Neural Single-Shot GHz FMCW Correlation Imaging (Supplementary Information)

This supplemental document provides additional information to support the findings in the main manuscript. Specifically, we discuss details of the supplementary code, experimental setup, network details, and additional experimental results.

1. CODE AND DATA

In addition to this supplementary document, we have provided code and data as supplementary files. We refer to the readme documents for for installation and details of the supplementary code.

2. DETAILED DESCRIPTION OF EXPERIMENTAL SETUP

In this section, we discuss the implementation details of the proposed experimental setup displayed in Fig. S1. We follow Baek et al.[1] to implement an all-optical correlation setup, that we, in our work, use for frequency modulation.

A. Illumination Module

In the illumination Module, a single transverse mode continuous wave laser (Laser Quantum Gem 532) at a wavelength of 532 nm with power of 3mW is coupled with a custom-design single mode high power optical fiber (OZ Optics QPMJ-A3AHPCA3AHPC-488-3.5/125-3AS-1-1), which removes the higher order modal light and produces a uniform Gaussian beam at the output of the fiber, maintaining the 20 to 30 percent laser output power. The light then enters an 2.5× inverse beam expander consists of a plano-convex lens and a plano-concave lens (Thorlabs LC1060-A and LA1608-A) that reduces the beam diameter from 1.25 mm down to 0.5 mm to be matched with the desired beam size to our EOM. The reduced light becomes horizontally linearly polarized by passing through the first polarizing beam-splitter (PBS, Thorlabs PBS101). Then, a pair of Half wave plate (HWP, Thorlabs WPMH05M-532) and quarter wave plate (QWP, Thorlabs WPQ05M-532) modulates the polarization state of the beam. The polarization-modulated light passes through the electro-optic modulators (EOM) that operates at the modulation frequency chirp with bandwidth of $B = 20$ MHz and chirp length of $T = 32.5$ μs.

The modulator required a custom resonant crystal which was fabricated by Qubig GmbH. The light is reflected by a mirror (Thorlabs PF10-03-P01), returning to the EOM, the QWP, the HWP, and the PBS. This procedure results in the GHz frequency modulation of light. The light then passes through a mirror (Thorlabs PF10-03-P01) and a non-polarizing beam-splitter (NBS, Thorlabs CCM1-BS013) dividing the incident beam into two beams of equal intensity. One beam is directed to the integrating sphere (Thorlabs S140C) followed by the reference photodiode (Thorlabs APD440A), which measure the intensity of emitted light for a as a reference signal. The purpose of this module is to calibrate intensity fluctuations from the laser by normalizing the signal incident on the detection module. The optical intensity modulation has higher frequency than the integration time of a few milliseconds, which allows compensation after the modulation without error. The other half of the beam is passed through another NBS (Thorlabs CCM1-BS013) and sent to a scene through a 2-axis galvo mirror system (Thorlabs GVS012) for spatial scanning.

B. Detection Module

After the frequency modulated light returns from a scene, it passes through the galvo mirror system and the mirror followed by a NBS which redirects the beam to the detection module. We use an 1.6× inverse beam expander, which consists of a plano-convex lens (Thorlabs LA1213-A) and a plano-concave lens (Thorlabs LC1060-A), resulting in a beam diameter of 0.5 mm. The collimated beam is then reflected off of a mirror (Thorlabs PF10-03-P01) and enters the detection EOM. Symmetric to the emission module, we mount a PBS, a HWP, a QWP, an EOM, and a mirror that constitute an optical demodulation of returned light from a scene. The frequency demodulated light is then captured by an avalanche photodiode (Thorlabs APD440A) with

a focusing lens (Thorlabs LA1951-A). We use a 1 MHz lowpass filter to filter out the higher frequency components of the demodulated signal and pass through the lower beat note $(\omega_r - \omega_p)$ as indicated in Eq. 3 and Eq. 4 in the main paper. The final correlation signal is then captured by an oscilloscope (R&S RTO6), which we pass to the frequency decoding network to recover the frequency that corresponds to each pixel depth in the scene.

In order to operate the EOMs with a frequency chirped sinusoidal voltage input, we use two custom RF drivers with a high-frequency DDS which are synchronized with an external clock source. We use a function generator (Siglent SDG2042X) to produce an external 10 MHz square

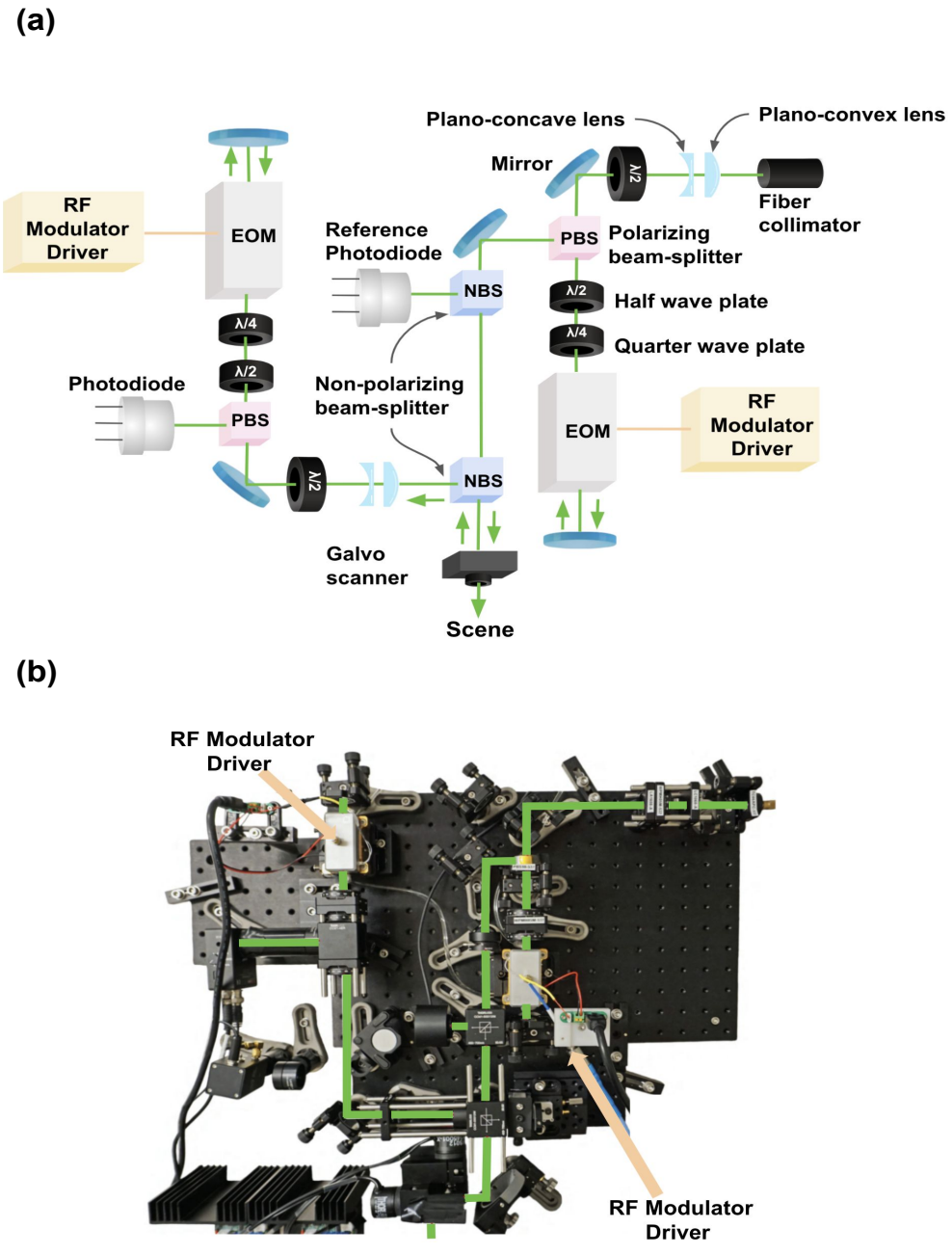

Fig. S1. (a) Schemetic illustration of our proposed all-optical FMCW prototype based on polarizing optics and EOMs. (b) Experimental setup photo with light paths highlighted in green. The EOMs generate GHz amplitude modulation, where frequency modulation is applied through an RF generator. See text for more details.

wave reference clock, which enables accurate controls the phase of the modulation signal ϕ. Our driver contains two RF modulators to output an RF signal provided to the EOMs. The RF driver performs frequency locking of an RF output signal to significantly increase the output power and reduce the frequency drifting in the EOM. We use an external RF generator (R&S SMW) for the FMCW operation to produce the RF chirped signals, which are input to the Qubig RF driver. The chirped signals have to be set to the correct chirp length and bandwidth so the EOMs can remained frequency locked, and not lose thermal equilibrium.

C. Electro-Optic Modulators

In the design of our custom resonant standing wave EOM, a housing incorporating an electro-optic crystalline material establishes a resonant cavity generating standing waves at a GHz modulating frequency by altering the polarization of light along its optical axis based on the applied voltage V. As indicated by the Jones matrix, this is accomplished by modifying the phase relationship between the perpendicular polarization components of light,

$$B(V) = \begin{bmatrix} e^{\frac{-i\Gamma(V)}{2}} & 0 \\ 0 & e^{\frac{i\Gamma(V)}{2}} \end{bmatrix}, \tag{S1}$$

where $\Gamma(V)$ is the voltage-dependent net birefringence:

$$\Gamma(V) = \frac{2/\pi}{\lambda} \Delta n^3 r V = \frac{\pi V}{V_\pi}. \tag{S2}$$

In this case, Δn is the birefringence, the difference between the ordinary and extaordinary refractive indices of the uniaxial crystal, and r is the element corresponding to index [3, 3] of the electro-optic tensor for the Lithium Niobate crystal, and V_π is the half wave voltage.

Employing a resonant circuit ensures precise impedance matching and reduces the necessary drive voltage. The LC tank configuration, a fundamental type of resonant circuit in EOMs, integrates a low-loss inductor and modulator crystal to create a series resonant circuit, acting as a small resistor at resonance. The resonant circuit's energy storage properties, combined with impedance matching and low-loss components, yield a capacitor voltage ten times greater than the input voltage. While effective in reducing half-wave voltages, the modulator's resonator is power-limited, limiting its modulation bandwidth to approximately 20 MHz when sampling a 7.15 GHz frequency.

3. FREQUENCY DECODING NETWORK ARCHITECTURE

In this section, we delve into the implementation details of our frequency decoding network, designed to estimate distances based on input signal arrays. We utilize a neural network, specifically a Multi-Layer Perceptron (MLP), with an 8-layer architecture, each containing 1024 neurons using the linear and softsign activation functions, which introduce non-linearity in hidden layers to capture more complex patterns in the data while still preserving some linear behavior [2]. Our primary aim is to predict the absolute depth d_p associated with a given pixel p based on its input signal array s_p. Our neural network model takes this input correlation signal array denoted as $s_p := \{s_p^1, s_p^2, ..., s_p^N\}$, where N denotes the length of the input signal array.

For training data, we acquired 7 sets of correlation signal data from a depth range of 0 mm to 1500 mm with 1 mm spacing using our measurement setup with a piezoelectric motion stage, which can travel along its axis in small increments and has a theoretical resolution of 50 nm. Prior to feeding the training signal array into the network, we employ a pivotal preprocessing phase. Here, we undergo the transformation of the raw time domain signal array (s_p) into the frequency domain, accomplished via the Fast Fourier Transform (FFT). We denote the transformed input as $s'_p = FFT(s_p)$, which is a signal array of the same length as s_p but in the frequency domain. That is

$$s'^k_p = \sum_{n=1}^{N} s_p^n \times e^{\frac{-2i\pi}{N} kn}. \tag{S3}$$

The primary role of this transformation is to capture the frequency-related characteristics within the input data. This preprocessing step assumes paramount importance as it equips the neural network with the capability to effectively glean insights from the frequency information embedded within the input array, thereby significantly bolstering the precision of depth inference. The detailed network architecture is reported below in Table S1.

For training, we employ the Adam optimizer and train the model for 5000 epochs with a batch size of 16 and a 10^{-4} learning rate. The model converges within 8 minutes in NVIDIA A100 GPU. We compute the discrepancy between the predicted depth and the ground truth and utilize the ℓ_1 loss function $\mathcal{L}_{FDN}$,

$$\mathcal{L}_{FDN}(s_p, d_p) = ||FDN(FFT(s_p)) - d_p||_1. \tag{S4}$$

By following these steps and utilizing the proposed network architecture and training setup, we have developed a robust frequency decoding network capable of accurately inferring depth based on the correlation signal array.

Table S1. Architecture description of the FDN (Frequency decoding network).

Layer	Type	In-Channel	Out-Channel
1st	Linear+Softsign	N	1024
2nd	Linear+Softsign	1024	1024
3rd	Linear+Softsign	1024	1024
4th	Linear+Softsign	1024	1024
5th	Linear+Softsign	1024	1024
6th	Linear+Softsign	1024	1024
7th	Linear+Softsign	1024	1024
8th	Linear+Softsign	1024	1

4. SIGNAL-TO-NOISE RATIO AND DEPTH DECODING ACCURACY

For indoor measurements, the signal-to-noise ratio (SNR) varies depending on the presence of environmental noise such as ambient lights. This section presents an analysis of varying SNR and its effect on the accuracy of FDN's wrap count prediction to validate the robustness of our system in the possible presence of a greater degree of noises. Fig. S2 presents the test performance of the Frequency-Decoding Network on signals with varying degrees of Gaussian noises. A quantitative comparison of wrap count accuracy in terms of mean Square Error (MSE) for the original signals with varying levels of Gaussian noises is presented in Tab. S2. Our network demonstrates great robustness in terms of wrap count prediction even with the presence of varying degrees of Gaussian noises. We note that in our experiments, the SNR for signals measured for the scenes appears to be consistent across every pixel. That is, the change in signal-to-noise ratio for varying spatial distances in an indoor environment is negligible.

5. ADDITIONAL RESULTS

In this section, we report additional qualitative reconstruction results on challenging scenes on Hypersim [3] RGB-D indoor scenes for the three discussed comparison methods: state-of-the-art single-frequency method PUMA [4], the kernel density based multi-frequency method KDE [5], the double-frequency neural phase unwrapping method Baek 2022 [1], and our proposed method. Qualitatively, these results further validate that with frequency decoding network, our method demonstrates superior performance in reconstruction of absolute depths with the presence of intricate geometric structures while maintaining spatial consistency of smooth surfaces with the aid of test-time optimization.

	Original	30 dB	15 dB	10 dB	5 dB
MSE	0.86	0.87	0.89	1.12	1.5

Table S2. Quantitative comparison of wrap count accuracy for the original signals with varying levels of Gaussian noises added in terms of mean Square Error (MSE). Our network demonstrates great robustness in terms of wrap count prediction even with the presence of varying degrees of Gaussian noises.

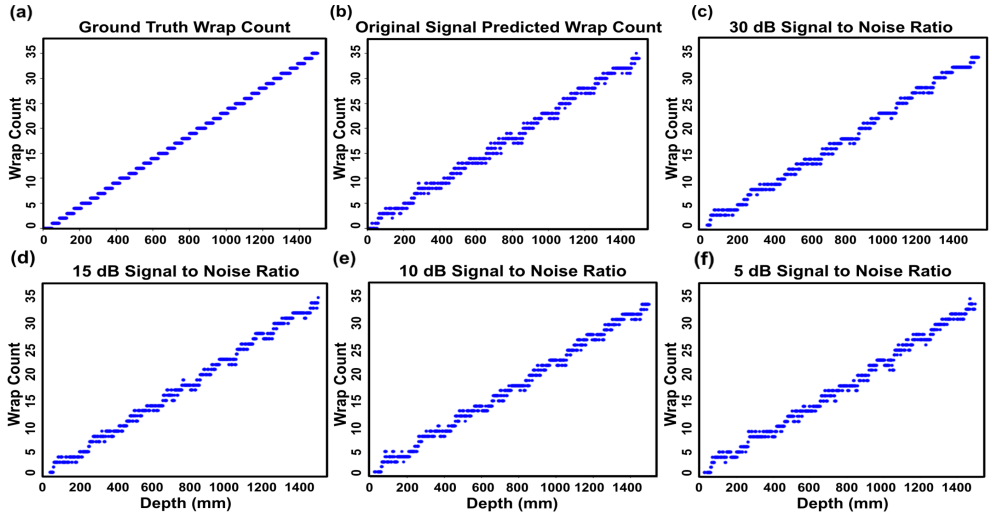

Fig. S2. Test Performance of the Frequency-Decoding Network on signals with varying degrees of Gaussian noises. (a) Displays the ground truth wrap count, and (b) shows the predicted wrap count over a depth range of 0 mm to 1500 mm using the frequency decoding network. (c) (d) (e) (f) display the predicted wrap count over the same depth range with an additional 30 dB, 15 dB, 10 dB, and 5 dB of noises with respect to the original signal.

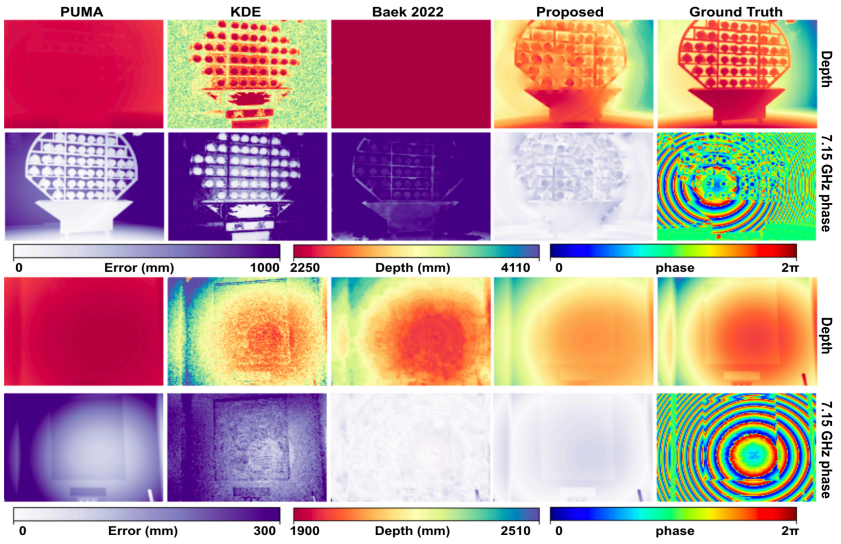

Fig. S3. The reconstruction results of Hypersim [3] RGB-D indoor scenes for the three discussed comparison methods: state-of-the-art single-frequency method PUMA [4], the kernel density based multi-frequency method KDE [5], the double-frequency neural phase unwrapping method Baek 2022 [1], and our proposed method.

REFERENCES

1. S.-H. Baek, N. Walsh, I. Chugunov, *et al.*, "Centimeter-wave free-space neural time-of-flight imaging," ACM Transactions on Graph. (TOG) (2022).

2. T. Szandała, *Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks* (2021), pp. 203–224.
3. M. Roberts, J. Ramapuram, A. Ranjan, *et al.*, "Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding," in *ICCV,* (2021).
4. J. M. Bioucas-Dias and G. Valadao, "Phase unwrapping via graph cuts," IEEE Transactions on Image Process. **16**, 698–709 (2007).
5. 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.