A 256x256 LiDAR imaging system based on a 200mW SPAD-based SoC with Micro-lens Array and Light-Weight RGB-guide Depth Completion Neural Network

Jier Wang , Jie Li , Yifan Wu , Hengwei Yu , Lebei Cui , Miao Sun * , Patrick Yin Chiang

Posted Date: 23 June 2023

doi: 10.20944/preprints202306.1664.v1

Keywords: LiDAR, 3D imaging, System on chip, Microlens array, Neural network, RGB-guided, Depth completion

Communication

A 256x256 LiDAR imaging system based on a 200mW SPAD-based SoC with Micro-lens Array and Light-Weight RGB-guide Depth Completion Neural Network

Jier Wang [1], Jie Li[1], Yifan Wu[2], Hengwei Yu[1], Lebei Cui[1], Miao Sun[1,*] and Patrick Yin Chiang [1]

[1] State Key Laboratory of ASIC and System, Fudan University, No. 825, Zhangheng Road, Shanghai 201203, China
[2] College of electronics and information engineering, Tongji University, Shanghai, China
* Correspondence:18112020006@fudan.edu.cn

Abstract: Light Detection and Ranging (LiDAR) technology, a cutting-edge advancement in mobile applications, presents a myriad of compelling use cases, including enhancing low-light photography, capturing and sharing 3D images of fascinating objects, and elevating the overall augmented reality (AR) experience. However, its widespread adoption has been hindered by the prohibitive costs and substantial power consumption associated with its implementation. To surmount these obstacles, this paper proposes a low-power, low-cost, SPAD-based system-on-chip (SoC) which packages the microlens arrays (MLA) and incorporates with a light-weight RGB-guided sparse depth imaging completion neural network for 3D LiDAR imaging. The proposed SoC integrates an 8x8 Single-Photon Avalanche Detectors (SPADs) macro pixel array with time-to-digital converters (TDC) and charge pump, fabricated using a 180nm bipolar-CMOS-DMOS (BCD) process. A random MLA-based homogenizing diffuser efficiently transforms Gaussian beams into flat-topped beams with a 45° field of view (FOV), enabling flash projection at the transmitter. To further enhance resolution and broaden application possibilities, a lightweight neural network employing RGB-guided sparse depth complementation is proposed, enabling a substantial expansion of image resolution from 8x8 to quarter video graphics array level (QVGA; 256x256). Experimental results demonstrate the effectiveness and stability of the hardware encompassing the SoC and optical system, as well as the lightweight features and accuracy of the algorithmic neural network. This integrated state-of-the-art hardware-software solution offers a promising and inspiring foundation for developing consumer-level 3D imaging applications.

Keywords: LiDAR, 3D imaging, System on chip, Microlens array, Neural network, RGB-guided, Depth completion

1. Introduction

The SPAD-based solid-state LiDAR systems, exhibiting a broad array of applications, outperforms traditional rotating LiDAR and Microelectro-Mechanical Systems (MEMS) LiDAR. This superiority arises from their remarkable mechanical stability, high-performance characteristics, and cost-effectiveness. In contrast to Indirect Time-of-Flight (iToF) systems[1, 2], which suffer from limited measurement distances (<20 m), substantial power consumption (tenfold that of Direct Time-of-Flight, dToF), and convoluted calibration algorithms, dToF imaging system offers a more promising solution. By combining the single-photon sensitivity of SPADs with the picosecond-level temporal resolution of TDCs, dToF systems enable long-range object measurement and centimeter-level depth resolution[3]. Consequently, dToF establishes itself as the predominant technological trajectory for the forthcoming generation of 3D imaging. The capability of measuring depth enables solid-state LiDAR to excel in numerous applications, from floor-sweeping robots and facial recognition for consumers to autonomous driving and 3D building modeling in the industrial domain[4].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1664.v1

As LiDAR technology advances towards more compact system designs, the increased SPAD imaging sensor pixels, and depth-RGB image fusion are anticipated to become primary issues to addressed.

3D depth imaging employing SPAD technology currently faces certain restrictions, including low spatial resolution, suboptimal effective optical efficiency (i.e., fill factor), and elevated costs. These limitations also impede the expansive array of applications for dToF LiDAR systems. Although recent research advancements indicate that SPAD arrays can achieve QVGA (320x240) resolutions and even higher (512x512), with prospective advancements targeting Video Graphics Array (VGA, 640x480) resolutions, the pixel number remains markedly inferior to traditional Contact Image Sensor (CIS) technology[5–8]. Apple LiDAR (256x192) and Huawei P30 (240x180) exemplify the potential of dToF technology in the consumer market, yet the resolution disparity persists. To address this challenge, the 3D stacking technique has been proposed, which positions the back-illuminated SPAD array on the top tier and the readout control circuits on the bottom tier[9]. This configuration allows researchers to enhance the resolution to 512x512[10]. However, the considerable costs associated with this technique have impeded its development and subsequent applications. At present, no effective solutions have been identified to surmount this obstacle in the realm of circuit design.

The inherent limitations of consumer-grade LIDAR technology, specifically its low resolution and high cost, have long been recognized within the industry. As 3D sensing technology advances, consumer-grade RGB-D cameras (e.g., Intel RealSense, Microsoft Kinect) have gained popularity owe to their ability to capture color images and depth information simultaneously, then recovering 3D scenes at a lower cost. Motivated by this development, researchers have explored RGB-guided depth-completion algorithms to reconstruct depth-density maps from sparse depth data obtained by dToF depth sensors and color images captured by RGB cameras, with the goal of predicting low-cost LiDAR-generated high-resolution scenes[11]. The authors[12] introduce a normalized convolution layer that enables unguided scene depth completion on highly sparse data using fewer parameters than related techniques. Their proposed method treats validity masks as a continuous confidence field and presents a new criterion for propagating confidences between CNN layers. This approach allows for the generation of point-wise continuous confidence maps for the network output. The authors[13] propose Convolutional Spatial Propagation Networks (CSPN), which demonstrate greater efficiency and accuracy in depth estimation compared to previous state-of-the-art propagation strategies, without sacrificing theoretical guarantees. Liu et al.[14] proposed an architecture, the differentiable kernel regression network, which consists of a CNN network that learns steering kernels to replace hand-crafted interpolation for performing the coarse depth prediction from the sparse input.

Gaussian beams play a significant role in various applications, including lidar, optical communication, and laser welding[15]. However, in flash LIDAR and laser TV projection, it is essential to homogenize Gaussian beams into flat-topped beams. Common techniques to generate flat-top beams include MLA, diffractive optics (DOE), and free-form mirror sets. Among these technologies, MLA-based beam homogenizers have garnered considerable interest, particularly in compact consumer devices, owing to their unique properties. MLA can divide a non-uniform laser into multiple beamlets, which can subsequently be superimposed onto a microdisplay with the assistance of an additional lens[16]. As a result, MLA diffusers display independence from the incident intensity profile and a wide spectral range. In 2016, Jin et al.[17] proposed enhancing homogeneity in the homogenizing system by substituting the first MLA with a free-form MLA. Each free-form surface within the MLA introduced suitable aberrations in the wavefront, redistributing the beam's irradiance. In the same year, Cao et al.[18] presented a laser beam homogenization method that employed a central off-axis random MLA. By adjusting the off-axis quantity of the center, the MLA's periodicity was disrupted, eliminating the periodic lattice effect on the target surface. In 2019, Liu et al. [19] designed a continuous profile MLA featuring sub-lenses

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1664.v1

with random apertures and arrangements. This innovation facilitated a breakthrough in beamlet coherence, achieving a simulated uniformity of 94.33%. Subsequently, in 2020, Xue et al.[20] proposed a monolithic random MLA for laser beam homogenization. During this homogenization process, the coherence between sub-beams was entirely disrupted, resulting in a homogenized spot with a high energy utilization rate.

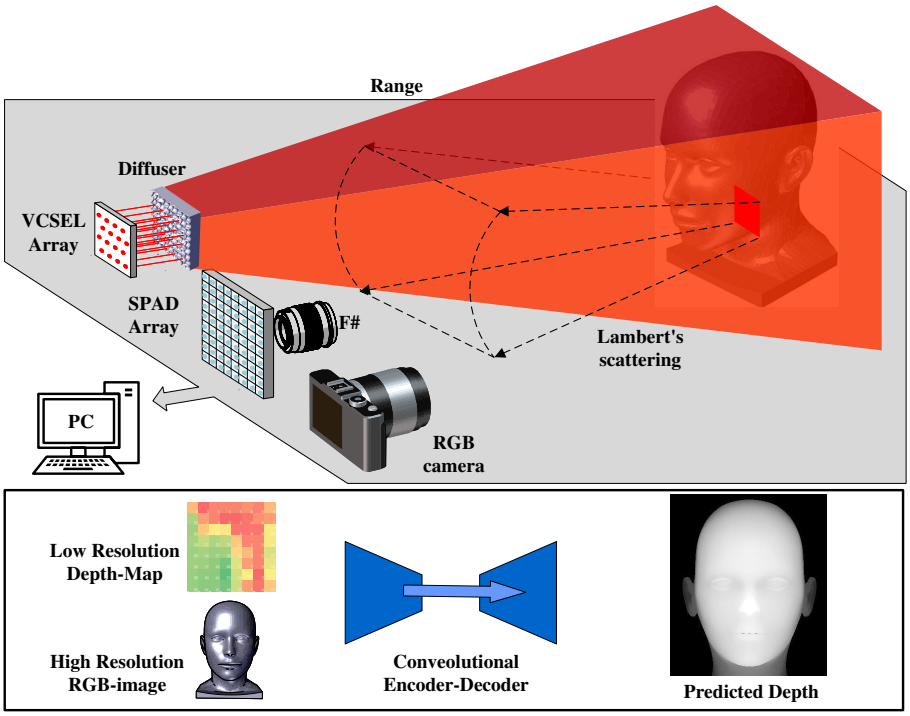

Figure 1. The proposed low-cost, light-weight SPAD-based SoC With micro-lens array and RGB-guide 256x256 depth completion for 3D LiDAR imaging.

In an effort to enhance the resolution of low-pixel depth sensors and broaden the application scope of sparse-pixel, low-cost ranging sensors, we propose a SPAD based LiDAR with micro-optics device and RGB-guided depth-complemented neural network to upsample the resolution from 8x8 to 256x256 pixels as shown in Fig. 1. This light-weight, SPAD-based dToF system is ideal for LiDAR applications.In this work, the developed sparse-pixel SoC is fabricated by the 130nm BCD process. The high Photon Detection Probability (PDP) SPAD and picosecond TDC ensure that the system effectively achieves millimeter-level measurement accuracy in indoor scenarios. The sensor integrates of a 16x16 SPAD array, which can be divided into 2x2 areas at high frame rates for ranging or 8x8 at low frame rates for imaging. Subsequently, we engineered an optical system to facilitate imaging at the hardware level. Micron-scale random MLA were employed on the transmitter to homogenize the Vertical-Cavity Surface-Emitting Laser (VCSEL) array's Gaussian beam into a flat-topped source with a 45° FOV. Free-form lenses were applied at the receiver to align the SPAD array with a 45° FOV and enhance the resolution for sparse imaging using the 8x8 array. To meet consumer-grade imaging requirements, an RGB-guided depth complementation neural network was integrated into the system, improving the resolution from 8x8 to 256x256 pixels (QVGA level). This cost-effective, light-weight imaging system has a wide range of applications in distance measurement, basic object recognition, and simple pose recognition. To validate the system, we conducted verification tests on the SPAD, TDC, and SoC, as well as FOV and resolution tests on micro-optical modules. Our self-developed RGB-guided Depth Completion neural network upsampled and complemented 8x8 depth information by x32 to map the real world. The paper is structured as follows: Section 2 presents the SoC implementation and key functional components design; Section 3 discusses the optical system simulation and MLA design;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1664.v1

Section 4 introduces the proposed RGB-to-Depth completion architecture; Section 5 reports the results and discussion of Sections 2-4; and Section 6 provides conclusions and future perspectives.

2. SOC Implement

Traditional LiDAR systems rely on a ranging framework in which the transmitter and receiver operate as separate entities. A short laser pulse is illuminated by the transmitter, which then undergoes diffuse reflection upon encountering an object before returning to the receiver. Here, the SPAD captures the corresponding photons and subsequently produces photon counts. This signal is conveyed through a multiplexed TDC, ultimately reaching the Digital Signal Processor (DSP) for processing and generation of a time-dependent histogram. However, the conventional architecture is encumbered by several limitations, such as the challenging task of achieving picosecond-level pulsing and the system's inherently poor signal-to-noise ratio (SNR). We propose a highly integrated SoC based on a 130nm BCD process, which integrates a SPAD readout circuit, high-voltage generation circuit, picosecond pulse width laser driver, and ranging algorithm processor, as illustrated in Fig. 2. This design offers superior temporal resolution and SNR compared to conventional architectures, as well as significant cost advantages. The ranging SoC integrates two SPAD pixel arrays, namely the signal SPAD array and reference SPAD array. Fig. 2 illustrates the signal path for both pixel arrays within the ranging SoC. The high voltage of all SPADs is generated in charge pump, which is linked to the HVDD. During pixel array testing, the HVDD should be powered off-chip, and the charge pump should be deactivated to mitigate its impact on the SPADs. In addition to the TDC array, the pixel array outputs are connected to a test mux. Employing the Low-Voltage Differential Signaling (LVDS) interface, high-speed pulse stop signals can be tested with appropriate configurations.

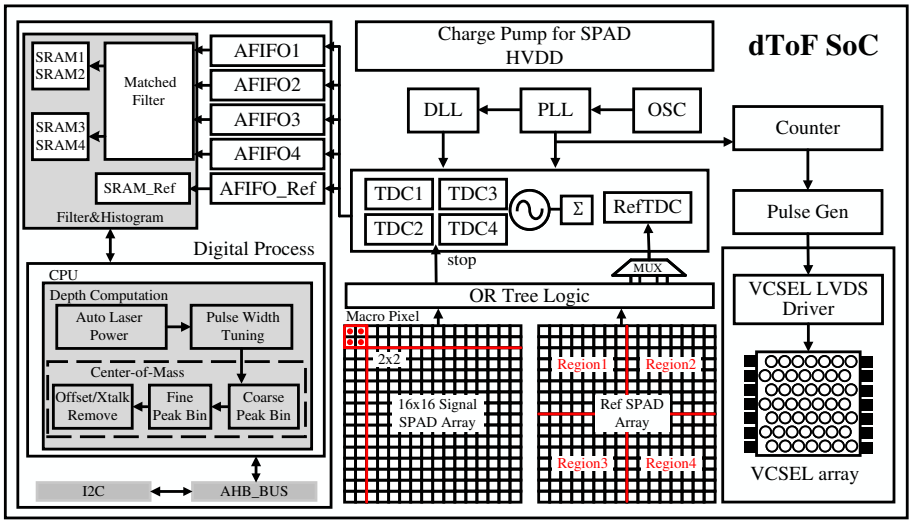

Figure 2. Architecture of dToF LiDAR SoC.

Fig. 3 presents the architecture of the signal SPAD array and its associated readout scheme. The signal SPAD array integrates a total of 256 SPAD pixels, organized in a 16x16 configuration, with each pixel exhibiting identical characteristics. As illustrated in Fig. 3(a), the entire array is partitioned into four distinct regions, wherein each region functioning as a macro-cell containing an 8x8 pixel array, connected through a logical tree. The macro-cell readout scheme is illustrated in Fig. 3(b), wherein the outputs of the 64 pixels are multiplexed collectively to generate a single stop signal. Consequently, the complete pixel array possesses four outputs, designated as stop<3:0>, which correspond to the regional definitions displayed in Fig. 3(a). An alternative simultaneous asynchronous approach can subdivide the 16x16 array into 8x8 macro-pixels, with the primary objective of attaining sparse imaging. However, this method results in a reduction of the system's frame rate.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1664.v1

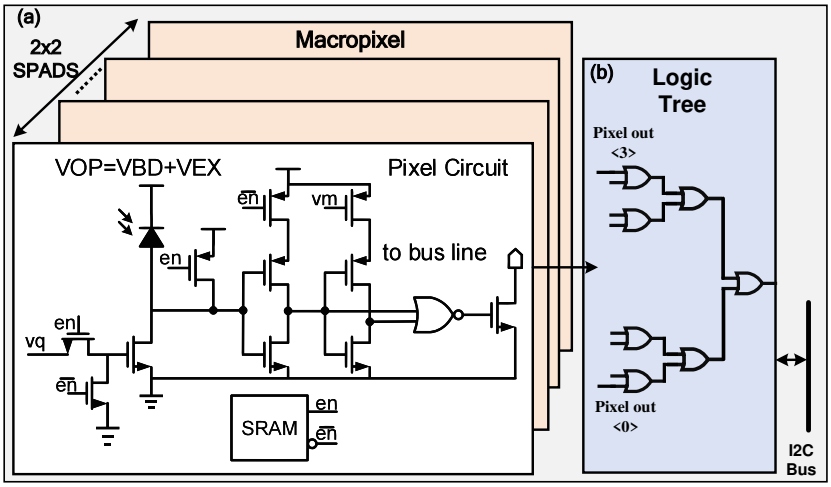

Figure 3. Schematic of SPAD's readout scheme. (**a**) Pixel circuit. (**b**) Logic of SPAD read out.

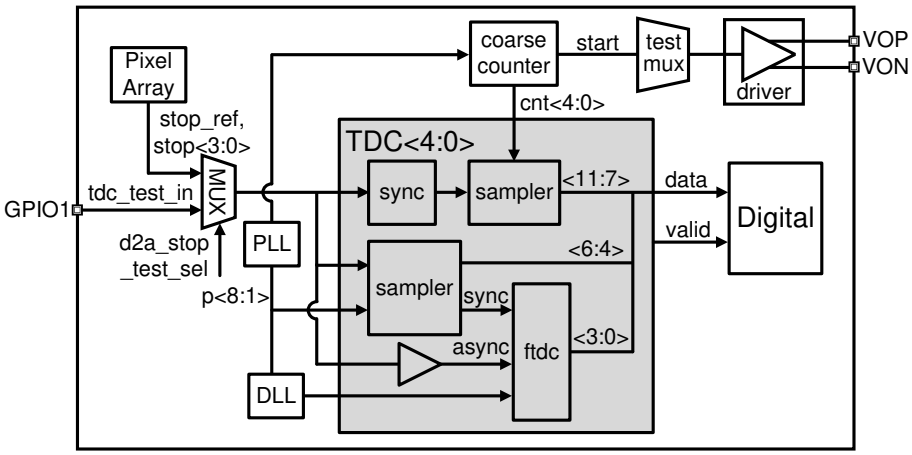

Figure 4. Simplified block diagram of TDC array and the signal path.

Fig. 4 presents the simplified block diagram of the TDC array and its corresponding signal path. In order to achieve a stable output, accounting for Process, Voltage, and Temperature (PVT) variations, the Least Significant Bit (LSB) of the TDC is ascertained by the Phase-Locked Loop (PLL). The TDC is partitioned into three distinct segments: Coarse, Medium, and Fine TDCs. The Coarse counter operates by counting the phase-8 (p<8>) clock of the PLL outputs, which is determined based on the setting of register d2a-tofclk-div<4:0>. Concurrently, the start signal that controls the VCSEL driver is generated at an identical frequency and synchronized with the counter. The stop signal samples the counter outputs, subsequently generating bits 11-7 of the tdc-data. The Medium TDC is implemented utilizing a sampler, which captures the output clock phases of the PLL, denoted as p<8:1>, employing the stop signal. The outputs are then decoded into bits 6-4 of the tdc-data. The stop signal, synchronized with the proximate phase by the Medium TDC (MTDC), generates a sync signal for the Fine TDC (FTDC). Lastly, the FTDC further quantizes the time difference between the sync and async (buffered stop signal) and generates bits 3-0 of the tdc-data. This cascading arrangement of the Coarse, Medium, and Fine TDCs ensures precise time measurement, contributing to the overall accuracy of the system.

3. Optical System and MLA

In active imaging applications, particularly for LIDAR systems, it is imperative to optimize not only the receiver components, such as lenses and sensors, but also the transmitter elements and the target object. This optimization process requires an in-depth understanding of various factors, including the laser emission mode (flash or dot matrix) and the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1664.v1

physical mechanism governing the object's reflection. The optical model is constructed by taking into account calculations derived from the optical power at the transmitter, the single-pixel optical power budget at the receiver, and the lens system. This model can be readily extended to encompass the entire array once the illumination pattern is known. In the standard scenario, the illumination pattern along the detector's field of view in both horizontal and vertical directions is designed to match the FOV angle at the transmitter. Employing complex physics can facilitate the incorporation of the optical model by calculating the optical power density on the target surface.

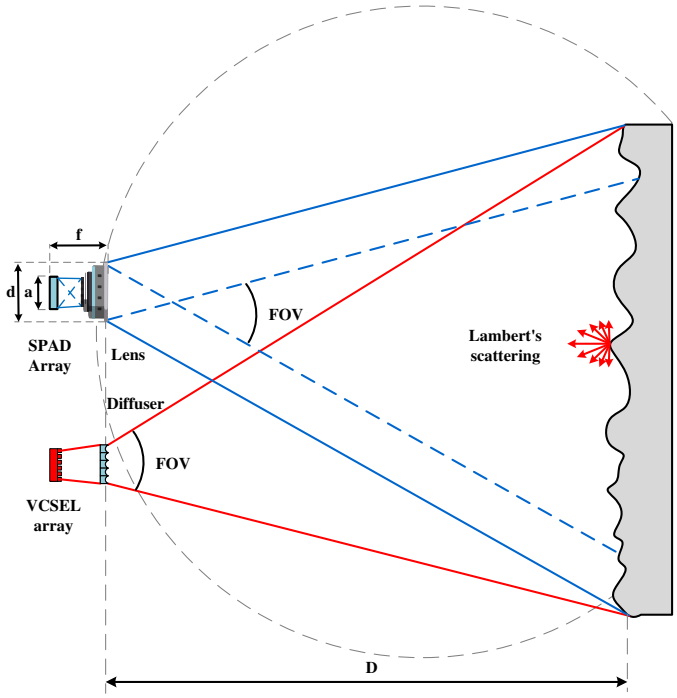

Figure 5. Typical active imaging system model, using a Lambetian reflectance as target.

Fig. 5 presents a generic flash LiDAR model, in which the scene of interest is uniformly illuminated by an active light source, typically a wide-angle pulsed laser beam. The coverage area in the target plane is contingent upon the FOV of the beam. All signal counting calculations are predicated on the lens focusing the reflected light energy back to the SPAD array. To maximize the efficiency of returned energy throughout the field of view, it is crucial to ensure a well-matched FOV between the transmitter and sensor. Regarding the returned light, the target is assumed to be a diffuse surface-i.e., a Lambertian reflector-which exhibits consistent apparent brightness to the observer, irrespective of the observer's viewpoint, and the surface brightness is isotropic. The active imaging system setup is illustrated in Fig. 5, with the distance to the target as D, the lens aperture as d, the focal length as f, the sensor height as a, and the area as A_{sensor}. The reflected light power, Ptarget, emanating from the target, is determined by the source power, P_{source}(D), and the reflectance, ρ_{target}, of the object. The lens component encompasses the f-number (f#=f/d) and FOV. Additionally, the light transmission rate, η_{lens}, of the lens and filter must be considered. When treating the illumination light as square, the optical power model can be articulated by Equation 1:

$$P_{received} = P_{source} \cdot \rho_{target} \cdot \left(\frac{d}{2 \cdot f}\right)^2 \cdot \eta_{lens} \cdot \frac{2 \cdot A_{sensor}}{\pi} \cdot \left(\frac{1}{2 \cdot D \cdot tan(FOV/2)}\right)^2 \quad (1)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1664.v1

Transitioning from the optical power model to a photon counting model proves to be more advantageous for system design and sensor evaluation. Consequently, the photon counting model can be formulated as presented in Equation 2:

$$N_{pulse} = P_{received} \cdot \frac{\lambda}{F_{laser} \cdot h \cdot v} \tag{2}$$

Where F_{laser} is the laser frequency, λ is the light source wavelength used, h is the Planck's constant ($6.62607004 \times 10^{34}$ J· s), and v is the speed of light (2.998×10^{8} m/s). This value defines the total number of photons per laser pulse that reaches pixel array.

The diffusion beam function of the laser diffuser is primarily achieved through the micro-concave spherical structure etched on its surface. This micro-concave spherical surface acts as a micro-concave lens, and the entire laser diffuser can be considered an array of micro-concave lenses. Essentially, the diffused laser surface light source is a superposition of the surface light sources emitted by each micro-concave lens. As illustrated in Fig. 6(d), the laser diffuser diffusion diagram demonstrates the laser incident vertically on the diffuser, with the laser beam diverging due to the influence of the micro-concave lens. This forms a surface light source with a specific diffusion angle, which is associated with the parameters of the micro-concave lens on the diffuser. In accordance with this principle, this paper proposes the preparation of two types of MLAs: the random MLA, illustrated in Fig. 6(b), suitable for generating a circular homogenized light spot; and the rectangular MLA, presented in Fig. 6(c), designed to produce a rectangular homogenized light spot.

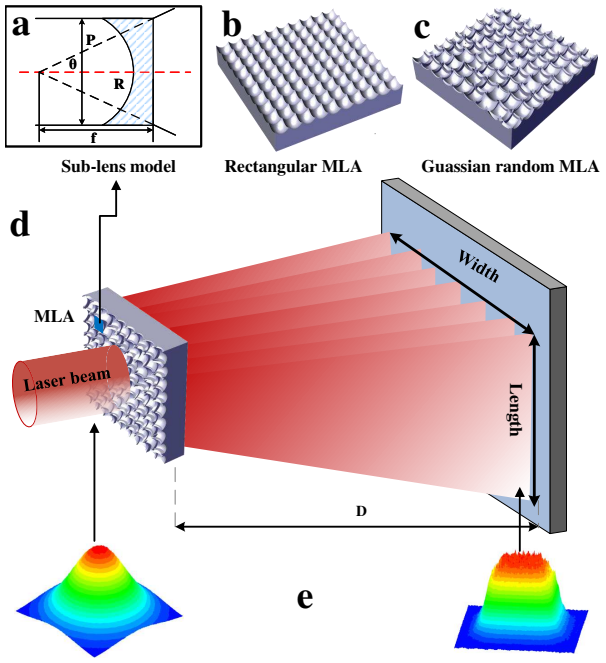

Figure 6. (**a**) structural parameters of microlens unit; (**b**) rectangular MLA; (**c**) Random MLA; (**d**) Schematic diagram of random microlens array; (**e**) Gaussian beams to flat-topped beams.

To elucidate the specific relationship between the characteristic parameters of the micro-concave spherical structure and the diffusion angle, a deductive argument will be presented from the perspective of geometrical optics. Ideally, the overall diffusion angle of the laser beam after traversing the diffuser is equivalent to the diffusion angle of α single sublens. To simplify the analysis of the diffusion angle in connection with the diffuser structure parameters, the sublens diffusion process is examined individually. Fig. 6(a) illustrates the diffusion diagram of the sub-lens, which involves two parameters: the hole P and the radius of curvature R. By employing the lens focal length formula and geometric

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1664.v1

principles, the set of equations (3) is derived. By further simplifying the system of equations, equation (4) can be obtained.

$$tan\theta = \frac{P}{2f}, |f| = \frac{R}{n-1} \tag{3}$$

$$\theta = arctan\frac{P(n-1)}{2R} \tag{4}$$

where n is the refractive index of the micro-concave lens material and f is the focal length of the micro-concave lens. If the parameters of the other sub-lenses are known, the diffusion angle of the sub-lenses can be derived, which in turn yields the diffusion angle of the laser diffuser. Equation (4) further reveals that the diffusion angle theta is directly proportional to the aperture P of the micro-concave lens and inversely proportional to the radius of curvature R.

4. Proposed RGB Guided Depth Completion Neural Network

The primary aim of our research is to accurately estimate a dense depth map, designated as $\hat{Y}$, derived from sparse LiDAR data (Y') in conjunction with the corresponding RGB image (I) serving as guidance. $\mathcal{W}$ denotes the network parameters. The task under consideration can be succinctly expressed through the following mathematical formulation:

$$\hat{Y} = F(Y', I; \mathcal{W}), \tag{5}$$

We implement an end-to-end network, denoted as F, to effectively execute the depth-completion task in our study. The optimization function can be concisely represented in the subsequent description:

$$\hat{\mathcal{W}} = \arg\min_{\mathcal{W}} L(\hat{Y}, Y'; \mathcal{W}), \tag{6}$$

L represents the loss function employed for the purpose.

In our proposed methodology, the network is divided into two distinct components: (1) a depth-completion network designed to generate a coarse depth map from the sparse depth input, and (2) a refinement network aimed at optimizing sharp edges while enhancing depth accuracy. We introduce an encoder-decoder based CNN network for estimating the coarse depth map, as previously demonstrated in [13,14]. However, the aforementioned networks encounter two primary challenges. First, networks such as [12,14] employ complex architectures to boost accuracy; nevertheless, their proposed networks are incompatible with most CNN accelerators or embedded DLAs. Second, networks like [13] utilize ResNet-50[21] or other similar networks as the CNN backbone, resulting in a substantial number of parameters (~240M) and posing implementation challenges on embedded hardware.

To make the network easier to implement on the embedded systems, we explore two optimization strategies:

1. We incorporate conv-bn-relu and conv-bn architectures, which are supported by the majority of CNN accelerators and embedded NPUs.
2. To reduce the number of parameters, we adopt depthwise separable convolution, as presented in [22], as the foundational convolution architecture.

4.1. Depthwise Separable Convolution

The architecture of depthwise separable convolution is illustrated in the Fig. 7. In contrast to the conventional 3x3 2D convolution, depthwise separable convolution employs a 3x3 group convolution, referred to as depthwise convolution, and a 1x1 standard convolution, known as pointwise convolution.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1664.v1

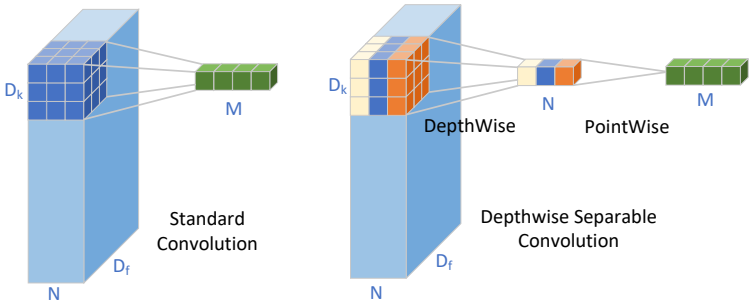

Figure 7. Architecture of depthwise separable convolution.

For a standard convolution with a filter K of dimensions $D_k \times D_k \times M \times N$ applied to an input feature map F of size $D_f \times D_f \times M$, the total number of parameters can be computed as:

$$D_k \times D_k \times D_f \times D_f \times M \times N, \tag{7}$$

In the case of depthwise separable convolution, the total parameters are calculated as follows:

$$D_f \times D_f \times D_k \times D_k \times M + M \times N \times D_k \times D_k, \tag{8}$$

Considering that the kernel size D_k for our network is 3 and the output channel N is considerably larger than $D_k \times D_k$, the standard convolution possesses a significantly higher number of parameters in comparison to depthwise separable convolution.

4.2. Network

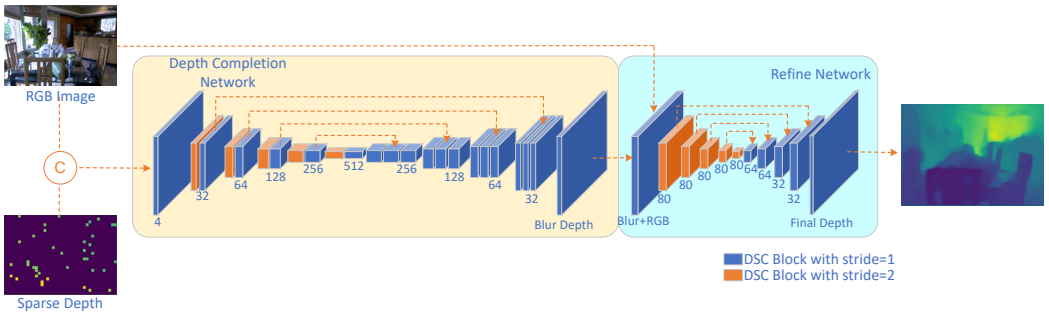

Figure 8. The propoesed RGB-guided depth completion network.

Different from the standard CNN network, we substitute the standard 3x3 convolution with our depthwise separable convolution. Our network architecture is depicted in Fig. 8. Within the depth-completion network, we employ an U-net [23] network as our backbone. The encoder consists of 5 stages, each containing two types of depthwise separable convolution blocks. One block performs downsampling by a factor of 2x using a stride of 2, while the other extracts features with a stride of 1. The decoder is also consists of 5 stages. In the first 4 stages, feature maps are upsampled by 2x, followed by a convolution operation with a stride of 1. These maps are then concatenated with the feature maps from the corresponding encoder stage outputs, and processed by two subsequent convolution blocks with a stride of 1. The final stage comprises an upsampling layer and a single convolution block with a stride of 1. For the refinement network, we adopt the architecture from [12], which has been demonstrated to possess the advantages of being light-weight and exhibiting high performance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1664.v1

4.3. Loss Function

We train our network in two steps: In Step 1, we train our network using the $L1$ error as our loss function. The loss function can be described as follows:

$$L_{1loss} = \frac{1}{M}\sum_{i,j}(Y_{i,j} > 0)|\hat{Y}_{i,j} - Y_{i,j}| \tag{9}$$

Where $Y_{i,j}$ denotes the ground truth at pixel (i, j), $\hat{Y}_{i,j}$ represents the predicted depth at pixel (i, j), and M stands for the total number of pixels with values greater than 0.

Upon completing the training of the network, we proceed to train the entire network using a new loss function to optimize the boundary regions performance, which can be described as follows:

$$L = w_1 L_1 + w_2 L_2 + w_3 L_{gard} \tag{10}$$

In this equation, w_1, w_2, and w_3 are hyperparameters. L_1 is defined similarly to Equation (5), L_2 represents the masked mean squared error (MSE), L_{gard} stands for the gradient error, which can be defined as:

$$L_{2loss} = \frac{1}{M}\sum_{i,j}(Y_{i,j} > 0)(\hat{Y}_{i,j} - Y_{i,j})^2 \tag{11}$$

To achieve improved boundary performance, we incorporate depth gradient information into our loss function, following [24]. The gradient loss function, $L_{gradloss}$, is described below and penalizes the disagreement of depth edges in both vertical and horizontal directions:

$$L_{gardloss} = \frac{1}{M}\sum_{i,j}|\nabla\hat{Y}_{i,j} - Y_{i,j}| \tag{12}$$

5. Results and discussion

This chapter will introduce the hardware, including the verification and testing of the system-on-a-chip and optical modules, and also do a systematic analysis and verification of the RGB-guided depth reconstruction imaging, which enables an efficient combination of hardware and algorithms to achieve a light-weight and low-cost system for depth imaging of indoor scenes.

5.1. System of Hardware

To accommodate the diverse scenarios and complex applications in indoor environments, the SOC based on micro-structured optical element packaging was tested using a Lambertian reflector plate with reflectivities of 85% and 10%, respectively. As illustrated in Fig. 9(a), under a 85% Lambertian reflector plate and an indoor lighting environment of 10 klux, the SOC achieves a maximum measurement distance of 6 m and a minimum measurement distance of 0.2 m, while maintaining good precision (10%) within the measurement range. Under the same conditions, the SOC with a 10% Lambertian reflector plate exhibits a maximum measurement distance of 4 m and a minimum measurement distance of 0.2 m, demonstrating good precision (15%) throughout the measurement range. These two scenarios encompass mainstream indoor applications and fully showcase the exceptional performance of SOCs packaged with micro-structured optical components at the hardware level. As illustrated in Fig. 9(c,d), the micro-structured optical elements encapsulated in the SOC can perform 8x8 sparse depth imaging of objects at a distance of approximately 2 m, without any loss of depth information at the edges. This feature proves to be highly beneficial for subsequent depth-completion algorithms.

Monte Carlo analysis and simulation were conducted to evaluate the performance of randomly distributed microlens array (MLA) homogenizers. In the optical simulation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1664.v1

model, the diffusion sheet dimensions are 4 × 4 mm with a 0.5 mm thickness. To ensure consistency between the diffusion half-angle a of the effective illumination area calculated by fluctuating optics and the micro-structural characteristics, the simulated microlens aperture P is set to 40.4 μm, the radius of curvature R to 50 μm, and the selected material is PMMA with a dielectric refractive index n=1.5. The detector is positioned at 100 mm downstream of the diffusion sheet, material absorption characteristics are neglected, and the diffusion effect is depicted in Fig. 9(b). From Fig. 9(b), it can be observed that the energy distribution of the surface light source, obtained after the laser beam's diffusion through the diffuser, remains strong at the center and weak at the edges. However, the illumination distribution curve is more uniform compared to the undiffused laser. Furthermore, an energy utilization rate of up to 90% can be derived from the total power captured on the detector. A diffusion half-angle of 22° in the full illumination region can also be determined from the coordinates of the zero position of the detector's Y-axis, which aligns with the geometric optical description.

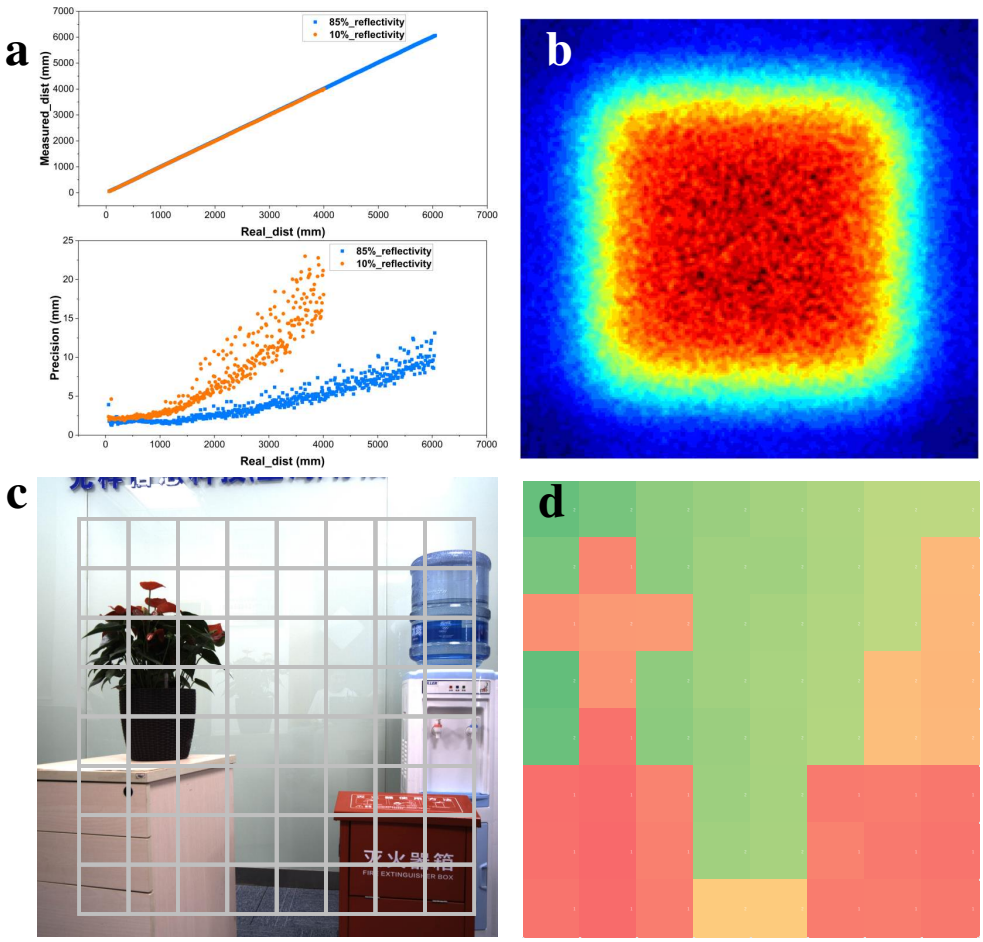

Figure 9. Verification of hardware. (a) Distance and precision at 85% and 10% reflectivity test conditions; (b) The output of intensity distribution on the screen radiated by the MLA diffuser; (c) Aligned depth sensor's zones and color image; (d) 8 × 8 depth map.

5.2. Comparison of network performance

5.2.1. Dataset

To adapt our method for indoor hardware systems, we evaluate its performance using the NYUv2 dataset [25], which consists of RGB and depth images gathered from 464 distinct indoor scenes. We employ the official split, utilizing 47k images for training and 654 images for evaluation. For the purpose of comparing our network's performance with existing methods, we adopt data processing techniques akin to those described in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1664.v1

[13,26]. The RGB-D images are subjected to downsampling and center-cropping, resulting in a resolution of 304x228. Moreover, we randomly sample 500 depth points to serve as the sparse depth input for our network. This strategy ensures a fair comparison with other methods while preserving the integrity of our evaluation.

5.2.2. Metrics

We adopt the same metrics and use their implementation in [26]. Given ground truth depth Y and predicted depth $\hat{Y}$, the metrics include:

1. RMSE:

$$\sqrt{\frac{1}{M}\sum_{i,j}(Y_{i,j}>0)(\hat{Y}_{i,j}-Y_{i,j})^2} \tag{13}$$

2. Abs Rel:

$$\frac{1}{M}\sum_{i,j}(Y_{i,j}>0)|\hat{Y}_{i,j}-Y_{i,j}|/Y_{i,j} \tag{14}$$

3. δ_t: % of $Y_{i,j}$, s.t.

$$max(\frac{\hat{Y}_{i,j}}{Y_{i,j}},\frac{Y_{i,j}}{\hat{Y}_{i,j}})<t, t\in(1.10,1.25,1.25^2,1.25^3) \tag{15}$$

5.2.3. Comparison

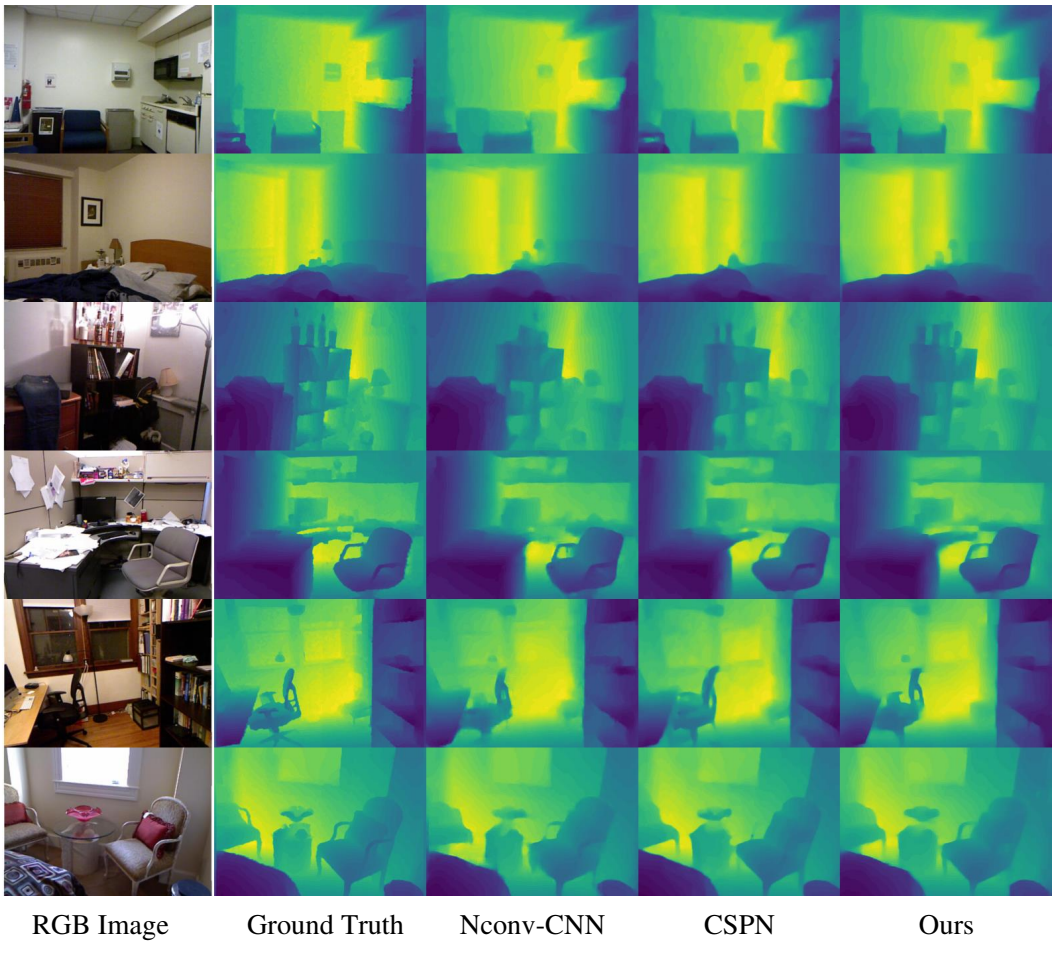

Figure 10. Qualitative Comparison on NYUv2. From left-to-right: Guidance RGB image, the ground truth depth map, results of Nconv-CNN[12], CSPN[13] and ours.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1664.v1

Table 1 presents a comparison of the qualitative results of our network and four state-of-the-art methods [4,12,13,25]. As illustrated, our network achieves superior performance while preserving a reduced number of parameters (~1M). This enhanced efficiency is attributed to the implementation of depthwise separable convolutions and the proposed CNN-based light-weight depth completion network. Consequently, our network can be readily deployed on CNN accelerators and Neural Processing Units (NPUs). A noteworthy feature of our network is its enhanced capacity to distinguish boundary regions, which is primarily owe to the employment of a gradient loss function.

Table 1. Summary of essential characteristics of existing RGB guided methods on the NYU-v2 dataset. For denoting loss functions, we omit the coefficient of each loss term for simplicity.

Method	Error				Accuracy		Parameters
	Rmse	Rel	$\delta_{1.10}$	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$	
Sparse-to-Dense[26]	0.230	0.044	92.6	97.1	99.4	99.8	28.4M
Unet+CSPN[13]	0.117	0.016	97.1	99.2	99.9	100.0	256M
KernelNet[14]	0.111	0.015	97.4	99.3	99.9	100.0	16.47M
Nconv-CNN[12]	0.125	0.017	96.7	99.1	99.8	100.0	484K
Ours	0.116	0.018	96.8	99.3	99.9	100.0	1.07M

5.3. Network implement on hardware system

As demonstrated in Section 5.2, our network exhibits state-of-the-art performance in RGB-guided depth completion tasks. For instance, the network achieves a Root Mean Square Error (RMSE) of 0.116 m on the NYUv2 training set, allowing for more detailed identification at the boundaries. In order to implement the proposed network within our hardware system, the 8x8 depth sensor is aligned with the RGB data(256x256). On the network side, we adapt the network input according to the hardware's RGB size and the aligned depth map coordinates, retrain the neural network, and integrate it with the hardware. The resulting completed depth map is illustrated in Figure 11, where the boundaries and shapes are clearly discernible, meeting the fundamental requirements for consumer-level imaging applications.

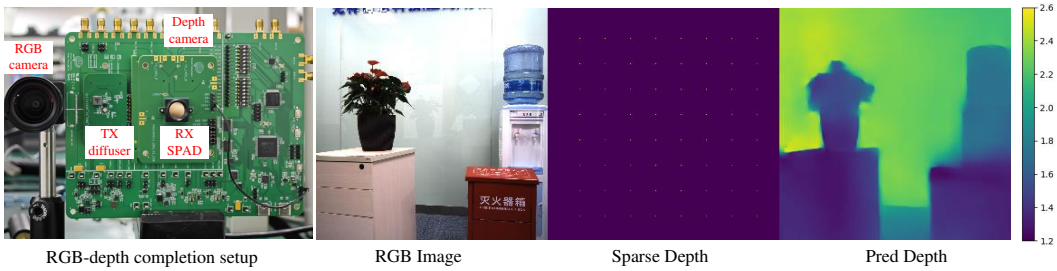

Figure 11. The result of our hardware system. From left-to-right: RGB-depth completion setup, Guidance RGB image, the sparse depth map and the predicted depth map from our network.

6. Conclusions

In this work, we propose a LiDAR system based on SPAD integrating micro-optical devices, incorporating RGB-guided 8x8 depth completion to 256x256 pixels with a lightweight neural network. To verify the effectiveness, we present a ranging SoC based on 130nm BCD process. The integrated SPAD and TDC enable the system to achieve millimeter-level measurement accuracy in indoor environments. The 16x16 pixels SPAD array can

be divided into 2x2 regions for ranging at high frame rates, or 8x8 for sparse imaging at lower frame rates with an energy efficiency of 200mW. Next, a micro-optical system is proposed to make imaging feasible at the hardware level. A micrometer-scale random MLA is used at the transmitting end to homogenize and expand the Gaussian beams from the VCSEL array into a flat-top light source with a 45° FOV. A freeform lens is utilized in the receiver, allowing the alignment of the SPAD array with the 45° FOV illumination and achieving 8x8 array sparse imaging at the optical level. To further increase the resolution to meet the requirements of consumer-level imaging, an RGB-guided depth completion neural network is integrated into the sparse depth imaging system, reaching a 256x256 resolution that matches QVGA standards. The low-cost, lightweight depth imaging system has widespread applications in distance measurement, simple object recognition, and basic pose recognition technology fields.

Author Contributions: All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing not applicable. Conflicts of Interest: The authors declare no conflict of interest.

References

1. David, R.; Allard, B.; Branca, X.; Joubert, C. Study and design of an integrated CMOS laser diode driver for an itof-based 3D image sensor. In Proceedings of the CIPS 2020; 11th International Conference on Integrated Power Electronics Systems. VDE, 2020, pp. 1–6.
2. Bamji, C.; Godbaz, J.; Oh, M.; Mehta, S.; Payne, A.; Ortiz, S.; Nagaraja, S.; Perry, T.; Thompson, B. A Review of Indirect Time-of-Flight Technologies. *IEEE Transactions on Electron Devices* **2022**.
3. Zhang, C.; Zhang, N.; Ma, Z.; Wang, L.; Qin, Y.; Jia, J.; Zang, K. A 240× 160 3D-stacked SPAD dToF image sensor with rolling shutter and in-pixel histogram for mobile devices. *IEEE Open Journal of the Solid-State Circuits Society* **2021**, *2*, 3–11.
4. Li, N.; Ho, C.P.; Xue, J.; Lim, L.W.; Chen, G.; Fu, Y.H.; Lee, L.Y.T. A Progress Review on Solid-State LiDAR and Nanophotonics-Based LiDAR Sensors. *Laser & Photonics Reviews* **2022**, *16*, 2100511.
5. Hutchings, S.W.; Johnston, N.; Gyongy, I.; Al Abbas, T.; Dutton, N.A.; Tyler, M.; Chan, S.; Leach, J.; Henderson, R.K. A reconfigurable 3-D-stacked SPAD imager with in-pixel histogramming for flash LIDAR or high-speed time-of-flight imaging. *IEEE Journal of Solid-State Circuits* **2019**, *54*, 2947–2956.
6. Ximenes, A.R.; Padmanabhan, P.; Lee, M.J.; Yamashita, Y.; Yaung, D.N.; Charbon, E. A 256× 256 45/65nm 3D-stacked SPAD-based direct TOF image sensor for LiDAR applications with optical polar modulation for up to 18.6 dB interference suppression. In Proceedings of the 2018 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2018, pp. 96–98.
7. Bantounos, K.; Smeeton, T.M.; Underwood, I. 24-3: Distinguished Student Paper: Towards a Solid-State LIDAR Using Holographic Illumination and a SPAD-Based Time-Of-Flight Image Sensor. In Proceedings of the SID Symposium Digest of Technical Papers. Wiley Online Library, 2022, Vol. 53, pp. 279–282.
8. Kumagai, O.; Ohmachi, J.; Matsumura, M.; Yagi, S.; Tayu, K.; Amagawa, K.; Matsukawa, T.; Ozawa, O.; Hirono, D.; Shinozuka, Y.; et al. 7.3 A 189× 600 back-illuminated stacked SPAD direct time-of-flight depth sensor for automotive LiDAR systems. In Proceedings of the 2021 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2021, Vol. 64, pp. 110–112.
9. Ximenes, A.R.; Padmanabhan, P.; Lee, M.J.; Yamashita, Y.; Yaung, D.N.; Charbon, E. A modular, direct time-of-flight depth sensor in 45/65-nm 3-D-stacked CMOS technology. *IEEE Journal of Solid-State Circuits* **2019**, *54*, 3203–3214.
10. Ulku, A.C.; Bruschini, C.; Antolović, I.M.; Kuo, Y.; Ankri, R.; Weiss, S.; Michalet, X.; Charbon, E. A 512× 512 SPAD image sensor with integrated gating for widefield FLIM. *IEEE Journal of Selected Topics in Quantum Electronics* **2018**, *25*, 1–12.
11. Hu, J.; Bao, C.; Ozay, M.; Fan, C.; Gao, Q.; Liu, H.; Lam, T.L. Deep Depth Completion from Extremely Sparse Data: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**.
12. Eldesokey, A.; Felsberg, M.; Khan, F.S. Confidence propagation through cnns for guided sparse depth regression. *IEEE transactions on pattern analysis and machine intelligence* **2019**, *42*, 2423–2436.
13. Cheng, X.; Wang, P.; Yang, R. Depth estimation via affinity learned with convolutional spatial propagation network. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 103–119.
14. Liu, L.; Liao, Y.; Wang, Y.; Geiger, A.; Liu, Y. Learning steering kernels for guided depth completion. *IEEE Transactions on Image Processing* **2021**, *30*, 2850–2861.
15. Zhu, J.; Xu, Z.; Fu, D.; Hu, C. Laser spot center detection and comparison test. *Photonic Sensors* **2019**, *9*, 49–52.
16. Yuan, W.; Xu, C.; Xue, L.; Pang, H.; Cao, A.; Fu, Y.; Deng, Q. Integrated double-sided random microlens array used for laser beam homogenization. *Micromachines* **2021**, *12*, 673.
17. Jin, Y.; Hassan, A.; Jiang, Y. Freeform microlens array homogenizer for excimer laser beam shaping. *Optics express* **2016**, *24*, 24846–24858.

18. Cao, A.; Shi, L.; Yu, J.; Pang, H.; Zhang, M.; Deng, Q. Laser Beam Homogenization Method Based on Random Microlens Array. Appl. *Laser* **2015**, *35*, 124–128.
19. Liu, Z.; Liu, H.; Lu, Z.; Li, Q.; Li, J. A beam homogenizer for digital micromirror device lithography system based on random freeform microlenses. *Optics Communications* **2019**, *443*, 211–215.
20. Xue, L.; Pang, Y.; Liu, W.; Liu, L.; Pang, H.; Cao, A.; Shi, L.; Fu, Y.; Deng, Q. Fabrication of random microlens array for laser beam homogenization with high efficiency. *Micromachines* **2020**, *11*, 338.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
22. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.
23. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 2015, pp. 234–241.
24. Hu, J.; Ozay, M.; Zhang, Y.; Okatani, T. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In Proceedings of the 2019 IEEE winter conference on applications of computer vision (WACV). IEEE, 2019, pp. 1043–1051.
25. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. *ECCV (5)* **2012**, *7576*, 746–760.
26. Ma, F.; Karaman, S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In Proceedings of the 2018 IEEE international conference on robotics and automation (ICRA). IEEE, 2018, pp. 4796–4803.