

A Tangent space of $\text{SL}(4)$

Here, we provide the explicit 15 generators, $\mathbf{G}_k \forall k : \{1 : 15\}$, of $\text{SL}(4)$, which allow us to relate the Lie algebra $\mathfrak{sl}(4)$ to the Lie group $\text{SL}(4)$.

The tangent space of $\text{SL}(4)$ consists of all 4×4 real matrices with zero trace. Thus, there are 15 generators, \mathbf{G}_k , where 12 of them are defined as \mathbf{E}_{ab} for $a \neq b$ where 1 is in the (a, b) entry and 0, elsewhere. The remaining three generators are $\mathbf{B}_1 = \text{diag}(1, -1, 0, 0)$, $\mathbf{B}_2 = \text{diag}(0, 1, -1, 0)$, $\mathbf{B}_3 = \text{diag}(0, 0, 1, -1)$. Explicitly, the generators are as follows:

$$\begin{aligned} \mathbf{G}_1 = \mathbf{E}_{01} &= \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & \mathbf{G}_2 = \mathbf{E}_{02} &= \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & \mathbf{G}_3 = \mathbf{E}_{03} &= \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \\ \mathbf{G}_4 = \mathbf{E}_{10} &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & \mathbf{G}_5 = \mathbf{E}_{12} &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & \mathbf{G}_6 = \mathbf{E}_{13} &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \\ \mathbf{G}_7 = \mathbf{E}_{20} &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & \mathbf{G}_8 = \mathbf{E}_{21} &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & \mathbf{G}_9 = \mathbf{E}_{23} &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \\ \mathbf{G}_{10} = \mathbf{E}_{30} &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, & \mathbf{G}_{11} = \mathbf{E}_{31} &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, & \mathbf{G}_{12} = \mathbf{E}_{32} &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \\ \mathbf{G}_{13} = \mathbf{B}_1 &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & \mathbf{G}_{14} = \mathbf{B}_2 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & \mathbf{G}_{15} = \mathbf{B}_3 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \end{aligned}$$

Thus, as briefly explained in Sec. 4.4 the relation between the Lie algebra, $\xi^\wedge \in \mathfrak{sl}(4)$, and the Lie group $\mathbf{H} \in \text{SL}(4)$ is given by:

$$\mathbf{H} = \exp(\xi^\wedge) = \exp\left(\sum_{k=1}^{15} \xi_k \mathbf{G}_k\right). \quad (6)$$

B Extra Quantitative Results

We provide addition results of evaluating on the 7-Scenes [60] and TUM RGB-D [65] datasets where we experiment with different submap sizes (Appendix B.1) and show the number of submaps and loop closures per scene (Appendix B.2).

B.1 Evaluation with different submap sizes

Here we show results for the 7-Scenes and TUM RGB-D datasets in Tables 5 and 6 with different submap sizes ($w = 8, 16, 32$). For 7-Scenes, we also include results for $w = 1$. Recall that w is the size of new images in the submap, so in the case of $w = 1$, each submap has one new image, one image from the prior submap, and up to one extra image from loop closures. For small submap size of $w = 1$, the backend becomes numerically unstable for some TUM scenes (consistently floor and 360) preventing an estimated alignment, and thus we do not include the $w = 1$ for TUM. This is due to reasons discussed in Sec. 6. Particularly, for the floor scene there are a large portion of images which only view a planar scene which makes the estimation of the full 15-DOF homography matrix

degenerate, and for the 360 scene, using a small submap size such as $w = 1$ is likely to encounter a pure rotation which can result in less accurate depth measurements from VGGT and hence reduced accuracy in our estimate of relative homography.

Table 5: Root mean square error (RMSE) of absolute trajectory error (ATE) on 7-Scenes [60] (unit: m). The * symbol indicates that the baseline is evaluated in the uncalibrated mode, all VGGT-SLAM configurations are evaluated in uncalibrated mode. Green is best and light green is second best.

	Method	Sequence							Avg
		chess	fire	heads	office	pumpkin	kitchen	stairs	
Uncalib.	DROID-SLAM* [67]	0.047	0.038	0.034	0.136	0.166	0.080	0.044	0.078
	MASt3R-SLAM* [46]	0.063	0.046	0.029	0.103	0.114	0.074	0.032	0.066
	Ours (Sim(3), $w = 1$)	0.047	0.025	0.032	0.113	0.138	0.050	0.083	0.070
	Ours (Sim(3), $w = 8$)	0.039	0.027	0.020	0.108	0.144	0.053	0.080	0.067
	Ours (Sim(3), $w = 16$)	0.037	0.027	0.021	0.107	0.135	0.051	0.093	0.067
	Ours (Sim(3), $w = 32$)	0.037	0.026	0.018	0.104	0.133	0.061	0.093	0.067
	Ours (SL(4), $w = 1$)	0.089	0.046	0.072	0.119	0.147	0.055	0.100	0.090
	Ours (SL(4), $w = 8$)	0.041	0.060	0.043	0.106	0.206	0.054	0.078	0.084
	Ours (SL(4), $w = 16$)	0.036	0.065	0.037	0.107	0.139	0.050	0.093	0.075
	Ours (SL(4), $w = 32$)	0.036	0.028	0.018	0.103	0.133	0.058	0.093	0.067

Table 6: Root mean square error (RMSE) of absolute trajectory error (ATE) on TUM RGB-D [65] (unit: m). The * symbol indicates that the baseline is evaluated in the uncalibrated mode, all VGGT-SLAM configurations are evaluated in uncalibrated mode. Green is best and light green is second best.

	Method	Sequence									Avg
		360	desk	desk2	floor	plant	room	rpy	teddy	xyz	
Uncalib.	DROID-SLAM* [67]	0.202	0.032	0.091	0.064	0.045	0.918	0.056	0.045	0.012	0.158
	MASt3R-SLAM* [46]	0.070	0.035	0.055	0.056	0.035	0.118	0.041	0.114	0.020	0.060
	Ours (Sim(3), $w = 8$)	0.070	0.026	0.030	0.048	0.026	0.081	0.024	0.035	0.015	0.040
	Ours (Sim(3), $w = 16$)	0.112	0.045	0.123	0.261	0.022	0.137	0.044	0.044	0.016	0.089
	Ours (Sim(3), $w = 32$)	0.123	0.040	0.055	0.254	0.022	0.088	0.041	0.032	0.016	0.074
	Ours (SL(4), $w = 8$)	0.179	0.046	0.095	0.210	0.033	0.272	0.038	0.130	0.031	0.115
	Ours (SL(4), $w = 16$)	0.147	0.032	0.087	0.158	0.027	0.150	0.037	0.069	0.035	0.083
	Ours (SL(4), $w = 32$)	0.071	0.025	0.040	0.141	0.023	0.102	0.030	0.034	0.014	0.053

Table 7: Dense reconstruction evaluation on 7-Scenes [60] (unit: m).

	Method	7-Scenes			
		ATE ↓	Acc. ↓	Comple. ↓	Chamfer ↓
Uncalib.	MASt3R-SLAM* [46]	0.066	0.068	0.045	0.056
	Ours (Sim(3), $w = 1$)	0.070	0.066	0.051	0.059
	Ours (Sim(3), $w = 8$)	0.067	0.054	0.056	0.055
	Ours (Sim(3), $w = 16$)	0.067	0.054	0.058	0.056
	Ours (Sim(3), $w = 32$)	0.067	0.052	0.062	0.057
	Ours (SL(4), $w = 1$)	0.090	0.080	0.068	0.074
	Ours (SL(4), $w = 8$)	0.084	0.067	0.065	0.066
	Ours (SL(4), $w = 16$)	0.075	0.061	0.063	0.060
	Ours (SL(4), $w = 32$)	0.067	0.052	0.058	0.055

B.2 Number of submaps per scene

As a reference, in Tables 8 and 9 we show the number of total submaps in each scene for 7-Scenes and TUM RGB-D for different values of experimented submap size, w , and also show the number of loop closures in each scene.

B.3 Evaluation of Focal length Consistency

To provide quantitative results showing that VGGT can produce an estimate of the scene which differs by more than a similarity transformation to the true scene, in this section we show inconsistencies in estimates of camera intrinsics from VGGT. Here, a single camera is used per scene and different scenes can use different cameras. We observe that even though the true intrinsics of the camera should be approximately constant within a scene, VGGT has a varying estimate of the intrinsics both inside a submap and across different submaps. This provides further demonstration that the VGGT reconstruction of a submap can differ from the true scene by more than a similarity transformation

Table 8: Window size w and corresponding submap and loop closure counts when $w_{\text{loop}} = 1$, shown as “# of submaps (# of loops)”.

Window size, w	Sequences in 7-Scenes [65]						
	chess	fire	heads	office	pumpkin	kitchen	stairs
1	29 (11)	50 (46)	62 (49)	58 (55)	43 (37)	43 (38)	14 (12)
8	4 (0)	7 (4)	8 (3)	8 (4)	6 (0)	6 (2)	2 (0)
16	2 (0)	4 (1)	4 (2)	4 (2)	3 (0)	3 (1)	1 (0)
32	1 (0)	2 (0)	2 (0)	2 (0)	2 (0)	2 (0)	1 (0)

Table 9: Window size w and corresponding submap and loop closure counts when $w_{\text{loop}} = 1$, shown as “# of submaps (# of loops)”.

Window size, w	Sequences in TUM-RGB-D [65]								
	360	desk	desk2	floor	plant	room	rpy	teddy	xyz
1	168 (151)	54 (42)	98 (84)	99 (87)	102 (92)	186 (162)	95 (89)	146 (125)	56 (54)
8	21 (4)	7 (4)	13 (7)	13 (3)	13 (5)	24 (7)	12 (10)	19 (9)	7 (5)
16	11 (2)	4 (2)	7 (4)	7 (2)	7 (2)	12 (4)	6 (4)	10 (4)	4 (2)
32	6 (1)	2 (0)	4 (2)	4 (1)	4 (2)	6 (2)	3 (1)	5 (2)	2 (0)

and contain affine and projective degrees of freedom which can be resolved using the homography alignment. In Table [10](#) we summarize the standard deviation, range, and average of all focal length estimates for four scenes. We observe that for both the office loop scene and 7-Scenes, our Sim(3) variant of VGGT-SLAM performs comparable to the SL(4) variant while SL(4) performs significantly better than Sim(3) on the Tabletop and Bollards scene. Consistent with this observation, in Table [10](#) we notice that the later two have much larger intrinsic error (larger standard deviation and larger range) than the former two.

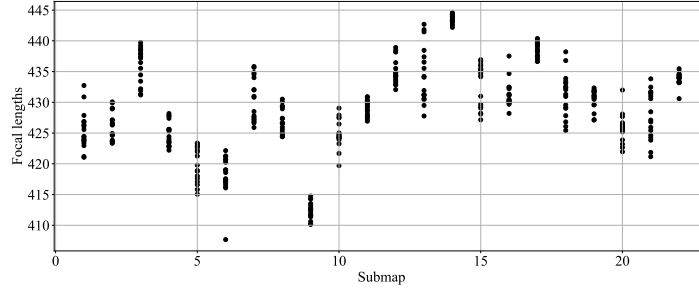


Figure 4: VGGT estimates of the focal length (fx) of every keyframe in the office loop scene from Fig. [2](#) for all 22 submaps.

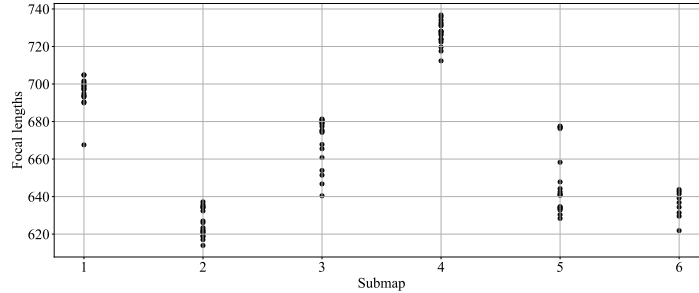


Figure 5: VGGT estimates of the focal length (fx) of every keyframe in the tabletop scene from Fig. [7](#) for all 6 submaps.

Table 10: Statistics of VGGT Focal length (fx) estimates. All values in pixels.

Scene	Std Dev	Range	Average
Office Loop (Fig. 2)	7.3	36.9	429.0
7-Scenes	9.0	59.7	435.1
Tabletop (Fig. 7)	37.1	122.8	669.1
Bollards (Fig. 8)	51.8	177.3	738.9

C Extra Qualitative Results

C.1 Extra examples of SL(4) versus Sim(3)

While we have mentioned that the Sim(3) version of VGGT-SLAM often provides high quality reconstructions, here we provide additional examples of cases where Sim(3) is not sufficient and SL(4) is necessary to achieve consistent alignment across submaps.

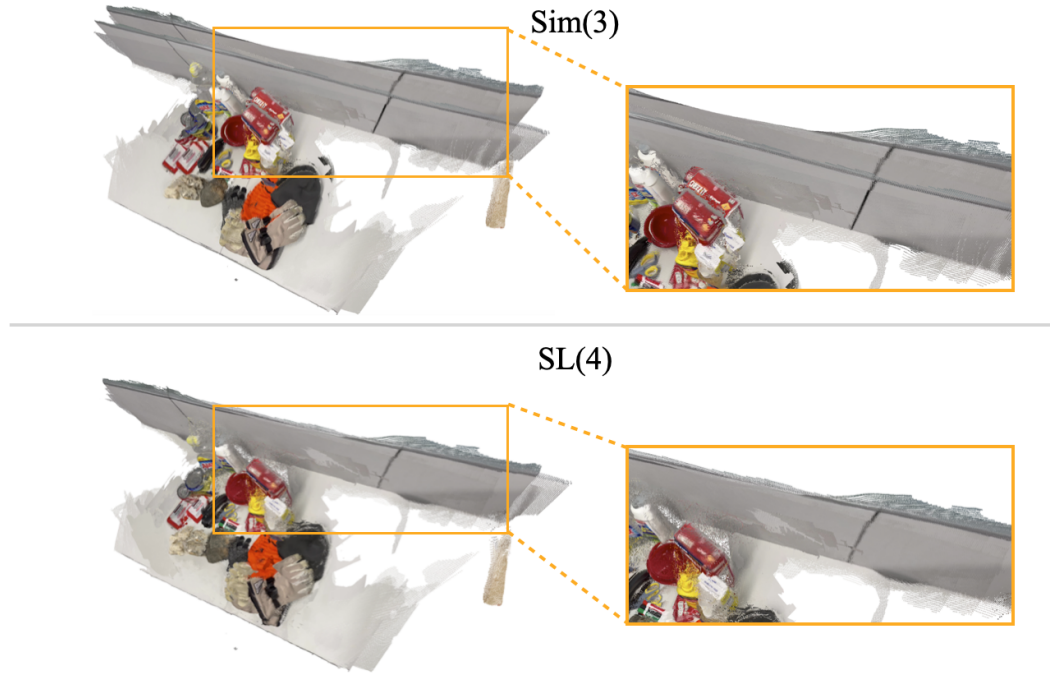


Figure 6: Example on a tabletop scene showing Sim(3) is unable to align the submaps while SL(4) is able to correct for projective ambiguity. Here $w = 32$ and $\tau_{\text{disparity}} = 50$.

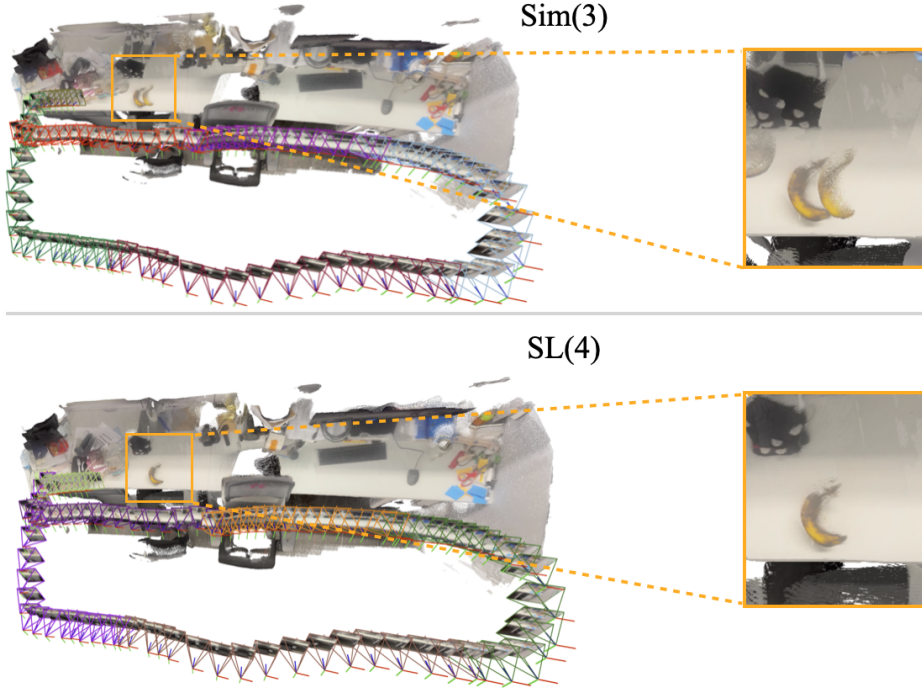


Figure 7: Example on a tabletop scene showing Sim(3) is unable to align the submaps while SL(4) is able to correct for projective ambiguity. The true scene only has one banana, but the Sim(3) reconstruction shows a hallucination of two caused by misalignment. Camera pose estimates are colored by submap. Here $w = 16$ and $\tau_{\text{disparity}} = 50$.

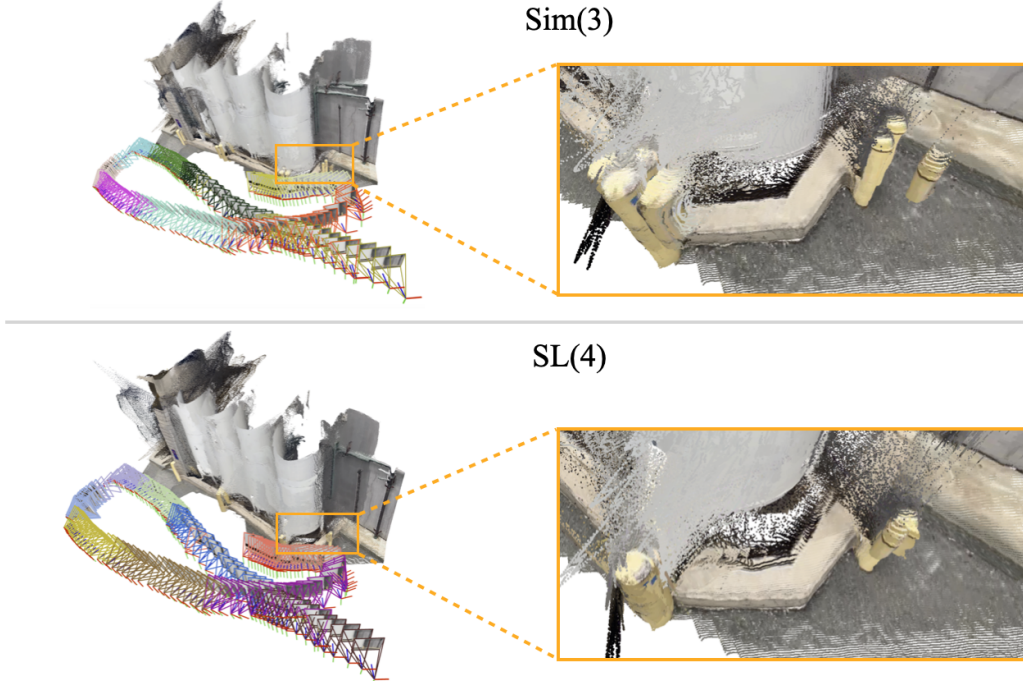


Figure 8: Example on an outdoor scene with yellow bollards surrounding tanks showing Sim(3) is unable to align the submaps while SL(4) is able to correct for projective ambiguity. The true scene has single bollards spaced around the tanks while the Sim(3) scene hallucinates clusters of bollards due to misalignment. Here $w = 16$ and $\tau_{\text{disparity}} = 25$.

C.2 7-Scenes Qualitative Results

Here we provide additional visualizations of scene reconstructions from the 7-Scenes dataset experiments for VGGT-SLAM with SL(4). We use the default parameters from Sec. 5.

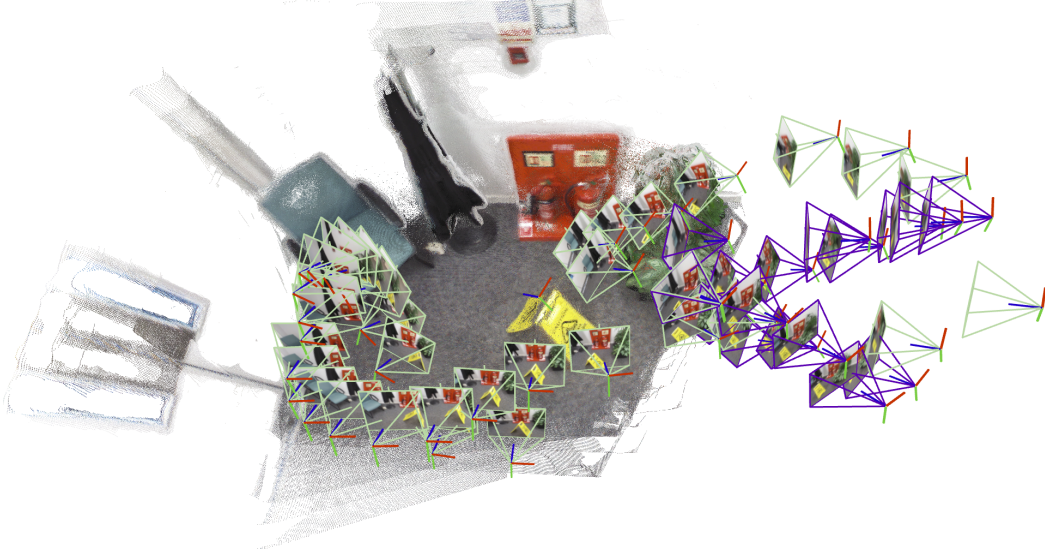


Figure 9: Visualization of reconstruction on 7-Scenes fire scene with 2 submaps. Camera pose estimates are colored by submap.

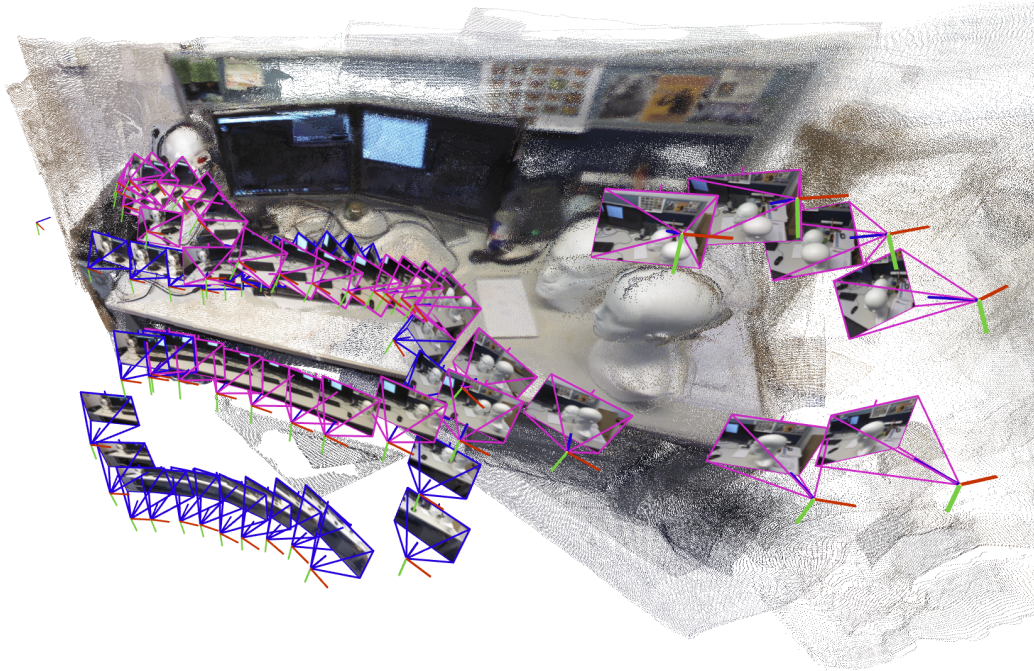


Figure 10: Visualization of reconstruction on 7-Scenes heads scene with 2 submaps. Camera pose estimates are colored by submap. Part of the scene is cropped for visual clarity.

C.3 TUM RGB-D Qualitative Results

Here we provide additional visualizations of scene reconstructions from the TUM RGB-D dataset experiments for VGGT-SLAM with SL(4). We use the default parameters from Sec. [5](#).

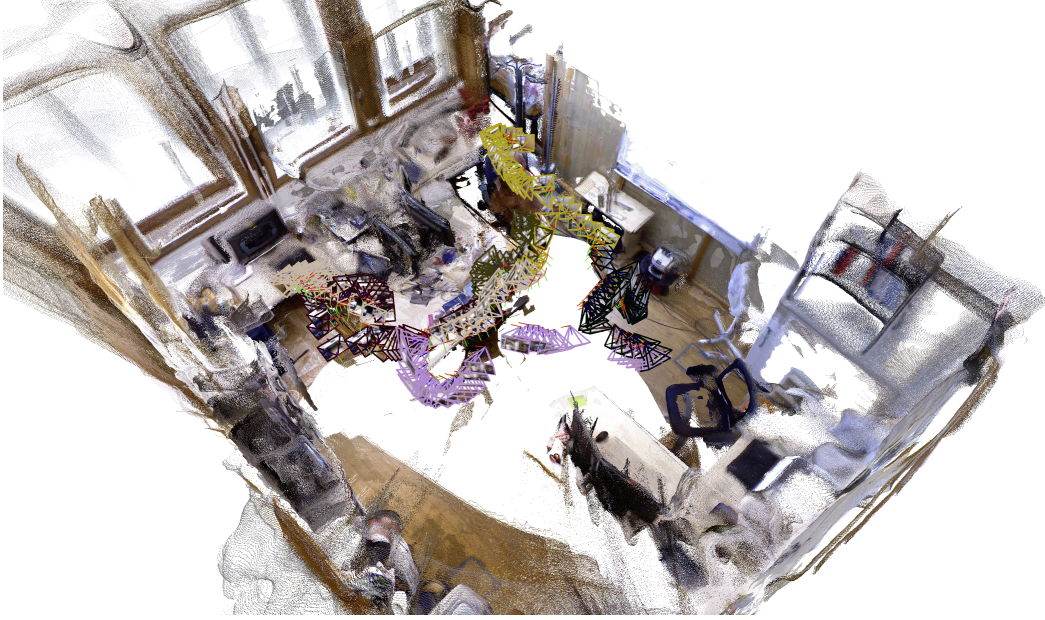


Figure 11: Visualization of reconstruction on TUM `room` scene with 6 submaps. Camera pose estimates are colored by submap. Part of the scene is cropped for visual clarity.



Figure 12: Visualization of reconstruction on TUM `360` scene with 6 submaps. Camera pose estimates are colored by submap. Part of the scene is cropped for visual clarity.

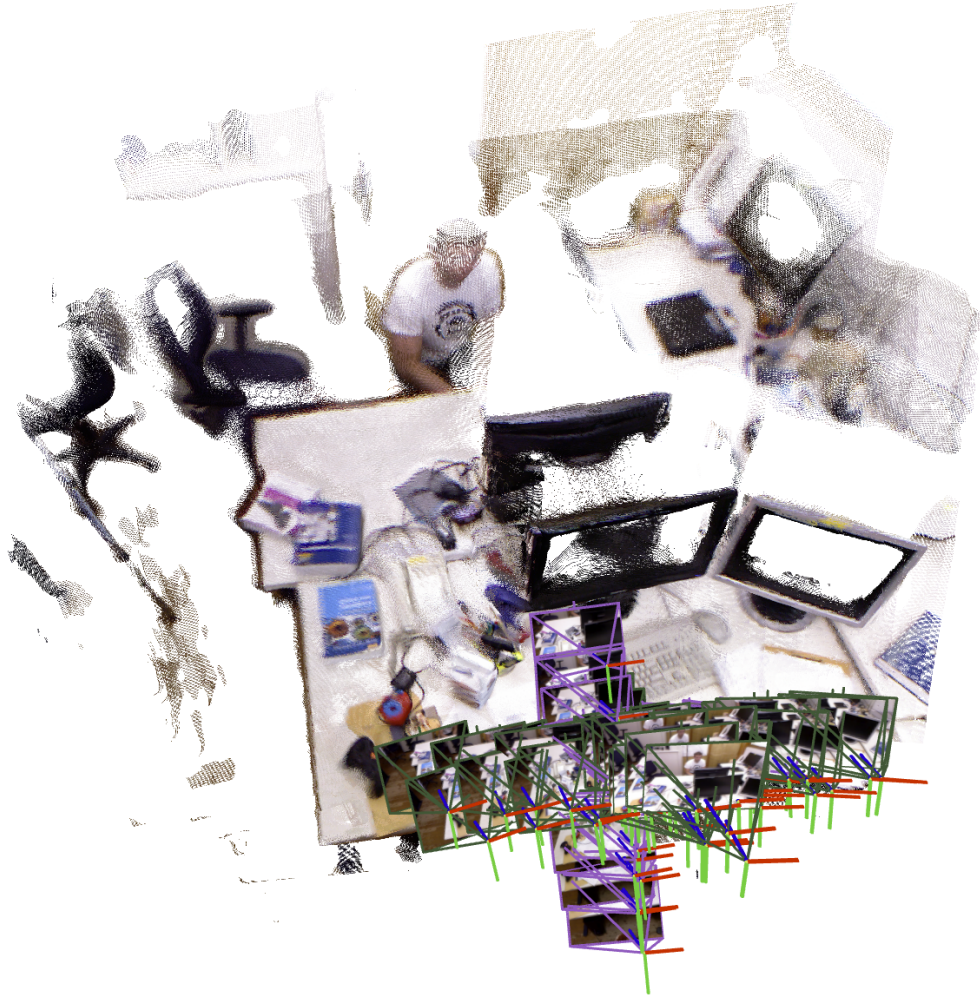


Figure 13: Visualization of reconstruction on TUM xyz scene with 2 submaps. Camera pose estimates are colored by submap.

C.4 Additional Outdoor Qualitative Results

While our method is primarily tested on indoor scenes, here we provide an additional example of VGGT-SLAM on an outdoor scene from the TartanAir dataset [74]. Here $w = 16$, $\tau_{\text{disparity}} = 50$, and $\tau_{\text{conf}} = 50$.

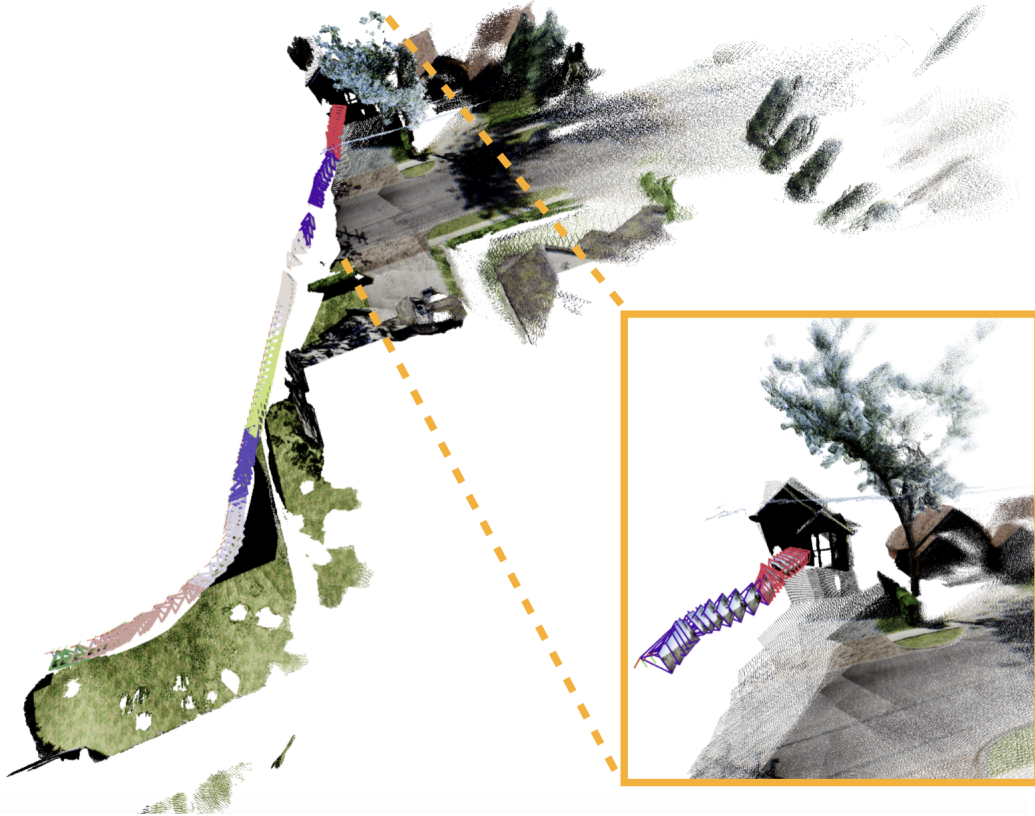


Figure 14: Visualization of reconstruction on TartanAir (scene *Neighborhood Easy*, *P005*, *left camera*) with 8 submaps. Camera pose estimates are colored by submap.