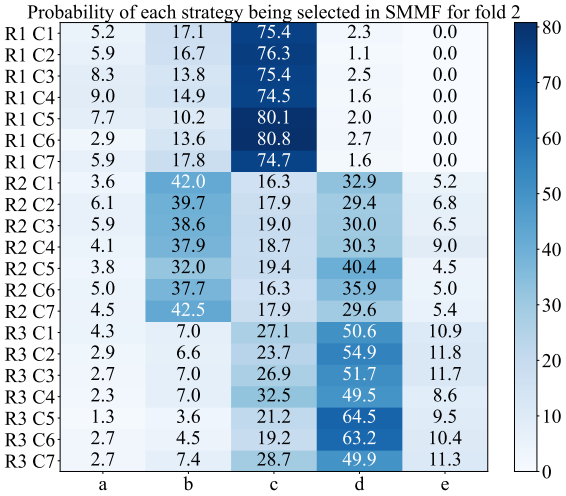
$$p_z = r \sin\phi, \tag{S-9}$$

where r, ϕ and θ is the same as above ones.

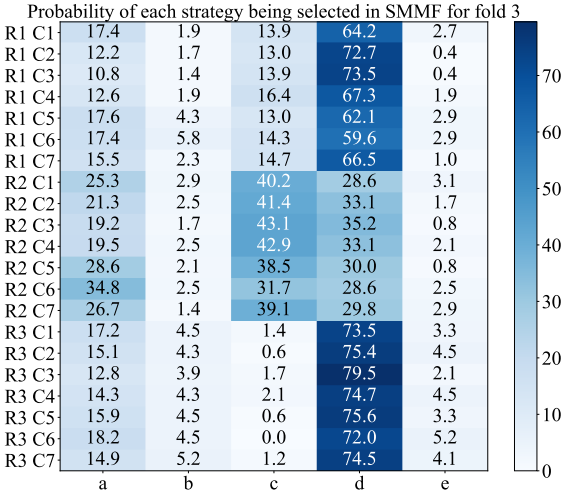
It can be observed that unlike HHA, REL is calculated without camera intrinsics.

B. The Detail Results Compared with SGAT4PASS

In this section, we show the detail performance on Stanford2D3D Panoramic datasets official fold 1 with SGA metrics. All 16 test situations are shown in both mIoU and pixel accuracy. From the results, we can know that variance / fluctuation of REL-SF4PASS is also about 30% / 60% of the SGAT4PASS ones from Tab. S-1, which is shown in Table 4 in the manuscript. It shows that REL-SF4PASS has the better 3D robustness.



(a) Fold 2.



(b) Fold 3.

Figure S-1. More visualization of the probability of each fusion strategy being selected for each region at inference. (a) is fold 2 of Stanford2D3D Panoramic datasets when (b) is fold 3. "RX CY" means the region in X^{th} row and Y^{th} column. The value is the probability (%) of choosing this fusion strategy throughout the inference process. a / b / c / d / e indicates $\mathbf{g} = [0, 0, 0, 0]$ / $[1, 0, 0, 0]$ / $[1, 1, 0, 0]$ / $[1, 1, 1, 0]$ / $[1, 1, 1, 1]$.

C. More Visualizations

In this section, we show more result visualizations for both SMMF consistency and different depth representation com-

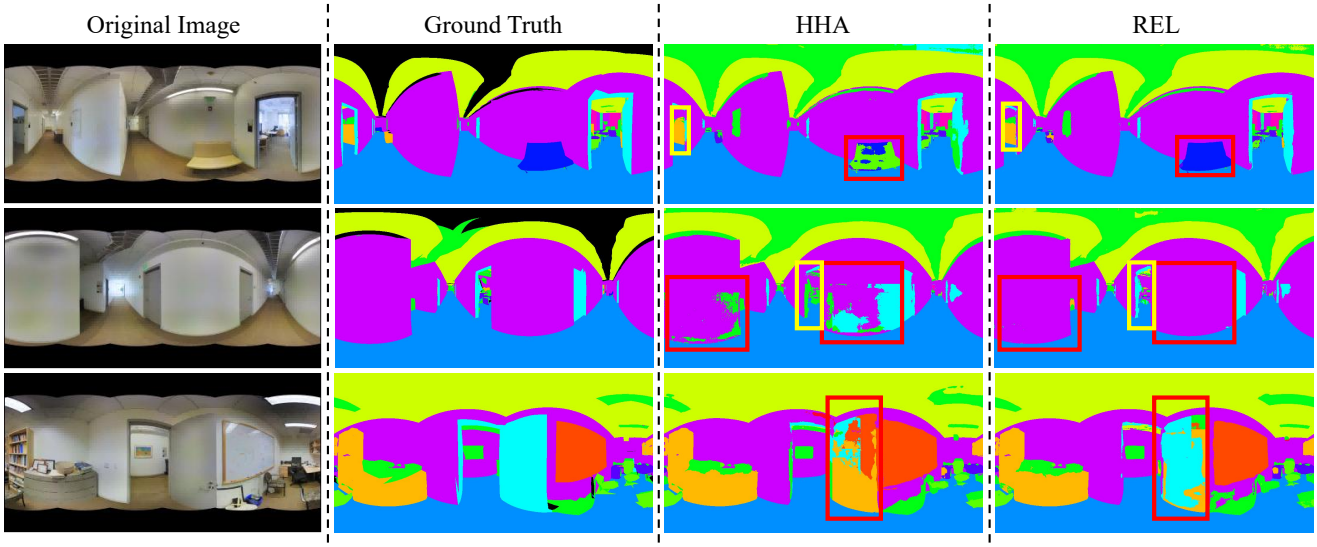


Figure S-2. More comparison of HHA and REL results. Similar to Figure 6 in the manuscript, REL reduces the noise in the red box when the details are clearer in the yellow box.

(β,γ,α) (°)	SGAT mIoU / PAcc Our mIoU / PAcc	(β,γ,α) (°)	SGAT mIoU / PAcc Our mIoU / PAcc	(β,γ,α) (°)	SGAT mIoU / PAcc Our mIoU / PAcc	(β,γ,α) (°)	SGAT mIoU / PAcc Our mIoU / PAcc
(0,0,0)	56.374 / 83.135 64.907 / 90.347	(0,5,0)	56.073 / 82.892 65.060 / 90.294	(5,0,0)	56.074 / 82.905 64.976 / 90.293	(5,5,0)	55.784 / 82.794 65.201 / 90.237
(0,0,90)	56.441 / 83.130 65.087 / 90.367	(0,5,90)	55.954 / 82.847 65.030 / 90.260	(5,0,90)	56.128 / 82.895 65.335 / 90.359	(5,5,90)	55.636 / 82.657 65.062 / 90.255
(0,0,180)	56.246 / 83.054 64.915 / 90.294	(0,5,180)	55.951 / 82.906 64.913 / 90.251	(5,0,180)	55.714 / 82.796 64.917 / 90.274	(5,5,180)	55.501 / 82.750 64.791 / 90.236
(0,0,270)	56.223 / 83.051 65.280 / 90.410	(0,5,270)	55.924 / 82.779 65.143 / 90.367	(5,0,270)	55.983 / 82.904 65.015 / 90.290	(5,5,270)	55.732 / 82.7011 65.011 / 90.312

Table S-1. Detail performance comparison with SGAT4PASS on Stanford2D3D Panoramic datasets fold 1 with SGA metrics. All 16 test situations are shown. "SGAT" indicates SGAT4PASS when "PAcc" means the pixel accuracy metric.

parison.

C.1. More Visualization of Different Depth Representation

In this part, we provide more visualization of HHA and REL. Similar to Figure 6 in the manuscript, as shown in Fig. S-2, results show that REL reduces the noise and makes details clearer.

C.2. More Results about SMMF Spherical Consistency

In Figure 7 in the manuscript, the probability of each strategy being selected in SMMF of Stanford2D3D Panoramic datasets fold 1 is shown. As shown in Fig. S-1, the results for the fold 2 and fold 3 of Stanford2D3D Panoramic datasets are similar to fold 1 shown in our manuscript.

D. Different m/n Influence

In this section, we show the influence of FLOPs and latency per image with a server with 8 NVIDIA GeForce RTX 3090 GPUs, which is shown in Tab. S-2. $m/n = 3/7$ (SMMF) corresponding to $m/n = 3/6$ (DyMM) or (CMX) is that one extra column is for seam-aware patches, and it has about 16 / 3 % more Flops / latency.

E. Performance on Other Datasets

We follow the Matterport3D setting in 360BEV [30] Table 3 (512*1024 input), the mIoU of HHA / REL+SMMF with CMX baseline is 47.37 / 48.05. Furthermore, we follow ToF-360 [19] Table 2 setting to evaluate on ToF-360 with 5792*2896 input, we use the same semantic-class mapping between Stanford2D3D Panoramic datasets and

m/n	mIoU	FLOPS	Latency	m/n	mIoU	FLOPS	Latency
3 / 6 (CMX)	64.47	4.61	0.399	3 / 5 (SMMF)	66.86	3.61	0.319
3 / 6 (DyMM)	66.73	4.35	0.405	3 / 9 (SMMF)	67.36	6.50	0.509
3 / 7 (SMMF)	67.37	5.06	0.416	5 / 11 (SMMF)	67.64	13.23	0.995

Table S-2. Cost comparison on Stanford2D3D Panoramic datasets fold 1 with RGB-HEL input and different m/n.

ToF360 as [19]. The mIoU of HHA / REL+SMMF with CMX is 21.83 / 25.61. Moreover, for ToF-360 [19] ParkingLot (outdoor scene), The mIoU of HHA / REL+SMMF with CMX is 24.38 / 38.42.