Energy-Efficient Adaptive 3D Sensing: Supplementary Technical Report

Brevin Tilmon[1*] Zhanghao Sun[2] Sanjeev J. Koppal[1] Yicheng Wu[3]
Georgios Evangelidis[3] Ramzi Zahreddine[3] Gurunandan Krishnan[3] Sizhuo Ma[3†] Jian Wang[3†]
[1]University of Florida [2]Stanford University [3]Snap Inc.

1. Derivation of Minimum Eye-Safety Distance

Here we give a detailed discussion on Eq. 9 in the main paper. We expand MPE based on definitions from ANSI Z136:

$$\frac{MPE(t_1)}{t_1} = \frac{C_\lambda t_1^{0.75} 10^{-3}\ (\mathrm{J} \cdot \mathrm{cm}^{-2})}{t_1} = k_e t_1^{-0.25}, \quad (1)$$

where k_e is a method-independent constant. According to ANSI Z136, this analytic expression for MPE we use holds for a wide range of wavelengths from visible light to NIR light (including commonly used $630nm$, $800nm$, $938nm$, and $1064nm$), as long as the temporal duration is longer than $18\mu s$. When SWIR laser such as $1550nm$ is used, eye-safety is not a concern, and our system's SNR gain still holds. When the temporal duration is shorter than $18\mu s$, which could happen for point scanning (*e.g.*, $10\mu s$ for each point for a 10fps $10k$-point system), we still use this expansion as an approximation for ease of derivation; it is easy to verify that the conclusion still holds with the exact formula. Rearranging Eq. 1,

$$\begin{aligned} l_{\text{min}} &= \frac{1}{k_e^{0.5}} P^{0.5} \frac{t_1^{0.125}}{a^{0.5}} = \frac{1}{k_e^{0.5}} P^{0.5} \frac{(\frac{T}{R_{t1}})^{0.125}}{(\frac{A}{R_a})^{0.5}} \\ &= \frac{1}{k_e^{0.5}} \frac{T^{0.125}}{A^{0.5}} P^{0.5} \frac{R_a^{0.5}}{R_{t1}^{0.125}}, \end{aligned} \quad (2)$$

and we have

$$l_{\text{min}} = k_l \cdot P^{0.5} R_a^{0.5} R_{t1}^{-0.125}, \quad (3)$$

where k_l is a method-independent constant.

2. Comparison with Other Sensor Designs

2.1. More Design Variations

In addition to the depth sensor designs mentioned in the main paper, in this section we discuss a few more existing or contrived design variations. A comparison between all design variations is given in Fig. 1.

Naive adaptive scanning. One naive way to implement adaptive sensing is to use a laser and 2D MEMS mirror to scan the ROI and use a 2D camera to capture the scene. We denote this method as adaptive V0. Although this method can scan the same maximum distance as proposed adaptive V1, the eye-safety distance is much longer.

Adaptive point scanning. An alternative way to implement adaptive sensing is to use a point scanner but scan the points in the ROI only. This can be achieved by a synchronized, co-located laser and single-pixel sensor to get N measurements for each depth map. This adaptive V2 method consumes less power but has a longer eye-safety distance. Another limitation is that the sensor itself cannot determine the ROI and needs an auxiliary camera running in parallel. Nevertheless, this is a competitive method and we expect it to be investigated further in future work.

Integration lens. Instead of a 2D camera, it is also possible to use a single-pixel sensor with an integration lens to sense the entire scene. This has been implemented in [5], which we call point scanning V3. From the analysis, it has exactly the same performance as point scanning V2. Similarly, adaptive V3 replaces the synchronized single-pixel sensor in adaptive V2 with a sensor with integration lens. Comparing to V2, this increases power consumption and is therefore not practical.

2.2. Advantage of Adaptive Sensing

To intuitively demonstrate the advantage of proposed adaptive sensing V1, we compare it with full-frame and line-scanning methods by assuming $N \sim 100$ to 1000, which is consistent with the spatial resolution of most concurrent 3D sensors. For high-resolution depth sensors with $N > 1000$, the gain is even greater.

Compare with full-frame projection (baseline):

- At identical maximum range, eye-safety distance is the same as our approach, and our sensor's power is N^{-1} (0.01 to 0.002) of the baseline.

Compare with line-scanning (SOTA):

- At identical power consumption, our proposed method can sense $N^{0.25}$ (3.16 to 4.73) longer distance than the

[*] Work done during internship at Snap Research.
[†] Co-corresponding authors

	Variations	$R_a/R_{t1}/R_{t2}$	Ambient Light Dominates			Read Noise Dominates	Evaluation
			SNR	Power P	Eye-Safety Distance l_{min}	SNR (readout noise only)	
Point scanning	V1: point laser + single pixel cam; N^2 readouts	$N^2/N^2/N^2$	$\boldsymbol{cN}$	$\boldsymbol{k_p d_{max}^2 N^{-1}}$	$k_{ld} d_{max} N^{0.25}$	e	✕ Low power, but not eye-safe
	V2: point laser + 2D cam; 1 image	$N^2/N^2/1$	c	$k_p d_{max}^2$	$k_{ld} d_{max} N^{0.75}$	e	✕ Worse P and l_{min} than V1
	V3: point laser + single pixel w/ integration lens; N^2 readouts	$N^2/N^2/1$(effective)	c	$k_p d_{max}^2$	$k_{ld} d_{max} N^{0.75}$	e	✕ Same problem as V2
Line scanning	V1: Episcan3d; N readouts	$N/N/N$	$cN^{0.5}$	$k_p d_{max}^2 N^{-0.5}$	$k_{ld} d_{max} N^{0.125}$	e	**SOTA**
	V2: line laser + 2D cam; 1 image	$N/N/1$	c	$k_p d_{max}^2$	$k_{ld} d_{max} N^{0.375}$	e	✕ Worse P and l_{min} than V1
Full pattern	Kinect, RealSense	1/1/1	c	$k_p d_{max}^2$	$\boldsymbol{k_{ld} d_{max}}$	e	**Baseline**
Adaptive	V0: point laser + 2D cam; 1 image	$N^2/N/1$	$\boldsymbol{cN}$	$\boldsymbol{k_p d_{max}^2 N^{-1}}$	$k_{ld} d_{max} N^{0.375}$	$\boldsymbol{eN}$	✕ Much longer l_{min} than V1
	V1: phase SLM or MEMS DOE + 2D cam; 1 image	$N/1/1$	$\boldsymbol{cN}$	$\boldsymbol{k_p d_{max}^2 N^{-1}}$	$\boldsymbol{k_{ld} d_{max}}$	$\boldsymbol{eN}$	✓
	V1-a: K ROIs; 1 image (Usually, 2 < K < 5)	$N/K/1$	cNK^{-1}	$k_p d_{max}^2 N^{-1} K$	$k_{ld} d_{max} K^{0.375}$	eNK^{-1}	✓ Slightly higher P and longer l_{min} than V1
	V1-b: K ROIs; K images (Usually, 2 < K < 5)	$N/K/K$	$cNK^{-0.5}$	$k_p d_{max}^2 N^{-1} K^{0.5}$	$k_{ld} d_{max} K^{0.125}$	eNK^{-1}	✓ But needs K × fps camera
	V2: point laser + receiver; N readouts	$N^2/N/N$	$\boldsymbol{cN^{1.5}}$	$\boldsymbol{k_p d_{max}^2 N^{-1.5}}$	$k_{ld} d_{max} N^{0.125}$	$\boldsymbol{eN}$	✓ Lower P but higher l_{min} than V1. Needs auxiliary camera
	V3: point laser + single pixel w/ integration lens; N readouts	$N^2/N/N^{-1}$(effective)	$cN^{0.5}$	$k_p d_{max}^2 N^{-0.5}$	$k_{ld} d_{max} N^{0.625}$	$\boldsymbol{eN}$	✕ Worse P and l_{min} than V1_3

Figure 1. **Comparison between different design variations.** Here we summarize all the design variations we have explored. Typically, $N = 100 \sim 1000$. Readers are referred to Tab. 1 and Tab. 2 for typical values of N^α and K^α. Our method (including the variations, V1, V1-a, V1-b) outperforms the traditional full-frame pattern and SOTA line scanning method.

	$N^{-0.5}$	$N^{-0.25}$	$N^{-0.125}$	$N^{0.125}$	$N^{0.25}$	$N^{0.375}$	$N^{0.5}$
$N = 100$	0.1	0.32	0.56	1.78	3.16	5.62	10
$N = 500$	0.04	0.21	0.46	2.17	4.73	10.28	22.36
$N = 1000$	0.03	0.18	0.42	2.37	5.62	13.34	31.62

Table 1. **Examples of N^α.**

SOTA line-scanning method, with eye-safety distance increased by $N^{0.125}$ (1.78 to 2.17).

- At identical eye-safety distance, our method consumes power at $N^{-0.25}$ (0.32 to 0.21) of SOTA, and our pro-

	$K^{-0.375}$	$K^{-0.125}$	$K^{0.25}$	$K^{0.375}$	$K^{0.75}$
$K = 1$	1	1	1	1	1
$K = 2$	0.77	0.92	1.19	1.3	1.68
$K = 3$	0.66	0.87	1.32	1.51	2.28
$K = 4$	0.59	0.84	1.41	1.68	2.83
$K = 5$	0.55	0.82	1.50	1.83	3.34
$K = 10$	0.42	0.75	1.78	2.37	5.62

Table 2. **Examples of K^a.**

posed method can sense $N^{0.125}$ (1.78 to 2.17) longer distance than SOTA.

- At identical maximum range, our power consumption is $N^{-0.5}$ (0.1 to 0.04) of SOTA, and our eye-safety distance is $N^{-0.125}$ (0.56 to 0.46) of SOTA.

2.3. Read-Noise Dominated Cases

Although our analysis so far focuses on the case where the noise is dominated by photon noise from the ambient light, we also analyze the SNR for different designs when the noise is dominated by the sensor read noise. Fig. 1 shows that in this case, all different variations of the proposed adaptive scheme perform similarly, and still outperforms existing point scanning, line scanning or full pattern designs. It is important to note that the images are dominated by read noise only when the scene is extremely dark and the laser power per area is extremely low, *i.e.* near the maximum sensing distance. In practice, the noise is likely to be a mixture of ambient photon noise, laser photon noise and read noise. If the specific range of ambient light, laser power budget and sensor read noise is known, it is possible to use Eq. 1 in the main paper directly compute the SNR at different depths for comparison.

3. Implementation Details

3.1. Hardware Prototypes

We use two FLIR BFS-U3-16S2C-CS cameras equipped with 20mm lenses as a stereo pair. Our SLM implementation uses a Holoeye GAEA LCoS (phase-only) SLM, which can display 4K phase maps at 30 frames per second. Our MEMS + DOE implementation uses a $0.8mm$ diameter bonded Mirrorcle MEMS Mirror. Since we did not find an off-the-shelf random dot DOE with a small FOV, we use a Holoeye DE-R 339 DOE that produces a periodic 6×6 dot pattern with 5° FOV. We tilt the DOE such that the pattern is still unique locally on the epipolar line.

Both projectors, lasers, and cameras are synchronized with a Teensy 4.0 microcontroller. The MEMS in our MEMS DOE prototype is controlled with a Mirrorcle PicoAmp 5.4 X200 digital analog converter. We use a Thorlabs L638P200 laser for the SLM implementation, and a Thorlabs HL6385DG laser for the MEMS + DOE implementation. We can drive the lasers from 40mW to 200mW.

For both hardware prototypes we collimate the lasers with a 40mm lens (Thorlabs AC254-040-A-ML). For the SLM, our laser diode is linearly polarized and we rotate it to match the requirements of the SLM incident light. After reflecting off the SLM we magnify the pattern FOV with a 75mm convex lens followed by a 100mm convex lens. After magnification our FOV is 35 degrees, but could easily be increased with a different optical design.

Pipeline. Following shows our system's pipeline:

- The Teensy triggers the cameras and the captured images are sent to the NVIDIA Jetson Nano for processing.
- The attention map is computed which determines where to project light.
- For our SLM implementation, a hologram is generated based on the attention map on the Jetson Nano. For our MEMS DOE prototype, digital voltage values are sent to the DAC based on the attention map so that the MEMS points the DOE pattern at the ROI.

This pipeline continues over the video sequence.

3.2. Calibration

We calibrate the intrinsics and extrinsics for each camera in the stereo pair. In our prototype, we place the SLM very close to the left camera in the pair such that the pixel correspondences between the SLM and left camera can be approximated via homography. It is possible to co-locate the camera and SLM with a beamsplitter to make the correspondences perfectly independent of depth.

We compute a homography between the SLM and closest camera by first generating a grid pattern hologram and projecting it to a plane. We then detect the grid points in the closest camera after undistorting and rectifying. Since we know the correspondence between the scene point and the slope of our mirror hologram generator, we can compute a homography to map between them. This enables selectively illuminating camera pixels by warping from camera space to SLM space based on the estimated homography.

Our projector lenses induce 3-5 pixel pincushion distortion at the far regions of the projector FOV to varying degrees in both the left and right stereo cameras, among other minor distortions. However, since we have both a homography and the ability to generate freeform projector patterns with the SLM, we simply compensate for this distortion by warping ideal patterns (such as straight lines for line scanning) from the rectified camera space to SLM space. For example, when emulating line scanning, this results in lines being straight in the camera and curved on the SLM, which makes stitching each line image together highly accurate. It is critical to warp from the undistorted and rectified camera space so that the generated patterns are on the same line of each camera so that we can stitch parallel lines from the left and right camera for stereo matching. This increases the number of possible lines we can use and improves energy efficiency gains since N from Fig. 1 increases. This precise calibration functionality is not possible in conventional line scanning systems like [4] since the projector line shape is fixed.

3.3. Stereo Depth Estimation

We choose semi global matching [3] as our depth algorithm for our experiments instead of recent, learning-based methods to clearly demonstrate the benefit of our theory. This is because it is well understood that block matching fails on textureless regions, while learning-based methods generate plausible but inaccurate depths on those regions. Based on recent work that shows active stereo improving learning based depth estimation [1, 2, 6], we believe it is likely our theory translates to learning-based depth estimation as well.

3.4. Real Time Demonstration

We provide the code used to simulate and run our phase SLM at https://github.com/btilmon/holoCu. The code enables simulating various 3D sensors and generating desired phase maps for a real SLM in real time through a fused CUDA kernel of Fresnel Holography. We achieve 30 frames per second using a 1080p hologram size with 100 dots, which enables the SLM to react to the environment in real time. In a loop, we first capture a passive stereo pair with no projector, then we randomly select pixel coordinates where depth uncertainty is highest. We then warp these coordinates to SLM space based on the estimated homography and compute the hologram to be displayed on the SLM. This process repeats for a video sequence.

4. Global Light Analysis

We summarize two major cases for multi-bounce in Fig. 2. In the case of scattering medium, both full-frame projection and line scanning suffers from the one-bounce path on the epipolar plane, while the proposed adaptive method has very little energy on the epipolar plane and therefore stray light mostly comes from paths with at least two bounces, which are much weaker than one-bounce paths. In the case of inter-reflection, both line scanning and adaptive method are robust to multi-bounce since fewer dots are projected and fewer multi-bounce paths exists. To summarize, the proposed method performs better than line scanning in scattering medium and performs comparably with line scanning when there is inter-reflection.

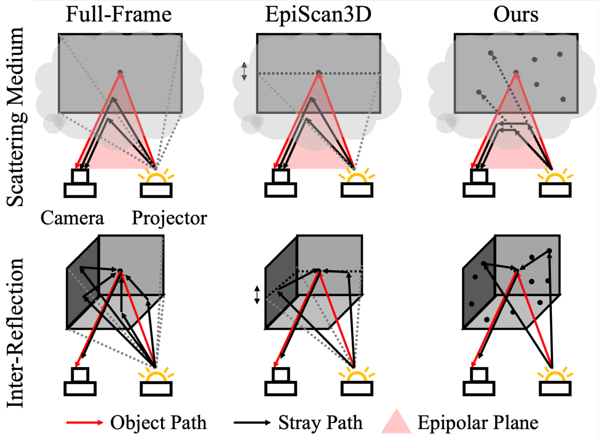

Figure 2. Global light analysis.

References

[1] Seung-Hwan Baek and Felix Heide. Polka lines: Learning structured illumination and reconstruction for active stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 4

[2] Sean Ryan Fanello, Julien Valentin, Christoph Rhemann, Adarsh Kowdle, Vladimir Tankovich, Philip Davidson, and Shahram Izadi. Ultrastereo: Efficient learning-based matching for active stereo systems. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6535–6544, 2017. 4

[3] Fixstars. libsgm. `https://github.com/fixstars/libSGM`. 4

[4] Matthew O'Toole, Supreeth Achar, Srinivasa G. Narasimhan, and Kiriakos N. Kutulakos. Homogeneous codes for energy-efficient illumination and imaging. *ACM Trans. Graph.*, 34(4), jul 2015. 4

[5] Francesco Pittaluga, Zaid Tasneem, Justin Folden, Brevin Tilmon, Ayan Chakrabarti, and Sanjeev J Koppal. Towards a mems-based adaptive lidar. In *2020 International Conference on 3D Vision (3DV)*, pages 1216–1226. IEEE, 2020. 1

[6] Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser, and Sean Fanello. Activestereonet: End-to-end self-supervised learning for active stereo systems. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII*, page 802–819, Berlin, Heidelberg, 2018. Springer-Verlag. 4