Energy-Efficient Adaptive 3D Sensing

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.

Abstract

Active depth sensing achieves robust depth estimation but is usually limited by the sensing range. Naively increasing the optical power can improve sensing range but induces eye-safety concerns for many applications, including autonomous robots and augmented reality. In this paper, we propose an adaptive active depth sensor that jointly optimizes range, power consumption, and eye-safety. The main observation is that we need not project light patterns to the entire scene but only to small regions of interest where depth is necessary for the application and passive stereo depth estimation fails. We theoretically compare this adaptive sensing scheme with other sensing strategies, such as full-frame projection, line scanning, and point scanning. We show that, to achieve the same maximum sensing distance, the proposed method consumes the least power while having the shortest (best) eye-safety distance. We implement this adaptive sensing scheme with two hardware prototypes, one with a phase-only spatial light modulator (SLM) and the other with a micro-electro-mechanical (MEMS) mirror and diffractive optical elements (DOE). Experimental results validate the advantage of our method and demonstrate its capability of acquiring higher quality geometry adaptively. Please see our project website for video results and code: https://btilmon.github.io/e3d.html.

1. Introduction

Active 3D depth sensors have diverse applications in augmented reality, navigation, and robotics. Recently, these sensor modules are widely used in consumer products, such as time-of-flight (*e.g.*, Lidar [15]), structured light (*e.g.*, Kinect V1 [18]) and others. In addition, many computer vision algorithms have been proposed to process the acquired data for downstream tasks such as 3D semantic understanding [29], object tracking [17], guided upsampling in SLAM [24], *etc*.

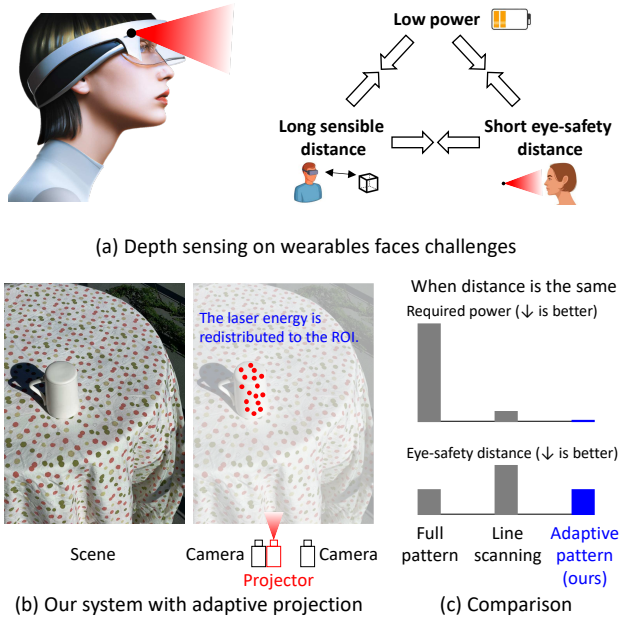


Figure 1. **Energy-efficient adaptive 3D sensing.** **(a)** Depth sensing devices have three key optimization goals: minimizing the power consumption and eye-safety distance while maximizing sensing distance. However, these goals are coupled to each other. **(b)** We propose a novel adaptive sensing method with an active stereo setup and a projector that can redistribute the light energy and project the pattern only to the required regions (e.g., textureless regions). **(c)** The proposed approach outperforms previous methods including full-frame pattern systems (like Intel RealSense) and line-scanning systems (like Episcan3D [22]): When the maximum sensing distance is the same, the required power is much less and the eye-safety distance is also shorter.

Unlike stereo cameras that only sense reflected ambient light passively, active depth sensors illuminate the scene with modulated light patterns, either spatially, temporally, or both. The illumination encodings allow robust estimation of scene depths. However, this also leads to three shortcomings: First, active depth sensors consume optical power, burdening wearable devices that are on a tight power budget. Second, the number of received photons reflected back from the scene drops with inverse-square relationship

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

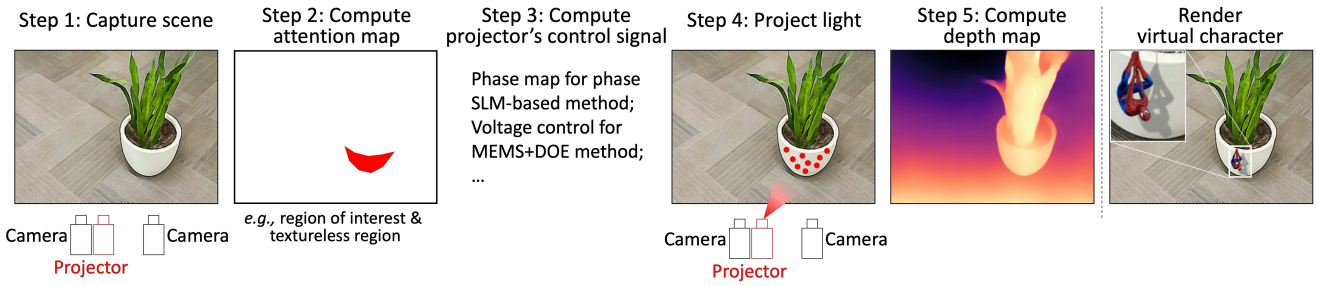


Figure 2. **Method overview.** Our system consists of a stereo-camera pair and an adaptive projector. The system first captures an image of the scene (Step 1) and then computes an attention map to determine the ROI (Step 2). This attention map is used to compute the control signal for the specific projector implementation (Step 3) such that light is only projected to the ROI (Step 4). A depth map is then computed from the new stereo images, which can be used for applications such as augmented reality (Step 5).

to scene depth. The maximum sensing distance is thereby limited by the received signal-to-noise ratio (SNR). Third, strong, active light sources on the device may unintentionally hurt the user or other people around. For consumer devices, this constraint can be as strict as ensuring safety when a baby accidentally stares at the light source directly. Interestingly, these three factors are often entangled with each other. For example, naively increasing range by raising optical power makes the device less eye-safe. An active 3d sensor would benefit from the joint optimization of these three goals, as illustrated in Fig. 1(a).

In this paper, we present an adaptive depth-sensing strategy. Our key idea is that the coded scene illumination need not be sent to the entire scene (Fig. 1(b)). Intuitively, by limiting the illumination samples, the optical power per sample is increased, therefore extending the maximum sensing distance. This idea of *adaptive sensing* is supported by three observations: First, illumination samples only need to be sent to parts of a scene where passive depth estimation fails (*e.g.*, due to lack of texture). Second, depth estimation is often application-driven, *e.g.*, accurate depths are only needed around AR objects to be inserted into the scene. Finally, for video applications, sending light to regions where depths are already available from previous frames is redundant. Based on these observations, we demonstrate this adaptive idea with a stereo-projector setup (*i.e.*, active stereo [4, 9, 34]), where an attention map is computed from the camera images for efficient light redistribution.

To quantitatively understand the benefits of our approach, we propose a sensor model that analytically characterizes various sensing strategies, including full-frame (*e.g.*, RealSense [19]), line-scanning (*e.g.*, Episcan3D [22]), point-scanning (*e.g.*, Lidar [25]) and proposed adaptive sensing. We establish, for the first time, a framework that jointly analyzes the power, range, and eye-safety of different strategies and demonstrates that, for the same maximum sensing distance, adaptive sensing consumes the least power while achieving the shortest (best) eye-safety distance.

Note that the realization of scene-adaptive illumination is not trivial: Common off-the-shelf projectors simply block part of the incident light, which wastes optical power. We propose two hardware implementations for adaptive illumination: One is inspired by digital holography, which uses Liquid Crystal on Silicon (LCoS) Spatial Light Modulators (SLM) to achieve free-form light projection. The other implementation uses diffractive optical elements (DOE) to generate dot patterns in a local region of interest (ROI), which is directed to different portions of the scene by a micro-electro-mechanical (MEMS) mirror.

Our contributions are summarized as follows:

- **Adaptive 3D sensing theory**: We propose adaptive 3D sensing and demonstrate its advantage in a theoretical framework that jointly considers range, power and eye-safety.
- **Hardware implementation**: We implement the proposed adaptive active stereo approach with two hardware prototypes based on SLM and MEMS + DOE.
- **Experimental validation**: Real-world experimental results validate that our sensor can adapt to the scene and outperform existing sensing strategies.

2. Related Work

Active 3D sensing with ambient light noise. Various techniques have been proposed to address photon noise due to strong ambient light (*e.g.*, sunlight), such as choosing a wavelength where sunlight is weak [23, 32], using a polarizing filter [23]. Gupta *et al.* [13] uses a theoretical model to show that instead of illuminating the full scene, concentrating light on different parts of a scene sequentially improves SNR for structured light, which is demonstrated with a rotating polygonal mirror. Based on similar principles, MC3D [21] uses a MEMS/galvo-driven laser and an event camera to achieve bandwidth-efficient scanning. Episcan3D [22] and EpiToF [1] use a line-scanning laser and a synchronized rolling-shutter camera to achieve fast

and efficient depth sensing. This paper further extends this line of work by showing that, with the freedom to adaptively illuminate part of the scene, a lower power budget is needed to achieve the same sensing range while being safer to the eyes.

Adaptive 3D sensing. Ideas from visual attention [6, 11] have influenced vision and robotics. Efficient estimation algorithms have been shown for adaptive sensing and point-zoom-tilt (PZT) cameras [7, 33]. In the 3D sensor space, 3D light-curtains [5,8,31] represent a flexible system where curtains can be adaptively placed in the scene for robotics, and other applications [2, 3, 26]. Full control of the MEMS mirror enables adaptive sampling for LIDAR [25, 27, 28] and adaptive passive camera resolution for monocular depth sensing [30]. While previous adaptive systems focus on different aspects such as flexibility, frame rate, *etc*., this work studies the interplay between range, power, and eye-safety.

3. Energy-Efficient Adaptive 3D Sensing

The workflow of the proposed adaptive 3D sensing is shown in Fig. 2. We use an active stereo design with two cameras and a projector. The device first captures an image of the scene and computes an attention map to determine the region of interest (ROI). Hardware-specific control signals are computed from the attention map such that the projector redistributes the light to the ROI. Finally, a high-quality depth map can be calculated from captured stereo images.

Before getting into details on how to implement this adaptive illumination in practice, let us assume an ideal *flexible* projector for a moment: If we can redistribute the optical power to an arbitrarily-shaped ROI, how well can it perform? We adopt a model to quantify its depth estimation performance and compare it to other existing or naively conceived active depth sensing strategies.

3.1. Sensor Model and SNR Analysis

The accuracy of various active depth sensors, including structured light, active stereo, continuous-wave time-of-flight (CW-ToF), *etc*., can be quantified by a single metric: the SNR of the measured light signal (projected by the active illumination source and reflected by the scene). The noise consists of the photon noise from both the signal itself and the ambient light, and the sensor readout noise, mathematically defined as follows [13, 14]:

$$SNR = \frac{\text{Signal}_{\text{projector}}}{\sqrt{N^2_{\text{photon\_ambient}} + N^2_{\text{photon\_projector}} + N^2_{\text{read}}}} = \frac{\frac{P}{d^2 a} t_1}{\sqrt{P_{\text{sun}} t_2 + \frac{P}{d^2 a} t_1 + N^2_{\text{read}}}} \quad (1)$$

where P is the optical power of the projector (assuming an albedo of 1 in the scene), a is the illuminated area at unit distance, d is the distance of the scene (thus the inverse-square fall-off), P_{sun} is the optical power of the ambient light, t_1 and t_2 are the duration when the laser is on and the camera is active, respectively, and N_{read} is the standard deviation of the read noise. Ambient light-induced photon noise dominates in outdoor scenarios and also indoors with power-limited devices, which is the major focus of the following analysis. In these situations, SNR can be simplified as:

$$SNR \approx \frac{\frac{P}{d^2 a} t_1}{\sqrt{P_{\text{sun}} t_2}}. \quad (2)$$

When readout noise dominates, which happens in a dark room or at night, SNR can be simplified as this

$$SNR \approx \frac{\frac{P}{d^2 a} t_1}{N_{\text{read}}}. \quad (3)$$

Analyzing different sensing strategies. We use this SNR model to compare the performance of different depth sensors. For a fair comparison, we assume all depth sensors have equal total optical power P, sensing at same depth d. Their performance is then uniquely determined by t_1, t_2 and a. For off-the-shelf *full-frame* projectors (Fig. 3(e)), we denote $a = A$ which corresponds to the entire FOV, and $t_1 = t_2 = T$ as both the sensor and the projector are active during the entire camera exposure T.

Previous work [1, 13, 21, 22] has shown that, instead of flood-illuminating the entire scene, focusing optical power on different parts of the scene sequentially can lead to higher SNR. To quantitatively analyze this effect, we represent the illuminated area, laser exposure and camera exposure as a division of the full-frame case:

$$a = A/R_a, \qquad t_1 = T/R_{t1}, \qquad t_2 = T/R_{t2}, \quad (4)$$

where R_a, R_{t1}, R_{t2} are defined as illuminated area divisor, laser exposure divisor, camera exposure divisor, respectively. SNR is then a function of these divisors:

$$SNR = \frac{\frac{P}{d^2 A/R_a} T/R_{t1}}{\sqrt{P_{\text{sun}} T/R_{t2}}} = \underbrace{\frac{P\sqrt{T}}{d^2 A \sqrt{P_{sun}}}}_{c} \underbrace{\frac{R_a \sqrt{R_{t2}}}{R_{t1}}}_{X}, \quad (5)$$

where c is the SNR of full-frame projection. X is a factor that describes how each method compares with the baseline full-frame projection. It is difficult to optimize X directly since not every combination of (R_a, R_{t1}, R_{t2}) is feasible in hardware. Nevertheless, it provides a useful tool to characterize different sensing strategies.

State-of-the-art systems such as Episcan3D [22] implement this idea as a *line scanning* scheme, as shown in Fig. 3(c). If we assume the total illuminated region of line scanning is the same as full-pattern, then $R_a = R_{t1} = R_{t2} = N$, where N is the number of scanlines (typically

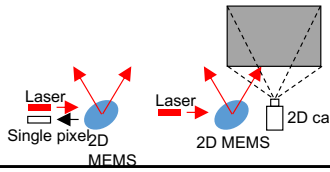
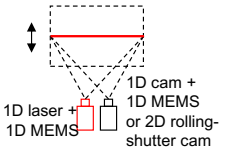
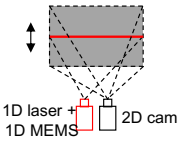
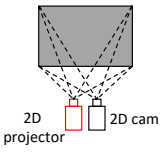
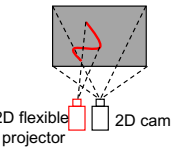
	Point scanning		Line scanning		(e) Full-frame pattern	(f) Adaptive **(Proposed)**
	(a) V1: Synced	(b) V2: Unsynced	(c) V1: Synced	(d) V2: Unsynced		
*Typically, $N = 10^2 \sim 10^3$						
Illuminated Area Divisor (R_a)	N^2	N^2	N	N	1	N
Laser Exposure Divisor (R_{t1})	N^2	N^2	N	N	1	1
Camera Exposure Divisor (R_{t2})	N^2	1	N	1	1	1
Same power, sensing at same distance:						
SNR $= c\frac{R_a\sqrt{R_{t2}}}{R_{t1}}$	$\boldsymbol{cN}$	c	$c\sqrt{N}$	c	c	$\boldsymbol{cN}$
Same maximum sensing distance d_{max}:						
Power P	$\boldsymbol{k_p d_{max}^2 N^{-1}}$	$k_p d_{max}^2$	$k_p d_{max}^2 N^{-0.5}$	$k_p d_{max}^2$	$k_p d_{max}^2$	$\boldsymbol{k_p d_{max}^2 N^{-1}}$
Eye-safety distance l_{min}	$k_{ld} d_{max} N^{0.25}$	$k_{ld} d_{max} N^{0.75}$	$k_{ld} d_{max} N^{0.125}$	$k_{ld} d_{max} N^{0.375}$	$\boldsymbol{k_{ld} d_{max}}$	$\boldsymbol{k_{ld} d_{max}}$

Figure 3. **Schematic diagrams and analysis of various sensing strategies.** When ambient light noise dominates, the proposed adaptive sensing achieves best SNR. To achieve the same maximum sensing distance, adaptive sensing consumes least power while having shortest eye-safety distance. **(Best performing methods in each row are highlighted in blue)**

$10^2 - 10^3$). By plugging these terms into Eq. 5, the SNR of line scanning is $X = \sqrt{N}$ times higher than the full-pattern.

One interesting question is: Can we push this idea further and scan a dot at a time? This *point scanning* idea can be implemented with a co-located laser and single-pixel sensor deflected by a 2D MEMS mirror. Using the same assumption, $R_a = R_{t1} = R_{t2} = N_p$, where N_p is the number of dots (typically $N_p = N^2 = 10^4 - 10^6$). Fig. 3(a) shows that dot scanning does offer a higher SNR and is $X = N$ times higher than the full-pattern. Notice that this SNR benefit comes from the fact that the laser and the sensor are *synchronized*: The sensor only receives light from the area illuminated by the laser at any instant. For their unsynchronized counterparts where the sensor is a 2D camera that captures the entire 2D FOV during the whole imaging time (easier to implement in hardware), their SNR is exactly the same as the full-pattern approach (Fig. 3(b,d)).

Adaptive sensing. Our *adaptive* sensor projects a static pattern that does not change during the entire exposure, *i.e.* $R_{t1} = R_{t2} = 1$. However, the optical power is concentrated to a small ROI, which we assume can be as small as one line in the line-scanning approach $R_a = N$. As shown in Fig. 3(f), our adaptive sensor has a N_l times higher SNR than the full-pattern approach. In summary, we observed that $SNR_{\text{adaptive}} = SNR_{\text{point}} \gg SNR_{\text{line}} \gg SNR_{\text{full}}$.

3.2. Comparison of Power, Range, and Eye-Safety

Sec. 3.1 analyzes the SNR for different sensors at the same depth. However, this analysis is insufficient, since increasing SNR and the maximum range implies a higher risk of eye injury. In this section, we discuss how this model can be extended to analyze the trade-off between power, range and eye-safety. We consider two key constraints: maximum sensing distance and minimum eye-safety distance.

Maximum sensing distance. We assume that for reliable estimation of the depth, the SNR must be greater than a minimum detection threshold $\text{SNR}_{\text{thres}}$. The equality holds when the maximum sensing distance $d = d_{max}$ is reached,

$$SNR_{\text{thres}} = \frac{P\sqrt{T}}{d_{\text{max}}^2 A\sqrt{P_{sun}}} X. \tag{6}$$

Rearranging this equation gives

$$\boxed{P = k_p \cdot d_{\text{max}}^2 X^{-1} = k_p \cdot d_{\text{max}}^2 R_a^{-1} R_{t1} R_{t2}^{-0.5},} \tag{7}$$

where k_p is a method-independent constant.

Minimum eye-safety distance. A minimum eye-safety distance can be defined when the maximum permissible exposure (MPE, defined in ANSI Z136) is reached:

$$\frac{P}{l_{\text{min}}^2 a} = \frac{MPE(t_1)}{t_1}, \tag{8}$$

It is considered dangerous for eyes to be exposed at a distance shorter than l_{min}. Intuitively, the shorter the minimal

eye-safety distance is, the more eye-safe the device is. We expand MPE based on definitions from ANSI Z136:

$$\frac{P}{l_{\text{min}}^2 a} = \frac{MPE(t_1)}{t_1} = \frac{C_\lambda t_1^{0.75} 10^{-3}\ (\text{J}\cdot\text{cm}^{-2})}{t_1} = k_e t_1^{-0.25}, \tag{9}$$

where k_e is a method-independent constant. **Please find a detailed discussion on this analytic expression in the supplementary report.** Plug in Eq. 4 and rearrange,

$$l_{\text{min}} = k_l \cdot P^{0.5} R_a^{0.5} R_{t1}^{-0.125}, \tag{10}$$

where k_l is a method-independent constant.

Comparing different sensors. From Eq. 7 and Eq. 10, it is clear that for a depth sensing method, specifying one quantity among P, d_{max} and l_{min} will also determine the other two. We thus focus on the following question: To reach the same maximum sensing distance d_{max}, what is the power consumption P and eye-safety distance l_{min} of each method? This is a key problem for consumer devices with limited power budget. Plug Eq. 7 into Eq. 10 and rearrange:

$$\boxed{l_{\text{min}} = k_{ld} \cdot d_{\text{max}} R_{t1}^{0.375} R_{t2}^{-0.25},} \tag{11}$$

where k_{ld} is a method-independent constant.

Fig. 3 summarizes the results derived from Eq. 7 and Eq. 11 for different sensing methods. Full-frame pattern method is the most eye-safe but consumes the most power. Conversely, point scanning (synced) consumes the least power but is also the least eye-safe, which highly limits its application in consumer devices (*e.g.*, laser projectors). Line scanning (synced) strikes the sweet middle ground, which extends the distance by a large margin while maintaining eye safety. Finally, by concentrating to a small ROI, the proposed adaptive method consumes the least power and achieves the best eye-safety.

To intuitively showcase this advantage, we assume $N \sim$ 100 to 500, 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. At the same maximum sensing distance, adaptive sensing:

- has the same eye-safety distance as full-frame sensors, while consuming N^{-1} (0.01 to 0.002)$\times$ lower power.
- has $N^{-0.125}$ (0.56 to 0.46)$\times$ shorter (better) eye-safety distance as line-scanning, while consuming $N^{-0.5}$ (0.1 to 0.04)$\times$ lower power.

It is important to mention that these calculations are based on the assumption that the illuminated area for adaptive sensing is the same as line scanning: $R_a = N$. In practice, this area may be larger depending on the scene and application. The adaptive projector (*e.g.*, SLM) may also have a limited light efficiency, which gives an effectively smaller R_a and thus lower SNR. Nonetheless, at the same maximum distance, adaptive sensing still has a power benefit as long as $R_a > \sqrt{N}$, and it always has a eye-safety benefit since l_{min} is independent of R_a. The theoretical analysis forms the foundation for the proposed adaptive 3D sensing. We validate the analysis in a real-world prototype in Sec. 5.

	SNR	P	l_{min}
V1	cN	$k_p d_{max}^2 N^{-1}$	$k_{ld} d_{max}$
V1-a	cNK^{-1}	$k_p d_{max}^2 N^{-1} K$	$k_{ld} d_{max} K^{0.375}$
V1-b	$cNK^{-0.5}$	$k_p d_{max}^2 N^{-1} K^{0.5}$	$k_{ld} d_{max} K^{0.125}$

Table 1. **Variations of adaptive sensing.**

Disjoint ROIs. So far, we assume an ideal flexible projector which can project light to arbitrarily-shaped, even disjoint ROIs simultaneously. In practice, certain hardware implementations do not have this capability (an example is discussed in Sec. 4.2). To this end, we propose a more flexible scanning strategy: During the camera exposure, the system scans K disjoint ROIs sequentially (typically $2 \leq K \leq 5$). This adaptive V1-a method consumes slightly more power and has a slightly longer eye-safety distance (Tab. 1). Another option is to divide the camera exposure into K shorter exposures, and the system scans a single ROI during each exposure. This adaptive V1-b method performs comparably as V1, but requires a K times higher camera frame rate. **Please refer to the supplementary technical report for more design variations and detailed comparisons.**

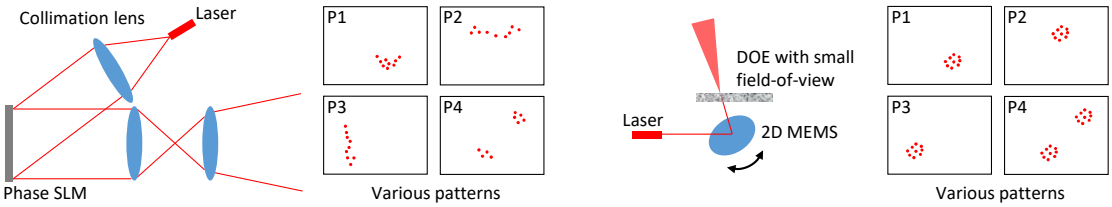
4. Implementation of Adaptive Illumination

Now that we have theoretically analyzed the benefit of the proposed adaptive illumination, how can the proposed adaptive illumination be realized? Notice that this is not a trivial problem. The hardware implementation must satisfy two criteria: (1) The system can redistribute the optical power to a small ROI (guided by an attention map), and (2) This ROI can be projected to different parts of the scene flexibly and in real-time (*e.g.*, 30Hz). A common LCD or DLP projector satisfies (2) but does not satisfy (1). In this section, we propose two hardware configurations that satisfy both conditions.

4.1. Implementation 1: Phase SLM

Fig. 4(a) shows our SLM-based implementation. Our holographic projection approach is inspired by recent work on holographic near-eye 3D displays [20]. Specifically, a hologram to be reproduced by the SLM is decomposed as a sum of *sub-holograms*, where each sub-hologram diffracts light to a single object point in the scene. In [20], each sub-hologram is created using a *lens phase function*:

$$f_n^{\text{lens}}(\mathbf{X}) = e^{j2\pi\sqrt{(X-x_n)^2+(Y-y_n)^2+(Z-z_n)^2}/\lambda}, \tag{12}$$



(a) Phase SLM-based implementation

(b) MEMS + DOE-based method

Figure 4. **Hardware implementations.** **(a)** Our first implementation is based on a phase-only SLM, which can redistribute light to arbitrarily-shaped, even disjointed ROIs. **(b)** Our second implementation is based on a DOE pattern generator with a 2D MEMS mirror. It has less flexibility and can only concentrate on one fixed-shape ROI at a time, but it benefits from low cost, simple optics and small form factor which makes it potentially easier to be miniaturized.

where (X, Y, Z) is the 3D position of each pixel in the sub-hologram, (x_n, y_n, z_n) is the 3D position of the n-th object point, and λ is the light wavelength. The full hologram is,

$$H(\mathbf{X}) = \sum_{n=1}^{N} f_n^{\text{lens}}. \tag{13}$$

This lens phase function mimics a lens that focuses light to the object point at the correct depth, which works well for near-eye displays. One limitation of this lens phase function approach is that light is only redistributed *locally*. This is because the SLM can only reproduce a smooth hologram due to the Nyquist frequency determined by the finite pixel pitch. However, f^{lens} varies rapidly for off-center pixels (*i.e.* X, Y far away from x_n, y_n), causing aliasing artifacts. Therefore, sub-holograms of much smaller sizes must be used, which greatly limits the light efficiency.

To alleviate this limitation, we propose the use of *mirror phase function*:

$$f_n^{\text{mirror}}(\mathbf{X}) = e^{j(X \cdot x_n + Y \cdot y_n)}, \; H(\mathbf{X}) = \sum_{n=1}^{N} f_n^{\text{mirror}}. \tag{14}$$

The mirror phase function corresponds to a smooth phase map linear in terms of X, Y, and can be implemented on the SLM without aliasing. It allows us to use each sub-hologram as a mirror that reflects light to the right direction. By taking the sum of sub-holograms that reflect to different directions, desired projection patterns can be achieved. **Please see the supplementary for detailed optical diagrams and discussion.**

Conversion to phase-only holograms. Notice that Eq. 14 creates a hologram with both amplitudes and phases being spatially-variant, which cannot be implemented on a phase-only SLM. Several approaches [12, 16] have been proposed to convert such a full hologram to a phase-only hologram. Fortunately, our goal is not to project a high-quality image, and simple amplitude discarding is sufficient to project unique texture to the scene:

$$H_{phase} = \text{Arg}[H]. \tag{15}$$

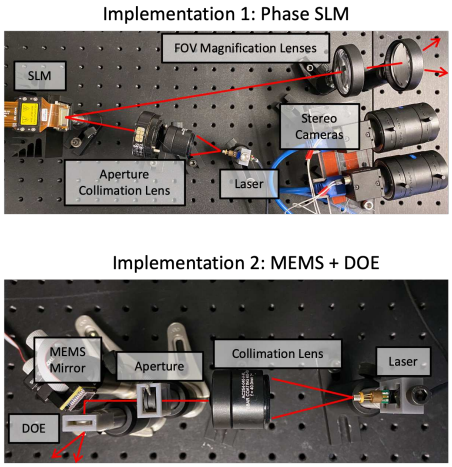


Figure 5. **Hardware prototypes.** (Upper) Spatial Light Modulator (SLM) implementation. (Lower) Micro-electromechanical (MEMS) mirror + diffractive optical element (DOE) implementation.

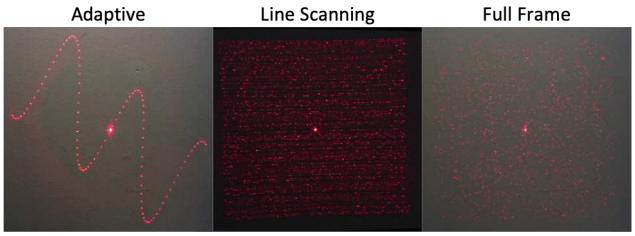


Figure 6. **Examples of emulating full frame, line scanning, and adaptive sensors on a phase only spatial light modulator.** Note that the full frame and adaptive patterns are captured within the full exposure time, while each line in the line scanning pattern is captured within $1/N$ of the time. We scale the intensities for visualization. Note that the bright center dot is caused by the SLM DC term, and can be mitigated with additional optical filters and manufacturing processes.

Efficient implementation. The mirror phase function consists of simple arithmetic operations on large matri-

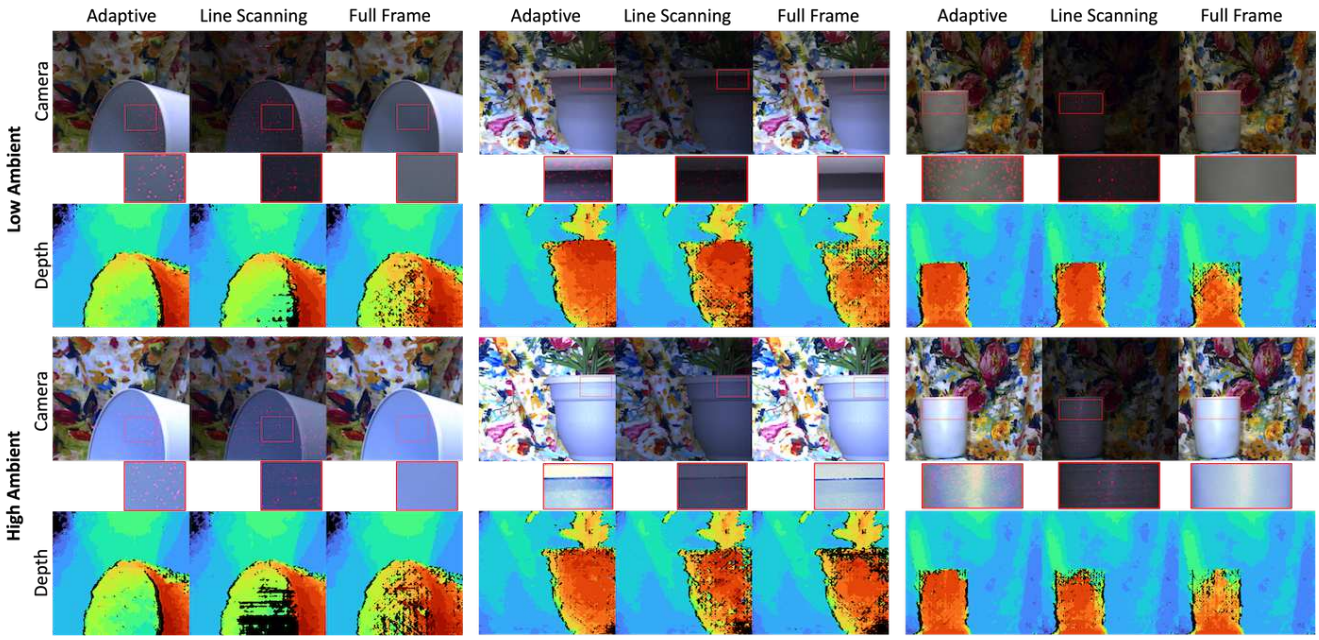


Figure 7. **Comparison between sensing strategies.** Our approach is more robust to ambient light than line scanning and full frame sensors; notice how our adaptive approach degrades less than line scanning and full frame from low ambient lighting (700 lux) to high ambient lighting (4000 lux). We keep sensor settings constant for each low ambient-high ambient pair. The total number of dots in the adaptive method (ours), line scanning and full frame is around 150, 108×150 and 150^2, respectively. Zoom in to see the red dots in the camera images.

ces, which can be implemented efficiently on a GPU. We implement our hologram generator in CUDA and render the resulting hologram phase from the framebuffer to the SLM using OpenGL-CUDA interoperability. On a NVIDIA Jetson Nano, an embedded system-on-module with a Tegra X1 Maxwell 256-core GPU and limited computing resources, we are able to generate 1080p holograms with around 100 points or less at 30 fps. Our implementation and simulator can be found at https://github.com/btilmon/holoCu.

4.2. Implementation 2: MEMS + DOE

Our second implementation is to adjust the beam incident angle of a diffractive optical element (DOE) with a MEMS mirror, as shown in Fig. 4(b). DOE offers a cheap, energy-efficient solution for random dot projection in single-shot structured light (Kinect V1) or active stereo (RealSense). While those systems use a DOE that covers the entire scene, we use a small FOV ($\approx 5°$) that only corresponds to a small ROI. By rotating the MEMS mirror, the deflected laser beam hits the DOE at different angles, thus generating a dot pattern at different ROIs of the scene.

Comparison with phase SLM. The MEMS + DOE implementation is less flexible than the SLM implementation since the hologram shape is fixed (determined by the DOE phase pattern). This is schematically shown in Fig. 4: While SLM can illuminate ROIs of various shapes (P1-P3), MEMS + DOE can only create the same shape shifted across the scene. Moreover, while the SLM can redistribute the optical power over two disjoint ROIs during the same camera exposure (P4), different ROIs are scanned and imaged sequentially by the MEMS mirror, which slightly decreases the SNR (see Sec. 3.2 for detailed analysis). Nevertheless, the MEMS + DOE approach benefits from low cost, simple optics and small form factor, which are important factors for mobile and wearable devices.

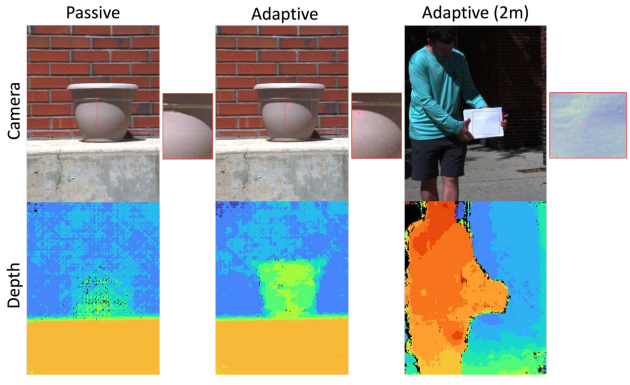


Figure 8. **Outdoor active depth sensing under 50 kilolux direct sunlight with Phase SLM.** Our sensor works outdoors under direct sunlight. We show results up to 2 meters but believe this range could be extended with further SLM optical engineering. Zoom in to see the red dots in the camera images.

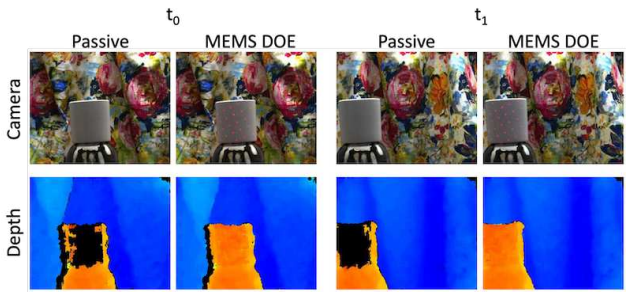


Figure 9. **MEMS + DOE implementation.** The MEMS voltages are updated to direct the dot pattern onto the predominant textureless region in both frames. Zoom in the camera images to see the red dots.

5. Experiments

Hardware prototypes. Fig. 5 shows both hardware prototypes of our proposed method. 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. A random dot DOE with a small FOV is preferred. Here, we used a Holoeye DE-R 339 DOE that produces a periodic 6x6 dot pattern with 5° FOV instead and we tilt the DOE such that the pattern is still unique locally on the epipolar line. **Please refer to the supplementary report for detailed setup and calibration procedures.**

Attention map and depth estimation. We adopt classical semi global block matching for depth estimation [10]. The attention map is determined by randomly choosing pixels that do not have a valid depth value from the depth map computed from passive images. In practice, the attention map can be conditioned by the application such that illumination is only needed within the regions where AR objects are inserted. Our goal is to present a general sensor that can fit into many different perception systems and improve active depth sensing.

Comparison between sensing strategies. We emulate full-frame and line-scanning strategies on our SLM implementation and compare them with our adaptive sensing strategy. An example of each emulated sensor can be found in Fig. 6. For line scanning, we compute and project the hologram of the dot pattern line-by-line. We capture the image for each line individually and stitch the corresponding camera rows together into a single image.

Fig. 7 shows the results for three different scenes. All scenes are illuminated with the same ambient lighting and laser power. Laser power, exposure time and illuminated area are chosen to ensure fair comparison. Due to the dominating photon noise from the ambient light, full-frame and line scanning methods have a low SNR. As a result, depth estimation fails in the textureless regions. Since the proposed adaptive sensing technique concentrates light to the textureless regions, it achieves much higher SNR and obtains higher-quality depth maps, which validates our theoretical analysis.

Outdoor depth sensing under direct sunlight. Fig. 8 demonstrates our Phase SLM prototype working outdoors under 50 kilolux direct sunlight. We can rely on passive stereo to compute depth for the majority of the scene and only project light where necessary, such as the white textureless pot. We also show a distance test of the Phase SLM prototype at 2 meters. We believe this distance could be increased with further SLM optical engineering in future work.

MEMS + DOE implementation. Fig. 9 demonstrates our MEMS + DOE prototype. The dot pattern is projected to a textureless object which improves the disparity compared to passive stereo. When the system moves to another location at t_1, it analyzes the new captured images and directs the ROI to the new position of the textureless object.

6. Limitations and Discussion

Optical power vs computation power. Although we do not explicitly compare the optical power savings from adaptive sensing with the additional computation power needed for computing the attention map and projector control signal (*e.g.*, phase map for SLM), we show that such computations consist of basic arithmetic operations and can be implemented on embedded systems like NVIDIA Jetson Nano, suggesting that our approach can be deployed on increasingly available mobile GPUs. Our system will have even higher benefits for outdoor applications where optical power dominates.

Learning-based attention map and depth estimation. In this work, we use simple, low-complexity texture analysis and semi global matching for attention map and depth estimation. It is possible to design neural networks to achieve better depth estimation, at the cost of higher computation. Our focus in on validating the proposed adaptive sensing as a promising novel sensing strategy, and we expect more practical algorithms to be developed in future work.

Other active depth sensing mechanisms. Although this paper only shows hardware implementations for active stereo, the adaptive sensing strategy can be applied to other depth sensing mechanisms such as single/multi-shot structured light, direct/indirect ToF, FMCW Lidar, *etc*., which can be a promising future direction.

Acknowledgements. The authors acknowledge partial funding from the Office of Naval Research through ONR N00014-18-1-2663 and the National Science Foundation through NSF 1942444.

References

[1] Supreeth Achar, Joseph R. Bartels, William L. 'Red' Whittaker, Kiriakos N. Kutulakos, and Srinivasa G. Narasimhan. Epipolar time-of-flight imaging. *ACM Transactions on Graphics*, 36(4):1–8, July 2017. 2, 3

[2] Siddharth Ancha, Gaurav Pathak, Srinivasa G Narasimhan, and David Held. Active safety envelopes using light curtains with probabilistic guarantees. *arXiv preprint arXiv:2107.04000*, 2021. 3

[3] Siddharth Ancha, Yaadhav Raaj, Peiyun Hu, Srinivasa G Narasimhan, and David Held. Active perception using light curtains for autonomous driving. In *European Conference on Computer Vision*, pages 751–766. Springer, 2020. 3

[4] 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. 2

[5] Joseph Bartels, Jian Wang, William Whittaker, and Srinivasa Narasimhan. Agile depth sensing using triangulation light curtains. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7899–7907, 2019. 3

[6] Neil Bruce and John Tsotsos. Attention based on information maximization. *Journal of Vision*, 7(9):950–950, 2007. 3

[7] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2016. 3

[8] Dorian Chan, Srinivasa G Narasimhan, and Matthew O'Toole. Holocurtains: Programming light curtains via binary holography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[9] James Davis, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–359. IEEE, 2003. 2

[10] Fixstars. libsgm. https://github.com/fixstars/libSGM. 8

[11] Simone Frintrop, Erich Rome, and Henrik I Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1):6, 2010. 3

[12] R W Gerchberg and W O Saxton. A Practical Algorithm for the Determination of Phase from Image and Diffraction Plane Pictures. *Optik*, 35:237–246, 1972. 6

[13] Mohit Gupta, Qi Yin, and Shree K. Nayar. Structured light in sunlight. In *2013 IEEE International Conference on Computer Vision*, pages 545–552, 2013. 2, 3

[14] Samuel W Hasinoff, Frédo Durand, and William T Freeman. Noise-optimal capture for high dynamic range photography. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 553–560. IEEE, 2010. 3

[15] Radu Horaud, Miles Hansard, Georgios Evangelidis, and Clément Ménier. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine Vision and Applications*, 27:1005–1020, Oct. 2016. 1

[16] C. K. Hsueh and A. A. Sawchuk. Computer-generated double-phase holograms. *Applied Optics*, 17(24):3874–3883, Dec. 1978. 6

[17] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krahenbuhl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5390–5399, 2019. 1

[18] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011. 1

[19] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras, 2017. 2

[20] Andrew Maimone, Andreas Georgiou, and Joel S. Kollin. Holographic near-eye displays for virtual and augmented reality. *ACM Trans. Graph.*, 36(4), jul 2017. 5

[21] Nathan Matsuda, Oliver Cossairt, and Mohit Gupta. MC3D: Motion Contrast 3D Scanning. In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–10, Houston, TX, USA, Apr. 2015. IEEE. 2, 3

[22] 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. 1, 2, 3

[23] Denise D. Padilla, Patrick A. Davidson Jr, Jeffrey J. Carlson, and David N. Novick. Advancements in sensing and perception using structured lighting techniques :an LDRD final report. Technical Report SAND2005-5935, 875617, Sept. 2005. 2

[24] Francesco Pittaluga, Sanjeev J Koppal, Sing Bing Kang, and Sudipta N Sinha. Revealing scenes by inverting structure from motion reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 145–154, 2019. 1

[25] 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. 2, 3

[26] Yaadhav Raaj, Siddharth Ancha, Robert Tamburo, David Held, and Srinivasa G Narasimhan. Exploiting & refining depth distributions with triangulation light curtains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7434–7442, 2021. 3

[27] Zhanghao Sun, Ronald Quan, and Olav Solgaard. Resonant scanning design and control for fast spatial sampling. *Scientific Reports*, 11(1):20011, 2021. 3

[28] Zaid Tasneem, Charuvahan Adhivarahan, Dingkang Wang, Huikai Xie, Karthik Dantu, and Sanjeev J Koppal. Adaptive

fovea for scanning depth sensors. *The International Journal of Robotics Research*, 39(7):837–855, 2020. 3

[29] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*, pages 537–547. IEEE, 2017. 1

[30] Brevin Tilmon and Sanjeev J. Koppal. Saccadecam: Adaptive visual attention for monocular depth sensing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6009–6018, October 2021. 3

[31] Jian Wang, Joseph Bartels, William Whittaker, Aswin C. Sankaranarayanan, and Srinivasa G. Narasimhan. Programmable triangulation light curtains. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3

[32] Jian Wang, Aswin C. Sankaranarayanan, Mohit Gupta, and Srinivasa G. Narasimhan. Dual Structured Light 3D Using a 1D Sensor. In *European Conference on Computer Vision (ECCV)*, volume 9910, pages 383–398, Cham, 2016. Springer International Publishing. 2

[33] Hongchuan Wei, Pingping Zhu, Miao Liu, Jonathan P How, and Silvia Ferrari. Automatic pan–tilt camera control for learning dirichlet process gaussian process (dpgp) mixture models of multiple moving targets. *IEEE Transactions on Automatic Control*, 64(1):159–173, 2018. 3

[34] Li Zhang, Brian Curless, and Steven M Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–367. IEEE, 2003. 2