Neural Inverse Rendering for High-Accuracy 3D Measurement of Moving Objects with Fewer Phase-Shifting Patterns: Supplementary Materials

Yuki Urakawa Yoshihiro Watanabe
Institute of Science Tokyo
watanabe@ict.eng.isct.ac.jp

1. Limitations in the Single-Camera Single-Projector Setup

Fig. 1 shows the results of applying the conventional TurboSL method [2] to a static object under a single-camera, single-projector setup with three phase-shifting patterns and a frequency of 32. As shown in the figure, errors similar to unwrapping failures can be observed. Similar errors were also found in the case of four phase-shifting patterns. These results indicate that reducing the number of patterns below four is challenging in a single-camera, single-projector setup when using standard phase-shifting patterns.

2. Network Architecture

For multiresolution hash encoding [3], the number of hash feature grids was 16, the feature size per level was 2, the resolution of the coarsest grid was 16, the scale factor per level was 1.45, and smoothstep was used for interpolation. The MLP consisted of two hidden layers, each with a width of 64, and used ReLU activation. Sigmoid activation was applied to the outputs for residual and reflectance, whereas no activation function was used for the outputs of the SDF and displacement field.

3. Training Time

Using 70 % of the CUDA cores on an NVIDIA RTX 6000 Ada GPU, training 10,000 iterations took approximately 16 minutes. We acknowledge that the training time is relatively long; however, our method requires only three input patterns and achieves accuracy that surpasses existing real-time methods, making it particularly valuable for offline applications such as detailed shape analysis on recorded video sequences.

To shorten the training time, we plan to initialize the network using shapes estimated from conventional phase-shifting techniques or learning-based monocular depth-estimation methods, which can accelerate convergence. For sequential inputs, training time can be further reduced by using the parameters from the previous frame as initialization. In addition, we plan to incorporate a coarse-to-fine ray-casting strategy to decrease the number of samples per ray and to improve per-ray efficiency by adopting lighter encoding schemes and more compact MLP architectures.

4. Overview of Comparison Methods

In addition to the explanations in Sec. 7 in the main paper, Table 1 summarizes the differences among the comparison methods used in the evaluation experiments in that section.

For the single-camera case, the experimentally optimal parameters were used: the L1 and L2 norms for $\mathcal{L}_{\text{pro2cam}}$ were set to 1.5 and 15, respectively.

When using standard phase-shifting patterns under the Cam1-TurboSL condition, the estimated shape exhibited large errors with both three and four patterns. As discussed in Sec. 1 of the supplementary material, this issue arises due to the limitations outlined there. Therefore, as an alternative sinusoidal pattern, we included a condition using four micro-phase-shifting (MPS) patterns [1].

Furthermore, for our two-camera setup (Cam2), we confirmed that high-accuracy shape estimation could be achieved with a minimum of three patterns. Therefore, we did not evaluate the four-pattern MPS method under this condition.

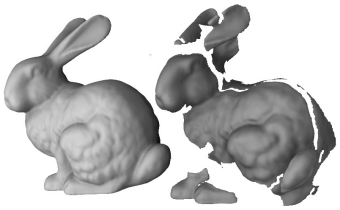
Additionally, in PS-Cam2-Standard, PS-Cam2-Weise, and PS8-Cam2-Zhang, the number of phase cycles $k(\boldsymbol{c})$ (shown in Eq. (3) in the main paper) is assumed to be obtained from geometric constraints using two cameras, as in previous studies [4, 5]. In our experiments, however, ground-truth values of $k(\boldsymbol{c})$ were directly provided instead of being estimated from geometric constraints.

5. Additional Comparisons

In the experiments described in Sec. 7 in the main paper, we conducted a quantitative evaluation of the objects undergoing uniform motion. In this section, we present additional

Table 1. Summary of the differences in comparison methods from Sec. 7. The numbers in parentheses under the 'Pattern' column indicate the number of patterns.

Name	Pattern	Number of cameras	Method	Motion compensation	Neural network
ALC3-Cam1-TurboSL	A La Carte (3)	1	TurboSL		✓
MPS4-Cam1-TurboSL	Micro phase shift (4)	1	TurboSL		✓
PS3-Cam2-Standard	Phase shift (3)	2	Standard phase shift		
PS3-Cam2-Weise	Phase shift (3)	2	Weise's phase shift	✓	
PS8-Cam2-Zhang	Phase shift (8)	2	Zhang's phase shift	✓	
ALC3-Cam1-w/ DF	A La Carte (3)	1	Ours (Partial)	✓	✓
MPS4-Cam1-w/ DF	Micro phase shift (4)	1	Ours (Partial)	✓	✓
ALC3-Cam2-w/o DF	A La Carte (3)	2	Ours (Partial)		✓
PS3-Cam2-w/o DF	Phase shift (3)	2	Ours (Partial)		✓
ALC3-Cam2-w/ DF	A La Carte (3)	2	Ours (Partial)	✓	✓
PS3-Cam2-w/ DF	Phase shift (3)	2	Ours (Full)	✓	✓
PS3-Cam3-w/ DF	Phase shift (3)	2	Ours (Full)	✓	✓



Ground Truth Measured 3D shape

Figure 1. Reconstructed 3D shape using three sinusoidal phase-shifting patterns with the conventional method [2].

results demonstrating the effectiveness of our method for objects with accelerated motion. Table 2 shows the results when the object's velocity in the n-th pattern is set to un mm/frame for a given parameter u. Even under this motion condition, the proposed method using standard phase-shifting patterns achieves high accuracy.

6. Additional real-world experiments

This section presents additional real-world experiments. Figs. 2–13 show the captured images from two cameras, along with the estimated results, including residuals, reflectance, displacement fields, and the reconstructed 3D shapes. Figs. 14–15 show the 3D shape reconstruction results for video sequences.

References

[1] Mohit Gupta and Shree K. Nayar. Micro Phase Shifting. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 813–820, 2012. 1

[2] Parsa Mirdehghan, Maxx Wu, Wenzheng Chen, David B. Lindell, and Kiriakos N. Kutulakos. TurboSL: Dense Accurate and Fast 3D by Neural Inverse Structured Light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25067–25076, 2024. 1, 2

[3] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics*, 41(4): 1–15, 2022. 1

[4] Thibaut Weise, Bastian Leibe, and Luc Van Gool. Fast 3D Scanning with Automatic Motion Compensation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 1

[5] Geyou Zhang, Ce Zhu, and Kai Liu. Binomial self-compensation for motion error in dynamic 3d scanning. In *European Conference on Computer Vision (ECCV)*, page 205–221, 2024. 1

Table 2. MAE for different accelerations. **Bold** and underline indicate the best and second-best results for each acceleration, respectively. The top five rows represent conventional methods, while the bottom seven rows correspond to the proposed methods.

Acceleration u (mm/frame^2)	-8	-4	-2	-1	0	1	2	4	8
ALC3-Cam1-TurboSL	21.3598	3.7988	0.9523	0.6810	0.5403	0.5919	0.8353	7.1506	24.0867
MPS4-Cam1-TurboSL	30.2778	13.6866	1.6048	1.0369	0.5031	0.7728	3.3795	2.9025	12.0904
PS3-Cam2-Standard	3.3261	1.8443	1.0295	0.6252	0.2724	0.4486	0.7576	1.6102	3.8133
PS3-Cam2-Weise	2.1688	1.4896	0.9177	0.5822	0.2606	0.2750	0.5619	0.9822	1.7980
PS8-Cam2-Zhang	9.9618	7.0769	3.1561	1.5522	0.2728	1.1814	2.7799	7.1996	10.0063
ALC3-Cam1-w/ DF	9.7854	0.8719	0.6656	0.5970	0.5824	0.6784	0.6588	0.7506	5.6192
MPS4-Cam1-w/ DF	19.5895	6.0639	0.9022	0.7263	0.6770	0.8686	0.9871	1.1934	24.4049
ALC3-Cam2-w/o DF	3.3463	1.7321	0.9617	0.6402	0.4195	0.5685	0.8191	1.4045	2.9924
PS3-Cam2-w/o DF	2.6758	1.4536	0.8168	0.4877	0.2351	0.4430	0.6043	0.8444	2.3782
ALC3-Cam2-w/ DF	**0.5710**	0.4495	0.3970	0.3940	0.4031	0.4347	0.3880	0.4729	0.5205
PS3-Cam2-w/ DF	1.5521	0.4233	**0.2456**	**0.2179**	**0.2208**	**0.2127**	0.2816	0.3142	0.5354
PS3-Cam3-w/ DF	1.0791	**0.3892**	0.2615	0.2294	0.2229	0.2313	**0.2512**	**0.2556**	**0.4211**

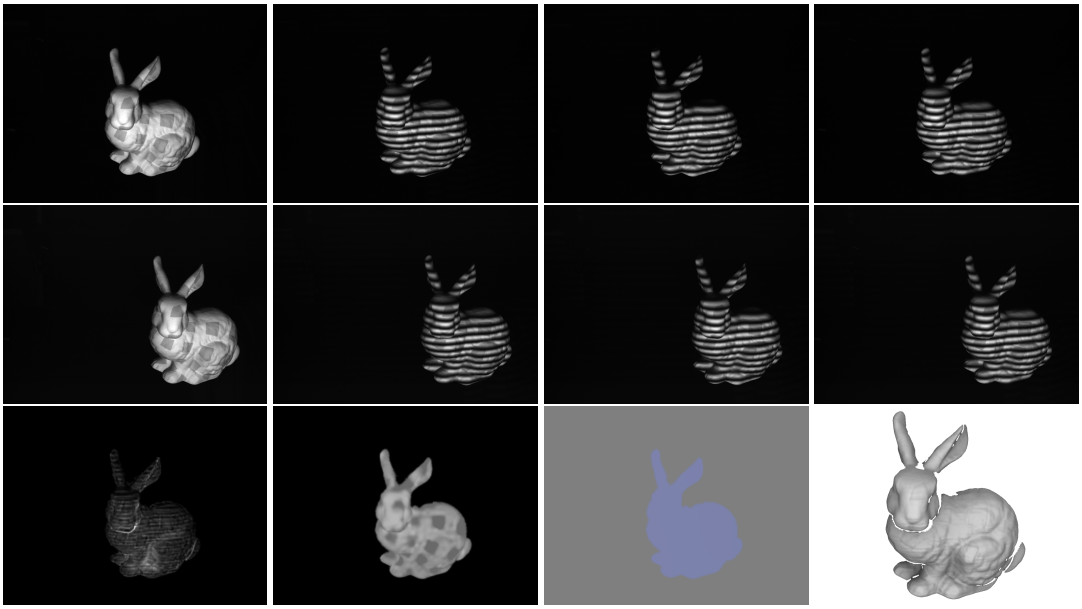


Figure 2. Additional results of the real-world experiment. Captured images from camera 1 and camera 2 under uniform lighting and three phase-shifting patterns (first and second rows), and residual (brightness ×5), reflectance, and displacement field projected to the camera 1 view and the reconstructed 3D shape after 10,000 training iterations (third row).

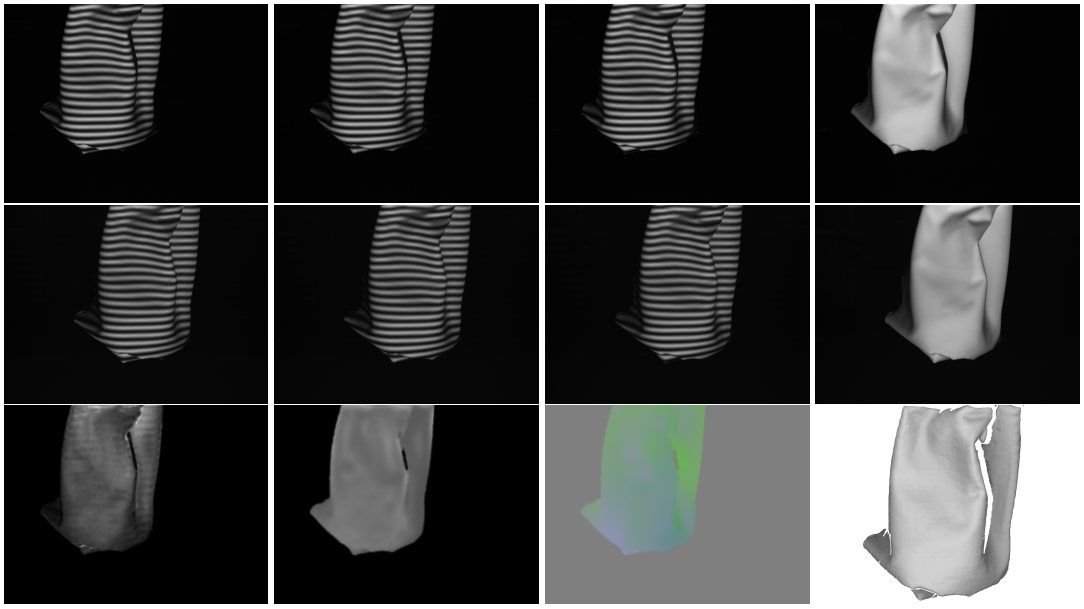


Figure 3. Additional results of the real-world experiment. Captured images from camera 1 and camera 2 under uniform lighting and three phase-shifting patterns (first and second rows), and residual (brightness ×5), reflectance, and displacement field projected to the camera 1 view and the reconstructed 3D shape after 10,000 training iterations (third row).

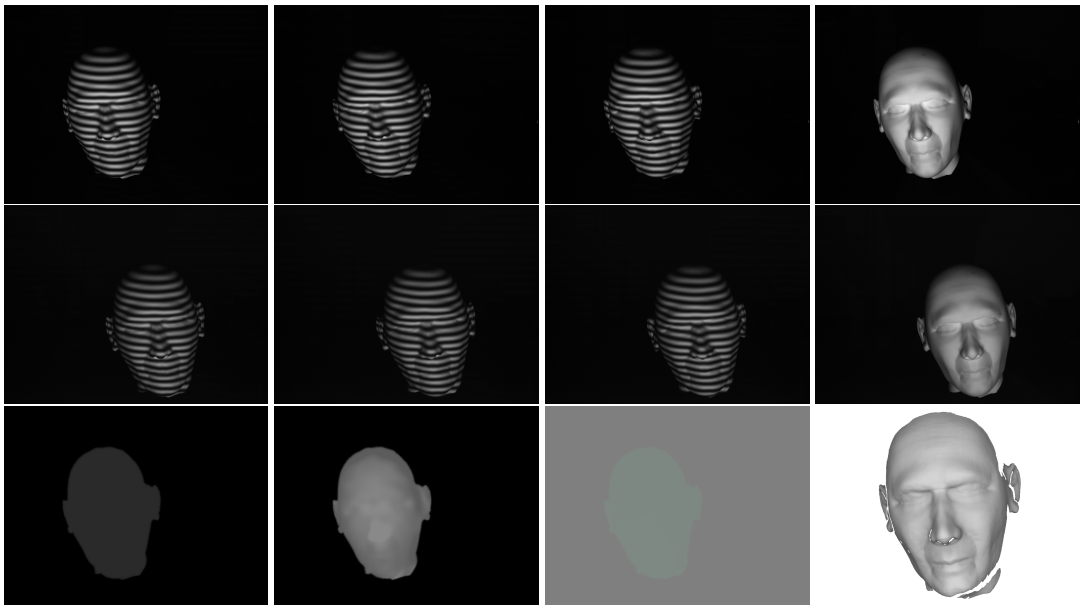


Figure 4. Additional results of the real-world experiment. Captured images from camera 1 and camera 2 under uniform lighting and three phase-shifting patterns (first and second rows), and residual (brightness ×5), reflectance, and displacement field projected to the camera 1 view and the reconstructed 3D shape after 10,000 training iterations (third row).

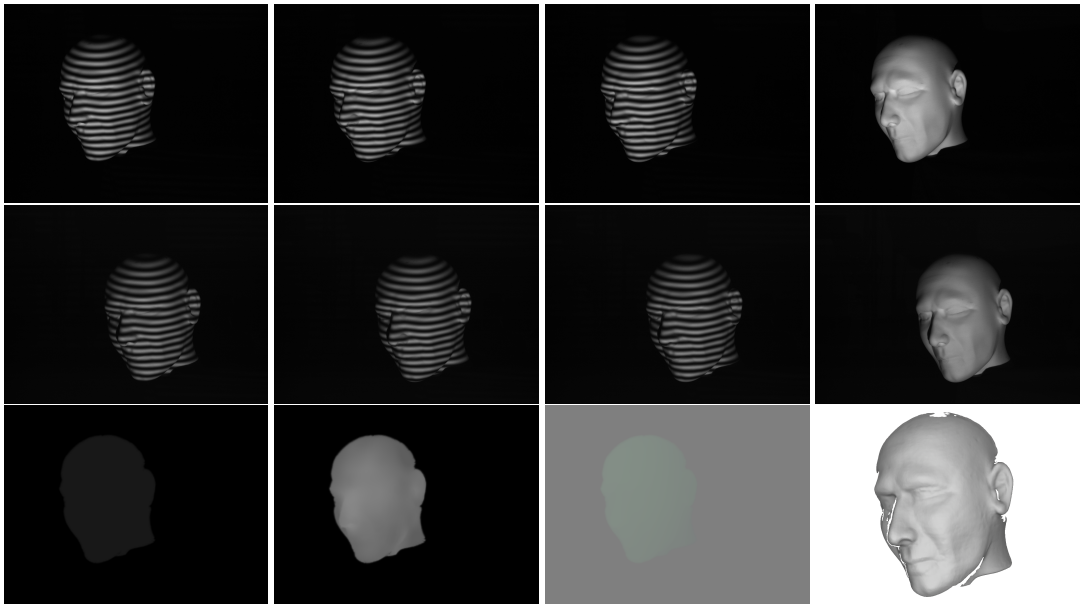


Figure 5. Additional results of the real-world experiment. Captured images from camera 1 and camera 2 under uniform lighting and three phase-shifting patterns (first and second rows), and residual (brightness ×5), reflectance, and displacement field projected to the camera 1 view and the reconstructed 3D shape after 10,000 training iterations (third row).

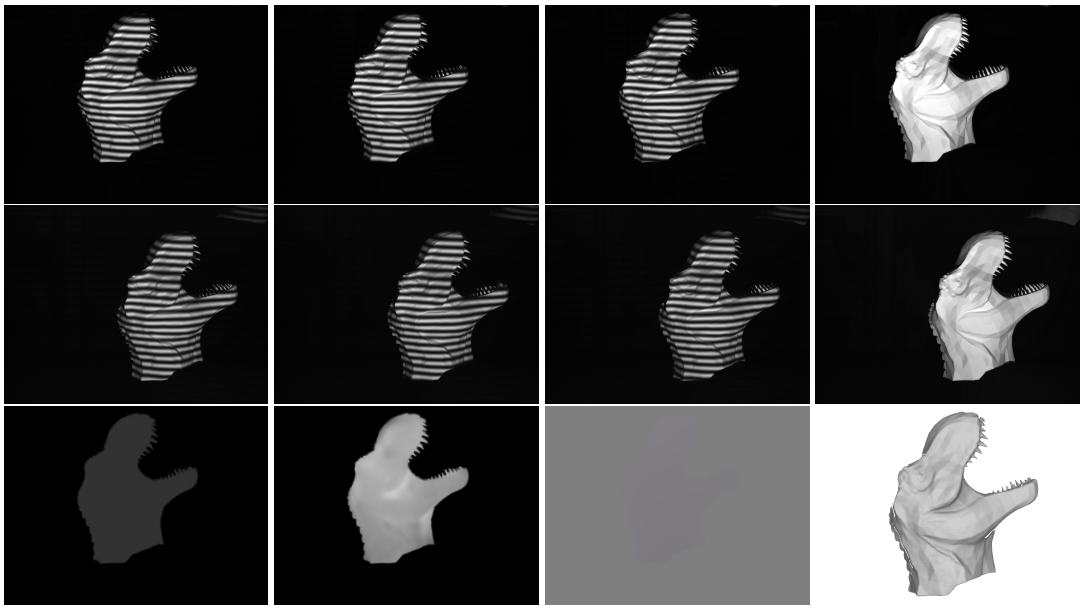


Figure 6. Additional results of the real-world experiment. Captured images from camera 1 and camera 2 under uniform lighting and three phase-shifting patterns (first and second rows), and residual (brightness ×5), reflectance, and displacement field projected to the camera 1 view and the reconstructed 3D shape after 10,000 training iterations (third row).

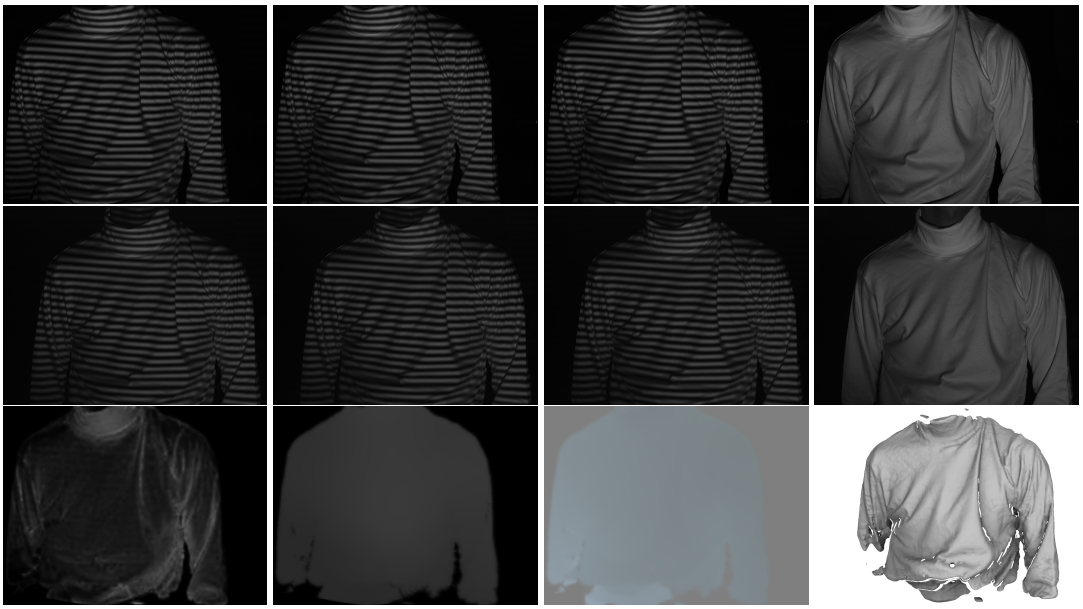


Figure 7. Additional results of the real-world experiment. Captured images from camera 1 and camera 2 under uniform lighting and three phase-shifting patterns (first and second rows), and residual (brightness ×5), reflectance, and displacement field projected to the camera 1 view and the reconstructed 3D shape after 10,000 training iterations (third row).

Figure 8. Additional results of the real-world experiment. Captured images from camera 1 and camera 2 under uniform lighting and three phase-shifting patterns (first and second rows), and residual (brightness ×5), reflectance, and displacement field projected to the camera 1 view and the reconstructed 3D shape after 10,000 and 20,000 training iterations (third and fourth rows).

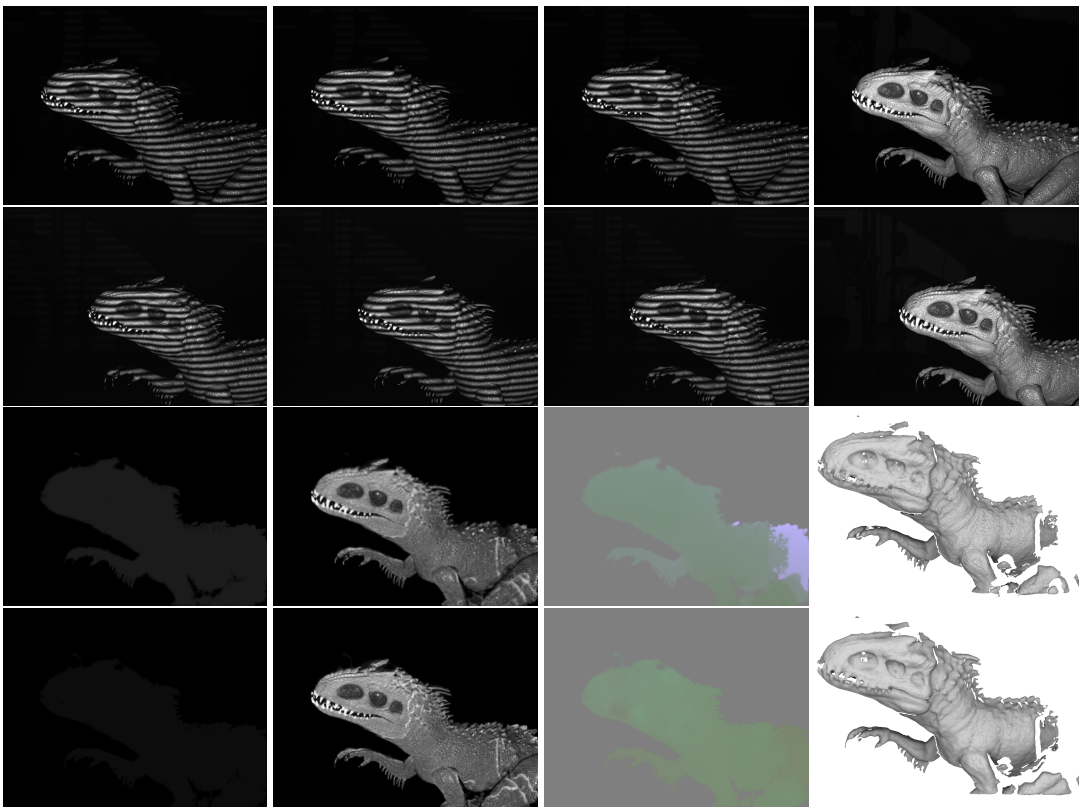


Figure 9. Additional results of the real-world experiment. Captured images from camera 1 and camera 2 under uniform lighting and three phase-shifting patterns (first and second rows), and residual (brightness ×5), reflectance, and displacement field projected to the camera 1 view and the reconstructed 3D shape after 10,000 and 20,000 training iterations (third and fourth rows).

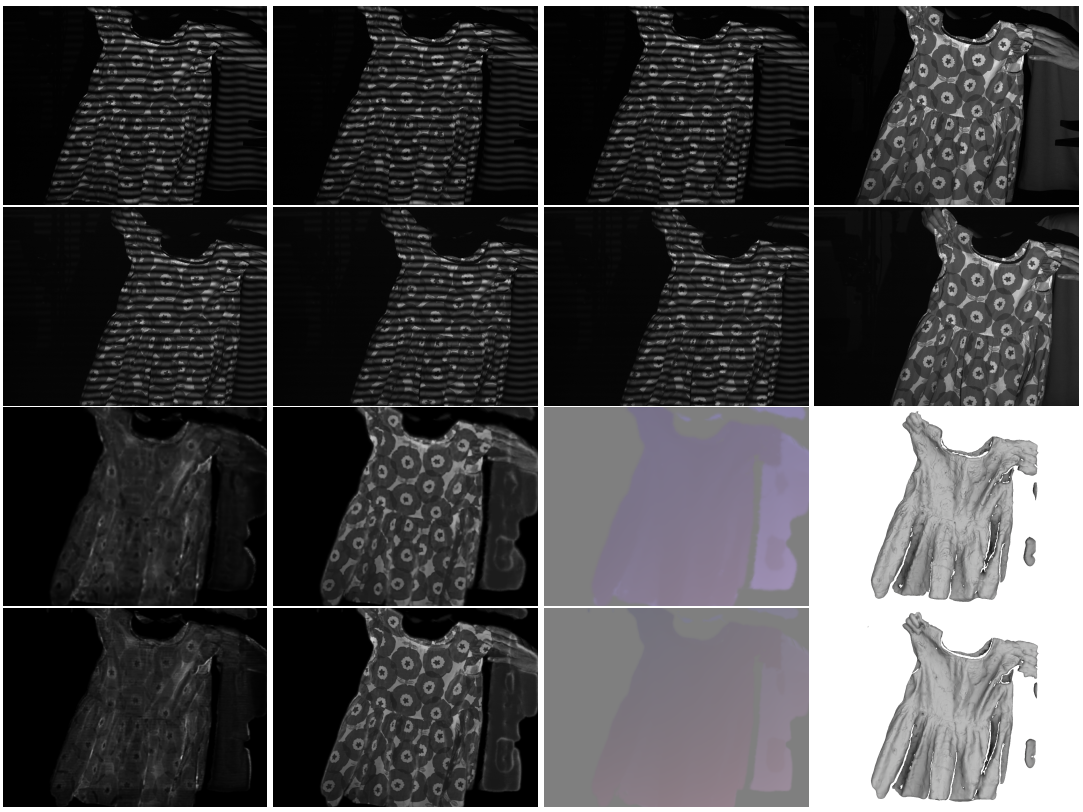


Figure 10. Additional results of the real-world experiment. Captured images from camera 1 and camera 2 under uniform lighting and three phase-shifting patterns (first and second rows), and residual (brightness ×5), reflectance, and displacement field projected to the camera 1 view and the reconstructed 3D shape after 10,000 and 20,000 training iterations (third and fourth rows).

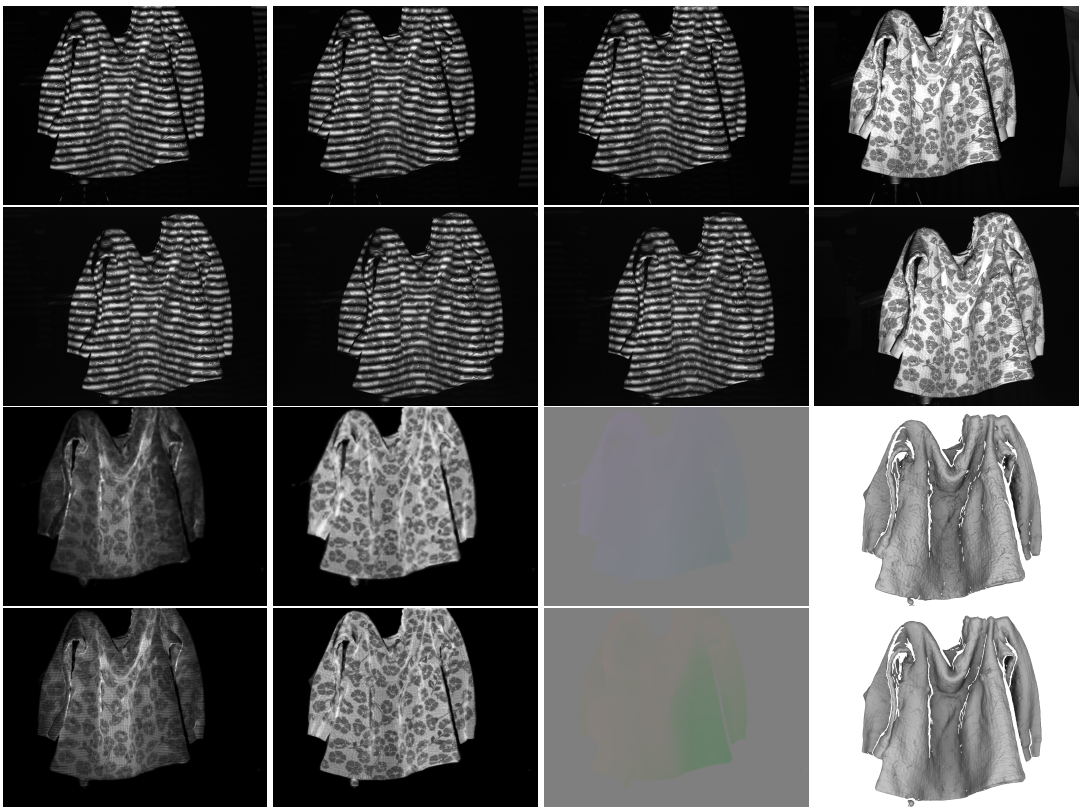


Figure 11. Additional results of the real-world experiment. Captured images from camera 1 and camera 2 under uniform lighting and three phase-shifting patterns (first and second rows), and residual (brightness ×5), reflectance, and displacement field projected to the camera 1 view and the reconstructed 3D shape after 10,000 and 20,000 training iterations (third and fourth rows).

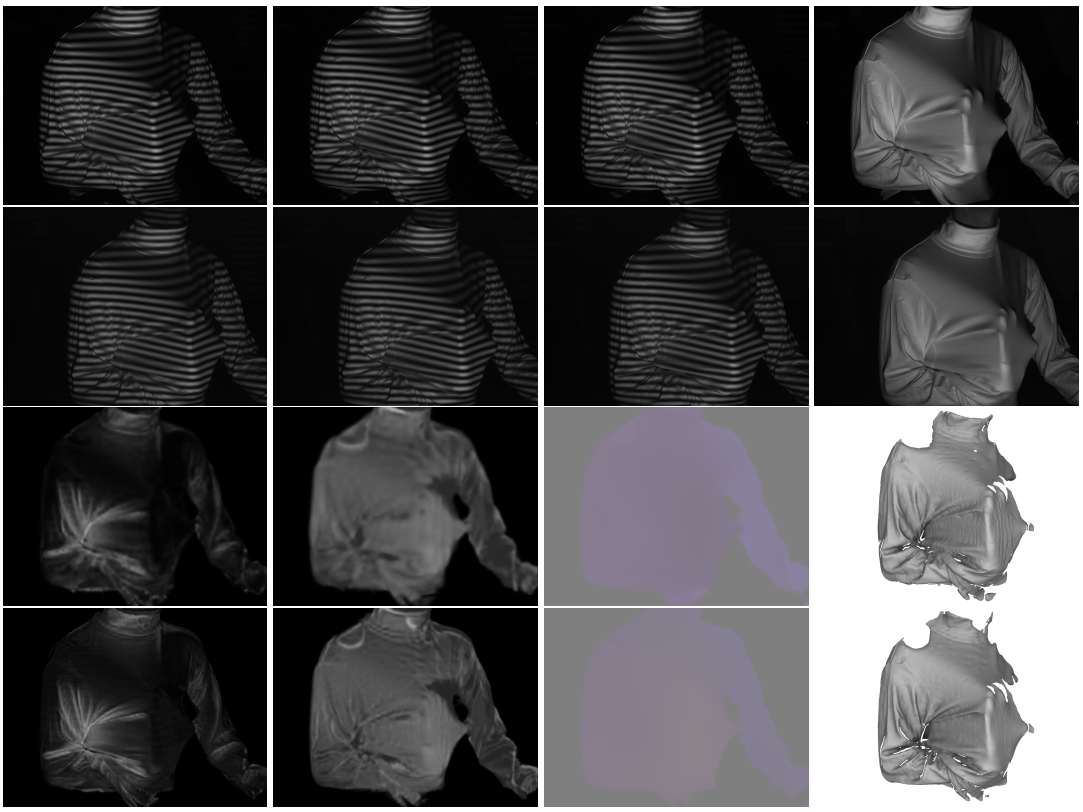


Figure 12. Additional results of the real-world experiment. Captured images from camera 1 and camera 2 under uniform lighting and three phase-shifting patterns (first and second rows), and residual (brightness ×5), reflectance, and displacement field projected to the camera 1 view and the reconstructed 3D shape after 10,000 and 20,000 training iterations (third and fourth rows).

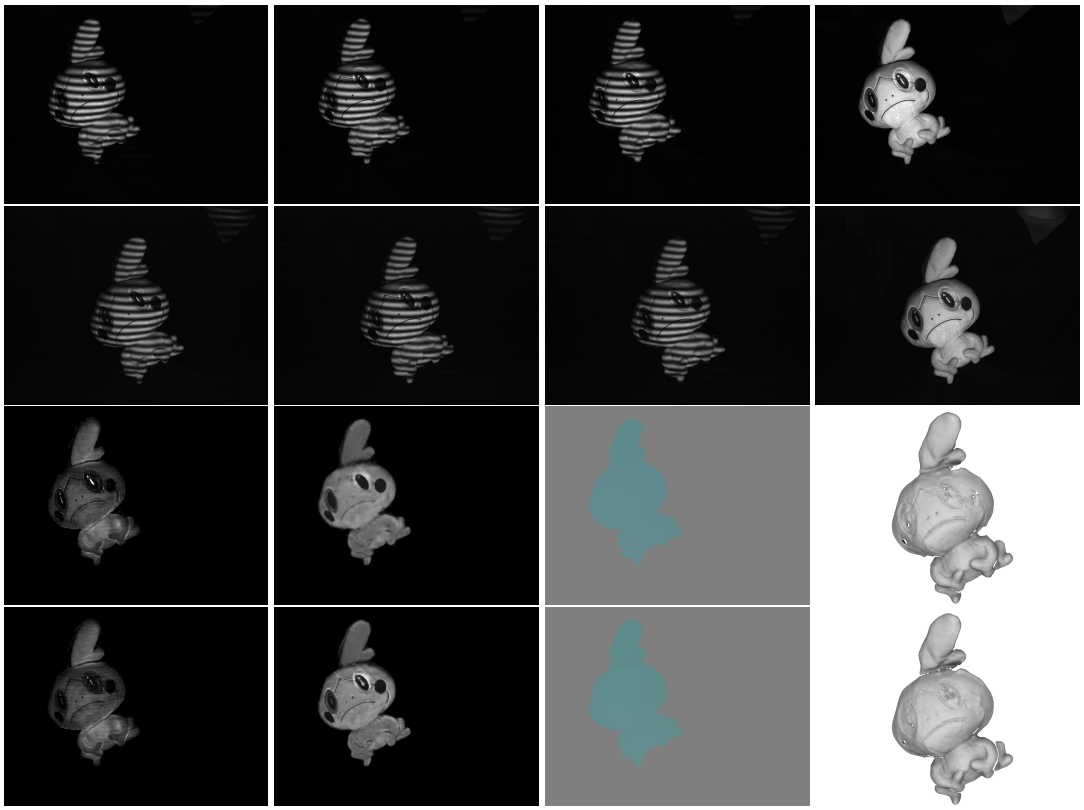


Figure 13. Additional results of the real-world experiment (Pokémon, Sobble). Captured images from camera 1 and camera 2 under uniform lighting and three phase-shifting patterns (first and second rows), and residual (brightness ×5), reflectance, and displacement field projected to the camera 1 view and the reconstructed 3D shape after 10,000 and 20,000 training iterations (third and fourth rows).

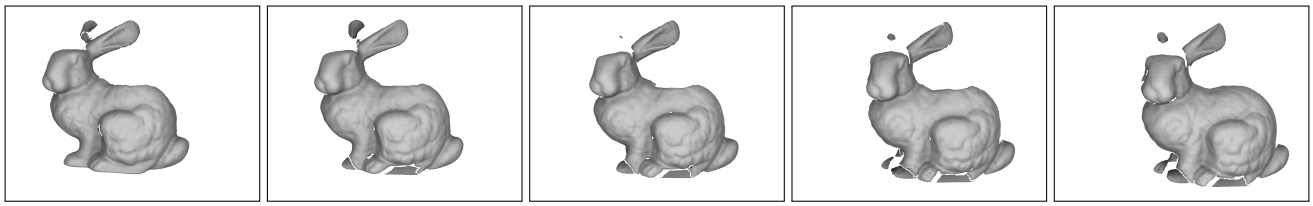


Figure 14. 3D shape reconstruction results for a video sequence, originally captured every 2 ms, with five frames presented at 72 ms intervals.

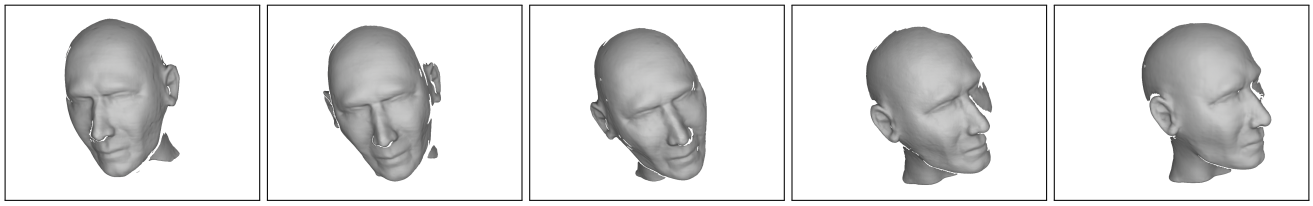


Figure 15. 3D shape reconstruction results for a video sequence, originally captured every 2 ms, with five frames presented at 72 ms intervals.