vMAP: Vectorised Object Mapping for Neural Field SLAM (Supplementary Material)

Xin Kong Shikun Liu Marwan Taher Andrew J. Davison
Dyson Robotics Lab, Imperial College London
{x.kong21, shikun.liu17, m.taher, a.davison}@imperial.ac.uk

1. Interactive Visualisation

We recommend readers to check out our project website https://kxhit.github.io/vMAP, showing the real-time scene-level and object-level reconstructions of some selected sequences.

2. Implement Details and Discussions

Depth-Guided Sampling As described in the main paper, we sampled more points near the object surface guided by the depth measurements. For rays that go through the 3D object bounding box but do not belong to the current instance, we then terminate these rays when they hit the object surface, to minimise the impact on the occluded objects, similar to ObjectNeRF [2]. A visualisation of depth guided sampling is shown in Fig. 1, and the sampled points are coloured by the measured depth.

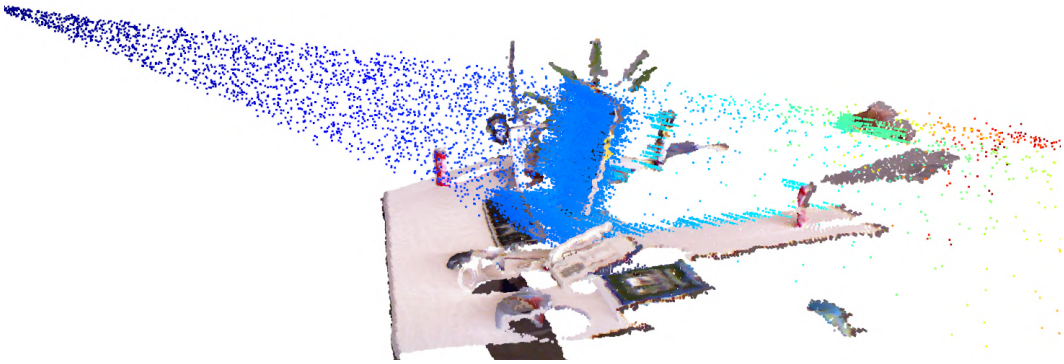

Figure 1. Visualisation of depth guided sampling.

Object-Level Positional Encoding Since object instances are different in size, the reconstruction quality can be maximised when trained with a suitable positional encoding frequency. Otherwise, the network training would be biased towards reconstructing large objects and overlook small objects or vice versa. To mitigate this scaling issue, we applied integrated positional encoding [1] and introduced an additional hyper-parameter, the scaling factor s, which is applied to all objects, such that they are bounded in a unit box within the range of $[-1, 1]$. We separately set this scaling factor slightly larger in the background model.

This scaling factor can be set as object specific if such object-specific prior is known, i.e. we can set a large s when training the object 'sofa', and a small s when training the object 'cup', because a sofa is typically larger than a cup. A visualisation of the object reconstruction with different choices of s is shown in Fig. 2. We can see a large scale s results a smoother geometry which is more suitable for reconstructing large objects like 'walls' and 'blankets', and a small s is more suitable for objects with complex geometries like 'chairs'.

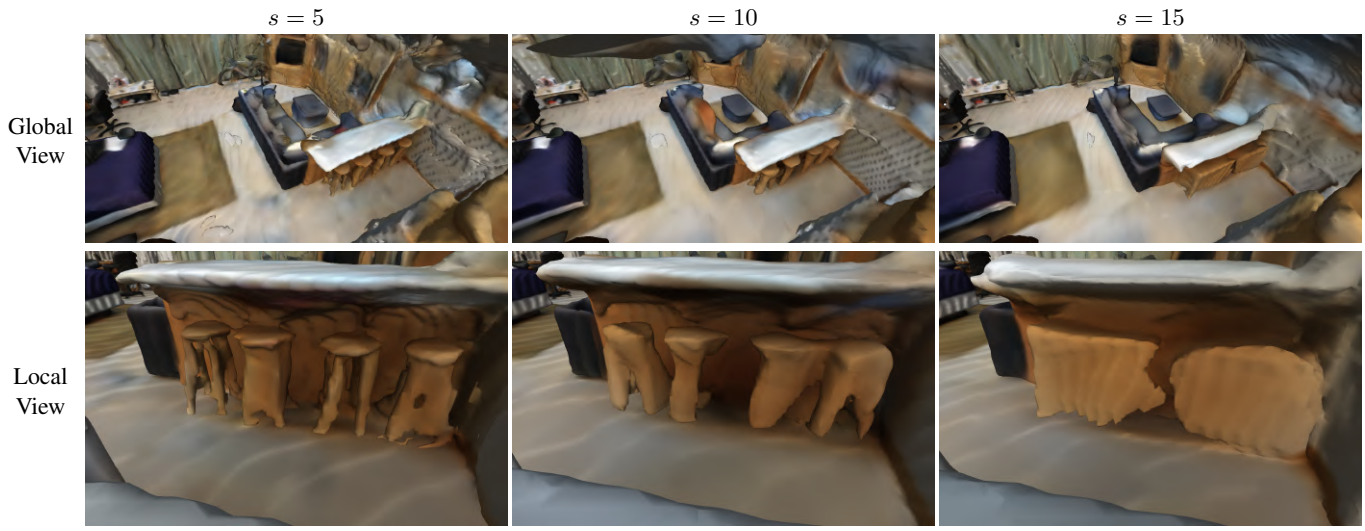

Figure 2. Visualisation of 3D object reconstructions trained with different scales.

3. Additional Experimental Results on Replica Scenes

In Tab. 1 and Tab. 2, we listed the detailed scene-level and object-level 3D reconstruction results for each sequence on Replica dataset.

		room-0	room-1	room-2	office-0	office-1	office-2	office-3	office-4	Avg.
TSDF-Fusion*	**Acc.** [cm] ↓	1.46	1.13	1.22	1.13	0.91	1.33	1.56	1.48	1.28
	Comp. [cm]	3.73	3.51	4.41	10.26	9.57	5.50	3.87	4.04	5.61
	Comp. Ratio [< 5cm %]	86.54	87.12	84.87	78.86	75.85	80.48	83.19	84.41	82.67
iMAP	**Acc.** [cm] ↓	3.58	3.69	4.68	5.87	3.71	4.81	4.27	4.83	4.43
	Comp. [cm] ↓	5.06	4.87	5.51	6.11	5.26	5.65	5.45	6.59	5.56
	Comp. Ratio [< 5cm %] ↑	83.91	83.45	75.53	77.71	79.64	77.22	77.34	77.63	79.06
iMAP*	**Acc.** [cm] ↓	2.06	1.65	1.92	2.36	1.94	2.61	2.41	2.23	2.15
	Comp. [cm] ↓	2.21	1.94	2.77	4.81	3.19	2.81	2.78	2.56	2.88
	Comp. Ratio [< 5cm %] ↑	94.93	94.87	90.78	86.85	87.79	89.61	90.54	91.46	90.85
NICE-SLAM	**Acc.** [cm] ↓	2.69	2.49	2.55	3.03	3.31	3.56	3.26	2.63	2.94
	Comp. [cm] ↓	2.92	2.33	2.96	8.34	5.18	3.35	3.37	3.68	4.02
	Comp. Ratio [< 5cm %] ↑	90.77	93.07	87.83	81.99	82.24	85.82	85.44	86.64	86.73
NICE-SLAM*	**Acc.** [cm] ↓	2.71	2.28	2.69	2.93	4.23	3.45	3.26	2.74	3.04
	Comp. [cm] ↓	2.84	2.23	3.02	7.54	4.52	3.31	3.58	3.64	3.84
	Comp. Ratio [< 5cm %] ↑	91.00	93.37	87.23	82.70	82.09	85.42	84.28	86.10	86.52
vMAP	**Acc.** [cm] ↓	2.77	3.87	1.83	4.82	3.51	3.35	3.19	2.26	3.20
	Comp. [cm] ↓	1.99	1.81	2.00	3.65	2.14	2.45	2.49	2.56	**2.39**
	Comp. Ratio [< 5cm %] ↑	97.10	96.59	95.72	87.53	85.08	94.70	93.65	93.56	**92.99**

Table 1. Scene-level reconstruction results on 8 indoor Replica scenes. $*$ represents the baselines we re-trained with ground-truth pose.

		room-0	room-1	room-2	office-0	office-1	office-2	office-3	office-4	Avg.
TSDF-Fusion*	**Acc.** [cm] ↓	0.43	0.45	0.45	0.49	0.43	0.41	0.45	0.52	**0.45**
	Comp. [cm] ↓	3.03	4.42	4.16	2.23	5.60	3.32	3.31	3.48	3.69
	Comp. Ratio [< 1cm %] ↑	62.27	59.44	53.57	68.16	55.73	67.34	63.35	63.75	61.70
	Comp. Ratio [< 5cm %] ↑	84.34	79.12	76.89	86.74	80.36	87.30	85.74	83.38	82.98
NICE-SLAM*	**Acc.** [cm] ↓	3.48	3.77	4.61	4.08	3.42	3.45	3.96	4.53	3.91
	Comp. [cm] ↓	2.51	2.82	3.19	3.05	3.29	3.47	3.61	4.23	3.27
	Comp. Ratio [< 1cm %] ↑	41.19	37.06	33.03	38.86	44.55	41.84	31.21	34.54	37.79
	Comp. Ratio [< 5cm %] ↑	86.88	86.43	83.96	84.54	89.08	83.77	79.40	77.64	83.96
iMAP*	**Acc.** [cm] ↓	3.02	3.35	4.50	3.84	2.62	3.22	3.58	4.43	3.57
	Comp. [cm] ↓	1.71	1.93	3.45	1.66	2.58	2.28	2.32	3.14	2.38
	Comp. Ratio [< 1cm %] ↑	52.57	43.56	45.06	48.16	48.93	53.59	51.07	39.36	47.79
	Comp. Ratio [< 5cm %] ↑	93.72	92.95	85.30	94.56	91.09	89.99	89.32	84.60	90.19
vMAP	**Acc.** [cm] ↓	2.18	3.46	2.01	2.37	2.27	1.75	1.90	1.93	2.23
	Comp. [cm] ↓	1.13	1.54	1.58	1.15	1.77	1.03	1.42	1.94	**1.44**
	Comp. Ratio [< 1cm %] ↑	74.09	68.51	66.81	67.00	65.24	77.98	68.62	65.56	**69.23**
	Comp. Ratio [< 5cm %] ↑	96.68	95.02	92.98	96.53	92.94	96.97	94.21	91.03	**94.55**

Table 2. Object-level reconstruction results on 8 indoor Replica scenes. ∗ represents the baselines we re-trained with ground-truth pose.

We generated a new / different sequence for each scene in Replica dataset. We performed 2D novel view synthesis and compared it to the ground-truth views from the generated sequence. We compared baselines in depth L1 error, PSNR, SSIM, and LPIPS in Tab. 3, and the 2D renderings for 3 selected scenes are shown in Fig. 3.

		room-0	room-1	room-2	office-0	office-1	office-2	office-3	office-4	Avg.
NICE-SLAM	**Depth L1.** [cm] ↓	1.99	1.57	2.72	12.50	7.37	3.03	2.39	2.18	4.22
	PSNR. ↑	24.11	23.43	23.48	23.91	22.69	23.78	23.78	26.00	23.90
	SSIM ↑	0.73	0.74	0.82	0.83	0.82	0.83	0.84	0.85	0.81
	LPIPS ↓	0.11	0.09	0.09	0.15	0.28	0.11	0.10	0.09	0.13
NICE-SLAM*	**Depth L1.** [cm] ↓	1.87	1.63	2.94	13.43	7.63	2.83	2.62	1.97	4.36
	PSNR. ↑	24.03	23.61	23.54	23.59	23.19	22.22	23.32	26.20	23.71
	SSIM ↑	0.73	0.75	0.82	0.83	0.84	0.85	0.84	0.86	0.82
	LPIPS ↓	0.11	0.09	0.09	0.16	0.26	0.10	0.10	0.09	0.13
iMAP*	**Depth L1.** [cm] ↓	1.23	2.16	2.53	13.29	5.14	2.31	1.77	1.44	3.73
	PSNR. ↑	25.83	25.51	25.22	24.17	23.94	24.02	25.45	29.13	**25.41**
	SSIM ↑	0.77	0.79	0.86	0.83	0.87	0.88	0.89	0.90	**0.85**
	LPIPS ↓	0.09	0.07	0.07	0.17	0.22	0.08	0.07	0.07	**0.11**
vMAP	**Depth L1.** [cm] ↓	1.68	1.57	2.37	7.73	6.60	2.50	2.30	1.85	**3.33**
	PSNR. ↑	25.23	25.27	24.31	23.78	23.59	23.10	23.83	27.91	24.63
	SSIM ↑	0.77	0.78	0.85	0.84	0.88	0.88	0.88	0.89	**0.85**
	LPIPS ↓	0.09	0.07	0.08	0.16	0.23	0.07	0.08	0.07	**0.11**

Table 3. 2D novel view synthesis rendering results on the Replica dataset.

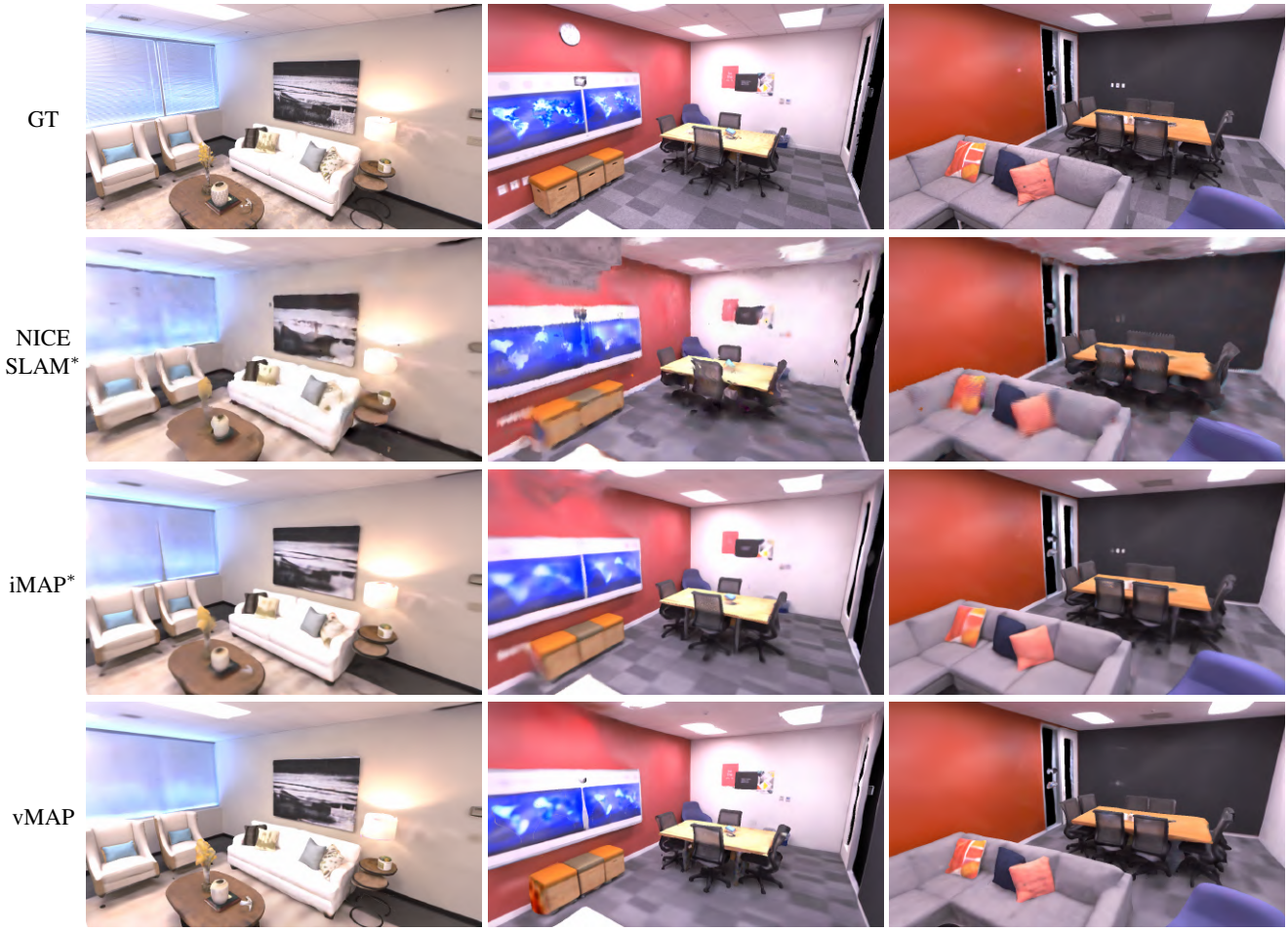

Figure 3. Visualisation of 2D novel view synthesis rendering results on the Replica dataset, better when zoomed.

4. Visualisation of Object-level Hole-filling

Compared to iMAP and NICE-SLAM, vMAP shows significantly better hole-filling capability in unobserved regions with visual consistency, thanks to the disentangled object representation design. As shown in Fig. 4, vMAP is able to generate smooth and natural geometries without requiring any other priors.

Figure 4. Visualisation of object-level hole-filling.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[2] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1