A ADDITIONAL RESULTS

A.1 Baseline Methods

We compared our method to several baseline camera pose estimation methods listed below.

- **OpenMVG+SIFT.** We used the reference implementation of panoramic camera pose estimation in OpenMVG with feature points extracted by the classic SIFT algorithm.
- **OpenMVG+LoFTR.** We also tried using a neural network-based LoFTR method to extract the feature points. Note that both SIFT and LoFTR are designed for planar domains (i.e., perspective cameras).
- **OpenMVG+SPHORB.** We tried using a state-of-the-art feature detection method, PHORB, that is designed specifically for spherical domains (i.e., panoramas).
- **LayoutLoc.** Recall that in the ZInD dataset, the authors described a simple algorithm, named "LayoutLoc", that automatically estimates a camera pose between two panoramas by matching room layouts (estimated by HorizonNet) in top-down views and other heuristics. As they did not provide code, we re-implemented the algorithm by ourselves.
- **CoVisPose.** This is a recent end-to-end neural method that jointly estimates the camera pose and room layout given two wide-baseline panoramas taken in the same room. They trained and tested on the ZInD dataset only. As they did not provide code for inference nor training, we directly used their reported performance on the ZInD dataset for comparison.
- **GPR-Net.** This is another recent neural method with very similar neural network designs and functions as CoVisPose. It was also trained and tested on the ZInD dataset only. Its reported performance was on pair with CoVisPose. The authors kindly provided code for training and inference and a pre-trained model on the ZInD dataset. However, as with CoVisPose, the method required ground truth room layouts for training, and the Matterport 3D dataset did not have them. Therefore, we used the ZInD pre-trained model on the Matterport 3D dataset for testing.

A.2 Automatic Selection Testing Results

We show how our approach would perform under the assumption that the best possible matching wall-wall pair is always selected. The testing results can be understood as the upper bounds for the performances that our approach may achieve given that the wall-wall pair selection problem is solved perfectly - either by interactive approaches or by a neural network that predicts the wall-wall selections in future work. Note that the user studies reported in Section A.3 demonstrated that the gap between the performances of human-based and automatic selections was narrow in practice.

Quantitative Results. Following the same evaluation methods in CoVisPose and GPR-Net, we compared the translation and rotation angular errors as well as the success rates versus several baseline methods. We omitted the comparisons of metric translation vectors because our method only calculates relative camera poses. Results on the ZInD dataset are shown in Table 3 and results on the M3D dataset are shown in Table 4. For the ZInD results, we further categorize the results by the gt co-visibility scores. For the M3D results, we show average results only (because co-visibility scores could not be calculated due to the lack of ground truth room layouts). To sum up, our method outperformed traditional (non-neural) methods by very large margins and performed similarly to CoVisPose and GPR-Net, both of which were trained on the ZInD dataset. However, on the M3D dataset that is unseen by both neural methods, our method outperformed GPR-Net (CoVisPose is unavailable for testing) by large margins. We consider the performance of GPR-Net on the M3D dataset as a proxy for the performance of CoVisPose on the M3D dataset.

Qualitative Results. We show qualitative comparisons of challenging cases in Figure 1 (M3D) and Figure 4 (both ZInD and M3D).

A.3 User Studies and Comparisons to Interactive Baseline

We asked 50 participants, who are mostly college students, to use our interactive system to conduct wall-wall matching selections. Every tester completed 200 cases. To prepare the cases, we randomly selected 100 panorama pairs from the ZInD dataset and calculated the room layouts estimated by HorizonNet and LGT-Net (100 cases each) while ensuring that every case has at least one good wall-wall matching. Our selected cases have diverse co-visibility scores. We recorded the user-selected wall-wall pairs and their times. Results are reported in Table 1. Overall, we observed that the users could select correct wall-wall matching with fairly good accuracy and the average time spent per case was low (less than 4 seconds).

We also implemented a straightforward *interactive baseline* method in which users manually select pairs of feature points by mouse clicks on the two panoramas. Tested by ourselves, it usually took tens of seconds to about 1 minute to conduct a feature point pairs selection session.

Avg. Time (min:sec)	Avg. Time poer case (sec)	Avg. Match Rate (%)
12:07	3.64	93.63%

Table 1: Statistics of the user study. We show the avg. time to finish 200 cases by the testers, the avg. time to finish one case (divide by 200), and the avg. ratio of selecting a correct wall-wall pair.

Method		Rotation			Translation angle		
Same height constraint	Wall normal extrusion	Mn(° ↓)	Med(° ↓)	2.5(° ↑)	Mn(° ↓)	Med(° ↓)	2.5(° ↑)
		10.91	6.78	24.53	12.10	7.94	20.39
V		9.72	6.22	24.90	11.77	7.75	20.68
	V	3.23	0.97	76.81	5.30	2.46	56.82
V	V	**2.48**	**0.66**	**88.89**	**4.37**	**1.80**	**64.42**

Table 2: Performances of alternative algorithm design choices to generate the feature point pairs given a selected wall-wall pair.

A.4 Ablation Studies

We compare the performances of alternative ways to generate the feature points pairs given a selected wall-wall pair in Table 2. We observed that the two key ideas of approach, including: 1) populating feature pairs along the wall normals and 2) fixing the heights to be the same, both significantly improved the performances.

Co-Vis.%	Method	Success(%↑)	Rotation			Translation angle		
			Mn(° ↓)	Med(° ↓)	2.5(° ↑)	Mn(° ↓)	Med(° ↓)	2.5(° ↑)
75-100	OpenMVG (SIFT)	75.14%	36.73	9.9	35.96	39.25	14.38	25.81
	OpenMVG (LoFTR)	71.75%	44.69	11.61	30.46	45.13	18.79	18.42
	OpenMVG (SPHORE)	73.59%	34.81	6.49	38.40	37.52	12.15	26.23
	Ours (HorizonNet)	100.00%	1.73	0.47	94.54	2.01	0.72	84.48
	Ours (LED2Net)	99.77%	2.07	0.53	93.49	3.42	1.22	72.52
	Ours (LGTNet)	100.00%	2.96	0.56	89.91	5.71	1.03	71.9
	LayouLoc	78.69%	13.13	0	70.19	14.86	1.46	51.12
	GPR-Net(Direct)	100.00%	*1.21	*0.78	*97.94	*6.86	*2.28	*53.95
	CoVisPose(Ransac)	99.73%	1.2	0.53	96.51	2.86	0.91	84.09
50-75	OpenMVG (SIFT)	61.28%	63.21	42.31	17.16	59.27	39.84	13.55
	OpenMVG (LoFTR)	51.25%	67.55	62.01	12.81	68.65	56.25	6.32
	OpenMVG (SPHORE)	56.57%	55.47	29.62	21.49	52.68	30.18	16.41
	Ours (HorizonNet)	100.00%	2.13	0.56	92.04	2.78	0.94	77.49
	Ours (LED2Net)	99.68%	5.53	0.79	85.99	11.91	2.82	46.93
	Ours (LGTNet)	100.00%	8.58	1.01	71.51	20.34	7.58	30.47
	LayouLoc	60.84%	41.64	0	40.17	38.57	4.26	26.18
	GPR-Net(Direct)	100.00%	*1.69	*0.76	*96.96	*4.01	*2.31	*53.96
	CoVisPose(Ransac)	99.22%	1.45	0.67	92.36	1.92	0.89	83.46
25-50	OpenMVG (SIFT)	52.56%	84.15	77.31	5.23	75.32	65.35	4.69
	OpenMVG (LoFTR)	47.19%	77.9	81.29	5.56	80.05	75.72	2.55
	OpenMVG (SPHORE)	43.84%	73.84	60.51	7.87	67.23	52.99	5.41
	Ours (HorizonNet)	100.00%	2.35	0.66	87.86	4.94	1.92	56.38
	Ours (LED2Net)	99.35%	5.89	0.87	83	17.72	7.99	25.82
	Ours (LGTNet)	100.00%	9.02	1.09	71.75	23.21	12.41	17.27
	LayouLoc	49.85%	77.39	90	18.57	63.4	50.52	8.27
	GPR-Net(Direct)	100.00%	*3.50	*0.77	*94.76	*5.33	*2.60	*48.43
	CoVisPose(Ransac)	96.42%	2.51	0.98	80.02	2.19	1.00	77.49
10-25	OpenMVG (SIFT)	49.44%	93.8	92.39	2.34	84.36	80.74	2.08
	OpenMVG (LoFTR)	45.03%	86.35	87.26	4.88	86.71	85.44	1.46
	OpenMVG (SPHORE)	33.63%	91.84	88.46	3.31	78.66	72.34	2.40
	Ours (HorizonNet)	100.00%	2.93	0.83	83.46	7.59	3.56	40.75
	Ours (LED2Net)	99.32%	4.87	0.93	82.94	20.24	13.28	15.55
	Ours (LGTNet)	100.00%	7.36	1.12	71.6	24.47	14.44	13.27
	LayouLoc	46.59%	91.3	90	11.85	77.21	70.11	2.19
	GPR-Net(Direct)	-	-	-	-	-	-	-
	CoVisPose(Ransac)	88.46%	6.18	1.78	54.36	4.82	1.59	57.66

Table 3: Performances of relative camera pose estimations stratified by co-visibility on the ZInD dataset. Same as in previous work, we report the mean ("Mn") and median ("Med") angular rotation and translation errors are reported in degrees, and the ratios of testing cases of for which the angular errors were less than 2.5 degrees. Highlights: 1st, 2nd and 3rd best results.

Region	Method	Success(%↑)	Rotation			Translation angle		
			Mn(° ↓)	Med(° ↓)	2.5(° ↑)	Mn(° ↓)	Med(° ↓)	2.5(° ↑)
Avg.	OpenMVG (SIFT)	61.28%	38.65	6.26	40.73	36.15	9.05	38.39
	OpenMVG (LoFTR)	59.52%	38.86	10.78	28.64	43.65	18.69	19.07
	OpenMVG (SPHORE)	56.51%	41.77	10.81	30.04	39.61	16.97	24.30
	Ours (HorizonNet)	100.00%	4.02	1.35	77.44	7.71	2.56	50.03
	Ours (LED2Net)	99.82%	4.02	1.41	76.13	7.13	2.66	48.94
	Ours (LGTNet)	100.00%	3.67	1.46	75.21	6.78	2.41	52.23
	LayouLoc	-	83.41	90	25.18	71.44	65.09	4.52
	GPR-Net	100.00%	37.73	19.1	6.94	73.35	76.3	3.65

Table 4: Quantitative results of relative pose estimation on the Matterport dataset. We use the same notations as in Table 3.