Neural Single-Shot GHz FMCW Correlation Imaging

CINDY (HSIN) PAN[1], NOAH WALSH[1], YUXUAN ZHANG[1], ZHENG SHI[1], AND FELIX HEIDE[1*]

[1] *Princeton University, USA*

[*] *Corresponding Author: fheide@cs.princeton.edu*

Abstract: Depth sensing is essential for 3D environmental perception across application domains, including autonomous driving, topographical mapping, and augmented and virtual reality (AR/VR). Traditional correlation time-of-flight (ToF) methods, while able to produce dense high-resolution depth maps, are plagued by phase wrapping artifacts which limit their effective depth range. Though multi-frequency methods can help reduce this problem by simultaneously solving for phase wrap counts in multiple wavelengths, this requires multiple measurements per pixel, necessitating additional hardware and imaging time. We introduce a 3D imaging method that requires a single per-point measurement by combining frequency-modulated continuous wave (FMCW) operation, all-optical correlation ToF imaging, and a specialized frequency-decoding network. Our system performs all-optical correlation imaging at GHz rates. The method is validated through both simulations and real-world experiments, comparing favorably to existing methods in all experiments.

1. Introduction

The field of image processing has witnessed significant advancements driven largely by the advent of large-scale image datasets, such as ImageNet [1], and the increase in computational power. For depth sensing, a parallel development has been sparked by affordable RGB-D depth cameras. This trend highlights the need for acquiring high-quality depth maps in large volumes, a key goal for a wide range of applications in 3D graphics and vision, ranging from autonomous driving [2] and topographical mapping [3] to gaming [4] and virtual reality [5]. For these applications, acquiring high-quality depth information is essential for accurate scene understanding and decision makeing. The quality of the captured depth hinges on not only the signal-to-noise ratio (SNR) of the hardware components [6] but also on the computational efficiency and capability of the subsequent processing algorithms [7].

Time-of-flight (ToF) methods stand out as some of the most effective techniques in active depth sensing. These approaches recover the distance between the scene and the detector either by directly measuring the round-trip travel time of light or by analyzing the interference patterns of different light paths. Direct ToF methods, such as flash-based direct-ToF cameras [8] and scanning-based LiDAR systems [9], generate point clouds by measuring travel distances. However, despite their compactness and cost-effectiveness, direct ToF techniques often suffer from low resolution. This is primarily due to the limitations of sensitive time-resolved detectors and the low photon-flux in reflected pulses, which in turn adversely affects the signal-to-noise ratio (SNR) [10, 11]. In contrast, correlation ToF methods overcome these constraints by utilizing the interference of continuously intensity-modulated signals between the emission and return paths. Depth is inferred by calculating the phase shift. Unlike direct ToF, correlation ToF does not require ultra-short pulse generation or extreme sampling rates. This obviates the need for time-tagging sparse photons, thus enabling significantly higher depth resolution [12].

In correlation ToF imaging, high modulation frequency is desirable for its ability to suppress signal perturbations and improve resolution. Holding sensor noise and measurement quantization constant, depth precision is directly proportional to phase contrast [13]. If the signal frequency is doubled, the depth change represented by a single bit flip in the phase measurement is correspondingly halved. As illustrated in Fig. 1, a 100 MHz ToF system might achieve cm-scale precision for a small object placed 1 meter away from the detector, whereas a 10 GHz system could resolve micron-scale textures. Building upon this concept, Baek et al. [13] propose the implementation of stable GHz modulation through electro-optic modulators (EOMs), polarizing optics, and integrated circuits, enabling all-optical correlation computations in free-space and bypassing the noise inherent in photon-electron conversion.

However, high-frequency modulation introduces a trade-off in range resolution. A 10 GHz signal, for example, has a wavelength of approximately 3 cm, translating to a range resolution of 3 cm per phase wrap. This becomes problematic in typical indoor settings where distances often surpass 3 cm, resulting in multiple phase wraps and the need for effective unwrapping mechanisms. Single-frequency phase unwrapping methods, while capable of recovering relative depths, encounter issues with reference ambiguity [14]. Without a zero wrap measurement, determining the starting point of the unwrapping process is challenging. These methods also face difficulties with phase discontinuities, where the exact count of phase wraps is ambiguous. To overcome these limitations, multi-frequency phase unwrapping algorithms have been proposed. For example, Gupta et al. [15] propose the use of look-up tables to discern phase numbers for micro-ToF unwrapping, where high temporal frequencies are used which have small (micro) periods. Additionally, Baek et al. [13] utilize double-frequency measurements combined with a trained-classification network to manage numerous phase wraps over larger distances. However, these multi-frequency approaches typically require significantly longer acquisition times and are less effective with narrow-bandwidth systems due to their sensitivity to noise.

In our work, we aim to retain the high contrast benefits of GHz modulation while significantly reducing capture times. We build upon the all-optical correlation approach from Baek et al. [13] and propose a frequency modulation capture scheme with a single measurement per point. Moving beyond merely modulating amplitude and measuring phases, this approach allows us to generate absolute depth information from one single-chirp measurement. Utilizing a specialized depth-decoding network, our proposed method can reconstruct absolute depth from a single measurement, thereby eliminating the need for multi-frequency measurements. Specifically, our contributions are as follows:

- We propose an depth estimation approach that integrates frequency modulated continuous wave (FMCW) operation with all-optical correlation ToF, enabling accurate absolute depth reconstruction in the GHz range from a single-chirp measurement, eliminating the need for multiple frequency measurements and effectively halving the capture time.
- We propose a trained frequency-decoding network that extends the FMCW range resolution beyond the traditional 12.5 m limitation, overcoming the constraints imposed by the EOMs with low 20 MHz modulation bandwidth.
- We validate our proposed system and frequency-decoding network, along with a inference-guided test-time optimization algorithm, in simulation and with an experimental prototype, demonstrating our capability for absolute depth imaging in a computationally effective and robust manner.

2. Related Work

In the following, we briefly review work related to the proposed method.

Correlation ToF. Correlation ToF involves illuminating a scene with periodically modulated light and determining distances by analyzing the phase shifts between transmitted and received signals. This depth sensing method, leveraging cost-effective CMOS sensors and standard laser diodes for capturing dense depth data [16], has been utilized in devices like the Microsoft Kinect.

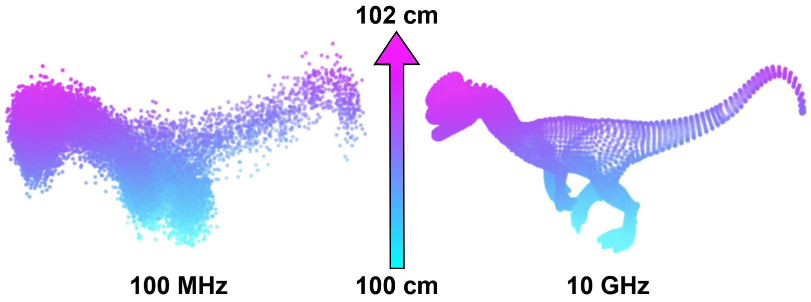

Fig. 1. Comparative simulated ToF measurements of a small object with height and width less than 2 cm at 1 m distance - 100 MHz system achieves cm-scale resolution, while 10 GHz system attains micron-scale. The higher modulation frequency of 10 GHz provides better resolution due to enhanced phase contrast, reducing unwanted signal perturbation.

Although flood illumination can lead to multi-path interference, significant research efforts have aimed to address this challenge, paving the way for diverse applications such as non-line-of-sight imaging, penetration through scattering media, and material classification [17–22]. Nevertheless, conventional methods are generally confined to modulation frequencies within the hundreds of MHz range due to the photon absorption depth in silicon, restricting depth resolution to millimeters or centimeters over several meters of range [23]. Overcoming limitations related to low modulation contrast and interferometry errors, which have impeded previous efforts to increase modulation frequency [24, 25], our work adopts an all-optical free space approach for correlation measurements [13]. This strategy circumvents the limitations imposed by photon absorption in silicon, facilitating operation at modulation frequencies beyond 10 gigahertz.

Phase Unwrapping. In high-frequency correlation ToF systems, the travel distance of the signal typically surpasses a single wavelength, inadvertently leading to phase shifts in correlation signals exceeding 2π. Accurately determining the phase offset to achieve absolute depth reconstruction necessitates the use of phase unwrapping algorithms. Current single-frequency phase unwrapping methods primarily recover relative depth and wrap count but require assumptions about the scene to infer absolute depths [26–28], and can only retrieve absolute depth to a limited extent. To circumvent this, multi-frequency phase unwrapping algorithms have been developed. These algorithms utilize lower-frequency signals to unwrap high-frequency phases and employ techniques like weighted Euclidean division or frequency-space lookup tables for wrap count retrieval [15, 29–32]. However, while achieving promising performance in MHz ToF imaging scenarios, these methods are prone to noise and ambient light interference, resulting in compromised performance when managing the numerous wrap counts encountered in GHz correlation imaging. To tackle this issue, our approach merges FMCW operation with all-optical correlation ToF, and decodes absolute depth information from the frequencies of the correlation signals. Combined with our advanced frequency-decoding network and test-time optimization algorithm, our method can reliably reconstruct absolute depths with high-fidelity geometric features from single-chirp measurement, even within the GHz range.

3. Methods

To efficiently acquire absolute depth information without the need for multiple frequency measurements, we introduce a single-chirp depth imaging pipeline which is illustrated in Fig. 2. This process begins by projecting a frequency-modulated signal onto the scene. The correlation

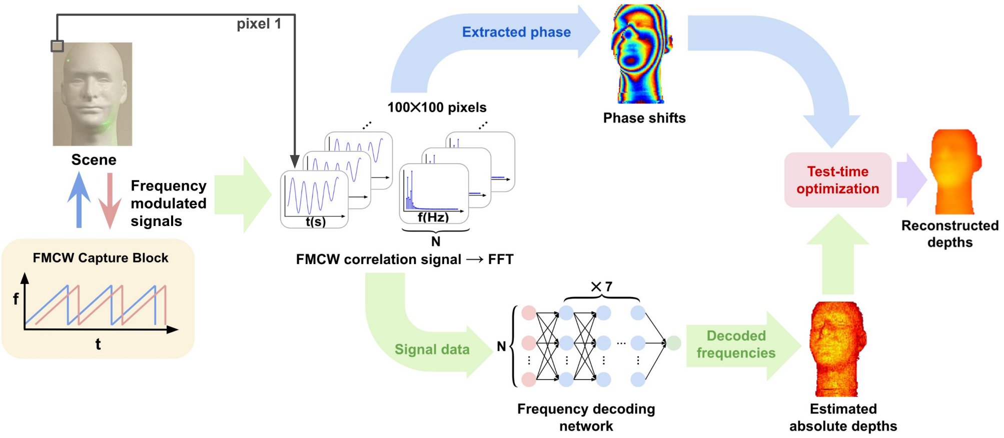

Fig. 2. Overview of the Single-chirp Depth Imaging Pipeline. The method begins with FMCW capture hardware emitting frequency-modulated signals into the scene. Reflected signals are captured, and the FMCW correlation signals are optically computed in the time domain. Phase shifts are then extracted before using FFT converting these signals to the frequency domain. Next, pixel-wise frequency decoding is applied to the frequency domain signals to determine absolute depths for each pixel. The final step involves a test-time optimization, which further refines the depth output by integrating the estimated depths with phase information.

signals generated are then optically computed and subsequently decoded on a pixel-by-pixel basis to compute the absolute depths for each pixel. Following this, the depth outputs are further refined through a test-time optimization process. Detailed description of each component in this pipeline are provided in the subsequent sections.

3.1. Frequency Modulated Continuous Wave

Next, we provide an overview of the proposed Frequency Modulated Continuous Wave (FMCW) ToF method. The method utilizes a signal, denoted as $p(t)$, which oscillates at a saw-tooth chirped frequency $\omega(t)$ with bandwidth B and chirp length T_s, as illustrated in Fig. 3. This signal, having an amplitude α and a DC offset β, is projected onto a scene, and can be expressed as:

$$p(t) = \alpha \cos(\omega(t)t) + \beta, \quad \omega(t) = 2\pi(f_0 + \frac{B}{T_s}t), \tag{1}$$

where $f_0 = f_c - \frac{B}{2}$. The light reflected back from the scene, denoted by $\tilde{p}(t)$, undergoes a time delay τ and oscillates at frequency $\omega_p = \omega(t+\tau)$, which introduces a phase shift ϕ and results in an attenuated amplitude $\tilde{\alpha}$ and offset $\tilde{\beta}$:

$$\tilde{p}(t+\tau) = \tilde{\alpha}\cos(\omega_p t + \phi) + \tilde{\beta}, \quad \phi = \omega_p \tau. \tag{2}$$

To extract the phase shift ϕ, the reflected signal $\tilde{p}(t)$ is mixed with a reference signal $r(t) = \cos(\omega_r t + \psi)$, where $\omega_r = \omega(t)$. The resulting correlation signal is:

$$\begin{aligned} \tilde{p}(t+\tau)r(t) &= \tfrac{\tilde{\alpha}}{2}\cos((\omega_r - \omega_p)t + \psi - \phi) \\ &+ \tfrac{\tilde{\alpha}}{2}\cos((\omega_r + \omega_p)t + \phi + \psi) + \tilde{\beta}\cos(\omega_r t + \psi). \end{aligned} \tag{3}$$

By integrating over the exposure time T, which acts as a low-pass filter when $T >> \frac{1}{\omega}$, the term $\frac{\tilde{\alpha}}{2}\cos((\omega_r - \omega_p)t + \psi - \phi)$ is isolated, enabling the decoding of depth information from the

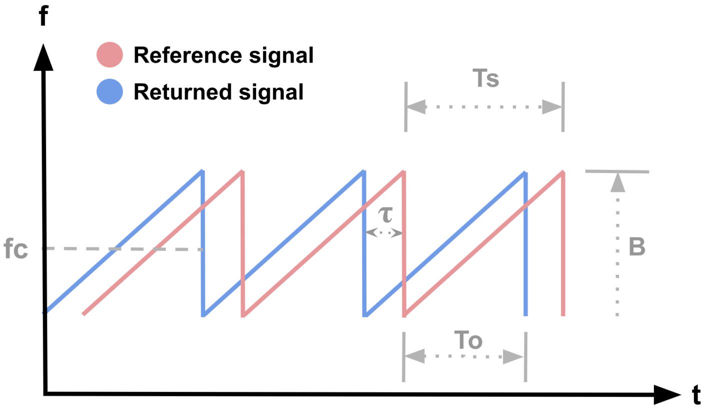

Fig. 3. Frequency-Time Plot showing a reference signal (red) with bandwidth B and chirp duration T_s, alongside the returned signal (blue). The observation period T_o and the time delay τ indicate the round-trip travel time between the detector and the scene.

phase shift ϕ:

$$\begin{aligned} C_\psi &= \int_0^T [\tilde{p}(t-\tau)r(t)]dt \\ &= \frac{\tilde{\alpha}}{2(\omega_r - \omega_p)} \sin((\omega_r - \omega_p)t + \psi - \phi) + TK. \end{aligned} \tag{4}$$

While the use of GHz-range modulation frequencies in our system enables ultra-high, mm-scale resolution, it also brings a phase unwrapping challenge due to the fact that the path length between the scene and the detector often surpasses a single wavelength of the modulated light, which is typically in the centimeter range. In standard meter-scale indoor scenes, this discrepancy results in dozens of phase wraps, posing a significant challenge in determining absolute depths accurately.

To address this challenge and accurately recover depths in the presence of phase wraps, we focus on the $\omega_b = \omega_r - \omega_p$ frequency component of the correlation signal, as defined in Eq. 4. The one-way travel distance, denoted as Δd, can be deduced from the beat frequency $f_b = \frac{\omega_b}{2\pi}$, in conjunction with the chirp slope S as follows:

$$f_b = \frac{S2\Delta d}{c}, \quad \Delta d = \frac{c f_b T_o}{2B_e}, \tag{5}$$

where c is the speed of light, $T_o = T_s - \tau$ is the observation time, and $B_e = B\frac{T_o}{T_s}$ is the effective bandwidth. Consequently, the correlation signal can be rewritten as:

$$S(t) = A\cos(2\pi f_b t + \phi). \tag{6}$$

In practice, the beat frequency f_b is typically estimated using Fast Fourier Transform (FFT). The ability to resolve distinct peaks in the frequency domain is constrained by the 3 dB width of the FFT sinc function centered at f_b, which inversely relates to T_o. This implies that two frequencies in the frequency domain are resolvable only if:

$$\Delta f > \frac{1}{T_o}. \tag{7}$$

Similarly, the range resolution Δr of the FMCW method can be defined as

$$\Delta r = \frac{c}{2B_e}. \tag{8}$$

Taking into account the aforementioned constraint, let us consider a system operating at a 7.15 GHz modulation frequency, which corresponds to a wavelength of 4.2 cm. In this scenario, for an indoor scene with a maximum depth of 2 meters, we encounter approximately $\frac{2\times 200}{4.2} \approx 100$ phase wraps that need to be resolved. To accurately determine the absolute distance for each wrap, an effective bandwidth of approximately 3.6 GHz is required.

However, while Electro-Optic Modulators (EOMs) capable of GHz-rate modulation in free space can be custom-designed, the practical limitation arises from the narrowed usable bandwidth of the tank resonant circuit, which in our case restricts the bandwidth to 20 MHz [33]. Given this bandwidth, the theoretical range resolution, calculated using Eq. 8, is approximately 12.5 m. This resolution is drastically below the 4.2 cm resolution necessary for effective unwrapping at a 7.15 GHz modulation frequency.

To address this limitation in range resolution, we introduce a two-step approach. First, a frequency decoding network is optimized to enhance the resolution limit from 12.5 meters to centimeter-scale. Second, a gradient-based test-time optimization algorithm is employed to further refine scene depth reconstruction to millimeter-scale resolution. In the following, we provide details of the experimental setup and the computational methods employed for depth reconstruction.

3.2. Frequency Decoding Network

We devise a neural network model where the input is the correlation signal array associated with a pixel to infer the absolute depth d_p from input signal arrays while elevating the resolution limit from 12.5 meters to the cm-scale. The input array is denoted as $s_p := s_p^1, s_p^2, ..., s_p^N$, where N is the length of the signal array. Prior to inputting the signal array into the network, we apply a Fast Fourier Transform (FFT) to the raw time-domain signal array s to transform it into the frequency domain. This transformation enhances feature extraction and representation, which helps the model to more effectively capture frequency characteristics essential for encoding absolute depth information. The transformed input, denoted as $s'_p = FFT(s_p)$, retains the same length as s_p but represents the frequency domain. Mathematically, this is expressed as:

$$s'^{k}_{p} = \sum_{n=1}^{N} s_p^n \times e^{\frac{-2i\pi}{N} kn}. \tag{9}$$

We employ a Multi-Layer Perceptron (MLP) for our frequency decoding network (FDN). It comprises 8 layers, each containing 1024 neurons and softsign activation function to introduce non-linearity and enable more intricate modeling capabilities. The deep architecture and wide layers enable the model to discern complex relationships among the signals.

We optimize the network to minimize the ℓ_1 loss $\mathcal{L}FDN$ between the predicted $FDN(FFT(s_p))$ and actual depths d_p,

$$\mathcal{L}FDN(s_p, d_p) = ||FDN(FFT(s_p)) - d_p||_1. \tag{10}$$

To train the network, we gathered 8 sets of correlation signal data, where 7 sets are used for training while 1 set is withheld for testing purpose. These measurements cover a depth range from 0 mm to 1500 mm, with 1 mm increments, using our measurement apparatus equipped with a piezoelectric motion stage. This stage is designed for minute axial adjustments, featuring a theoretical resolution of 50 nm. We use the Adam optimizer [34], and train the model for 5000 epochs with a batch size of 16 and a learning rate of 10^{-4}. The model achieves convergence in approximately 8 minutes on an NVIDIA A100 GPU. Further details of the network are elaborated in the supplemental document.

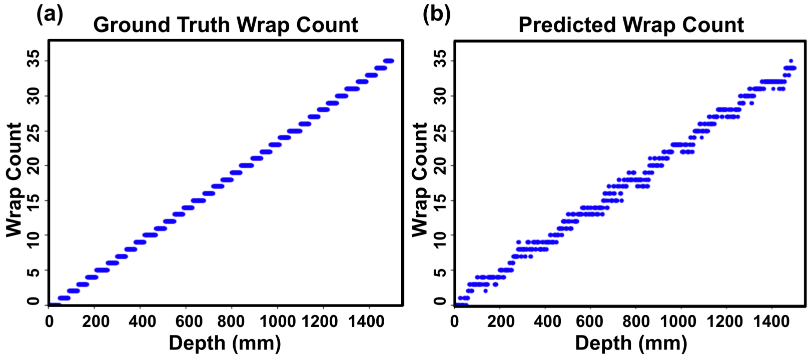

Fig. 4. Test Performance of the Frequency-Decoding Network. (a) Displays the ground truth wrap count, and (b) shows the predicted wrap count over a depth range of 0 mm to 1500 mm using the frequency decoding network. Notably, conventional baseline methods are unable to unwrap this 1D data due to their reliance on spatial context.

As demonstrated qualitatively in Fig. 4, our tests on withheld measurements reveal a mean-square error of 0.86. This deviation in wrap count is further minimized when both phase measurements and spatial information are incorporated during the test-time optimization step, which will be described in the subsequent section. It is also important to note that conventional phase unwrapping methods are unable to unwrap this 1D data efficiently, largely due to their reliance on spatial context. Please refer to the supplemental document for more detailed analysis of the impact of signal-to-noise ratio on measurement accuracy.

3.3. Test-Time Optimization

To further refine the depth output from the pixel-wise frequency decoding network, we implement a test-time optimization approach. This process incorporates spatial information from measured phase shifts, optimizing the wrap count to align the spatial gradient of the refined depth output with that of the high-resolution features provided by the GHz phase contours.

To extract pseudo-ground-truth gradients, the phase map is first rescaled so that a change of 2π in phase directly corresponds to a change in depth of one wavelength, $\lambda = 4.2$ cm, represented by the equation

$$\phi_d(x, y) = \frac{\phi(x, y)}{2\pi}\lambda, \tag{11}$$

where x and y are the coordinates in the pixel grid. We then filter out the contour lines that appear between phase wraps. This is achieved by substituting pixels in the rescaled phase map, specifically those with spatial gradient values equal to λ or $-\lambda$, with spatial gradients derived from the estimated absolute depths provided by the frequency decoding network. Together, the psedo-ground-truth gradient δ_h and δ_v can be expressed as

$$\delta_h = \begin{cases} \frac{\partial \phi_d}{\partial x}, & \frac{\partial \phi_d}{\partial x} < \lambda \\ \frac{\partial d_{FDN}}{\partial x}, & \frac{\partial \phi_d}{\partial x} \geq \lambda \end{cases}, \quad \delta_v = \begin{cases} \frac{\partial \phi_d}{\partial y}, & \frac{\partial \phi_d}{\partial y} < \lambda \\ \frac{\partial d_{FDN}}{\partial y}, & \frac{\partial \phi_d}{\partial y} \geq \lambda \end{cases}, \tag{12}$$

where $\frac{\partial d_{FDN}}{\partial x}$ and $\frac{\partial d_{FDN}}{\partial y}$ are the gradients of the frequency decoding network output.

The optimization algorithm iteratively updates the depth output, starting with the initial estimates from the frequency decoding network and progressively minimizing the per-pixel mean squared error (MSE) loss between the psedo-ground-truth horizontal and vertical gradient δ_h

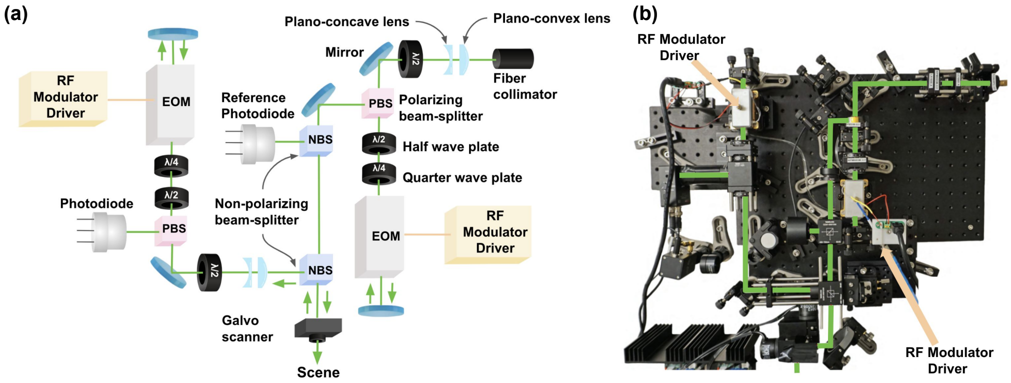

Fig. 5. (a) Schematic illustration of our all-optical FMCW prototype, utilizing polarizing optics and Electro-Optic Modulators (EOMs). (b) Photograph of the experimental setup, with light paths marked in green. The EOMs are responsible for generating GHz amplitude modulation, while frequency modulation is achieved via an RF generator. See Sec. 4 for more details.

and δ_v, and that of the reconstructed depth, represented as δ'_h and δ'_v:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{MSE}}^{\delta_h} + \mathcal{L}_{\text{MSE}}^{\delta_v} \\ &= \frac{1}{p}\sum_{n=1}^{p}(\delta'_h - \delta_h)^2 + \frac{1}{p}\sum_{n=1}^{p}(\delta'_v - \delta_v)^2, \end{aligned} \tag{13}$$

where p represents the total number of pixels in the image. We have implemented this optimization using PyTorch, configured to run for 200 epochs. The learning rate is set to begin at 1 and reduces at a rate of 0.95 per epoch. The optimization utilizes the ADAM optimizer [34] and converges within 2 minutes on an NVIDIA A100 GPU. Please refer to the supplementary file Code File 1 for more implementation details.

4. Experimental Prototype

In this section, we provide an overview of our all-optical FMCW ToF measurement setup which builds on Baek et al. [13]. Fig. 5 provides a schematic illustration of our prototype, and a photograph of the setup, with light paths highlighted.

The process starts in the illumination module, where a polarizing beam-splitter (PBS) receives a 3mW, 532 nm wavelength light beam and converts this light into vertically linearly polarized light, which can be represented by:

$$E_0 = AL_v, \quad L_v = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \tag{14}$$

where A is the amplitude of the incoming light and L_v is the Jones matrix for vertical linear polarizer.

Following this, the light undergoes modulation through a sequence of optical elements: a half wave plate (HWP), a quarter wave plate (QWP), and an Electro-Optic Modulator (EOM). This sequence is repeated in reverse after the light reflects off a mirror. The modulation imparted by the HWP, oriented at $\theta_{\text{HWP}} = 11.25°$ and QWP, oriented at $\theta_{\text{QWP}} = 45°$, can be defined using their respective Jones matrices:

$$\begin{aligned}
&H(\theta_{\text{HWP}}) \\
&= e^{\frac{-i\pi}{2}} \begin{bmatrix} \cos^2\theta_{\text{HWP}} - \sin^2\theta_{\text{HWP}} & 2\cos\theta_{\text{HWP}}\sin\theta_{\text{HWP}} \\ 2\cos\theta_{\text{HWP}}\sin\theta_{\text{HWP}} & \sin^2\theta_{\text{HWP}} - \cos^2\theta_{\text{HWP}} \end{bmatrix}, \\
&Q(\theta_{\text{QWP}}) \\
&= e^{\frac{-i\pi}{4}} \begin{bmatrix} \cos^2\theta_{\text{QWP}} + i\sin^2\theta_{\text{QWP}} & (1-i)\cos\theta_{\text{QWP}}\sin\theta_{\text{QWP}} \\ (1-i)\cos\theta_{\text{QWP}}\sin\theta_{\text{QWP}} & \sin^2\theta_{\text{QWP}} + i\cos^2\theta_{\text{QWP}} \end{bmatrix}.
\end{aligned} \tag{15}$$

We employ an external RF generator (R&S SMW) to input a frequency chirped sinusoidal voltage with a center frequency of 7.15 GHz and a bandwidth of 20 MHz to the RF drivers of our EOMs. This GHz modulation within the EOMs is characterized using a Jones matrix $B(V)$, which captures the phase relationship between the light's perpendicular polarization components:

$$B(V) = \begin{bmatrix} e^{\frac{-i\Gamma(V)}{2}} & 0 \\ 0 & e^{\frac{i\Gamma(V)}{2}} \end{bmatrix}, \tag{16}$$

where $\Gamma(V)$ is the net birefringence and V is an oscillating voltage at frequency $\omega(t)$. For more details of the custom operation of our EOMs, please refer to the supplemental document.

The polarization state of the light, modulated through the previously described sequence of HWP, QWP, and EOM in both forward and backward directions, can be expressed as:

$$E_1 = L_h \overbrace{H(-\theta_{\text{HWP}})Q(-\theta_{\text{QWP}})B(V)}^{\text{backward pass}}\, M\, \overbrace{B(V)Q(\theta_{\text{QWP}})H(\theta_{\text{HWP}})}^{\text{forward pass}} E_0, \tag{17}$$

where M and L_h are the Jones matrix of a mirror and a horizontal linear polarizer,

$$M = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad L_h = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{18}$$

By substituting the corresponding Jones matrices into Eq. 17, we can express the detected signal E_1 as a function of voltage V,

$$E_1 = A \begin{bmatrix} \frac{i(\cos V - \sin V)}{\sqrt{2}} \\ 0 \end{bmatrix}, \tag{19}$$

and thereby obtain the signal intensity

$$I(V) = |E_1^2| = \frac{A^2}{2}(1 - \sin 2V). \tag{20}$$

When applying a voltage oscillating at a GHz modulation frequency ω_r, the signal $I_r(t)$ detected by the reference photodiode takes the form as shown below. This expression can be further simplified using a Taylor expansion:

$$\begin{aligned}
I_r(t) &= \frac{A^2}{2}(1 - \sin(2\alpha \sin(\omega_r t + \phi))) \\
&\approx -A^2\alpha \sin(\omega_r t + \phi),
\end{aligned} \tag{21}$$

where a small modulation power α is assumed.

This signal propagates in free space towards the scene, completing the illumination module's role. Once reflected back from the scene, the time-delayed signal enters the detection module, which mirrors the structure of the illumination module. This detection stage, demodulating the returned light, comprises a HWP, a QWP, and an EOM, synchronized with its counterpart in the illumination module via an external clock from a function generator (Siglent SDG2042X), and a mirror. This demodulation process is akin to mixing the time-delayed signal with a reference signal. The demodulated signal is then passed through a 1 MHz lowpass filter to extract the correlation signal, featuring the lower beat note $(\omega_r - \omega_p)$, as outlined in Eq. 3 and Eq. 4.

5. Assessment

In this section, we validate the proposed neural single-chirp depth imaging approach using both simulated and experimental data. Specifically, we first perform quantitative evaluation of our proposed method on the synthetic Hypersim dataset [35], where our method is compared against representative existing approaches. We then experimentally validate our hardware prototype and depth reconstruction pipeline on unseen real-world measurements, where our method is qualitatively compared against the state-of-the-art single-frequency phase unwrapping method.

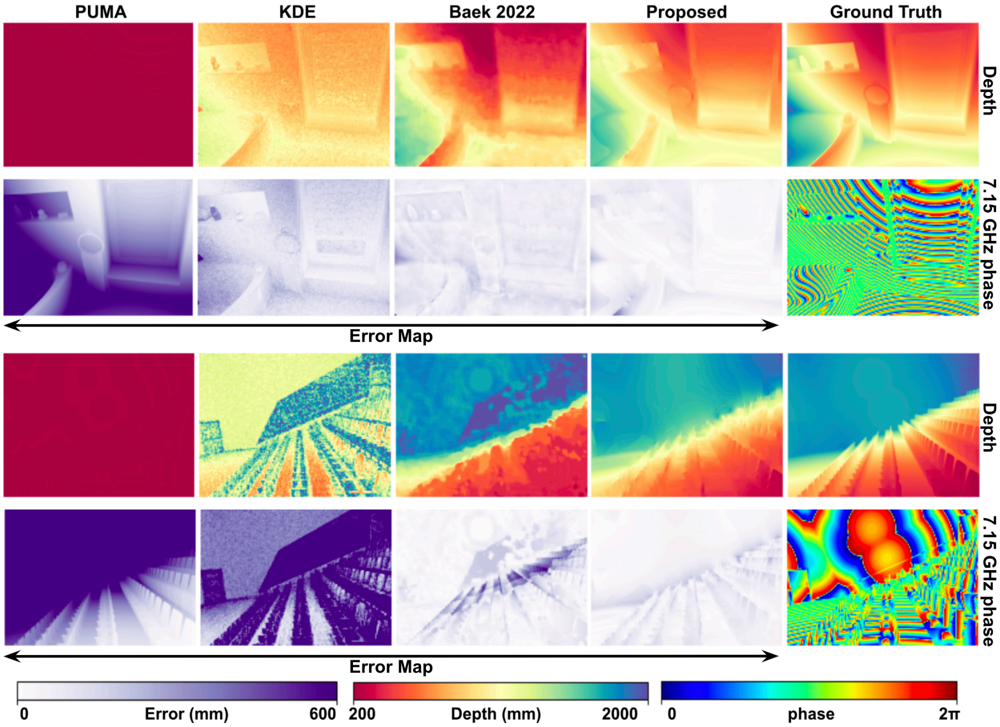

Fig. 6. Depth reconstruction results and corresponding error maps on selected Hypersim [35] RGB-D indoor scenes: we present a qualitative comparison between conventional, learned methods, and our proposed approach. From left to right, the methods displayed are: state-of-the-art single-frequency method PUMA [27], the kernel density based multi-frequency method KDE [32], the double-frequency neural phase unwrapping method Baek et al. [13], and our proposed method.

	RMSE (mm)	MAE (mm)	RE (1)
PUMA [27]	572.71	487.70	0.30
KDE [32]	542.62	419.86	0.31
Baek et al. [13]	373.49	346.27	0.14
Proposed	**88.17**	**71.79**	**0.04**

Table 1. Quantitative comparison of baseline methods and the proposed method on a set of synthetic test scenes evaluated in Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Relative Error (RE). Our method produces results with significantly lower error comparing to the baseline methods.

5.1. Synthetic Experiments

In our evaluation, we consider two types of existing phase unwrapping methods as baselines. The first type is the traditional single-phase unwrapping method, represented by Phase Unwrapping Maximum Flow (PUMA) [27], which reconstructs relative depth information from a single measurement. The second type is multi-frequency phase unwrapping methods, aimed at reconstructing absolute depth from at least two measurements. Representative methods include the traditional kernel density estimation (KDE) method [32], used in Kinect V2 software, and the recent neural unwrapping method by Baek et al. [13], the most pertinent to our proposed approach.

Qualitative and quantitative comparisons are reported in Fig. 6 and Tab. 1, respectively, while additional qualitative results are presented in the Supplemental Document. For these comparisons, we utilize the Hypersim RGB-D dataset [35], which contains 77,400 synthesized indoor scenes, each comes with ground truth depth maps and RGB images. For our simulations, we select scenes within a 0 to 2 meters depth range and simulate synthetic captures at a frequency of 7.15 GHz. For the baseline methods requiring multiple frequency measurements, we additionally simulate captures at a higher frequency of 14.32 GHz.

PUMA [27] is an energy minimization framework for single-frequency phase unwrapping. In this framework, the objective functions are modeled as first-order Markov random fields and a minimization process is then performed through a series of max-flow/min-cut calculations. While it is appealing that PUMA provides an exact solution of the energy minimization problem using the graph-cuts, it encounters limitations when faced with scenarios involving over a hundred phase wraps, and often defaults to predicting a uniform wrap count across the entire image.

Both KDE [32] and Baek et al. [13] are dual-frequency phase unwrapping methods that are capable of deriving absolute depth information from phase measurements at two different frequencies. KDE, which is an estimation approach based on neighborhoods of phase wrap hypotheses that favors spacial consistencies. However, this method often struggles to capture detailed surface features or handle discontinuities in phase wraps, which is particularly problematic when there are over a hundred phase wraps. Baek et al. is a double-frequency neural phase unwrapping method optimized for GHz frequency operation. As shown both qualitatively and quantitatively, Baek et al. offers a substantial improvement in depth estimation accuracy compared to the previously discussed baselines. However, its performance struggles when numerous objects having sharp edges and complex geometric structures are present. In contrast, our method not only reduces the error to less than one-third across all metrics and is able to preserve fine geometric details in large-scale indoor scenes.

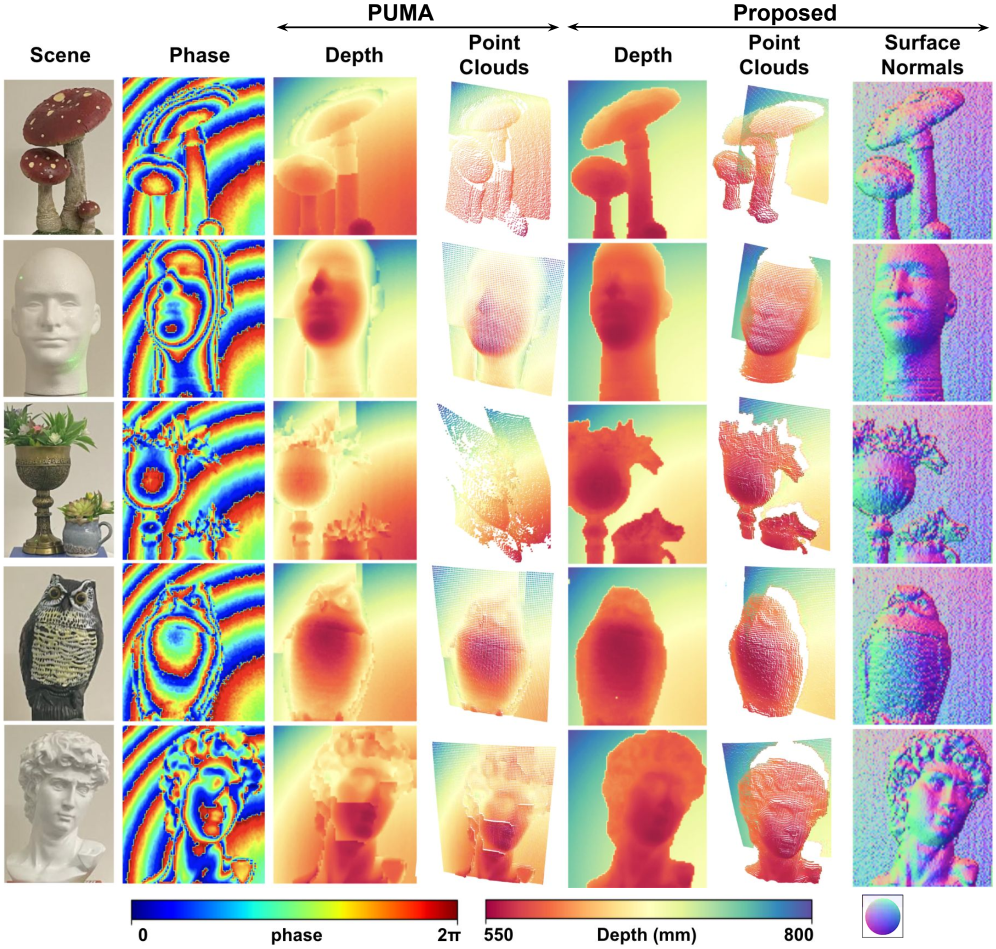

Fig. 7. Our experiments validate the proposed method across five real-world scenes, featuring diverse geometric structures and surface textures, under indoor ambient lighting. We also assess the performance of existing single-frequency phase-unwrapping method PUMA [27] on GHz phase measurements. Since PUMA reconstructs only the relative wrap count, we adjusted its output to align with the depth range of our test scenes. PUMA often merges objects with the background, leading to failures in most scenes. In contrast, our method effectively reconstructs absolute depth information for target objects, showcasing its robust capability across a wide range of surface materials, from low-reflectance, dark-colored surfaces to glossy ones.

5.2. Real-World Experiments

We further validate the proposed system on real-world scenes containing target objects with different materials and complex geometric structures. In each case, we gathered 100×100 pixel measurements using our experimental prototype, assisted by a galvo scanning system. Sample measurements for these scenes are shown in Fig. 7, and we additionally compared to the single-frequency phase unwrapping method PUMA using the same 7.15 GHz measurement. Given that PUMA exclusively reconstructs the relative wrap count, we modified its output to correspond with the depth range of the test scenes. PUMA performs significantly better when dealing with just a few phase wraps as opposed to the hundred wraps in the synthetic experiments. However, constrained by the inherently ill-posed nature of unwrapping with single-frequency input, it struggles to handle phase discontinuity, often resulting in the foreground object blending into the background or background being teared. The proposed method successfully tackles this challenge, enabling accurate reconstruction of absolute depth information across a broad spectrum of surface materials, including low-reflectance, dark-colored surfaces as well as glossy ones. The outcomes are comprehensively visualized through depth maps, point clouds, and surface normals in Fig. 7.

6. Conclusion

In this work, we proposed a novel depth sensing system capable of reconstructing absolute depth information from single-chirp measurements. Our approach, leveraging frequency-modulated continuous-wave optics combined with a frequency-decoding network and test-time optimization, effectively halves the capture time compared to the traditional multi-frequency measurement-based methods commonly used for absolute depth inference. Using all-optical GHz Time-of-Flight methods, our method *improves the range resolution from 12.5 meters, a limit set by the 20 MHz modulation bandwidth, to 4.2 centimeters*. As demonstrated in Figs. 6 and 7, our method showcases exceptional depth reconstruction capabilities in both synthetic and real-world scenarios, surpassing current state-of-the-art techniques.

While our approach is generally resilient across various material types, from matte to glossy surfaces, we acknowledge that the precision of our frequency decoding network can be affected by noise variations in the correlation signals of low reflectance materials. Objects with low reflectance may absorb more of the laser energy, resulting in weaker return signals and lower signal-to-noise ratio. This can lead to reduced precision and accuracy in depth detection for such objects [36, 37]. In the future, this limitation may be addressed by retraining the network with a broader range of material data, thereby enhancing its robustness. Additionally, the use of narrowband spectral filters can further refine precision, particularly in environments with strong ambient light.

Looking ahead, we see the proposed method as a building block for diverse computational imaging challenges, including non-line-of-sight imaging, single-shot ultrafast optical imaging, and single-photon ToF imaging. Beyond indoor imaging, our method also holds the potential for large-scale applications such as autonomous driving and wireless radio systems communications, underscoring its versatility and potential impact across various fields.

Acknowledgment. This work was supported by the AI- SEE project with funding from the FFG, BMBF, and NRC-IRA. Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Disney Research Award, a Sony Young Faculty Award, a Project X Innovation Award, and an Amazon Science Research Award.

Disclosures. The authors declare no conflicts of interest.

Supplemental Information and Data. See Supplement 1 for supporting content.

Data availability. Data underlying the results presented in this paper are available in Ref. [35].

References

1. J. Deng, W. Dong, R. Socher, *et al.*, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition,* (Ieee, 2009), pp. 248–255.
2. Y. Xiao, F. Codevilla, A. Gurram, *et al.*, "Multimodal end-to-end autonomous driving," IEEE Trans. on Intell. Transp. Syst. **23**, 537–547 (2022).
3. K. Shin, D. Kim, H. Park, *et al.*, "Artificial tactile sensor with pin-type module for depth profile and surface topography detection," IEEE Trans. on Ind. Electron. **67**, 637–646 (2020).
4. G. Yahav, G. J. Iddan, and D. Mandelboum, "3d imaging camera for gaming application," in *2007 Digest of Technical Papers International Conference on Consumer Electronics,* (2007), pp. 1–2.
5. K. H. Sing and W. Xie, "Garden: A mixed reality experience combining virtual reality and 3d reconstruction," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems,* (Association for Computing Machinery, New York, NY, USA, 2016), CHI EA '16, p. 180–183.

6. Q. Hao, K. Zhou, J. Yang, *et al.*, "High signal-to-noise ratio reconstruction of low bit-depth optical coherence tomography using deep learning," J. Biomed. Opt. **25**, 123702 (2020).
7. L.-K. Liu, S. H. Chan, and T. Q. Nguyen, "Depth reconstruction from sparse samples: Representation, algorithm, and sampling," IEEE Trans. on Image Process. **24**, 1983–1996 (2015).
8. I. Vornicu, R. Carmona-Galán, and A. Rodríguez-Vázquez, "Photon counting and direct tof camera prototype based on cmos spads," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS),* (2017), pp. 1–4.
9. B. Schwarz, "Lidar: Mapping the world in 3D," Nat. Photonics **4**, 429–430 (2010).
10. I. Gyongy, N. A. W. Dutton, and R. K. Henderson, "Direct time-of-flight single-photon imaging," IEEE Trans. on Electron Devices **69**, 2794–2805 (2022).
11. D. Bronzi, Y. Zou, F. A. Villa, *et al.*, "Automotive three-dimensional vision through a single-photon counting SPAD camera," IEEE Trans. Intell. Transp. Syst. **17**, 782–795 (2016).
12. F. Heide, S. Diamond, D. Lindell, and G. Wetzstein, "Sub-picosecond photon-efficient 3d imaging using single-photon sensors," Sci. Reports **8** (2018).
13. S.-H. Baek, N. Walsh, I. Chugunov, *et al.*, "Centimeter-wave free-space neural time-of-flight imaging," ACM Trans. on Graph. (TOG) (2022).
14. M. A. Herráez, D. R. Burton, M. J. Lalor, and M. A. Gdeisat, "Fast two-dimensional phase-unwrapping algorithm based on sorting by reliability following a noncontinuous path," Appl. Opt. **41**, 7437–7444 (2002).
15. M. Gupta, S. K. Nayar, M. B. Hullin, and J. Martin, "Phasor imaging: A generalization of correlation-based time-of-flight imaging," ACM Trans. Graph. **34** (2015).
16. M. Tölgyessy, M. Dekan, t. Chovanec, and P. Hubinský, "Evaluation of the azure kinect and its comparison to kinect v1 and kinect v2," Sensors **21** (2021).
17. S. Achar, J. R. Bartels, W. L. R. Whittaker, *et al.*, "Epipolar time-of-flight imaging," ACM Trans. Graph. **36** (2017).
18. N. Naik, A. Kadambi, C. Rhemann, *et al.*, "A light transport model for mitigating multipath interference in time-of-flight sensors," (2015), pp. 73–81.
19. D. Jimenez, D. Pizarro, M. Mazo, and S. Palazuelos-Cagigas, "Modeling and correction of multipath interference in time of flight cameras," Image Vis. Comput. **32**, 1–13 (2014).
20. F. Heide, L. Xiao, A. Kolb, *et al.*, "Imaging in scattering media using correlation image sensors and sparse convolutional coding," Opt. Express **22**, 26338–26350 (2014).
21. A. Kadambi, H. Zhao, B. Shi, and R. Raskar, "Occluded imaging with time-of-flight sensors," ACM Trans. Graph. **35** (2016).
22. S. Su, F. Heide, R. Swanson, *et al.*, "Material classification using raw time-of-flight measurements," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* (2016).
23. R. Lange and P. Seitz, "Solid-state time-of-flight range camera," Quantum Electron. IEEE J. **37**, 390 – 397 (2001).
24. M. Gupta, A. Velten, S. Nayar, and E. Breitbach, "What are optimal coding functions for time-of-flight imaging?" ACM Trans. on Graph. **37**, 1–18 (2018).
25. F. Li, F. Willomitzer, P. Rangarajan, *et al.*, "Sh-tof: Micro resolution time-of-flight imaging with superheterodyne interferometry," (2018).
26. J. Bioucas-Dias, V. Katkovnik, J. Astola, and K. Egiazarian, "Absolute phase estimation: adaptive local denoising and global unwrapping," Appl. optics **47**, 5358–69 (2008).
27. J. M. Bioucas-Dias and G. Valadao, "Phase unwrapping via graph cuts," IEEE Trans. on Image Process. **16**, 698–709 (2007).
28. R. Crabb and R. Manduchi, "Fast single-frequency time-of-flight range imaging," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops,* (2015).
29. X. Dun, H. Ikoma, G. Wetzstein, *et al.*, "Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging," Optica **7**, 913–922 (2020).
30. *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 18-22, 2010, Taipei, Taiwan* (IEEE, 2010).
31. J. Bioucas-Dias, V. Katkovnik, J. Astola, and K. Egiazarian, "Multi-frequency phase unwrapping from noisy data: Adaptive local maximum likelihood approach." (2009), pp. 310–320.
32. F. Järemo Lawin, P.-E. Forssén, and H. Ovrén, "Efficient multi-frequency phase unwrapping using kernel density estimation," in *Computer Vision – ECCV 2016,* B. Leibe, J. Matas, N. Sebe, and M. Welling, eds. (Springer International Publishing, Cham, 2016), pp. 170–185.
33. A. Yariv and P. Yeh, *Photonics: Optical Electronics in Modern Communications* (Oxford University Press, 2007).
34. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," (2017).
35. M. Roberts, J. Ramapuram, A. Ranjan, *et al.*, "Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding," in *ICCV,* (2021).
36. N. Csanyi and C. Toth, "Improvement of lidar data accuracy using lidar-specific ground targets," Photogramm. Eng. Remote. Sens. **73**, 385–396 (2007).
37. H.-G. Maas, "On the use of pulse reflectance data for laserscanner strip adjustment," INTERNATIONAL ARCHIVES OF PHOTOGRAMMETRY REMOTE SENSING AND SPATIAL INFORMATION SCIENCES **34**, 53–56 (2001).