Supplementary Material for Neural Active Structure-from-Motion in Dark and Textureless Environment

Kazuto Ichimaru[1,2], Diego Thomas[1], Takafumi Iwaguchi[1], and Hiroshi Kawasaki[1]

[1] Kyushu University, Japan
https://www.cvg.ait.kyushu-u.ac.jp/
[2] Fujitsu Defense & National Security Limited, Japan
https://www.fujitsu.com/jp/group/fdns/

1 Procedure of system pose noise addition

In this section, we explain the detailed procedure of system pose noise addition. Assume we have ground-truth system rotation $\boldsymbol{r}_i^{gt}$ and translation $\boldsymbol{t}_i^{gt}$, which convert the camera coordinate system into the world coordinate system (*i.e.*, $\boldsymbol{t}_i^{gt}$ are the true camera positions in world coordinate system). To determine the scale of the scene, we compute baseline T, as follows:

$$T = \frac{1}{N-1} \sum_{i=0}^{N-1} \left| \boldsymbol{t}_i^{gt} - \boldsymbol{t}_{i-1}^{gt} \right|_2 , \tag{1}$$

where N is the number of the images. Once rotational noise scale δ_r and translational noise scale δ_t are specified, we compute initial system rotation $\boldsymbol{r}_i^{init}$ and translation $\boldsymbol{t}_i^{init}$ as follows:

$$\begin{aligned}
r &= U(-\delta_r/2, +\delta_r/2), \\
p &= U(-\delta_r/2, +\delta_r/2), \\
y &= U(-\delta_r/2, +\delta_r/2), \\
\boldsymbol{r}_i^{noise} &= euler(r, p, y), \\
\boldsymbol{r}_i^{init} &= SO3(so3(\boldsymbol{r}_i^{noise}) \cdot so3(\boldsymbol{t}_i^{gt})), \\
x &= U(-\delta_t/2, +\delta_t/2) \cdot T, \\
y &= U(-\delta_t/2, +\delta_t/2) \cdot T, \\
z &= U(-\delta_t/2, +\delta_t/2) \cdot T, \\
\boldsymbol{t}_i^{init} &= \boldsymbol{t}_i^{gt} + (x, y, z)^T,
\end{aligned} \tag{2}$$

where *euler* makes rotation vector from roll, pitch, yaw, $so3$ makes rotation matrix from rotation vector, $SO3$ makes rotation vector from rotation matrix, and $U(a, b)$ is uniform noise with range of a to b.

2 Evaluation metrics

As mentioned in the main text, we used Chamfer distance (reconstructed shape to GT shape) to measure the accuracy of the reconstructed shapes. However, we

noticed too large values were obtained from the reconstructed shapes with many outliers, such as Light-sectioning. Thus, we added the following pre-processes before computing Chamfer distance for fair comparison.

1. Adjust rotation, translation and scale of the estimated poses to fit the GT poses to remove ambiguity on global transformation.
2. Compute the point-to-point distances between the reconstructed shapes and the GT shapes, and remove points with minimum distance larger than 5% of the scene boundary as outliers.

3 Evaluation on various pose noise scales

In this section, we show the results of additional experiments on various range of pose noise scales in Table 1 and Table 2. Specifically, we tried $10°, 20°$ for rotation, and $50\%, 100\%$ for translation.

As the results show, the proposed method consistently outperforms the comparative methods. Light-sectioning sometimes surpasses the proposed method for larger pose noises, which is because the NeRF-based methods tend to fail to converge when the scale of noise is large, which is a limitation of NeRF-based localization.

Interestingly, reconstruction accuracy is higher in no illumination with larger pose noises, despite lack of textural information. We consider it is because textural information is beneficial only when there is enough overlap between frames, otherwise, it produces local optima. On the other hand, the projected laser maintains a nearly constant color in the image, almost always intersecting with the other lasers in adjacent frames at some point. This indicates new advantage of the proposed method.

4 Comparison to traditional SfM, state-of-the-art Neural SDF, and 3DGS

We conducted an additional experiment using COLMAP (SfM and MVS), Neuralangelo [17] (state-of-the-art Neural SDF) and 3DGS [14] on Lego (NeRF-Synthetic), Stone (BlendedMVS), and Real (controlled sequence) scenes. The results still show ours is advantageous in no illumination scenes, but the comparative methods achieved remarkable accuracy in normal illumination scenes. As for 3DGS, since it is highly sensitive to initialization, it did not converge in Stone scene (Table 3).

5 Pattern sparsity and scene scale

We conducted an additional experiment for more thorough analysis. We changed the number of lasers and applied the proposed method to Lego and Stone scenes. The results show the proposed method works consistently regardless of the sparsity (Table 4).

Table 1: Chamfer distances [mm] of the reconstructed shapes. Ours consistently outperforms the comparative methods. "N/A" indicates no shape reconstructed. Legend: **Best**, Second best.

Illum	Rot[°]	Trans[%]	Method	NeRF-Synthetic				BlendedMVS			
				Lego	Chair	Hotdog	Mic	Stone	Dog	Bear	Sculpture
Normal	10	50	Light-sectioning	39.88	40.56	43.72	39.90	96.97	25.31	81.32	20.80
			NeuS+Pose estim.	13.73	50.97	29.10	N/A	59.51	28.04	**23.17**	28.57
			NeuS+SL	42.59	46.98	48.36	35.13	85.06	29.80	46.00	27.64
			ActiveSfM (Ours)	**13.07**	**36.93**	**22.42**	**12.13**	**31.84**	**19.89**	23.40	**10.75**
	10	100	Light-sectioning	40.50	37.90	49.99	40.47	103.44	**25.46**	92.35	19.63
			NeuS+Pose estim.	32.6	N/A	48.6	N/A	76.62	44.99	**27.38**	**8.71**
			NeuS+SL	41.47	49.33	51.36	50.34	90.85	40.74	55.49	39.04
			ActiveSfM (Ours)	**25.10**	**32.59**	**44.64**	**33.66**	**40.12**	26.05	30.85	11.18
	20	50	Light-sectioning	42.30	42.80	50.00	43.93	93.45	30.40	81.27	24.61
			NeuS+Pose estim.	**32.6**	N/A	48.60	N/A	76.62	44.99	27.38	**8.71**
			NeuS+SL	41.54	53.98	48.15	**36.43**	106.52	35.21	N/A	45.85
			ActiveSfM (Ours)	41.41	**39.35**	**45.73**	40.60	**41.02**	**28.16**	**23.63**	10.35
	20	100	Light-sectioning	41.99	44.81	48.77	44.39	100.23	**26.29**	79.68	**28.93**
			NeuS+Pose estim.	N/A	N/A	**28.84**	N/A	N/A	39.64	N/A	N/A
			NeuS+SL	40.96	54.74	55.8	**4.15**	97.41	41.12	61.98	40.01
			ActiveSfM (Ours)	**38.76**	**42.63**	46.12	48.78	**39.46**	38.91	**24.20**	N/A
No	10	50	Light-sectioning	39.88	40.56	43.72	39.90	96.97	25.31	81.32	20.80
			NeuS+Pose estim.	59.75	48.45	61.15	43.94	N/A	29.31	81.40	32.26
			NeuS+SL	35.04	44.76	52.76	N/A	N/A	50.64	N/A	N/A
			ActiveSfM (Ours)	**18.50**	**18.60**	**28.47**	**6.37**	**35.36**	**21.67**	**22.99**	**11.20**
	10	100	Light-sectioning	40.05	**37.90**	49.99	40.47	103.44	**25.46**	92.35	19.63
			NeuS+Pose estim.	62.87	52.60	45.24	**36.50**	112.05	30.29	N/A	N/A
			NeuS+SL	35.04	44.76	52.76	N/A	N/A	50.64	N/A	N/A
			ActiveSfM (Ours)	**31.26**	40.19	**44.99**	45.28	**36.10**	28.57	**22.50**	**10.68**
	20	50	Light-sectioning	42.30	42.80	50.00	43.93	93.45	**30.40**	81.27	24.61
			NeuS+Pose estim.	32.74	48.71	N/A	37.67	N/A	49.41	85.97	27.52
			NeuS+SL	28.16	51.67	N/A	N/A	N/A	43.90	112.48	N/A
			ActiveSfM (Ours)	**18.45**	**17.64**	**28.07**	**32.84**	**36.21**	32.83	**26.33**	**10.10**
	20	100	Light-sectioning	41.99	44.81	48.77	44.39	100.23	**26.29**	79.68	28.93
			NeuS+Pose estim.	**31.70**	54.47	56.56	**22.69**	110.47	48.86	60.32	N/A
			NeuS+SL	30.19	48.67	N/A	N/A	N/A	48.58	90.25	N/A
			ActiveSfM (Ours)	41.64	**34.28**	**47.81**	44.06	**38.49**	35.62	**26.30**	**9.77**

6 Evaluation on projector pose refinement

We also conducted another evaluation on projector pose refinement. We empirically observed that large projector pose error leads to severe collapse of the reconstructed shapes, but small errors can be refined by defining the projector poses as learnable parameters. Therefore, we added $0, 1, 4°$ rotational noise and $5, 10\%$ translational noise on the projector poses in addition to the camera poses with $10°$ rotational noise and 50% translational noise. Note that since we assume projectors are fixed relative to the camera, we have only 6 parameters to refine per projector. To verify the effectiveness of the proposed method, we also trained the pipeline without projector pose refinement.

Figure 1 and Table 5 shows the qualitative and quantitative results of the evaluation. As the results show, reconstruction and pose estimation quality are consistently improved by projector pose refinement. We noticed the projector pose estimation accuracy is considerably bad despite the drastic improvements in the reconstruction quality. This is presumed to be because the pattern projected by the projectors was a cross-laser pattern, which left degrees of freedom in the pose of the projector, especially in the no illumination where textural information is not available. It is rather interesting the training was performed to maintain the consistency of the scene and the reconstruction results were accurate, even in such an under-constrained problem.

Table 2: Mean L1 errors of the estimated poses (rotation and translation). "N/A" indicates no shape reconstructed. Legend: **Best**.

Illum	Rot[°]	Trans[%]	Metric	Method	NeRF-Synthetic				BlendedMVS			
					Lego	Chair	Hotdog	Mic	Stone	Dog	Bear	Sculpture
Normal	10	50	Rot[°]	NeuS+Pose estim.	0.84	3.90	1.50	N/A	0.60	3.02	0.96	2.40
				ActiveSfM (Ours)	**0.56**	**1.18**	**1.29**	**0.88**	**0.52**	**0.76**	**0.38**	**0.67**
			Trans[%]	NeuS+Pose estim.	4.25	25.63	9.61	N/A	1.37	7.83	1.53	4.44
				ActiveSfM (Ours)	**1.43**	**7.29**	**5.59**	**3.38**	**0.09**	**0.62**	**0.43**	**0.56**
	10	100	Rot[°]	NeuS+Pose estim.	3.82	N/A	3.67	N/A	3.74	4.64	**0.85**	2.15
				ActiveSfM (Ours)	**1.95**	**1.38**	**4.04**	**3.80**	**0.79**	**3.91**	3.98	**0.71**
			Trans[%]	NeuS+Pose estim.	28.98	N/A	**41.24**	N/A	7.65	52.78	**4.25**	3.64
				ActiveSfM (Ours)	**14.30**	**8.13**	44.62	**46.59**	**0.08**	**8.16**	5.07	**0.70**
	20	50	Rot[°]	NeuS+Pose estim.	7.07	N/A	6.08	N/A	4.07	6.83	N/A	8.66
				ActiveSfM (Ours)	**4.91**	**4.87**	**5.37**	**7.07**	**1.03**	**3.95**	**0.42**	**1.07**
			Trans[%]	NeuS+Pose estim.	**25.16**	N/A	36.99	N/A	4.40	55.78	N/A	15.25
				ActiveSfM (Ours)	25.36	**25.20**	**25.14**	**25.11**	**0.09**	**4.86**	**0.39**	**1.02**
	20	100	Rot[°]	NeuS+Pose estim.	N/A	N/A	7.42	N/A	N/A	**7.87**	N/A	N/A
				ActiveSfM (Ours)	**6.27**	**7.05**	**6.01**	**6.78**	**1.26**	7.93	**1.71**	N/A
			Trans[%]	NeuS+Pose estim.	N/A	N/A	49.1	N/A	N/A	20.31	N/A	N/A
				ActiveSfM (Ours)	**41.2**	**43.89**	**43.37**	**48.79**	**0.07**	**13.72**	**2.07**	N/A
No	10	50	Rot[°]	NeuS+Pose estim.	6.79	7.18	7.86	7.46	N/A	7.99	7.57	5.80
				ActiveSfM (Ours)	**0.74**	**1.07**	**1.02**	**0.63**	**0.53**	**1.01**	**0.42**	**0.90**
			Trans[%]	NeuS+Pose estim.	25.80	23.86	25.20	25.95	N/A	10.71	8.60	17.77
				ActiveSfM (Ours)	**1.48**	**3.70**	**2.97**	**2.32**	**0.41**	**1.78**	**0.48**	**0.71**
	10	100	Rot[°]	NeuS+Pose estim.	7.68	8.88	3.22	4.98	3.58	8.99	N/A	N/A
				ActiveSfM (Ours)	**1.19**	**1.24**	**1.82**	**1.85**	**0.65**	**3.08**	**0.48**	**1.18**
			Trans[%]	NeuS+Pose estim.	47.16	47.26	68.04	37.25	12.79	12.11	N/A	N/A
				ActiveSfM (Ours)	**1.82**	**4.53**	**6.31**	**16.18**	**0.44**	**8.04**	**0.50**	**0.79**
	20	50	Rot[°]	NeuS+Pose estim.	**0.51**	8.65	N/A	9.14	N/A	10.09	10.16	3.58
				ActiveSfM (Ours)	1.05	**1.02**	**1.36**	**3.39**	**0.65**	**7.11**	**0.55**	**1.28**
			Trans[%]	NeuS+Pose estim.	20.4	25.76	N/A	**26.22**	N/A	10.82	8.80	7.03
				ActiveSfM (Ours)	**3.38**	**3.59**	**4.17**	26.43	**0.40**	**10.60**	**0.59**	**1.25**
	20	100	Rot[°]	NeuS+Pose estim.	4.31	9.64	9.45	7.63	11.06	9.99	3.67	N/A
				ActiveSfM (Ours)	**2.49**	**2.02**	**5.36**	**7.20**	**0.69**	**8.52**	**0.97**	**1.69**
			Trans[%]	NeuS+Pose estim.	24.35	49.26	45.24	**52.39**	28.18	**12.51**	6.56	N/A
				ActiveSfM (Ours)	**1.43**	**7.42**	**39.41**	56.34	**0.43**	14.02	**0.80**	**1.81**

7 Evaluation without mask loss

During the experiments, we used mask loss to separate NeRF-Synthetic, BlendedMVS, and controlled sequence scenes. However, some may wonder if object mask is not available in dark environments, and in that case, using mask loss does not make sense. Therefore, we conducted another experiment to clarify how the proposed method works in the dark environment without mask loss. Specifically, we ran evaluations on shape reconstruction accuracy and pose estimation accuracy on NeRF-Synthetic and BlendedMVS as same as the main text, but without mask loss. Figure 2 shows the qualitative results and Table 6 shows the quantitative results. From the results, we can say it is possible to run ActiveSfM without mask loss, but its accuracy is suboptimal compared to that with mask loss, and enhancing the robustness is the future work.

8 Details of the experimental system configuration

In Table 7, we briefly explain the details of the experimental system configuration used in section 5. As for the devices, we used a Basler camera and commercial cross laser projectors.

Pose noise	Method	Normal illum.			No illum.		
		Lego	Stone	Real	Lego	Stone	Real
No	COLMAP	6.19	**12.04**	**7.12**	N/A	N/A	N/A
	Light-sectioning	21.08	17.40	13.33	21.08	**17.40**	13.33
	NeRF	23.04	44.31	18.28	24.75	63.99	19.44
	3DGS	**5.50**	132.85	21.66	N/A	N/A	N/A
	Neuralangelo	16.85	26.15	12.59	N/A	N/A	N/A
	Ours	11.62	30.92	9.55	**14.97**	35.35	**9.39**
Yes	COLMAP	**6.19**	**12.04**	**7.12**	N/A	N/A	N/A
	Light-sectioning	39.88	96.97	23.45	39.88	96.97	23.45
	NeRF	56.65	115.54	25.73	59.16	121.60	26.18
	3DGS	19.65	105.73	23.40	N/A	N/A	N/A
	Neuralangelo	48.90	100.18	30.56	N/A	N/A	N/A
	Ours	13.07	31.84	9.27	**18.50**	**35.36**	**12.29**

Table 3: Comparison results of traditional SfM, SL, NeRF, 3DGS, Neuralangelo. "N/A" indicates no shape reconstructed.

Method	Normal illum.		No illum.	
	Lego	Stone	Lego	Stone
NeuS + Pose estim.	13.73	59.51	59.75	N/A
Ours (2 lasers)	12.96	30.63	16.91	38.70
Ours (4 lasers)	13.07	31.84	18.50	35.36
Ours (6 lasers)	12.84	31.68	17.39	23.61

Table 4: Results of pattern sparsity analysis. "N/A" indicates no shape reconstructed.

9 Procedure of synthetic image generation

Here, we explain the procedure of the synthetic image generation used in subsection 5.2. Assume we have the original images $J(\boldsymbol{p})$ and GT depth images $D(\boldsymbol{p})$. For each 2D point $\boldsymbol{p}$, we back-project it to the world coordinate system to obtain 3D point $\boldsymbol{P}_s$ as follows,

$$\boldsymbol{P}_s = \boldsymbol{R}_i \cdot K_c^{-1} \boldsymbol{p}^T D(\boldsymbol{p}) + \boldsymbol{t}_i, \tag{3}$$

then, re-project it to the k-th projector screen coordinate system to obtain the projected color onto the point $\boldsymbol{P}_s$,

$$Q_k(\boldsymbol{P}_s) = I^k(K^k(R^k \boldsymbol{P}_s + \boldsymbol{t}^k)^T), \tag{4}$$

finally, pixel color of the synthesized image J' on $\boldsymbol{p}$ is computed as follows,

$$J'(\boldsymbol{p}) = J(\boldsymbol{p}) + \sum_{k=1}^{N_p} o(\boldsymbol{P}_s, k)(i_r^{GT} J(\boldsymbol{p}) + i_b^{GT}) Q_k(\boldsymbol{P}_s), \tag{5}$$

where i_r^{GT} and i_b^{GT} are GT illumination parameters, and $o(\boldsymbol{P}_s, k)$ returns 1 if $\boldsymbol{P}_s$ is not occluded by other points when viewed from the k-th projector, otherwise returns 0. Occlusion is computed using Z-buffer in our implementation.

Table 5: Quantitative results of evaluation on projector pose refinement. Numbers in parenthesis are the results without projector pose refinement. "failed" indicates they did not produce any mesh.

Illum	Proj noise		Camera		Projector		Shape
	Rot [°]	Trans [%]	Rot [°]	Trans [%]	Rot [°]	Trans [%]	
Normal	0°	5%	**0.33** (0.43)	**2.48** (7.59)	3.16 (-)	5.28 (-)	**5.19** (12.73)
	0°	10%	**0.34** (0.55)	**2.07** (4.06)	1.66 (-)	44.40 (-)	**5.09** (11.02)
	2°	5%	**0.31** (0.47)	**1.87** (3.04)	7.22 (-)	6.66 (-)	**5.48** (5.70)
	2°	10%	**0.33** (0.45)	**1.93** (3.05)	4.34 (-)	10.02 (-)	**5.32** (5.50)
	4°	5%	**0.29** (0.97)	**2.10** (6.58)	7.52 (-)	6.46 (-)	**5.18** (7.83)
	4°	10%	**0.38** (0.70)	**3.23** (8.17)	11.88 (-)	10.13 (-)	**5.73** (7.07)
No	0°	5%	**0.41** (0.44)	**17.43** (25.80)	0.41 (-)	5.40 (-)	**14.78** (18.62)
	0°	10%	**0.44** (0.52)	**14.05** (32.11)	0.60 (-)	9.79 (-)	**11.18** (19.19)
	2°	5%	**0.51** (5.84)	**11.50** (42.12)	4.83 (-)	6.80 (-)	**9.39** (failed)
	2°	10%	**0.42** (5.83)	**9.98** (39.61)	0.10 (-)	49.11 (-)	**9.07** (failed)
	4°	5%	**0.54** (0.65)	**9.24** (11.18)	2.96 (-)	6.69 (-)	**15.34** (17.20)
	4°	10%	**0.43** (0.67)	**7.53** (9.71)	2.69 (-)	10.40 (-)	**12.44** (17.48)

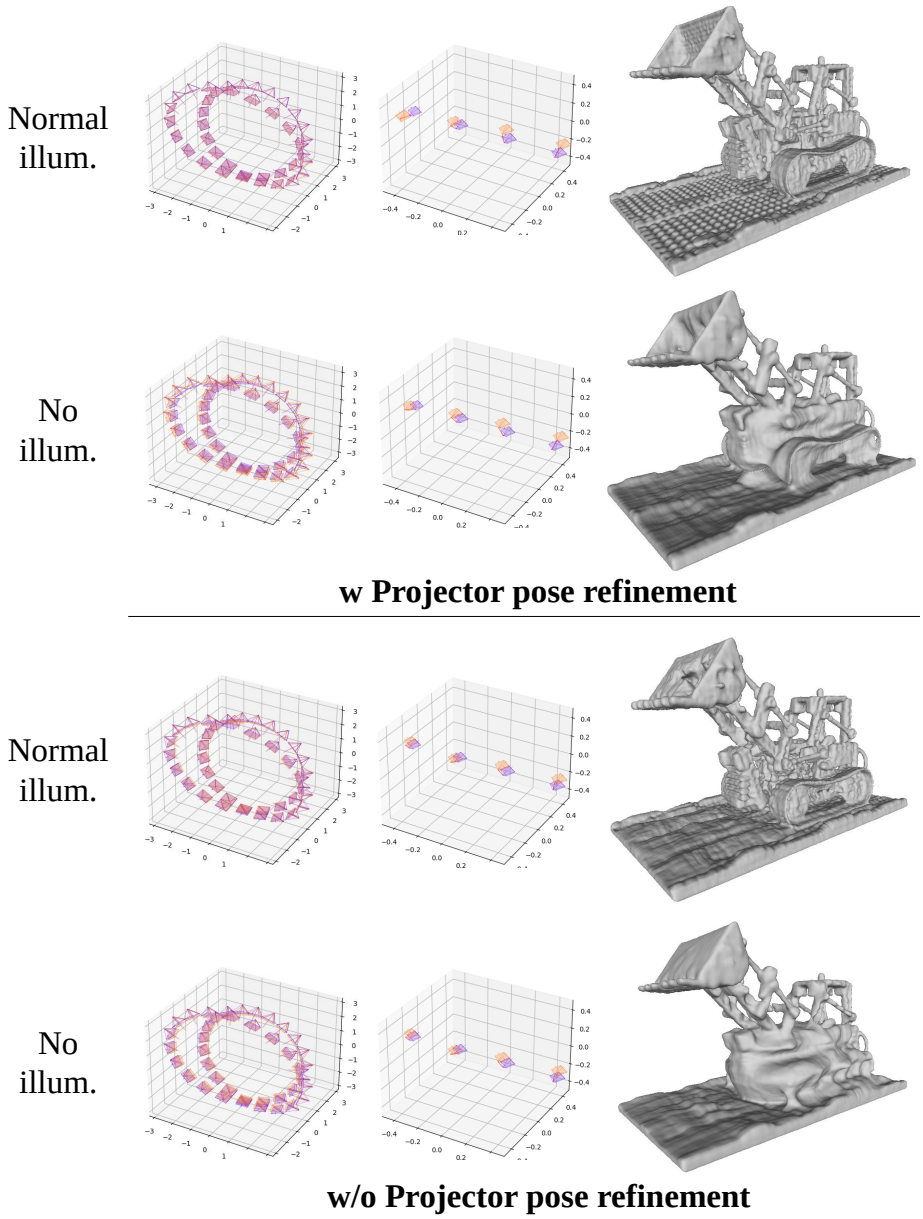

Fig. 1: Qualitative results of projector pose refinement on rotational noise 4° and translational noise 10% case. **Left to Right**: Camera poses, projector poses, reconstructed shapes.

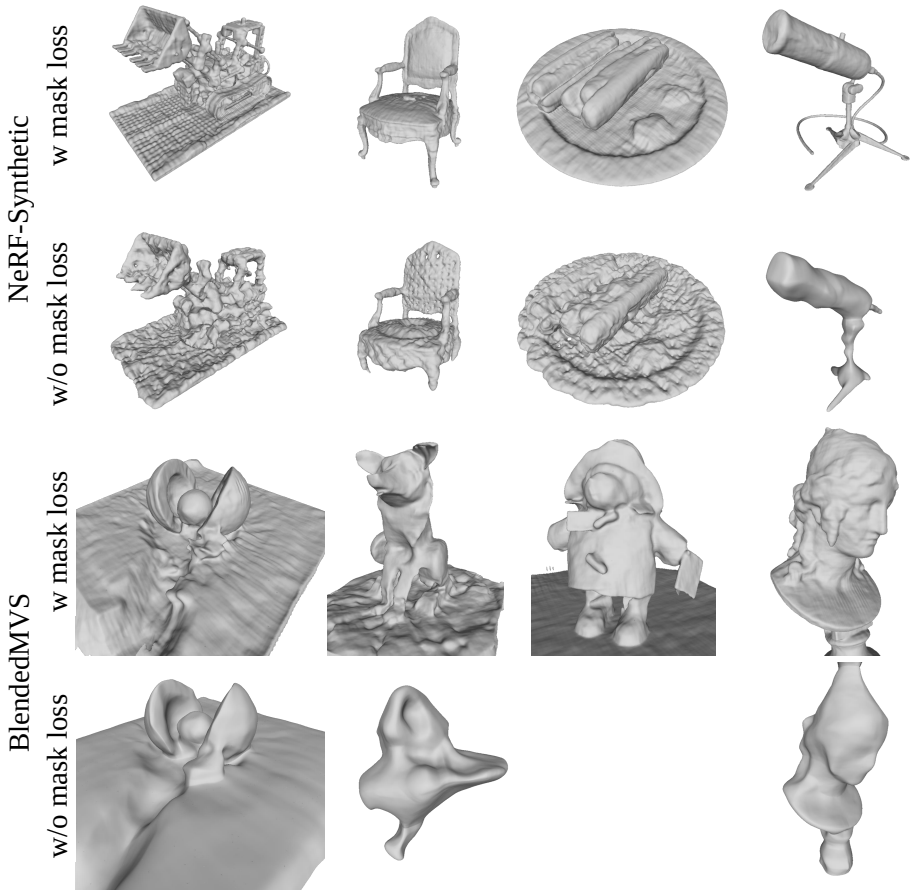

Fig. 2: Qualitative comparison with and without mask loss.

Method	NeRF-Synthetic				BlendedMVS			
	Lego	Chair	Hotdog	Mic	Stone	Dog	Bear	Sculpture
ActiveSfM (w mask loss)	18.50	18.60	28.47	6.37	35.36	21.67	22.99	11.20
ActiveSfM (w/o mask loss)	25.47	14.58	25.46	9.10	31.88	10.40	N/A	12.39

Table 6: Quantitative comparison in Chamfer distance (mm) with and without mask loss. N/A indicates no shape reconstructed.

Image resolution	800x800 (NeRF-Synthetic) 768x586 (BlendedMVS) 1280x960 (Real)
Laser wavelength	$520 \pm 10nm$
Line thickness	3mm at 1m away

Table 7: Details of the experimental system configuration.

10 Procedure of dark image synthesis

As for the dark image synthesis used in subsection 5.3, we simply scaled the intensities of the images to simulate quantization error in 8-bit image format. Other types of noise are the future work, like high ISO noise or motion blur.

11 On GNSS & IMU accuracy

Some people may have a doubt if GNSS or IMU accuracy is really insufficient for 3D measurement. The accuracy of common GPS is said to be 1-3 meters, and that of common IMU is said to be 2-5 centimeters and 0.005-0.015 degrees, which are covered in the experiments. This may be acceptable for large scale capturing, while it may be impossible to recover a meaningful trajectory in micro-baseline settings. Our structured light system is intended for the latter use-cases, such as underwater facility inspection or endoscopic surgery.

12 Implementation details of comparative methods

In general, we used official implementations for the comparative methods in the experiments. However, we had to modify a little bit to acquire the results, which is described below.

Light-sectioning Light-sectioning is a general principle for 3D shape reconstruction with active projection, and there are no specific implementation. Therefore, we implemented light-sectioning by ourselves from scratch.

LLNeRF [34] Since the official implementation does not provide a feature for mesh reconstruction, we implemented mesh reconstruction feature by marching cubes with the obtained density values from MLP. As for the hypermarameter, we does not change them except near / far clip and data loss function. Because the original rawnerf loss function did not converge well with our data, we replaced it with Mean-Squared-Error (mse).

NeuS [35] We used the official implementation with few modifications. Specifically, we increased resolution for mesh reconstruction to obtain finer results and re-implemented camera parameters as learnable parameters for NeuS+Pose estimation. Note, NeuS does not update camera parameters by itself (no gradients are computed).

NeuS+Pose estimation We added the pose estimation pipeline identical to ours on NeuS. We can also say it is SDF-based NeRF– [37].

NeuS+SL NeuS+SL is identical to the proposed method without pose estimation, as mentioned in the main text.