Graph-CoVis: GNN-based Multi-view Panorama Global Pose Estimation: Supplementary Material

Negar Nejatishahidin*†[1], Will Hutchcroft*[2], Manjunath Narayana[2], Ivaylo Boyadzhiev[2], Yuguang Li[2], Naji Khosravan[2], Jana Košecká[1], Sing Bing Kang[2]

[1]George Mason University [2]Zillow Group

1. Group size and connectivity

To further investigate the results, we evaluate the performance of the model on the data set separated into group size (number of panoramas) and percentage of co-visible connections between the panoramas. Table 1 shows results of this analysis. The main takeaway is that the rotation and translation errors are low when there are a large number of training examples. For datasets with more than 500 training examples, the maximum mean translation error for Graph-CoVis is $0.13m$, whereas the best performing baseline has a maximum error of $0.11m$. For datasets with fewer than 100 training examples, the minimum mean translation error for Graph-CoVis is $0.341m$, whereas the best performing baseline has a minimum error of only $0.14m$. Thus, Graph-CoVis performs worse for datasets with very few training examples.

Figure 1 provides a visual representation of the ATE and ARE as a function of connectivity percentage for the three groups. As the connectivity percentage increases, we observe an improvement in the performance of Graph-CoVis as well as the baselines.

Figure 2 shows that the number of training examples is larger for higher connectivity percentages. While the baselines use an optimization step to obtain global poses, Graph-CoVis must learn the global poses in an end-to-end fashion, requiring sufficient training data across the spectrum of input cases. As such, we believe that increasing the number of training examples among the low connectivity percentage sets will benefit Graph-CoVis and further improve its performance.

2. Qualitative examples

Figures 3 and 4 show some qualitative examples of where our system performs significantly better and moderately better than baseline approaches. Fig 5 shows examples when our system performs worse than the baselines.

3. Pose Graph Optimization

We perform Pose Graph Optimization (PGO) using GT-SAM [1]. We apply a diagonal Gaussian noise model on the prior constraint to specify the origin node, with standard deviations of 20 cm and 0.1 radians for translation and rotation, respectively. We also apply the same model for the odometry noise, with standard deviations of 30cm and .3 radians. We use the Levenberg-Marquardt optimizer with 1000 iterations, with a relative error tolerance of 1×10^{-5} for the convergence criteria.

4. Other Baselines

A fair criticism may be made of our paper in that it could be possible to compare to a more extensive set of baselines based on additional recently published papers on multi-view image and panorama pose estimation. Notably, three recent works PoGO-Net [3], SALVe [2], and Extreme SfM [4] are relevant.

As explained in the paper, Extreme SfM and SALVe solve the problem of extreme-wide baseline pose estimation, subject to little-to-no visual overlap, to estimate floor level reconstruction of indoor spaces. Extreme SfM in particular focuses on the difficult cases where a single panorama is captured per room and "*seeks to align images from different rooms by exploiting the regularities of room arrangement at a house-scale*". Since our problem consists of multiple panoramas within the *same large space and not at a house-scale*, we do not consider Extreme SfM as a directly comparable baseline approach.

SALVe is similar in application to Extreme SfM in its end goal of floor plan reconstruction, but by the nature of the method and input data distribution, it is a stronger candidate baseline. It handles all the panoramas captured in a floor of a home, which include multiple visually connected

[*] Equal contribution.

[†] Done during Negar Nejatishahidin's internship at Zillow.

Group-Size	%Connection	#Test	#Train	Methods	Rotation			Translation		
					Mn (° ↓)	Med (° ↓)	Std(° ↓)	Mn (m. ↓)	Med (m. ↓)	Std(m. ↓)
Three	66%	52	108	CoVisPose + Greedy	7.849	1.991	18.743	**0.308**	**0.095**	**0.641**
				CoVisPose + PGO	15.744	7.971	21.691	0.685	0.283	1.218
				Graph-CoVis	**5.362**	**1.364**	**15.923**	0.340	0.124	0.993
	100%	1203	2886	CoVisPose + Greedy	2.423	1.007	10.944	0.084	0.051	0.205
				CoVisPose + PGO	2.612	0.948	11.386	0.084	0.046	0.228
				Graph-CoVis	**1.856**	**0.833**	**8.706**	**0.069**	**0.037**	**0.208**
Four	50%	3	10	CoVisPose + Greedy	53.348	34.067	60.672	1.908	1.642	**1.856**
				CoVisPose + PGO	51.777	**25.058**	59.321	**1.614**	**0.867**	1.936
				Graph-CoVis	**45.675**	27.136	**57.675**	2.638	2.113	2.478
	66%	38	73	CoVisPose + Greedy	**7.048**	**1.920**	**22.248**	**0.248**	**0.099**	0.598
				CoVisPose + PGO	21.626	8.032	28.184	0.671	0.277	1.156
				Graph-CoVis	10.694	2.099	25.795	0.391	0.166	**0.567**
	83%	67	153	CoVisPose + Greedy	**4.536**	1.413	**13.568**	**0.204**	**0.081**	**0.418**
				CoVisPose + PGO	12.131	4.962	19.478	0.490	0.180	0.910
				Graph-CoVis	6.411	**1.204**	20.784	0.282	0.110	0.603
	100%	437	1160	CoVisPose + Greedy	3.199	1.069	15.071	0.111	0.064	0.256
				CoVisPose + PGO	3.429	1.045	13.949	0.127	0.056	0.319
				Graph-CoVis	**1.754**	**0.870**	**7.397**	**0.095**	**0.052**	**0.226**
Five	50%	3	2	CoVisPose + Greedy	**2.739**	**2.472**	**2.053**	**0.140**	**0.115**	**0.083**
				CoVisPose + PGO	26.961	5.718	35.364	0.830	0.559	0.777
				Graph-CoVis	8.262	1.046	10.001	0.504	0.418	0.412
	60%	14	39	CoVisPose + Greedy	10.203	**2.961**	18.762	0.539	**0.145**	0.994
				CoVisPose + PGO	27.103	12.960	34.467	0.810	0.478	0.956
				Graph-CoVis	**9.439**	3.542	**18.216**	**0.528**	0.217	**0.924**
	70%	26	64	CoVisPose + Greedy	**8.096**	2.114	26.835	**0.196**	**0.132**	**0.212**
				CoVisPose + PGO	20.244	11.380	23.232	0.721	0.267	1.190
				Graph-CoVis	8.303	**1.889**	**17.971**	0.341	0.162	0.627
	80%	44	117	CoVisPose + Greedy	4.155	**1.211**	14.433	**0.169**	**0.080**	0.285
				CoVisPose + PGO	18.725	8.783	24.781	0.471	0.232	0.658
				Graph-CoVis	**3.311**	1.236	**13.439**	0.181	0.106	**0.218**
	90%	46	149	CoVisPose + Greedy	2.611	1.516	7.474	0.160	**0.078**	0.408
				CoVisPose + PGO	7.975	3.258	13.829	0.339	0.138	0.639
				Graph-CoVis	**2.354**	**1.022**	**6.865**	**0.151**	0.098	**0.181**
	100%	219	609	CoVisPose + Greedy	2.584	1.107	11.986	**0.120**	0.070	**0.244**
				CoVisPose + PGO	3.368	1.028	12.599	0.139	**0.063**	0.367
				Graph-CoVis	**2.433**	**0.948**	**10.915**	0.128	0.064	0.319

Table 1. Mean rotation and translation error for different group sizes, separated into sub-sets based on connectivity (percentage) between panoramas. The number of training and test examples are shown for each sub-set.

panoramas in a single space and visually not-connected panoramas across rooms. The former case is the same as our multi-view setting. SALVe uses separately trained depth and room layout estimation networks followed by a geometric alignment of the top-down projections of the room layouts. A deep network then verifies if the top-down projections are plausible. Finally a pose graph optimization algorithm estimates the global poses of all the panoramas that were connected by the alignment and verification steps. In theory, one could apply required modifications to the SALVe system to run on smaller sub-sets of panoramas once the SALVe code becomes available.

Finally, PoGO-Net is a GNN-based alternative to pose graph optimization. Applied to perspective images, it requires a preprocessing step of generating an initial view-graph that is subsequently refined and filtered by a GNN to estimate the final poses. A possible preprocessing approach is to estimate two-view poses using state-of-the-art for two-view panorama pose (which is CoVisPose) followed by pose graph optimization. This is indeed one of the chosen comparison baselines in our paper. Upon PoGO-Net becoming publicly available, it's accuracy can be directly compared to Graph-Covis by applying it on the view-graph that results from running the CoVisPose + Greedy baseline.

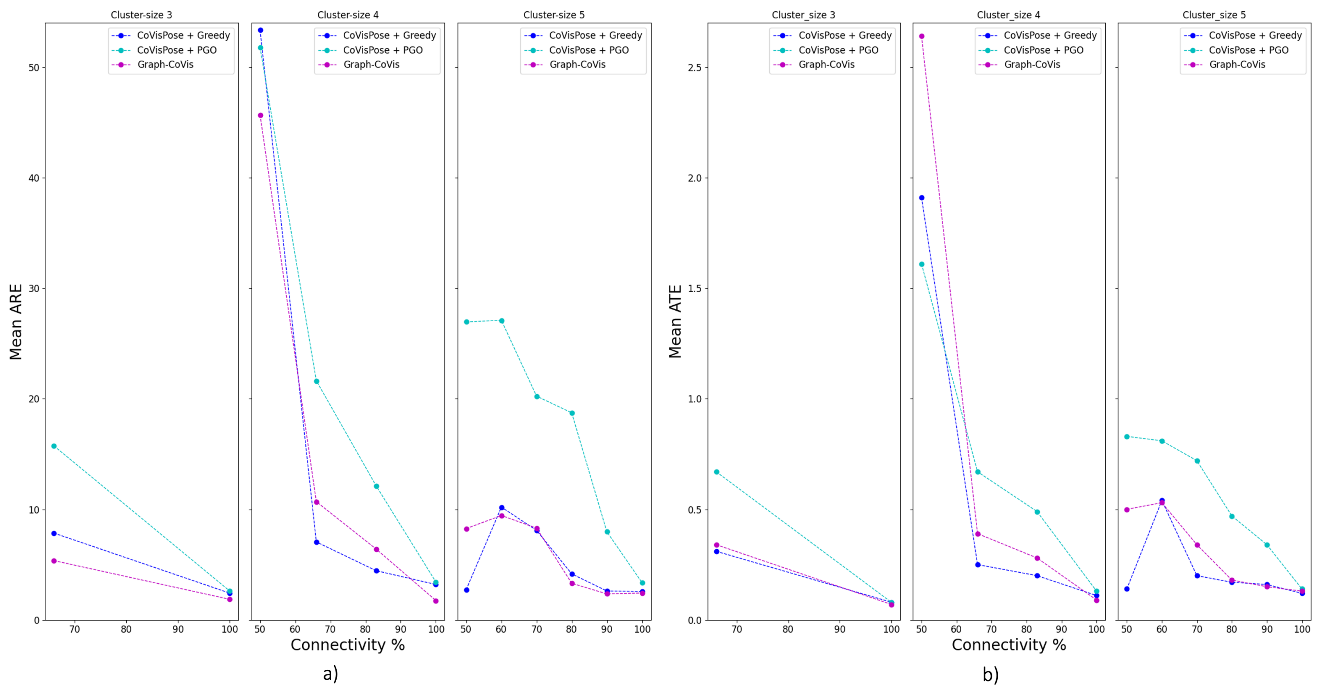

Figure 1. Connectivity percentage plotted against a) Mean ARE and b) Mean ATE in the graph for group size of three, four, and five.

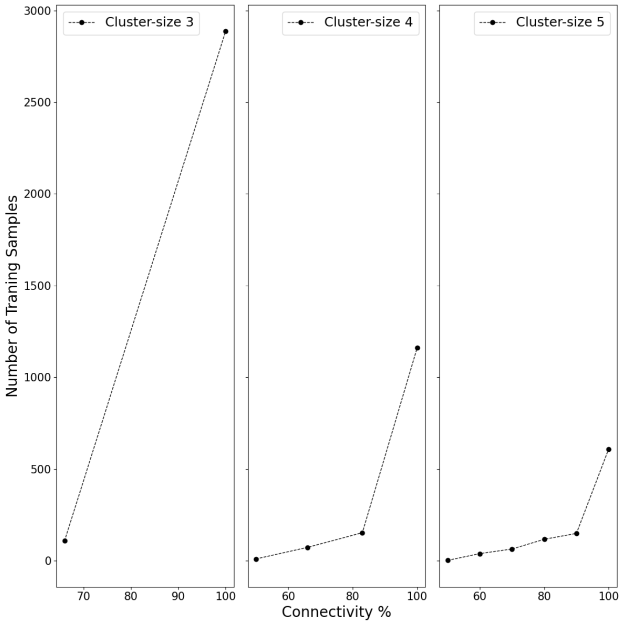

Figure 2. Connectivity percentage plotted against the number of training samples for the trained model, separated by group size of three, four, and five.

References

[1] Frank Dellaert. Factor graphs and GTSAM: A hands-on introduction. Technical report, Georgia Institute of Technology, 2012. 1

[2] John Lambert, Yuguang Li, Ivaylo Boyadzhiev, Lambert E. Wixson, Manjunath Narayana, Will Hutchcroft, James Hays, Frank Dellaert, and Sing Bing Kang. Salve: Semantic alignment verification for floorplan reconstruction from sparse panoramas. In *ECCV*, 2022. 1

[3] Xinyi Li and Haibin Ling. Pogo-net: Pose graph optimization with graph neural networks. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5875–5885, 2021. 1

[4] Mohammad Amin Shabani, Weilian Song, Makoto Odamaki, Hirochika Fujiki, and Yasutaka Furukawa. Extreme structure from motion for indoor panoramas without visual overlaps. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5683–5691, 2021. 1

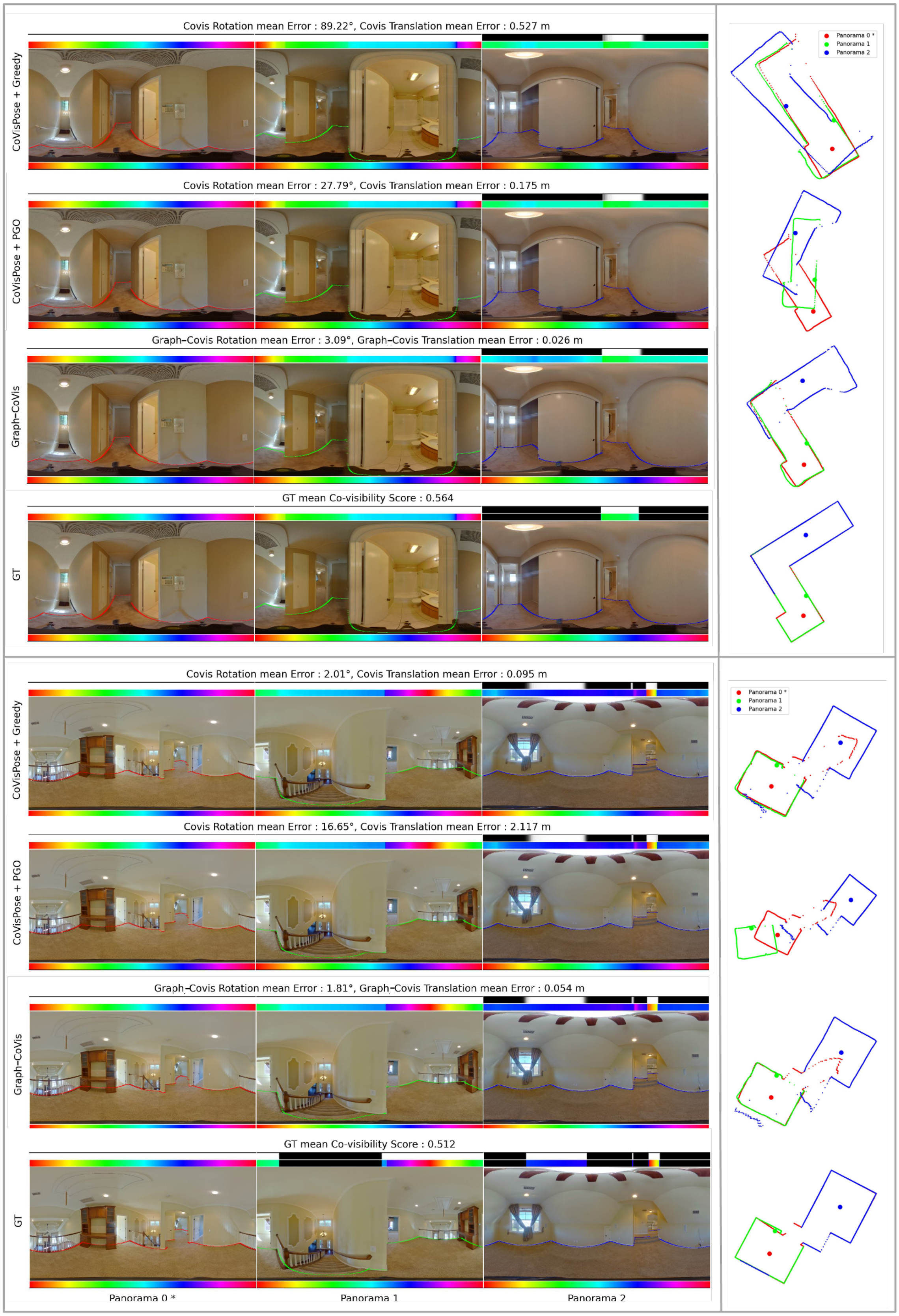
Covis Rotation mean Error : 89.22°, Covis Translation mean Error : 0.527 m
CoVisPose + Greedy
Covis Rotation mean Error : 27.79°, Covis Translation mean Error : 0.175 m
CoVisPose + PGO
Graph–Covis Rotation mean Error : 3.09°, Graph–Covis Translation mean Error : 0.026 m
Graph–CoVis
GT mean Co-visibility Score : 0.564
GT
Panorama 0 *
Panorama 1
Panorama 2
Covis Rotation mean Error : 2.01°, Covis Translation mean Error : 0.095 m
CoVisPose + Greedy
Covis Rotation mean Error : 16.65°, Covis Translation mean Error : 2.117 m
CoVisPose + PGO
Graph–Covis Rotation mean Error : 1.81°, Graph–Covis Translation mean Error : 0.054 m
Graph–CoVis
GT mean Co-visibility Score : 0.512
GT
Panorama 0 *
Panorama 1
Panorama 2
Panorama 0 *
Panorama 1
Panorama 2

Figure 3. Two examples where Graph-CoVis performs significantly better than baselines.

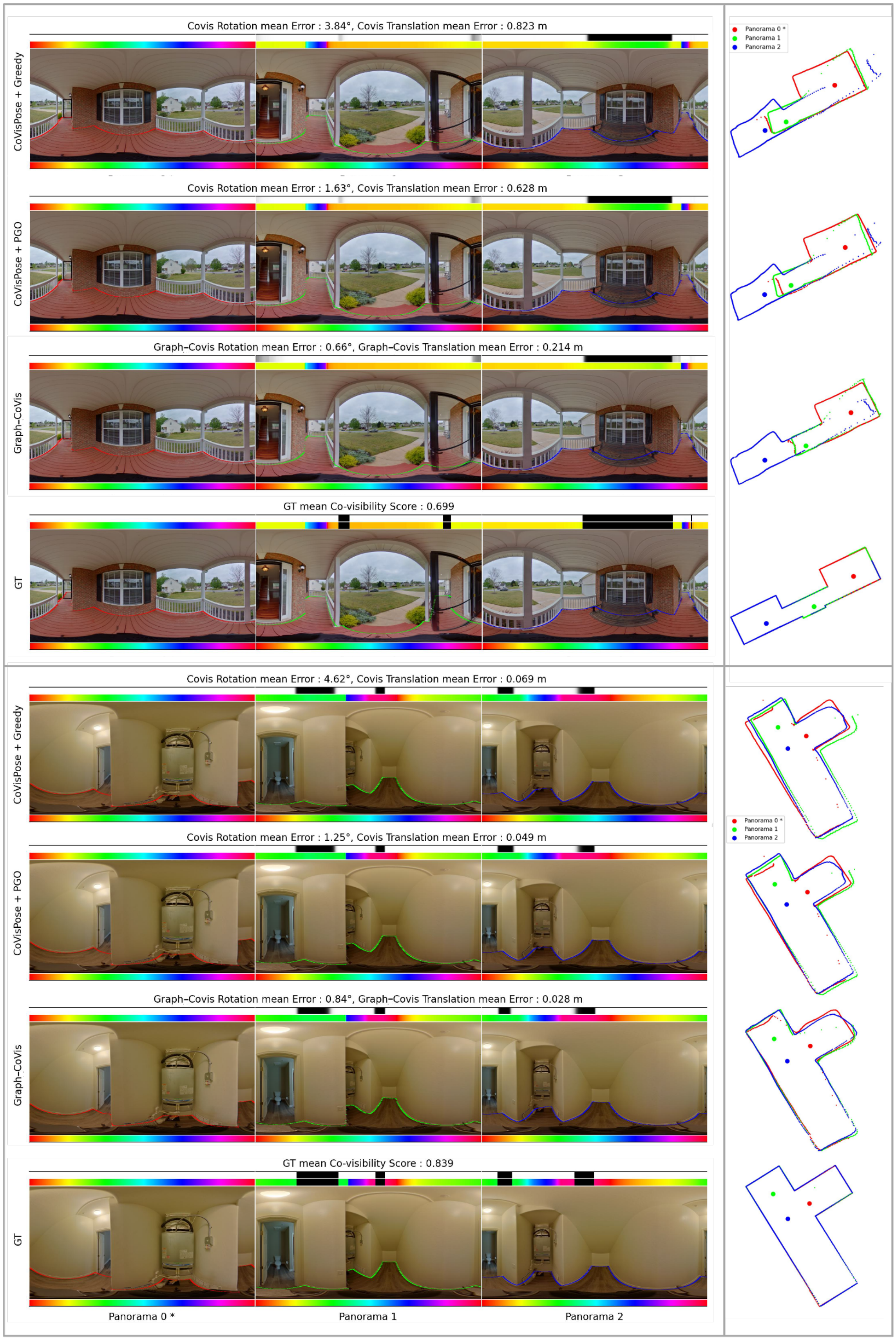

Figure 4. Two examples where Graph-CoVis performs moderately better than baselines

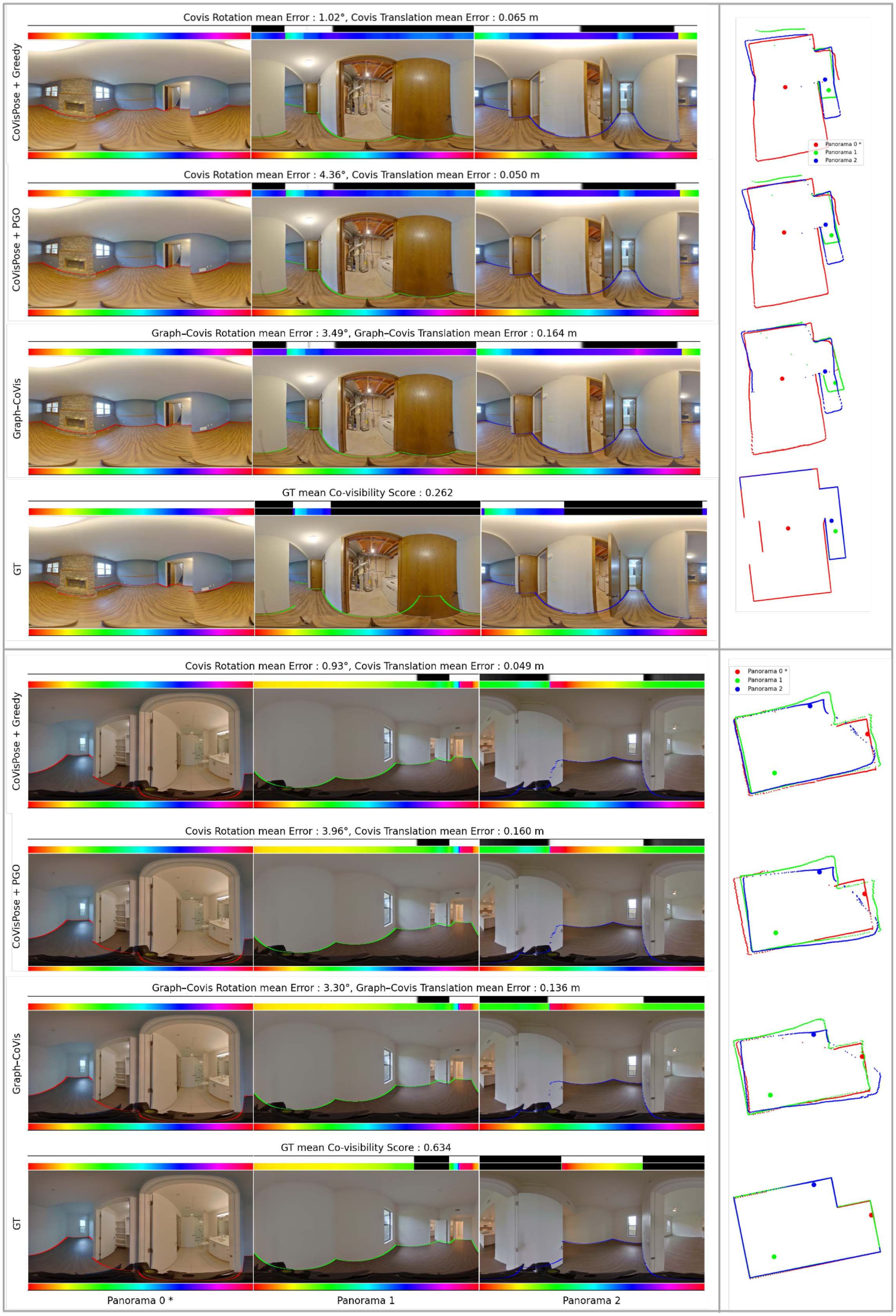
Covis Rotation mean Error : 1.02°, Covis Translation mean Error : 0.065 m
CoVisPose + Greedy
Covis Rotation mean Error : 4.36°, Covis Translation mean Error : 0.050 m
CoVisPose + PGO
Graph–Covis Rotation mean Error : 3.49°, Graph–Covis Translation mean Error : 0.164 m
Graph–CoVis
GT mean Co-visibility Score : 0.262
GT
Panorama 0 *
Panorama 1
Panorama 2
Covis Rotation mean Error : 0.93°, Covis Translation mean Error : 0.049 m
CoVisPose + Greedy
Covis Rotation mean Error : 3.96°, Covis Translation mean Error : 0.160 m
CoVisPose + PGO
Graph–Covis Rotation mean Error : 3.30°, Graph–Covis Translation mean Error : 0.136 m
Graph–CoVis
GT mean Co-visibility Score : 0.634
GT
Panorama 0 *
Panorama 1
Panorama 2
Panorama 0 *
Panorama 1
Panorama 2

Figure 5. Two examples where Graph-CoVis performs worse than baselines.