Supplemental Material for Neural SDF for Shadow-aware Unsupervised Structured Light

Kazuto Ichimaru Diego Thomas Takafumi Iwaguchi Hiroshi Kawasaki
Kyushu University, Japan
https://www.cvg.ait.kyushu-u.ac.jp/index.html

Pattern	Method	Synthetic data (Lego)		Real data (Dog)	
		w ambient	w/o ambient	w ambient	w/o ambient
No	NeuS	6.91	8.92 (+2.01)	7.62	8.45 (+0.83)
Random dot	USSL	6.47	7.07 (+0.60)	7.36	7.79 (+0.43)
Hamming	USSL	6.84	7.31 (+0.47)	6.70	6.38 (-0.32)

Table 1. Results of quantitative evaluation with and without ambient illumination.

1. Effects of ambient light

As mentioned in the paper, we assume that the scenes are dark. However, we did not show the effects of ambient light in the paper due to page limitations. In fact, we tested our method under several ambient light conditions, including complete darkness (*i.e.*, no illumination), to answer the question that is texture information actually dominant with SL only making a small contribution, or can SL alone achieve sufficient accuracy?

Specifically, we synthesized a dataset without ambient illumination for synthetic evaluation and also captured real data without room light for real-world evaluation. We then ran the evaluations as described in the main paper. As a comparative method, we also evaluated NeuS without ambient illumination, *i.e.*, with completely black images, where the mask supervision was the only clue.

The results are shown in Figure 1 and Table 1. As expected, our method maintained almost the same accuracy across all ambient conditions, whereas NeuS's accuracy drastically decreased as it got darker. Surprisingly, the accuracy of the proposed method on real data was a little improved without ambient light compared to with ambient light. We believe this is because the projected pattern was observed with higher contrast. These results encourage us to utilize the proposed system in extreme environments like undersea, where ambient illumination is missing as sunlight is heavily attenuated.

2. Limitation on few-shot case

Through our experiments, we observed that when the number of viewpoints is extremely limited, such as with only five views, USSL tends to fail in accurately reconstructing shapes, whereas ActiveNeuS remains relatively stable, as illustrated in Figure 2. We assume this is due to the number of unknowns exceeding the constraints in such few-shot scenarios. Let N represent the number of 3D points in the scene and I represent the number of viewpoints. In a conventional structured light (SL) context, the unknowns are N correspondences along the epipolar lines, and the constraints are IN. Thus, a single image is sufficient for reconstruction if the projected pattern successfully reduces ambiguity in correspondence search. However, if information about the projected pattern is absent, the unknowns increase to $(3+1)N$ (the RGB values of the projected pattern), while the constraints remain the same, requiring at least four viewpoints. We assume the number of required viewpoints increases in the Neural SDF context, as we model reflection using a MLP.

To address this issue, we experimented with a weakly-supervised SL (WSSL) approach, where information about the projected pattern is implicitly utilized by concatenating the RGB values of the projected pattern with the pattern feature vector. As shown in Figure 2 (right), WSSL significantly improves qualitative reconstruction accuracy in few-shot scenarios, as anticipated. Although we proposed USSL (completely without pattern information) for theoretical interest, WSSL may be advantageous in few-shot scenes in practice.

3. Convergence speed

Given that USSL is expected to converge faster than NeuS, we compared their convergence speeds on the NeRF-Synthetic (Lego) dataset using the Hamming pattern. Specifically, we evaluated the reconstructed meshes at various stages up to 150k training steps out of the total 300k steps (Figure 4). Qualitatively, there is little noticeable difference between NeuS and USSL in terms of convergence

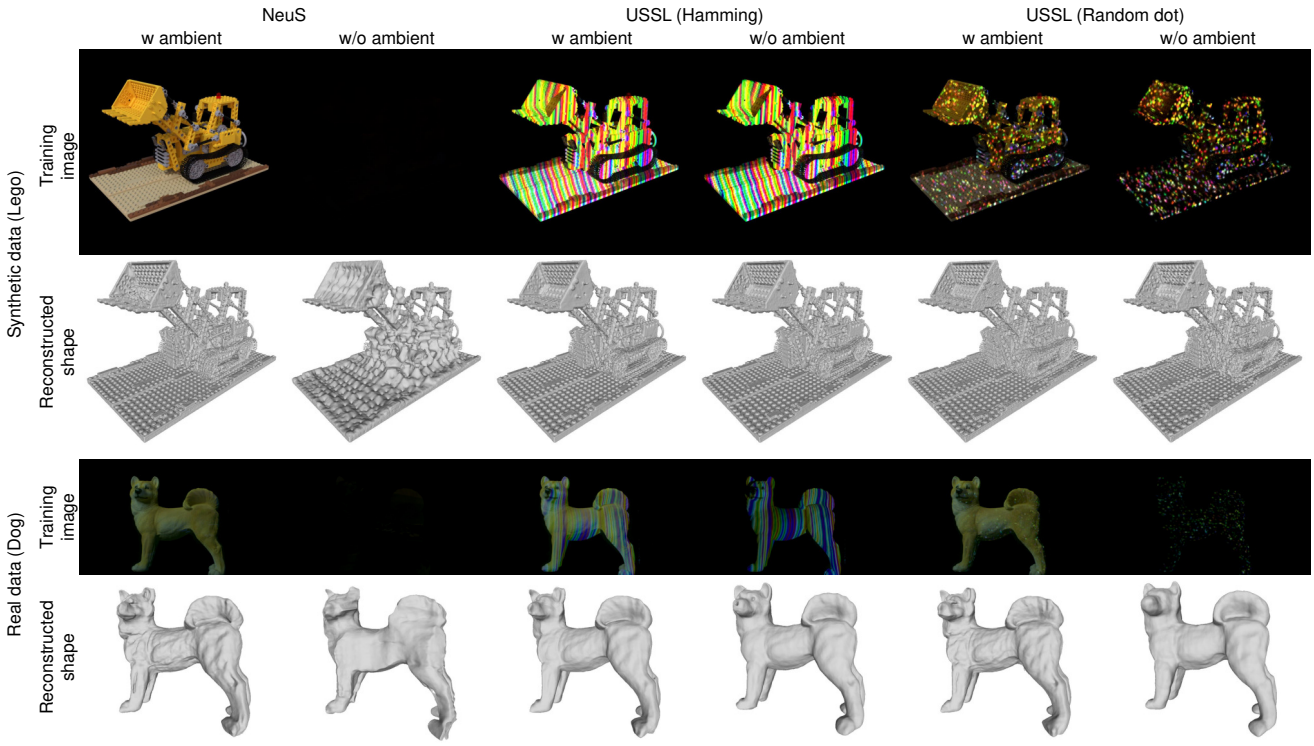

Figure 1. Results of qualitative evaluation with and without ambient illumination.

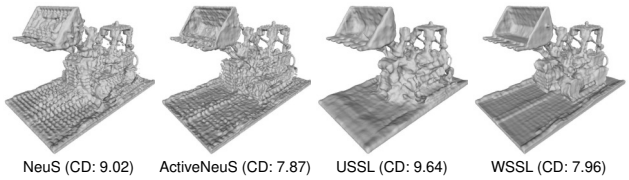

Figure 2. Evaluation of WSSL compared to NeuS, ActiveNeuS, and USSL.

speed per training step. However, USSL requires significantly more computation, resulting in approximately three times slower training in wall clock time. Note that, both USSL and NeuS use hashgrid encoding for their SDF MLP, thus, NeuS is almost identical to Neuralangelo without curvature loss.

Despite this, quantitative metrics reveal a consistent advantage of USSL over NeuS, particularly evident shortly after the training begins, as shown in Figure 3. These results suggest that while USSL demonstrates superior convergence speed, but improving its computational efficiency remains a critical challenge.

Additionally, it should be noted that running an experiment without shadow ray pruning was not feasible due to VRAM limitations.

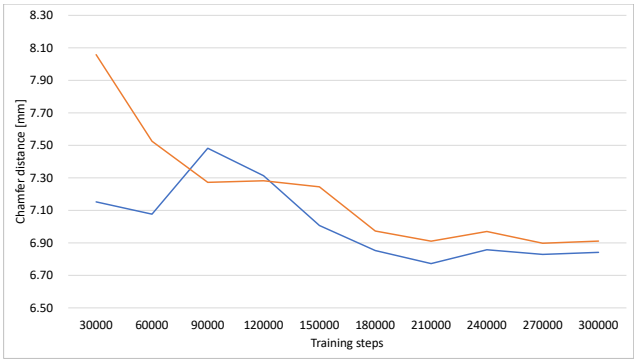

Figure 3. Quantitative difference in convergence speed between NeuS and USSL in NeRF-Synthetic (Lego).

4. Scene texture recovery

Scene texture recovery is another critical task related to SL, as users typically seek the original scene texture without the interference of pattern projection. In the proposed pipeline, the outputs from the color MLP and reflection MLP are blended with learnable parameters during training to explicitly account for external illumination beyond the original ambient lighting. Consequently, it may be possible to recover the scene texture by disabling the blending mechanism and using only the output from the color MLP to render the image.

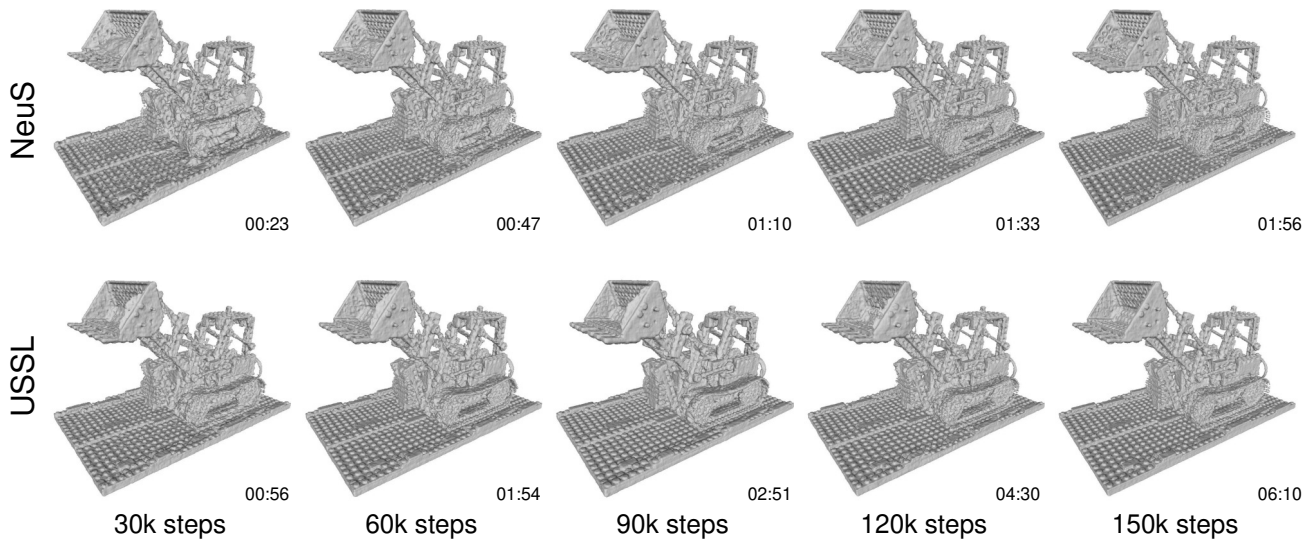

Figure 4. Qualitative difference in convergence speed between NeuS and USSL in NeRF-Synthetic (Lego). Right bottom numbers are elapsed time from the training start.

We evaluated scene texture recovery by following this approach, rendering images in two modes: "with pattern" and "pattern removal." The "with pattern" mode is a standard rendering process without any modifications to the proposed pipeline, where the outputs are expected to closely match the training images. In contrast, the "pattern removal" mode renders images without the blending mechanism, aiming to recover the original scene texture.

Figure 5 shows the qualitative results, while Table 2 presents the quantitative results in terms of PSNR compared to the GT. From these results, it is evident that the images rendered in "with pattern" mode are highly accurate. For the "pattern removal" mode, the quality varies depending on the scene. The chair, mic, and block scenes produce images very close to the GT, while the lego and mannequin body scenes appear flat, lacking ambient illumination shading. In the dog and mannequin head scenes, the renderings are quite noisy, highlighting potential challenges in accurately recovering scene textures from in-the-wild images.

An interesting case is the hotdog scene, where the rendering of the hotdogs is relatively accurate, but the brightness of the dish differs significantly. This discrepancy is likely due to severe inter-reflection, where all illumination information on the dish is absorbed by the reflection MLP, including ambient illumination.

In conclusion, while the proposed pipeline shows great potential for scene texture recovery, further improvements are necessary to handle noisy in-the-wild images effectively.

5. Extension of the proposed method

We would like to highlight that the proposed method has the potential to be extended to various tasks, including self-calibration of the structured light (SL) system, system pose refinement, and robust shape reconstruction in the presence of environmental disturbances, among others. While this paper focuses on learning the projected pattern, a similar approach could be applied to estimate parameters such as the camera-to-projector pose, affine distortion of the projector screen, and projector color temperature, by defining these factors as learnable parameters.

Furthermore, it is important to note that the proposed method may demonstrate robustness against various types of noise and could be applicable to in-the-wild images, unlike conventional SL techniques that typically require precise capture configurations. This is analogous to how methods like NeRF or Neural SDF are known for their noise resilience.

6. A big-picture of this line of work

Some may wonder if really there are situations where only the projected pattern is unknown, while the relative transformation between the camera and the projector, and the system's transformation, are all known. We consider this possible in scenarios where the camera configuration is incorrect and there are significant color space changes or defocus blur, when a special projection device (such as a diffractive optical element) does not project a pattern as intended, or when a light source unintentionally projects a certain pattern due to a crack or similar issue. Usually, we can simply reconfigure, remake, or replace the devices, but

Pattern	Mode	NeRF-Synthetic (40 views)				Real dataset (36 views)			
		Lego	Chair	Hotdog	Mic	Dog	Block	Head	Body
Random dot	w pattern	36.22	33.02	36.97	35.00	34.71	40.55	43.26	-
	pattern removal	34.42	33.43	27.40	34.04	29.84	28.24	29.64	-
Hamming	w pattern	28.23	27.69	30.54	32.88	31.70	38.08	40.51	-
	pattern removal	29.63	30.17	25.98	32.21	27.73	29.37	30.06	-
Cross-lasers	w pattern	37.99	36.42	37.61	35.79	-	-	-	29.84
	pattern removal	27.49	36.61	21.85	34.63	-	-	-	22.14

Table 2. Results of quantitative evaluation of scene texture recovery.

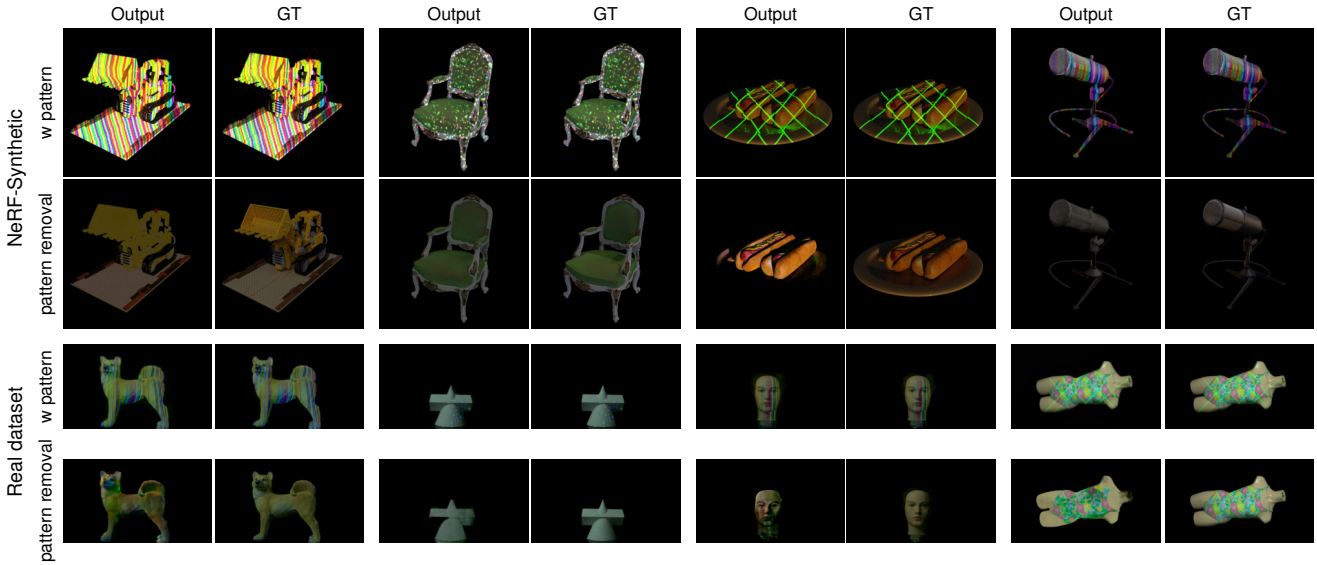

Figure 5. Example images of scene texture recovery.

this might not be possible in certain situations, such as with unmanned underwater vehicles (UUV) in the deep sea.

Furthermore, we are not solely focused on unknown patterns; we are also interested in refining the projector pose, intrinsic parameters, and system pose, and this paper addresses one aspect of this broader picture. Ultimately, we aim to develop a fully calibration-free SL system for simultaneous localization and mapping (SLAM) in extreme environments.