MAROON: A Dataset for the Joint Characterization of Near-Field High-Resolution Radio-Frequency and Optical Depth Imaging Techniques

VANESSA WIRTH, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
JOHANNA BRÄUNIG, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
NIKOLAI HOFMANN, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
MARTIN VOSSIEK, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
TIM WEYRICH*, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany and University College London, UK
MARC STAMMINGER*, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

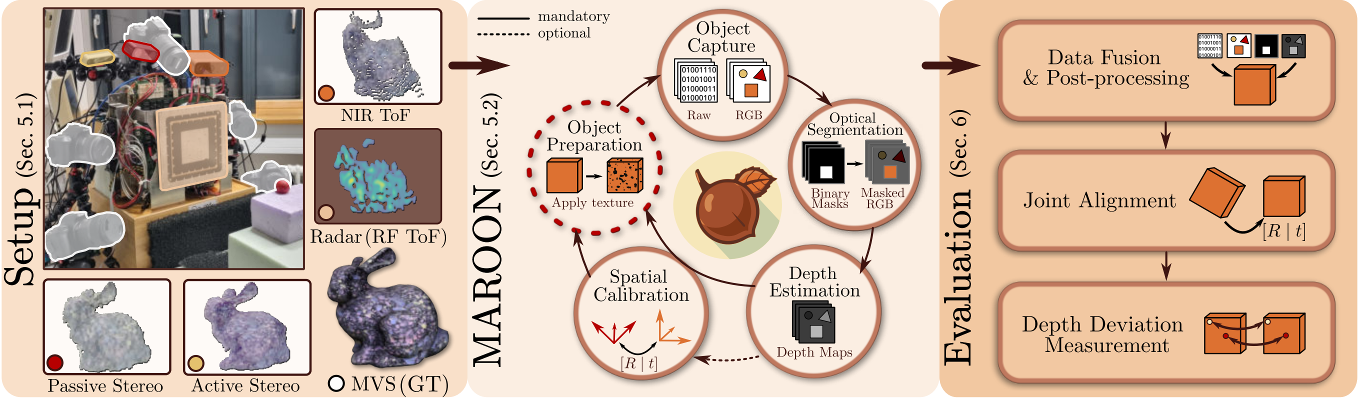


Fig. 1. Recent developments for near-field imaging radars enabled the acquisition of high-resolution depth images, and the sensors are now increasingly gaining attention as complementary modalities to optical depth sensing. Direct comparisons from our MAROON dataset, however, highlight significant differences between radar and optical reconstructions. This work employs the collected multimodal data of four depth imagers, depicted on the *left*, to systematically characterize these fundamental differences together with sensor-specific findings in a joint evaluation framework.

Utilizing the complementary strengths of wavelength-specific range or depth sensors is crucial for robust computer-assisted tasks such as autonomous driving. Despite this, there is still little research done at the intersection of optical depth sensors and radars operating close range, where the target is decimeters away from the sensors. Together with a growing interest in high-resolution imaging radars operating in the near field, the question arises how these sensors behave in comparison to their traditional optical counterparts. In this work, we take on the unique challenge of jointly characterizing depth imagers from both, the optical and radio-frequency domain using a multimodal spatial calibration. We collect data from four depth imagers, with three optical sensors of varying operation principle and an imaging radar. We provide a comprehensive evaluation of their depth measurements with respect to distinct object materials, geometries, and object-to-sensor distances. Specifically, we reveal scattering effects of partially transmissive materials and investigate the response of radio-frequency signals. All object measurements will be made public in form of a multimodal dataset, called MAROON.

Additional Key Words and Phrases: Time-of-Flight, Radar Imaging, Radio Frequency, mmWave Imaging, MIMO Radar, Depth Camera, Spatial Calibration, Multimodal Sensor Fusion, Dataset

1 Introduction

Real-world computer-assisted tasks, for instance in robotics and tracking applications, frequently require the immediate assessment of spatial information to accurately reason about the environment at a specific point in time, which has led to the development of several single-view range and depth sensors. For autonomous driving, it has been shown that utilizing multimodal depth sensing techniques from both the optical (lidar) and radio-frequency (radar) domain can lead to superior performance and robustness in computer-assisted tasks [Velasco-Hernandez et al. 2020]. Due to its environment, the autonomous driving industry has traditionally concentrated on far-field range sensing, with an unambiguous range of several meters and beyond. As recent high-resolution radio-frequency technologies utilize the concept of *radar imaging* to produce 3D information in form of a depth map — similar to optical depth or RGB-D cameras — they also become more popular in close range, where the target of interest is up to a few decimeters away from the sensor; however, a comprehensive and detailed characterization of these radar imaging technologies, which frequently operate in the radar's near field, is yet to be realized.

As part of this work, we devised a dataset **MAROON** (**M**ultimodal **A**ligned **R**adio and **O**ptical frequency **O**bject Reconstructions in

[*] Both authors contributed equally to this research.

the **N**ear Field) (cf. Section 5) that enables studying of different sensor modalities in direct comparison. As is immediately visible in Figure 1 (*left*), the reconstructions of near-field imaging radars appear fundamentally different in comparison to their well-researched counterparts in the optical domain.

A key advantage of radar is that it is insensitive to environmental light and can penetrate, for instance, fabric and dust. Following the success of Google's project Soli [Lien et al. 2016] for gesture sensing, radars were utilized in close range for the detection of vital signs [Vilesov et al. 2022], activity recognition [Braeunig et al. 2023], people tracking [Zewge et al. 2019] and human body reconstruction [Chen et al. 2023, 2022]. With the growing trend towards larger antenna apertures to achieve high-resolution imaging [Chen et al. 2023; Schwarz et al. 2022], radars will more frequently operate in the near field, as determined by the Fraunhofer boundary condition [Selvan and Janaswamy 2017]. At the same time, characteristics of near-field radar are generally under-researched.

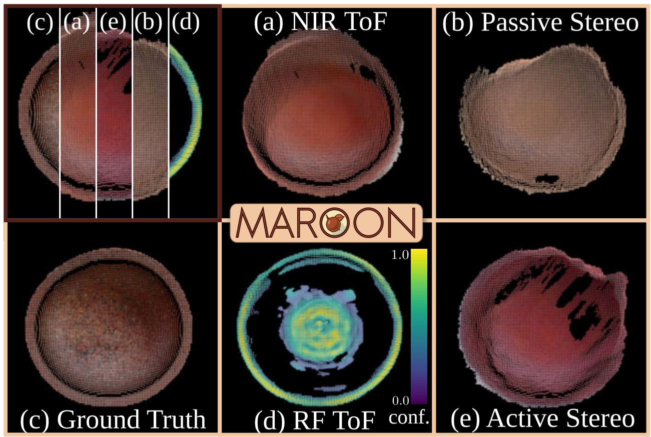


Fig. 2. Example data of the *Plunger* object from the MAROON dataset. In the *upper left*, all reconstructions are spatially aligned with respect to the RF ToF coordinate system. The RF ToF colorscale encodes the normalized reconstruction confidence (cf. Section 4.2.2).

Drawing on prior research about wavelength-specific strengths and weaknesses, this paper addresses the unique challenge of characterizing various optical depth-imaging techniques alongside a high-resolution multiple-input multiple-output (MIMO) imaging radar in the near field. The latter is interchangeably referred to as a radio-frequency (RF) Time-of-Flight (ToF) depth imager. To this end, we mutually calibrated sensors of four different depth sensing technologies, that is active and passive stereo, near-infrared (NIR) amplitude-modulated continuous wave (AMCW) ToF, and RF frequency-stepped continuous wave (FSCW) ToF in the millimeter-wave range.

There is a notable lack of multimodal datasets suitable for close-range applications, and, to our knowledge, this work is the first to incorporate imaging radars in this research area. With this in mind, we captured the MAROON dataset of various household objects and construction materials, of which example data is shown in Figure 2. Utilizing a high-resolution MIMO imaging radar, with a spatial resolution currently far beyond prevalent RF imaging sensors, we captured this dataset with a multitude of key objectives:

(1) To evaluate sensor-specific reconstructions, considering various object materials, geometries, and distances to the sensors.
(2) To establish a public data base for multimodal reconstruction research in close-range applications, bridging the radio-frequency and optical domains.
(3) To characterize the under-researched effects of millimeter waves in near-field imaging radars, e.g. object materials in the radio-frequency domain, akin to studies on bi-directional reflectance distribution functions (BRDFs) in optics.
(4) To improve radio-frequency signal simulations by supplying data for lower-resolution radar architectures and comparing synthetic signals with real measurements and a ground truth.

Together with the dataset, we developed a joint sensor evaluation framework that measures reconstruction differences between sensors and a ground truth using different metrics, providing supplementary visualizations tailored to identify sensor-specific trends across multiple objects. By analyzing these trends, we identified ToF scattering effects in partially transmissive materials and examined RF ToF reconstructions, which are typically less complete than those from optical sensors.

Moreover, we utilize the multimodal data of MAROON in two example applications: first, we show that the dataset serves as foundation for addressing inverse rendering problems in the radio-frequency domain. Taking up on concurrent work [Hofmann et al. 2025], we determine object-specific material properties, which are crucial in high-fidelity radar simulation Second, the variety of challenging objects in our dataset serves as a benchmark for developing novel multimodal reconstruction algorithms, as demonstrated in [Wirth et al. 2025]. We extend this benchmark by additional experiments, varying the sensor configuration.
To summarize, our contributions include:

- A novel multimodal dataset, MAROON, comprising common objects in the near field, captured using a jointly calibrated setup of three optical depth sensors, a high-resolution imaging radar, and high-quality multi-view reconstruction for ground-truth geometry. We are releasing this dataset alongside the raw radar measurements to facilitate exploration of various signal reconstruction and filtering techniques.
- A detailed analysis of trends and sensor-specific effects emerging from that dataset. This includes aspects of different object materials, geometries, and distances to the sensors, signal response and reconstruction quality of imaging radars, as well as ToF scattering effects of partially transmissive materials.
- Two applications of the dataset: inverse rendering for material characterization and high-speed multimodal reconstruction.

2 Related Work

While a considerable amount of literature exists on optical and RF depth sensors in isolation, no directly related work on the joint characterization of these two domains has been identified. Instead, the first two sections comprise an overview of existing research about sensor characteristics self-contained within a single frequency domain. We further address the sensor fusion of optical and RF sensors, since in that research direction the complementary strengths of the sensors are utilized as well.

2.1 Optical Depth Sensing

Depth cameras have been characterized with respect to a number of different aspects, and related work can be broadly classified into three categories: the sensor technologies, the capture environments, and the methods of comparison used to evaluate their performance.

Considering the sensor technologies, metrological research has been conducted in terms of optical ToF [Xiong et al. 2017; Zanuttigh et al. 2016] and active stereo [Giancola et al. 2018; Wang and Shih 2021]. Furthermore, working principles of passive stereo sensors have been widely addressed in computer vision algorithms [Szeliski 2022]. Similar to our work, Chiu et al. [2019] and Halmetschlager-Funek et al. [2019] jointly characterize ToF and active stereo.

With respect to the capture environments, related work examined the effects of object material [Giancola et al. 2018; Halmetschlager-Funek et al. 2019; Hansard et al. 2012; Xiong et al. 2017], color [Giancola et al. 2018; Hansard et al. 2012; Xiong et al. 2017], texture [Hansard et al. 2012; Xiong et al. 2017] and distance to objects [Halmetschlager-Funek et al. 2019]. Furthermore, environmental lighting conditions [Halmetschlager-Funek et al. 2019] and multi-path effects [Giancola et al. 2018] were investigated. Specifically for ToF sensors, Wu et al. [2012] analyze multi-path effects originating from subsurface scattering and interreflections.

Moreover, we discuss related frameworks for jointly characterizing sensors. Halmetschlager-Funek et al. [2019] compare individually estimated depth values against manual measurements. Chiu et al. [2019] and Giancola et al. [2018] align the 3D data captured from sensors with real or synthetic ground-truth data, respectively. Most similar to ours, Hansard et al. [2012] analyze ToF and structured light sensors using a spatial calibration and investigated object material, color, geometry, and texture using ground-truth data obtained from a structured light scanner.

2.2 Radio Wave Propagation and Range Sensing

So far, radio-frequency depth sensors (radar) were characterized in isolation. A more fundamental research direction examines the propagation of electromagnetic waves, which is the basis for RF ToF sensors. The general RF propagation behavior under different materials and geometries is measured by a parameter known as the radar cross section (RCS) [Knott et al. 2004]. The RCS approximates the returned ratio of a transmitted radio signal and was measured in relation to a variety of materials [Knott et al. 2004; Semkin et al. 2020], as well as in the context of humans [Deep et al. 2020; Marchetti et al. 2018]. Orthogonal research of Zhadobov et al. [2011] investigates the interaction of radio waves and human skin with respect to electromagnetic, thermal and biological aspects.

Moreover, studies of individual radar technologies have been conducted. Čopič Pucihar et al. [2022] evaluate the recognition of hand gestures using millimeter-wave radars in the presence of various materials. Wei et al. [2021] characterize imaging radars with respect to the geometry of metal objects in the context of security scanning. Furthermore, Bhutani et al. [2022] examine millimeter-wave radars at different frequencies, whereas Jha et al. [2018] analyze differences in their radiation between the near field and the far field. Sun et al. [2020] provide an overview of MIMO radars for autonomous driving, together with the characterization of their wave forms. Lastly, Ahmed [2021] presents millimeter-wave MIMO radar imaging systems in the context of security screening. To the best of our knowledge, no comprehensive characterization in conjunction with optical technologies has been done so far. Additionally, the existing efforts have been limited in scope with regard to RF depth sensing in the near field.

2.3 Fusion of RF and Optical Sensors in Close Range

Knowledge about complementary strengths is important for both, sensor characterization and sensor fusion. While significant research efforts have been devoted to the field of autonomous driving — where radar sensors primarily operate in their respective far field — research on multimodal sensor fusion in close range is very limited and mostly focused on capturing humans.

Zewge et al. [2019] perform people tracking with a 4×3 MIMO radar and an active stereo camera. Similarly, Lee et al. [2023] propose a method for human pose estimation, which utilizes the data acquired from two 4×3 MIMO radars synchronized with a monocular RGB camera. Both works do not utilize radar imaging methods due to the limited resolution. More similar to ours, Chen et al. [2023] use a high resolution 48×48 MIMO radar and an RGB camera for human body reconstruction.

Furthermore, we address related datasets. Lim et al. [2021] introduce RaDICaL, an indoor and outdoor dataset of multiple people and objects, captured with a 4×3 MIMO imaging radar and an active stereo camera. In the context of human body reconstruction, Chen et al. [2022] propose the mmBody benchmark that was captured with a 48×48 MIMO radar and an RGB camera.

3 Preliminaries

As different research communities partially differ in their terminology, this paper pursues a unified terminology, summarized in the table below and used in the remainder of this paper.

Depth Imager. A sensor that, directly or indirectly, captures a depth image $\boldsymbol{D}$ of resolution $W \times H$, where each pixel (u, v) contains a depth value d measured along the axis perpendicular to the image plane. The depth may be indirectly measured from range and pixel position. We show the difference between range and depth in Figure 3.

Transmitter and Receiver. Optical receivers are small cells of image sensors, with a direct mapping to pixels. Transmitters are commonly LEDs or projectors. RF sensors have transmitting (TX) and receiving (RX) antennas.

Sensor. Describes all physical parts required for depth sensing and their spatial arrangement. Optical sensors typically consist of one or two cameras. Active sensors contain an additional illumination unit. RF sensors usually have one or more antenna arrays in different arrangements.

Depth Image Resolution W × H. The number of depth samples computed from the incoming signal. In cameras, the depth samples are directly computed for each pixel, i.e., each receiver. MIMO imaging radars apply signal post-processing to compute depth from the signal diversity at different receiver positions. Hence, the image resolution is not directly affected by the number of receivers but by the signal processing parameters such as the voxel density (cf. Section 4.2.2). While, in theory, the depth image resolution can be indefinitely high, in practice it is limited by the spatial resolution.

Spatial Resolution δ. The term resolution has several definitions. Here, we explicitly refer to spatial resolution as the minimum distance between two points in space that can be resolved from the received signal. Lower spatial resolution means higher minimum resolvable distance, so more incorrect measurements are made when points become inseparable, as seen in Figure 3. Spatial resolution is a theoretical measure, and external factors such as the sensor design can affect its *effective* resolution. Literature about traditional multi-static RF ToF sensors divides spatial resolution into range resolution (δ_r), and cross-range resolution along the horizontal (δ_h) and vertical (δ_v) axes, respectively. While δ_h and δ_v are measured at a specific range, a more general formulation is the angular resolution, often referred to as azimuth (ω_h) and elevation (ω_v) resolution [Willis and Griffiths 2007]. Literature from the optical domain shares a similar definition, however, with a different terminology and refers to depth resolution [Zanuttigh et al. 2016] (δ_z) and pixel resolution (δ_h, δ_v) instead. Similar to traditional RF ToF sensors, optical sensors postulate that the target is situated in the far field, where depth is assumed to be approximately the same as range and, hence, $\delta_z \approx \delta_r$. We note that this assumption is not accurate in the near field. Due to the concept of an expanded antenna aperture with multiple transmitters and receivers, the resolution of near-field MIMO imaging radars is defined with respect to the three orthogonal axes x, y, z. In these, z refers to the depth axis and x, y are parallel to the antenna aperture. Contrary to their difference in definition, they share the same terminology as far-field RF ToF sensors such that δ_z is referred to as range resolution and $\delta_{x,y}$ is the cross-range [Ahmed 2014] or lateral [Ahmed 2021] resolution, respectively. We illustrate the respective definitions that are used for optical (δ_h, δ_v, δ_r) and RF sensors (δ_x, δ_y, δ_z) in Figure 3. For simplicity, we define spatial resolution only for the sensor center, where depth and range are approximately the same, such that $\delta_z \approx \delta_r$ and $\delta_{x,y} \approx \delta_{h,v}$.

4 Working Principles of Depth Imagers

In order to gain insight into the fundamental differences between optical and RF sensors, the first section characterizes wavelength-specific signal propagation. This is followed by an outline of the hardware design choices that are made for optical and RF depth imagers. After this, the working principles of the sensor technologies that are used in our experiments are discussed.

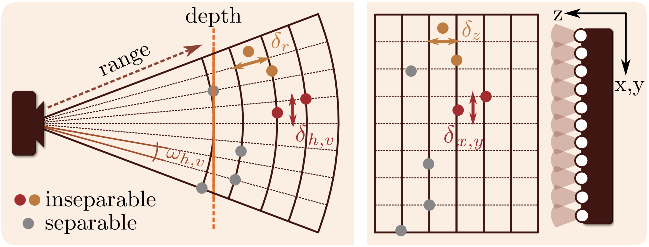


Fig. 3. Visualization of the effects caused by limited spatial resolution for multiple point targets. Optical sensors (*left*) have similar definitions as far-field RF ToF sensors and divide spatial resolution into depth, δ_r, and pixel resolution, $\delta_{h,v}$. Contrary to that, near-field imaging radars (*right*) refer to range δ_z and cross-range $\delta_{x,y}$ resolution. We assume $\delta_z \approx \delta_r$ and $\delta_{x,y} \approx \delta_{h,v}$ for the sensor center, yet emphasize the conceptual difference between range and depth.

4.1 Characterization of Wavelength

Depth imagers are susceptible to the received, and optionally transmitted, signal wavelength. The wavelength affects both, the interaction of the signal with matter and the design of the sensor hardware and depth sensing algorithms. In this section we elaborate on both aspects, with a particular focus on the near-infrared light spectrum and the millimeter-wave (mmWave) radio-frequency spectrum.

Signal Interaction. NIR signals have a wavelength in the nanometer range. Given their high energy and strong interaction with matter, signal reflection or absorption is common, with scattering and non-diffractive phenomena dominating across most materials. Indirect effects on interactions with matter, therefore, often play a subordinate role, such that short propagation paths can be expected. Moreover, NIR light is pervasive in the environment, rendering optical technologies susceptible to external interference.

As suggested by their name, the wavelength of mmWave signals is longer by comparison. The low energy and reduced interaction with matter result in lower absorption and reflection, while there is a higher chance of a signal being transmitted through material. Specifically, the penetration depth of millimeter waves through matter is dependent on material parameters, such as, the resistivity and permittivity. For instance, the signals of security scanners can penetrate fabric but are primarily reflected on contact with metal objects [Ahmed 2021]. Furthermore, diffraction is more common with millimeter waves. This allows waves to bend around objects. Due to the aforementioned phenomena, the propagation paths of signals from active RF sensors are typically longer than of signals from optical sensors. Lastly, mmWave depth imagers operate with reduced external interference, as there are few natural microwave sources in the environment.

Wavelength-specific Hardware. Due to the wavelength, the sensor design of RF sensors is inherently different from that of optical sensors. As stated by the general formulation of the Rayleigh criterion, the focus capacity and, hence, angular resolution ω of a sensor

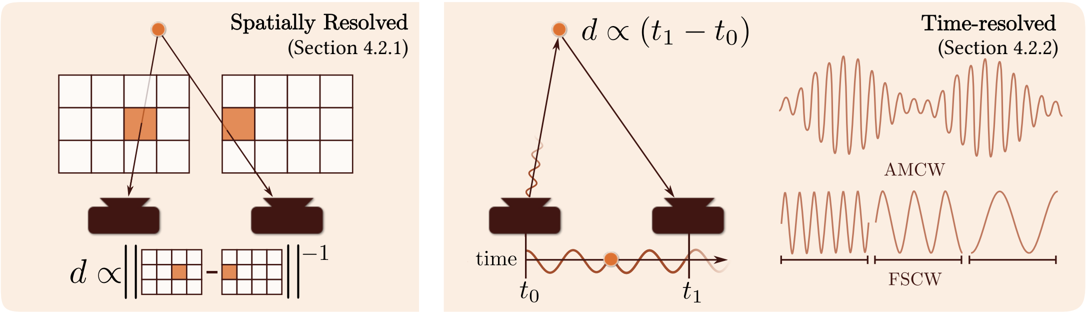


Fig. 4. Overview of the two depth sensing categories considered in this work. Spatially resolved methods compute the depth from disparity in the pixel positions. Time-resolved methods measure the depth through the round-trip propagation time of the received continuous wave (CW) signal. The types of wave forms utilized in our experiments are amplitude-modulated continuous wave (AMCW) and frequency-stepped continuous wave (FSCW).

is limited by the signal wavelength λ and the size of the sensor's aperture L [Hasch et al. 2012]:

$$\omega_{x,y} = 1.22 \frac{\lambda}{L_{x,y}} \; . \tag{1}$$

Optical sensors utilize camera lenses to refract the received signal, which enable a precise focus onto nanometer-sized pixels and, at the same time, exhibit a high angular resolution. In the context of the mmWave domain, a camera analogue can be conceptualized as a single-input multiple-output (SIMO) radar, i.e., a sensor comprising a single transmitter and multiple receivers. As indicated by Equation 1, mmWave sensors have a considerably lower angular resolution than optical sensors. Thus, high-resolution SIMO radars require comparably large antenna arrays with large lenses, which has proven to be impractical. Instead, high-resolution RF imaging sensors often are synthetic aperture radars (SAR), which use digital beamforming to focus. They utilize the angle diversity from distinct transmitter and receiver positions, which form a virtual aperture of size L, to increase the angular resolution [Bliss and Forsythe 2003] and require fewer antennas compared to SIMO systems. The majority of near-field SAR radars is implemented with MIMO arrays, that is, with multiple transmitters and receivers.

4.2 Depth Sensing Methods

In this section, we address the working principle of both optical and RF-based depth sensing methods used in our experiments. The content is organized in two categories: spatially resolved and time-resolved depth sensing, which are both depicted in Figure 4.

4.2.1 Spatially Resolved Depth Sensing. Spatially resolved depth imagers compute the point-wise depth from the respective pixel position in the image. In the following, we particularly address passive or active stereo(scopy) sensors.

Passive stereo sensors commonly utilize two cameras with a known relative spatial position to identify surface points in their respective images, a process known as *correspondence* or *stereo matching* [Szeliski 2022]. Given a correspondence pair of two pixels, the respective depth of this surface point is computed from their disparity. The quality of the correspondence matches affects the depth and accuracy of the results. Ambiguities in correspondence can arise due to textureless regions, poor lighting, motion or lens blur. Similarly, stereo matching can fail in terms of view-dependent effects or partial surface occlusions from one of the two receivers.

Active stereo sensors assist correspondence finding with an illumination unit that projects a pattern onto the target, usually in the NIR range, captured by the two cameras. The signal-multiplexed [Zanuttigh et al. 2016] pattern supports epipolar correspondence matching in addition to shading and texture cues, improving depth quality in textureless regions and low light. However, challenges include pattern distortions and signal oversaturation at the NIR receiver in bright conditions. Further details on spatially resolved depth sensors are provided in the supplementary material.

4.2.2 Time-Resolved Sensors (Time-of-Flight). Time-of-Flight is an active depth sensing method, in which depth is derived from the round-trip propagation time that it takes for a signal to be transmitted and received. The majority of ToF sensors utilized in the near field employ continuous wave (CW) signal modulations, which measure time based on the relative phase shift $\Delta\varphi$ between the transmitted and received signal. The depth is derived from the range r, which is measured by [Zanuttigh et al. 2016]:

$$r = \mathrm{c} \frac{\Delta\varphi}{4\pi f} \; . \tag{2}$$

The signal frequency is denoted as f, and c is the speed of light in vacuum, which closely matches that of light in air. For further details about the operating principle, we refer to the supplementary material. ToF technologies employ a simplified model for range sensing, which assumes that targets are weak scatterers [Ahmed 2014], with each signal reflecting directly from the first target. As a result, these technologies are sensitive to multi-path interference. In Section 7.2, we identify partially transmissive materials as a major cause of such interference. Furthermore, the unambiguous range, in which $\Delta\varphi$ can be correctly resolved, is limited to the periodicity of the sinusoidal CW signal. To extend this range, the carrier signal can be modulated over time. Noting that various modulation schemes exist, e.g., frequency-modulated continuous wave (FMCW) modulation, we use ToF sensors with AMCW and FSCW signal modulations, which are illustrated in Figure 4. Up next, we will discuss the operating principles of these depth sensing methods.

NIR AMCW Time-of-Flight. AMCW ToF algorithms modulate the amplitude A of a carrier signal over time t using a repetitive modulation signal s_m such that the transmitted signal s_t can be described as:

$$s_\mathrm{t}(t) = \underbrace{s_\mathrm{m}(t)}_{A} \cdot \cos(2\pi t f + \phi_\mathrm{c}) \ . \tag{3}$$

A constant phase offset is described by ϕ_c. To extract the phase shift from the received signal, it is demodulated to yield m_r and cross-correlated with a so-called signal hypothesis s_h [Zanuttigh et al. 2016]:

$$c_\mathrm{r}(t) = \int_0^{T_\mathrm{m}} m_\mathrm{r}(t) s_\mathrm{h}(t + t')dt' \ . \tag{4}$$

T_m is the period of the modulation signal. Commonly, s_h is chosen as the currently transmitted signal s_t such that c_r describes the signal similarity from which the relative phase shift to m_r is inferred. Extracting this shift requires solving a multivariable equation system with parameters such as the received amplitude and external illumination. To achieve this, c_r and, consequently, m_r are commonly sampled at four points within T_m (four-bucket-method) [Giancola et al. 2018]. During the acquisition of those samples, AMCW ToF is affected by environmental changes, such as varying external NIR illumination and motion. Moreover, oversaturation of the NIR receiver may cause invalid signal responses.

MIMO FSCW Time-of-Flight. FSCW ToF sensors model the frequency of the transmitting signal as a function of time. Given the frequency band $\mathrm{b} = f_\mathrm{max} - f_\mathrm{min}$, they iteratively send N_f signals of one frequency in steps of $\Delta f = \mathrm{b}/(\mathrm{N}_f - 1)$ [Bräunig et al. 2023]. More specifically, the transmitted signal s_t of one capture can be described as:

$$s_\mathrm{t}(t) = A \cdot \cos(2\pi t f_\mathrm{m}(n) + \phi_\mathrm{c}) \text{ with } n = \lfloor t/\Delta t \rfloor \tag{5}$$

$$f_\mathrm{m}(n) = f_\mathrm{min} + \left(n \bmod \mathrm{N}_f\right) \Delta f \ . \tag{6}$$

We denote the time window of one frequency step as Δt. Time-division multiplexing (TDM) avoids signal interference and facilitates the separation of the received signal into its originating transmitter and frequency components. SAR signal processing computes the depth d and the pixel position (u, v) from the phase shift and the angular diversity originating from multiple transmitting and receiving positions. For MIMO imaging radars, the state-of-the-art algorithm of *backprojection* (BP) [Ahmed 2021; Wolf 1969] computes confidence values about a target's presence in 3D space. This is achieved on the basis of local feature distributions within a volume based on the integrated signal of each RX-TX antenna pair. Similar to the four-bucket-method for AMCW ToF, the BP algorithm computes a correlation between the received demodulated signal m_r and a signal hypothesis s_h:

$$c_\mathrm{BP}\underbrace{(x, y, z)}_{\boldsymbol{p}} = \sum_{n=1}^{\mathrm{N}_f} \sum_{i=1}^{\mathrm{N}_\mathrm{RX}} \sum_{j=1}^{\mathrm{N}_\mathrm{TX}} m_\mathrm{r}(f_n, \boldsymbol{r}_i, \boldsymbol{t}_j) s_\mathrm{h}(f_n, \boldsymbol{r}_i, \boldsymbol{t}_j, \boldsymbol{p}) \ . \tag{7}$$

A hypothesis is made on the basis of the transmitted signal, which is assumed to reflect at a point $\boldsymbol{p} \in \mathbb{R}^3$ in the sensor coordinate system, commonly sampled from a voxel grid of size $N_v = N_x \times N_y \times N_z$. The demodulated received signal, m_r, varies in transmit frequency $f_n = f_\mathrm{m}(n)$, transmitter position $\boldsymbol{t}_j \in \mathbb{R}^3$, and receiver position $\boldsymbol{r}_i \in \mathbb{R}^3$. The numbers of transmitters and receivers are denoted as N_TX, and N_RX, respectively. Generally, hypotheses are made by assuming the signal propagation is following the Born approximation [Ahmed 2014]. The result of the above equation is a complex phasor c_BP, calculated from m_r and s_h, which are analytic signals in complex notation. To compute a 2D depth map from the 3D voxel grid, an orthogonal *maximum (intensity) projection* [Bräunig et al. 2023] is performed for each pixel $(u, v) = (x, y)$ along the cross-depth axes of the voxel grid:

$$d(u, v) = \underset{z}{\operatorname{argmax}} \|c_\mathrm{BP}(x, y, z)\|_2 = \underset{z}{\operatorname{argmax}} \ \kappa(x, y, z) \ . \tag{8}$$

The letter κ denotes the so-called *confidence* of a target's presence, as visualized in Figure 2. Besides projection, the confidence values are used as thresholds for depth filtering, that is, to distinguish target depth from sidelobes and background noise. As κ directly relates to c_BP, it depends on both, the received phase and amplitude. Besides the depth, reasons for varying amplitude and phase over different object materials and geometries are manifold, and further insights will be given in Section 7.3. As a result, it is challenging to generalize the depth filtering process for unknown objects. Conversely, if a point $\boldsymbol{p}$ on the target is partially occluded for multiple RX-TX antenna pairs, resulting in a decrease in its confidence value — potentially to the level of background noise — it may be filtered out.

Similar to NIR AMCW ToF, a MIMO FSCW ToF sensor is sensitive to environmental changes while capturing multiple signal samples.

5 The MAROON Dataset

The capture of the MAROON dataset allows for a comprehensive analysis with respect to the characteristics of the four previously described depth sensing techniques.

To accomplish this, we collected a diverse set of common household and construction objects. We ensured having a broad variety of materials and geometries, with varying complexity, which we selected based on prior knowledge of the sensor operating principles (see Section 4.2). The selection aimed to identify challenging objects for reconstruction, highlighting the limitations of current depth imagers and providing a valuable data resource for improving these technologies, e.g. through integration of multimodal sensor data.

In the further course of this section, we outline the capture setup and data acquisition pipeline, depicted in Figure 1, to aid future research on the publicly accessible data. Four single-view depth sensors are used in our experiments: Stereolabs ZED X [Stereolabs 2023] (Passive Stereo), Intel Realsense D435i [Intel 2023] (Active Stereo), Microsoft Azure Kinect [Microsoft 2022] (NIR ToF), and a submodule of Rohde & Schwarz's QAR50 [Rohde & Schwarz 2023] (RF ToF).

5.1 Sensor Setup

Our sensor setup consists of four mounted single-view depth sensors and a ground-truth (GT) optical multi-view stereo (MVS) system comprising five calibrated DSLR cameras, which are depicted on the *left* in Figure 1. While all single-view sensors are designed to achieve an optimal balance between depth quality and acquisition

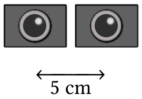
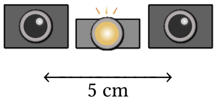
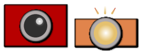
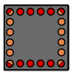
		ZED X Mini (2.2 mm) [Stereolabs 2023]	**Realsense D435i** [Intel 2023]	**Azure Kinect** [Microsoft 2022]	**QAR50 (Submodule)** [Rohde & Schwarz 2023]
Manufacturer		Stereolabs	Intel	Microsoft	Rohde & Schwarz
Depth Sensing Technology		Passive Stereoscopy	Active Stereoscopy	Time-of-Flight (NIR)	Time-of-Flight (RF)
Arrangement				 SIMO	 MIMO (square)
Capture Frame Rate		30 fps	30 fps	30 fps	≈ 70 fps*
Depth Processing Time*		< 33 ms	< 33 ms	< 33 ms	≈ 78 s
Transmitters	Type	–	Laser Projector	LED Array	TDM Antenna Array
	Array Size	–	–	–	2×47 TX ↕
	Wavelength	–	840–860 nm	860 nm	3.6-4.2 mm
	Frequency	–	≈ 353 THz	≈ 353 THz	72-82 GHz
	Signal Modulation	–	Spatial Multiplexing	AMCW	FSCW (N_f = 128)
Receivers	Type	Image sensor	Image sensor	Image Sensor	Antenna Array
	Array Size	2×1928×1208 px	2×1280×800 px	1024×1024 px	2×47 RX ↔
	Spatial Size*	2×5.8×3.6 mm	2×3.8×2.4 mm	3.6×3.6 mm	2×13.8 cm
	Field of view	110° × 80°	87° × 58°	75° × 65°	≈ 53° × 53° *
Depth Image Resolution		1920×1080 px	1280×720 px	640×576 px	301×301 px
Spatial Resolution* $\delta_x \times \delta_y \times \delta_z$	30 cm	0.30×0.39×1.34 mm	0.36×0.42×0.21 mm	0.61×0.59× ≤ 2.0 mm	4.08×4.08×11.08 mm
	40 cm	0.40×0.52×2.38 mm	0.47×0.56×0.38 mm	0.82×0.79× ≤ 2.0 mm	5.38×5.38×12.44 mm
	50 cm	0.50×0.65×3.72 mm	0.59×0.70×0.59 mm	1.02×0.98× ≤ 2.0 mm	6.69×6.69×13.23 mm

Table 1. Overview of the sensors and their parameters used in our experiments. Rows with * indicate derived information not directly given by the manufacturer. Depth processing times were computed on a system with an NVIDIA GeForce RTX 3080 graphics card (10GB VRAM) and an Intel Xeon W-1390P (3.50 GHz) processor. Note that due to its fundamentally different operating principle, modeling the field of view of the QAR50 similar to a camera is a very simplified approximation, and we refer to the supplementary material for further details. δ_x and δ_y of camera-based systems is approximately determined from the per-pixel field of view. Spatial resolution formulae are provided in the supplementary material. Due to missing data for the Azure Kinect, δ_z is assumed to be theoretically higher than the depth accuracy given in [Bamji et al. 2018].

time, the MVS system employs an offline reconstruction process that is specifically optimized for depth quality. In summary, eight cameras are mounted on tripods and arranged around the MIMO imaging radar on a desk, thereby maximizing the area of intersection of each sensor's field of view, to ensure similar object visibility. All sensors and the GT system are time-synchronized, either through hardware or software, to capture the object at the same moment.

Prior to capture, the object is positioned in the center of the squared radar aperture, and approximately at the center of the joint field-of-view intersection, propped up with boards crafted from styrofoam — a material that is considered to be nearly fully penetrated by the RF signal — to prevent external interference of radio waves from other sources in the vicinity, apart from the object of interest. For similar reasons, absorbers are placed behind the object of interest. Similarly, for optical sensors a loose black cloth, which is penetrated by RF signals, is suspended in front of the absorbers to visually occlude the room's background. The sensor settings are chosen with respect to a trade-off between fair sensor comparability and practical applicability (see supp. mat.). An overview of the chosen settings, together with relevant sensor parameters, is given in Table 1.

5.2 Data Acquisition Pipeline

The MAROON dataset comprises static and *quasi-static*, i.e., with slow, minimal motion as in case for human hands, targets of differing materials and geometries, captured at multiple distances from all sensors simultaneously. With respect to the order of steps described in Figure 1 (*middle*), we will now continue to elaborate on the details of the acquisition pipeline.

Spatial Calibration. We spatially aligned the coordinate systems of each depth imager using the calibration method by Wirth et al. [2024]. In this method, four respective spherical objects of styrofoam and metal, tailored to the visibility of optical and RF sensors, are captured. In the sensor-specific reconstructions, these spheres are automatically located and jointly aligned using spatial registration. This approach enables a direct comparison of the object reconstructions in a metrical space. Calibration errors are expected to be in 1–2 mm range with respect to the Chamfer distance, in analogy with the evaluation scheme used in Wirth et al. [2024].

The five DSLR cameras of the MVS system are treated as a unified sensor with a common coordinate system, which is spatially calibrated with that of the depth imagers. The camera extrinsic and intrinsic parameters of the MVS system are determined from images capturing a conventional optical calibration target with a checkerboard pattern. For this, we use the commercial software provided by Agisoft Metashape. Remaining calibration errors exhibit a root mean square reprojection error of 0.38 px, averaged over all camera calibrations performed during the dataset capture.

Object Preparation and Capture. The reconstruction method of MVS is similar to passive stereo imagers. Hence, inaccurate reconstructions can be the result when dealing with textureless and view-dependent object materials. To circumvent this limitation, we implement a distinct capture process for a subset of particularly challenging objects to generate more reliable GT reconstructions. After the object has been captured once by all sensors (including MVS), a thin multicolored speckle pattern is applied using water colors that assists the correspondence finding of the subsequent, additional MVS-only capture. In order to ensure exact alignment between that GT reconstruction and other imaging modalities, the speckle is applied in situ without moving the object.

In total, each object is recorded at three different distances to the MIMO imaging radar of 30 cm, 40 cm, and 50 cm, respectively. The remaining depth imagers are situated behind the radar. Their corresponding object-to-sensor distance is determined from the distance to the radar and from the relative position between each optical sensor and the imaging radar, which is given by the calibration parameters. Based on the Euclidean norm of the mean translation across all calibrations conducted, we report an additional object-to-sensor distance of +8 cm (Azure Kinect), +6 cm (Realsense D435i), and +5 cm (ZED X Mini), respectively. We record 20 frames for each optical sensor and 10 radar frames. In total, we capture 45 objects and list further statistics about the dataset in Table 2.

Optical Segmentation. To perform an accurate object-centric sensor evaluation, it is essential to isolate the estimated object depth from the background. For optical systems, we acquire segmentation masks by performing a semi-automatic foreground-background segmentation. Given that all depth imagers capture RGB images — either for depth estimation or via a separate calibrated camera — we first segment the RGB images using manually defined object labels in conjunction with Grounded-SAM [Ren et al. 2024]. This generates a binary segmentation mask of the object, $\boldsymbol{M}$, where all valid pixels (u, v) are included in $\boldsymbol{M}^{+}(u, v) = \{\boldsymbol{M}(u, v) > 0\}$. We then manually correct failure cases in the resulting segmentation masks. The same procedure is employed to MVS images to produce masked GT reconstructions. For the imaging radar, the voxel volume of the BP algorithm (Equation 7) is constrained to enclose only the object of interest. In this way, segmentation masks are automatically determined from the valid pixels remaining after depth estimation.

Reconstruction and Depth Estimation. MAROON offers raw sensor data, along with intermediate and final reconstruction output, stored in various data representations depending on each depth imager.

For optical depth sensors, we store RGB images, auxiliary data such as infrared measurements, and depth maps, which are obtained using the corresponding signal processing algorithms provided by the manufacturer.

The imaging radar captures raw measurements in form of a tensor of $N_{\text{RX}} \times N_{\text{TX}} \times N_f$ complex numbers, where $N_{\text{RX}} = N_{\text{TX}} = 94$ and $N_f = 128$. They are stored alongside the volumetric output produced after backprojection, as well as post-processed 2D depth and confidence maps. Using the raw tensor, we perform the BP algorithm on a $301 \times 301 \times 201$ voxel grid, with voxel centers uniformly sampled within a $30 \times 30 \times 20\ \text{cm}^3$ volume around the object center, yielding volumetric data that is stored as intermediate output. Subsequently, we apply maximum projection (Equation 8) to acquire a 2D projection of the depth as well as a 2D confidence map. Using the latter, we filter out depth values according to a threshold of −14 dB relative to the maximum value. As mentioned in Section 4.2.2, such thresholding is challenging for unknown objects. We chose this threshold empirically over all objects in the dataset, aiming at a good balance between pruning of noise and retention of object details, and provide an ablation study with different thresholds in the supplementary material. We encourage interested readers to experiment with different thresholds, using the raw radar data available in our dataset. After thresholding, the filtered result is stored as an orthographic depth map.

The ground-truth MVS setup captures five RGB images, which are stored alongside post-processed depth images and a mesh representation of the object, after performing reconstruction using Agisoft Metashape. Metashape (formerly Photoscan) commonly has a reconstruction accuracy in sub-millimeter range for similar capture environments [Mousavi et al. 2018; Remondino et al. 2014]. After reconstruction, we finally apply Laplacian smoothing.

Statistics	MAROON
# objects	45
# static objects	41
# quasi-static objects	4
# prepared speckled objects	14
# captures (# objects × 3 distances)	135
# total / unique optical depth frames	8100 / 405
# total / unique RF depth frames	1350 / 135

Table 2. Statistics of the MAROON dataset. Assuming that all captured objects are static, the number of total frames include duplicate captures, possibly varying in random depth noise, while the unique frames only contain one capture per object of each sensor.

6 Evaluation

In this section, we compare the reconstructions produced by the four presented depth imagers with a ground-truth reconstruction in a common metric space and describe the metrics used in this process. Subsequently, the results of these methods are presented.

We note that in this section the results are objectively presented, reserving further interpretations for Section 7, where they will be discussed with specific attention to partially transmissive media (Section 7.2) and focusing on the RF signal response (Section 7.3).

6.1 Metrics

First, we average valid depth values of each sensor across 10 frames for static objects to incorporate temporal characteristics and reduce random noise. We do not average quasi-static objects, of which their reconstruction did not require the application of speckles, and instead take the first frame, as it is closest to the point in time where the GT captures without the speckle pattern have been taken.

Using the extrinsic calibration parameters (see supp. mat.), we subsequently transform the masked GT reconstruction, $\boldsymbol{R}_g$, into each sensor space s, yielding $\boldsymbol{R}_g^s$. We use the notation $\boldsymbol{R}^*$ to indicate a transformation to sensor space $*$.

Next, for each object, we compute the point-wise deviation between a sensor and the transformed GT reconstruction with respect to two metrics: one-sided Chamfer distance and one-sided projective error. The one-sided Chamfer distance, C, is computed per point $\boldsymbol{p} \in \mathbb{R}^3$ in the source point cloud $\boldsymbol{P} \in \mathbb{R}^{N\times 3}$ with respect to the distance to the nearest point $\boldsymbol{q} \in \mathbb{R}^3$ in the destination point cloud $\boldsymbol{Q} \in \mathbb{R}^{M\times 3}$:

$$\mathrm{C}_{\boldsymbol{p}}(\boldsymbol{Q}) = \min_{\boldsymbol{q}\in\boldsymbol{Q}} \|\boldsymbol{p} - \boldsymbol{q}\|_2 \ . \tag{9}$$

The one-sided projective depth error P is computed per pixel (u, v) of two depth maps $\boldsymbol{D} \in \mathbb{R}^{W\times H}$ and $\boldsymbol{F} \in \mathbb{R}^{W\times H}$ of a common image plane:

$$\mathrm{P}_{u,v}(\boldsymbol{D}, \boldsymbol{F}) = |\boldsymbol{D}(u, v) - \boldsymbol{F}(u, v)| \ . \tag{10}$$

For both, C and P, the respective subscripts $\boldsymbol{p}$ and u, v are used as placeholders to denote the points and pixels used in the metric computation.

Since the one-sided Chamfer distances are sensitive to the point cloud density, we uniformly sample the points from the sensor and the GT with respect to a common image pixel grid. We achieve this by, first, computing a simulated depth map $\widehat{\boldsymbol{D}}_g^s$ in the image space of the sensor s and, second, reconstructing $\widehat{\boldsymbol{R}}_g^s$ from this depth map, using the inverse camera parameters to project it into 3D (see supp. mat.). The simulated depth map is computed by rasterizing a triangulated representation of $\boldsymbol{R}_g^s$ with respect to $\boldsymbol{T}_s$. In this way, we also discard points in optical sensors $\boldsymbol{R}_g^s$ that are not visible in the view of s. The resulting depth map $\widehat{\boldsymbol{D}}_g^s$ is additionally used to measure the projective error. To summarize, we compute:

$\mathrm{C_g}$ **Chamfer distance ground truth.** $\forall \boldsymbol{g} \in \widehat{\boldsymbol{R}}_g^s : \mathrm{C}_{\boldsymbol{g}}(\boldsymbol{R}_s)$

$\mathrm{C_s}$ **Chamfer distance sensor.** $\forall s \in \boldsymbol{R}_s : \mathrm{C}_s(\widehat{\boldsymbol{R}}_g^s)$

P **Projective error.** $\forall (u, v) \in \boldsymbol{M}^+(u, v) : \mathrm{P}_{u,v}(\boldsymbol{D}_s, \widehat{\boldsymbol{D}}_g^s)$ for $\boldsymbol{M} = \boldsymbol{M}_s \cap \widehat{\boldsymbol{M}}_g^s$, where $\boldsymbol{M}_s = \{\boldsymbol{D}_s > 0\}$ and $\widehat{\boldsymbol{M}}_g^s = \{\widehat{\boldsymbol{D}}_g^s > 0\}$ describe the intersection masks of valid pixels from the sensor and the projected GT, respectively.

$\mathbf{P_e}$ **Projective error with erosion.** $\forall (u, v) \in \boldsymbol{M}_e^+(u, v) : \mathrm{P}_{u,v}(\boldsymbol{D}_s, \widehat{\boldsymbol{D}}_g^s)$ for $\boldsymbol{M}_e = \boldsymbol{M}_s \cap f(\widehat{\boldsymbol{M}}_g^s)$, where $f(\widehat{\boldsymbol{M}}_g^s)$ is a function performing mask erosion using a kernel size of $K \times K$ pixels. The size $K \in [0, 20]$ is semi-manually selected for each object and sensor, and included together with other evaluation metadata in the release of our dataset.

Metric Type	Silhouette Noise	Missing Surfaces	3D Error	Depth Error
$\mathbf{C_g}$	–	✓	✓	–
$\mathbf{C_s}$	✓	–	✓	–
P	✓	–	–	✓
$\mathbf{P_e}$	–	–	–	✓

Table 3. Categorization of the presented metrics with respect to their sensitivities. In addition to depth, a 3D error evaluates errors along the cross-depth axes.

6.2 Results

In this section, we first present the evaluation results, quantified using four complementary metrics: $\mathrm{C_g}$, $\mathrm{C_s}$, P, and $\mathrm{P_e}$. Each metric is sensitive to different aspects, as detailed in Table 3. We begin by presenting the depth deviation in relation to various objects and different object-to-sensor distances. Given that near-field imaging radars are less explored compared to optical depth sensors, we dedicate the latter part of this section to the radio-frequency signal response.

6.2.1 Depth Deviation. In Table 4, we list the mean μ and standard deviation σ of each metric type with respect to 12 selected objects from the MAROON dataset. These objects were positioned at an object-to-sensor distance of 30 cm. We provide object images and a comprehensive evaluation of all 45 objects in the supplementary material. For completeness, we give a brief overview of the overall statistics by investigating the number of best and worst results across all objects for μ, respectively:

- RF ToF: performs worst in $\mathrm{C_g}$
- NIR ToF: performs worst in $\mathrm{C_s}$, P, $\mathrm{P_e}$
- Active Stereo: performs best in $\mathrm{C_g}$, $\mathrm{C_s}$, P, $\mathrm{P_e}$
- Passive Stereo: performs neither best or worst

To give an intuition on relative depth deviations between sensors, we present the median, mean, and standard deviation, denoted as $\widetilde{\mu}/\mu$ $(\pm\sigma)$, calculated across the differences in metric values for all pairwise sensor combinations. The results for each metric type are:

- $\mathrm{C_g}$: 0.23 cm/0.48 cm (±0.61 cm)
- $\mathrm{C_s}$: 0.19 cm/1.06 cm (±3.53 cm)
- P: 0.34 cm/1.58 cm (±4.36 cm)
- $\mathrm{P_e}$: 0.31 cm/1.78 cm (±5.14 cm)

Additionally, we illustrate the distribution of the depth deviation across all objects for varying placement distances of 30 cm, 40 cm, and 50 cm in the supplementary material.

6.2.2 RF Signal Response. The point-wise confidence of the backprojection algorithm for RF ToF is considerably affected by the signal amplitude; however, disentangling amplitude from phase in the presence of signal interference imposes the same challenges that arise from recovering the phase shift itself. To avoid inducing additional bias through the assumptions made in signal processing, we investigate the unprocessed signal response of the MIMO imaging radar across multiple objects. For each object at 30 cm object-to-sensor distance, we compute the mean absolute value out

	Metric Type	Cardboard	Sponge	Scrubber	Plushie	Tape Dispenser	Statue
RF ToF	C_g	0.13 (± 0.06)	1.52 (± 0.97)	0.58 (± 0.29)	0.81 (± 0.47)	0.31 (± 0.23)	0.27 (± 0.25)
NIR ToF		0.10 (± 0.06)	0.79 (± 0.45)	0.64 (± 0.34)	0.49 (± 0.21)	0.84 (± 0.30)	0.32 (± 0.28)
Active Stereo		**0.08** (± **0.05**)	**0.18** (± 0.17)	0.20 (± 0.16)	**0.13** (± **0.14**)	0.16 (± 0.14)	0.16 (± **0.13**)
Passive Stereo		0.24 (± 0.11)	0.26 (± **0.15**)	**0.14** (± **0.11**)	0.19 (± 0.21)	**0.15** (± **0.11**)	**0.13** (± **0.13**)
RF ToF	C_s	0.15 (± 0.08)	0.54 (± 0.40)	0.97 (± 0.61)	2.21 (± 2.05)	0.55 (± 0.77)	**0.17** (± **0.11**)
NIR ToF		0.16 (± 0.18)	0.79 (± 0.42)	0.59 (± 0.30)	0.53 (± 0.28)	1.16 (± 0.60)	0.43 (± 0.42)
Active Stereo		**0.08** (± **0.05**)	**0.17** (± **0.15**)	**0.17** (± **0.11**)	**0.12** (± 0.38)	**0.17** (± **0.17**)	0.19 (± 0.76)
Passive Stereo		0.30 (± 0.13)	0.33 (± 0.18)	0.19 (± 0.13)	0.23 (± **0.27**)	0.22 (± 0.20)	0.18 (± 0.69)
RF ToF	P	0.13 (± 0.14)	2.73 (± 2.47)	1.28 (± 0.76)	3.64 (± 3.23)	0.67 (± 1.08)	**0.20** (± **0.26**)
NIR ToF		0.19 (± 0.25)	1.32 (± 0.58)	0.91 (± 0.47)	0.85 (± **0.52**)	1.57 (± 0.70)	0.77 (± 3.13)
Active Stereo		**0.10** (± **0.12**)	**0.29** (± **0.42**)	**0.27** (± **0.34**)	**0.24** (± 1.25)	**0.24** (± **0.35**)	0.90 (± 5.17)
Passive Stereo		0.36 (± 0.17)	0.47 (± 0.51)	0.29 (± 0.38)	0.35 (± 0.54)	0.29 (± **0.35**)	1.43 (± 7.09)
RF ToF	P_e	0.12 (± 0.13)	2.93 (± 2.60)	1.35 (± 0.66)	3.61 (± 3.23)	0.66 (± 1.07)	0.20 (± 0.27)
NIR ToF		0.08 (± 0.10)	1.69 (± **0.23**)	1.16 (± **0.27**)	0.82 (± 0.38)	1.70 (± 0.89)	0.25 (± **0.13**)
Active Stereo		**0.06** (± **0.08**)	**0.42** (± 0.50)	**0.22** (± **0.27**)	**0.16** (± 0.30)	**0.15** (± **0.21**)	0.16 (± 0.21)
Passive Stereo		0.38 (± 0.14)	0.55 (± 0.56)	0.24 (± 0.29)	0.18 (± **0.26**)	0.20 (± 0.22)	**0.10** (± **0.13**)

	Metric Type	S1 Hand Open	Hand Printed Flat	Mirror	Candle	Flowerpot (Transparent)	V1 Metal Plate
RF ToF	C_g	0.36 (± 0.38)	0.71 (± 0.78)	**0.87** (± **0.26**)	1.50 (± 1.12)	1.31 (± 1.21)	0.12 (± **0.05**)
NIR ToF		0.31 (± 0.14)	0.25 (± 0.12)	3.77 (± 1.97)	2.04 (± 0.40)	2.73 (± 1.03)	0.77 (± 0.42)
Active Stereo		**0.12** (± **0.09**)	**0.09** (± **0.07**)	2.13 (± 1.52)	**0.26** (± **0.29**)	**0.74** (± **0.53**)	**0.08** (± 0.06)
Passive Stereo		0.20 (± 0.16)	0.21 (± 0.37)	2.31 (± 1.61)	1.64 (± 0.78)	2.01 (± 0.83)	0.13 (± 0.07)
RF ToF	C_s	0.22 (± 0.15)	0.17 (± 0.13)	**0.91** (± **0.14**)	5.57 (± 2.78)	1.86 (± 2.41)	0.13 (± **0.06**)
NIR ToF		0.38 (± 0.26)	0.29 (± 0.20)	33.31 (± 9.07)	1.71 (± 0.49)	3.10 (± 1.22)	0.81 (± 0.43)
Active Stereo		**0.13** (± **0.10**)	**0.09** (± **0.06**)	30.21 (± 14.59)	**0.25** (± **0.26**)	**1.27** (± 1.78)	**0.09** (± 0.07)
Passive Stereo		0.26 (± 0.22)	0.18 (± 0.34)	27.02 (± 11.33)	1.28 (± 0.65)	1.86 (± **0.93**)	0.16 (± 0.11)
RF ToF	P	**0.22** (± **0.25**)	**0.16** (± **0.20**)	**0.93** (± **0.12**)	7.41 (± 3.79)	2.74 (± 3.66)	0.11 (± **0.12**)
NIR ToF		0.52 (± 0.43)	0.33 (± 0.29)	37.84 (± 14.84)	2.78 (± **0.35**)	5.24 (± 2.04)	0.95 (± 0.48)
Active Stereo		**0.22** (± 1.25)	**0.16** (± 1.30)	39.66 (± 24.75)	**0.42** (± 0.49)	**2.08** (± 2.30)	**0.10** (± 0.13)
Passive Stereo		0.35 (± 0.41)	1.73 (± 8.95)	30.82 (± 14.01)	2.10 (± 0.98)	3.50 (± **1.37**)	0.19 (± 0.15)
RF ToF	P_e	0.22 (± 0.25)	0.16 (± 0.20)	**0.93** (± **0.12**)	7.37 (± 3.85)	2.76 (± 3.66)	0.10 (± 0.11)
NIR ToF		0.51 (± 0.27)	0.30 (± **0.09**)	39.68 (± 6.57)	2.75 (± **0.15**)	6.18 (± 1.79)	0.79 (± 0.39)
Active Stereo		**0.16** (± **0.24**)	**0.08** (± 0.10)	43.84 (± 20.28)	**0.31** (± 0.44)	**2.52** (± 1.77)	**0.07** (± **0.09**)
Passive Stereo		0.25 (± 0.34)	0.17 (± 0.15)	35.96 (± 7.83)	2.15 (± 0.65)	4.36 (± **0.71**)	0.15 (± **0.09**)

Table 4. We measure the depth deviation with respect to C_g, C_s, P, and P_e, which we list in the form $(\mu \pm \sigma)$, consisting of the mean μ and standard deviation σ in centimeters, computed over the entire metric domain, respectively. The best results among all sensors of one metric type are highlighted in **bold** and the worst results are underlined. The results are discussed in Section 15.

of all complex phasors received from the raw signal, averaged over 10 frames. We refer to this quantity as *signal (phasor) magnitude*, emphasizing the difference from signal amplitude.

Differentiating between signal response and reconstruction quality, we examine their relationships with respect to object material, geometry, and size. Our findings are shown on the *left* of Figure 5, where we visualize these relationships for signal magnitude (*upper row*) and mean depth deviation (*bottom row*) in isolation. On the *right*, we further investigate correlations between signal magnitude and mean depth deviation.

Object materials are categorized into six classes, with detailed information available in the supplementary material. The goal of this classification is to highlight material differences on a coarse level, noting the large object variety that still persists within one material class.

The object geometry is quantified by the median angle in degrees between the point-wise surface normals of the GT reconstruction and the depth direction (along the *z*-axis) of the imaging radar. As we positioned the objects to align their primary orientation with the viewing direction of the planar square-shaped antenna aperture — which particularly becomes important for flat objects — the median angle mainly reflects geometric complexity, with objects having a higher surface incidence angle showing larger portions oriented away. An extension of Figure 5 (*left*) is available in the

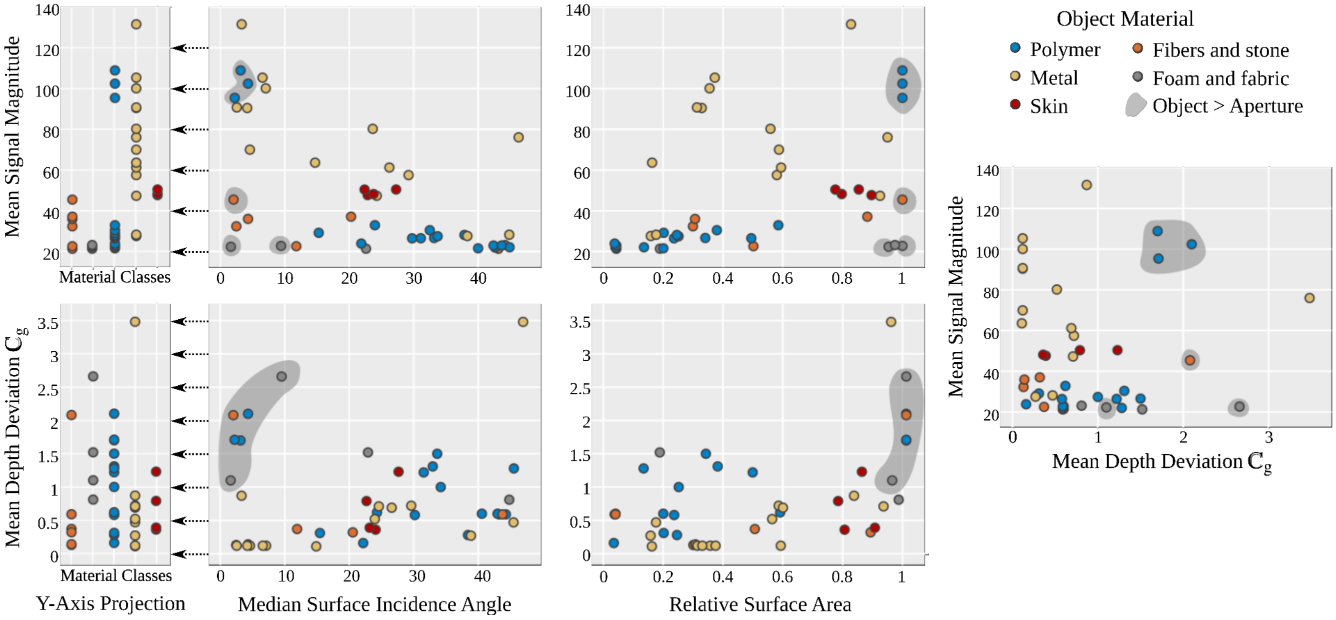


Fig. 5. On the *left*, object material, geometry (median surface incidence angle), and size (relative surface area) are put in relation to received signal response (mean signal magnitude, *top row*) and mean depth deviation (*bottom row*). On the *right*, both quantities are directly compared to each other. Measurements, where large objects appear outside the radar's antenna aperture, are highlighted in gray regions, as they exhibit higher depth deviations compared to the ground-truth reconstructions, which may extend beyond this aperture; this is attributed to the comparably small field of view and the surface reflection characteristics with respect to radio waves (see supp. mat.). The results are discussed in Section 7.3.

supplementary material, visualizing the correlation of the median angle as well as per-angle measurements with respect to the depth deviation of all four sensors.

Object size is determined by the relative surface area compared to the radar antenna aperture. It is computed from the fraction of the object's 2D axis-aligned bounding box A (in the x- and y-axis) inside of the 2D axis-aligned bounding box B of the antenna array by using the formula: $(A \cap B)/B$.

7 Discussion

In this section, we provide a general discussion of the previously reported results related to depth deviation, followed by two focused discussions of time-of-flight sensor effects that offer complementary perspectives on these results. Regarding the latter, we investigate effects of partially transmissive media and explore RF ToF as a particularly under-explored sensor technology, focusing on the received signal response in relation to depth deviation.

We note that a comprehensive sensor characterization, highlighting common trends across all 45 objects, is provided in the supplementary material, where we discuss the interpretation of metrics, the depth deviation over varying distances, as well as relative depth deviations between sensors.

7.1 Discussion of Object-specific Depth Deviation

The following section will analyze the objects in Table 4 in regard to their relative depth deviations over one or multiple sensors.

Radio-Frequency Time-of-Flight. For RF ToF, we find that the least deviation relative to the mean of all metrics occurs with planar object geometries (*V1 Metal Plate*, *Cardboard*), followed by more complex shapes (*Statue*, *S1 Hand Open*, *Hand Printed Flat*). For a deeper discussion, see Section 7.3.

Objects made of foam (*Sponge*), thin plastic (*Scrubber*, *Flowerpot*), fabric (*Plushie*), and paraffin wax (*Candle*), exhibit the highest depth deviations due to a large fraction of the transmitted RF signal not being immediately reflected. In the case of *Mirror*, RF penetrates the first (glass) surface and images the silver coating behind, leading to an offset in the depth reconstruction.

Near-infrared Time-of-Flight. For similar reasons, NIR ToF shows large depth deviations for visually transparent objects like *Flowerpot*, *Candle*, *Sponge*, and *Tape Dispenser*. Both RF ToF and NIR ToF are susceptible to multi-path effects; however, our experiments suggest these effects do not occur for the same objects. Further examination of wavelength-specific multi-path effects, with a particular focus on partially transmissive materials, will be discussed in Section 7.2.

Additional sources of high depth deviation for NIR ToF include thin structures (*Scrubber*), which reduce the sensor's effective spatial resolution. Highly reflective objects (*Metal Plate*) may cause sensor oversaturation, while perfectly specular materials (*Mirror*) yield depth values from the first weak scatterer after perfect reflection.

Active and Passive Stereo Sensors. For the active stereo sensor, we observe higher depth deviations for textureless and partially transmissive materials (*Sponge*, *Candle*). Similar to NIR ToF, the

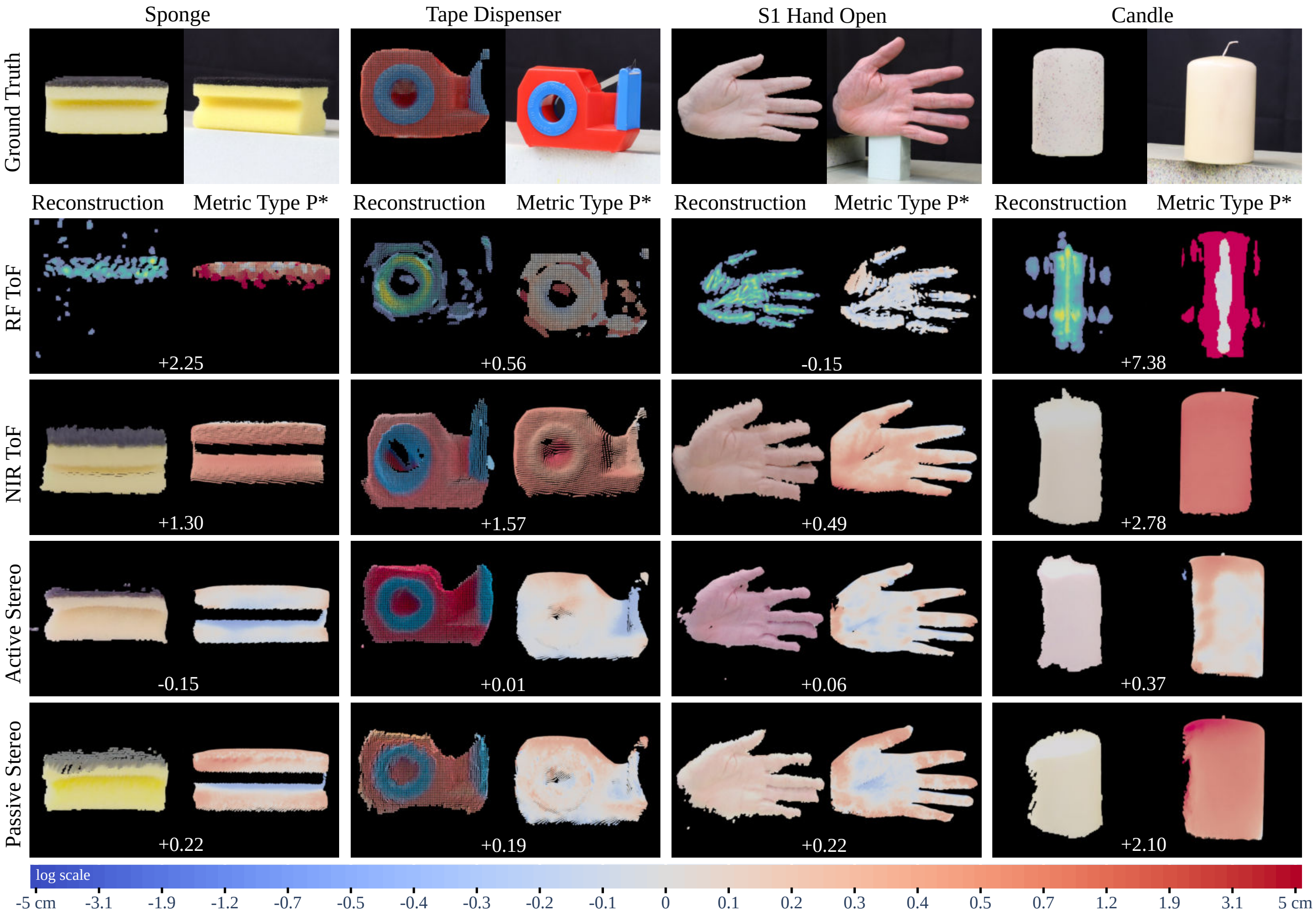


Fig. 6. For selected objects, we show the reconstructed point clouds (*left*) next to their deviation from to the MVS reconstruction (*right*). The signed depth deviation P^* is given for each pixel (u, v) in centimeters. All measurements in the domain $\boldsymbol{M}^+(u, v)$ are projected onto the GT reconstruction and mapped to color using a combination of a symmetrical logarithmic scale and linear mapping between $[-0.5, 0.5]$ centimeters. The mean deviation of P^* is quantified in centimeters below each sensor measurement.

uniqueness of the active NIR light pattern can be compromised by multi-path effects. The passive stereo camera is particularly sensitive to textureless objects (*Sponge*, *Candle*).

7.2 Discussion of Time-of-Flight Sensors: Partially Transmissive Media

As previously discussed in Section 4.2.2, both NIR and RF ToF sensors assume direct reflection and thus are susceptible to internal reflections, such that multi-path effects within the scene may lead to missing or incorrect reconstructions. In this analysis, following the nomenclature by Nayar et al. [2006], we classify radiance transport that involves a single signal bounce between sender and receiver as *direct* (as, within the sensor's spatial resolution, it interacts with the scene at one surface point only), and all other types of transport as *global* (involving multiple scattering or diffraction events within and between objects). Due to their significant difference in wavelength relative to scene features, global radiance transport takes very different forms for each modality. In the case of NIR, representative forms of global transport include inter-reflections, half-transparent surfaces, and subsurface scattering within the object material. Global transport at radio frequencies, on the other hand, is dominated by diffraction and reflections that reshape and redirect the wave front as it interacts with multiple scene elements, and by multiple superimposed responses akin partial transmittance at different depths.

In the remainder, we will now study the four selected objects in Figure 6. In addition to Table 4, this figure visualizes depth deviations using a signed version P^* of metric P, color-encoded on a symmetrical logarithmic scale (`SymLogNorm`[1]), with a linear mapping between $[-0.5, 0.5]$ centimeters. The supplementary material includes signed versions of P and P_e for all MAROON objects.

Near-infrared Time-of-Flight. In the NIR domain, the most prominent effect of global transport occurs for objects with strong internal

[1] https://matplotlib.org/3.8.4/api/_as_gen/matplotlib.colors.SymLogNorm.html

scattering. Here, the ToF reconstructions exhibit systematic depth deviations of P*, generally biased toward larger distances than the ground truth. This is consistent with the light traveling an additional distance due to scattering within the object before being remitted again, so that the observed propagation time of the actively transmitted signal is consistently longer than for a direct (local) reflection at the object surface. Examples in Figure 6 for internal scattering include subsurface scattering (*S1 Hand Open*, *Sponge*, *Candle*) and inter-reflections within hollow objects (*Tape Dispenser*). Extended path length due to subsurface scattering is an established effect, systematically measured by Lukinsone et al. [2020]. For human skin (e.g. *S1 Hand Open*), and for points of incidence and exitance one millimeter apart, Lukinsone et al. observe effective sub-surface path lengths of up to 26 ± 3 mm at 800 nm wavelength, which — in the context of a ToF sensor — would result in a systematic depth deviation of half that path length (≈ +13 mm). At the same time, however, for human skin a significant portion of the total remitted light stays very close to the point of incidence [Jensen et al. 2001], suggesting that the bulk of the received signal experiences even smaller path length extensions, lending plausibility to our measured systematic depth deviation of +4.9 mm for *S1 Hand Open* to be due to subsurface scattering.

Radio-Frequency Time-of-Flight. For RF ToF, only one object (*Candle*) showed a systematic path length extension, suggesting that optical subsurface scattering cannot fully model RF interactions. In contrast to the NIR ToF measurements, the depth deviation for the *Candle* object is non-uniform, with higher values near the edges due to variations in surface position and orientation that affect radiance transport.

Where the *Candle* surface faces the antenna array, the received signal is dominated by direct reflections; where direct reflections reflect away from the array (nearer to the candle's silhouettes), mostly global transport is observed. In accordance with the results by Álvarez López et al. [2018] the depth reconstruction in the parts with little direct reflection appear more distant than ground truth, which the authors attribute to the high relative permittivity $\varepsilon_r \approx 2.6$ of paraffin wax that extends the inferred path length under the assumption of speed of light in vacuum.

In summary, objects composed of partially transmissive media primarily yield systematic bias in ToF reconstructions, with estimated depths biased toward larger values than the ground truth. Nevertheless, the factors causing these distortions vary between optical and RF modalities.

7.3 Discussion of MIMO Radar: Signal Response and Depth Deviation

Our observations in Section 15 suggest that RF ToF reconstructions are generally less complete than those of optical sensors, as illustrated in the *second row* of Figure 6, where depth deviations are lowest when surface orientations align with the antenna aperture.

Initially, this seems to contradict the expectation that larger antenna apertures should capture more surface compared to cameras, given the variety of positions and viewing angles from the individual RX-TX antenna pairs; this advantage, however, seems to be mitigated by the fact that most object surface reflections appear to be specular [Lu et al. 2013]. This means that reflections at surfaces oblique to the aperture are only received by a small fraction of antennas, thereby weakly contributing to the overall signal response, potentially at the same level as noise.

We discuss further sources of incomplete RF ToF reconstructions in the next sections, where we first analyze the raw signal response — without inducing additional bias from the reconstruction algorithms — and subsequently relate it to the quality of the measured depth after reconstruction.

7.3.1 *Radio-Frequency Signal Response.*

In Figure 5 (*top left*), we presented the received signal magnitude across objects with varying material, geometry and size. Following the scatter plot order from left to right, we will now discuss common trends, noting that it remains challenging to disentangle the presented quantities, as the large variability across objects prevents us from isolating one quantity while keeping the others constant.

Influence of Material. Metal and metallic-coated objects generally show higher signal magnitudes, which is consistent with previous studies [Ahmed 2014, 2021]. With a considerably lower spread, large magnitudes are also observed for captures of human skin, which is highly reflective due to its rich water content [Ahmed 2014]. Object materials made of polymers, fibers and stone, or foam generally respond with much smaller signal magnitude.

Regarding *object geometry* and *size*, no significant global trends are observed; however, consistent patterns emerge within subsets of the same material, particularly in the metal and polymer classes, which have the highest number of samples. We will discuss these patterns within the next paragraphs.

Influence of Geometry. For objects of more complex geometry, with a median surface incidence angle greater than 10°, large portions of their surface area face away from the antenna array, resulting in decreased signal responses compared to planar objects aligned closely with the antenna aperture (<10°). The reflection direction of the transmitted signal depends on the surface normal's orientation. As the angle between this normal and the depth axis increases, the solid angle of the object relative to the planar square-shaped antenna aperture (cf. Table 1) decreases. In other words, a decreasing area around the hemisphere of outgoing reflection directions is aligned with the approximate, 53° field of view of the RX antennas, resulting in reduced signal energy reception, and thus radar cross-section [Knott et al. 2004]. Superficially this resembles the well-known cosine law in radiometry, but the exact quantitative relationship depends on the object's location relative to the individual RX and TX antennas and is further modulated by the non-trivial radiation and signal lobes of the antennas.

Influence of Size. Aside from object geometry, the received signal magnitude also appears to increase with object size for non-metal materials. Within these material classes, the highest signal magnitude is achieved for objects close to or even larger than the antenna aperture. As the latter also typically exhibits a low median surface incidence angle, it remains questionable, whether this observation can be attributed to object size or object geometry. To address this, we additionally visualize the relation between the two quantities in the supplementary material. Assuming that signal magnitude is proportional to the received energy, our findings correspond to the fact that, the reflected signal energy received at the RX antennas directly depends on the surface area of the irradiated object, in case the energy density is constant.

7.3.2 Radio-Frequency Depth Deviation. Following the previous section, we now summarize the results of Figure 5 (*bottom left*), where we relate the depth deviation to varying object material, size, and geometry in their respective scatter plots.

Influence of Material. Similar to our findings for signal response, metal objects generally exhibit the lowest depth deviation with a relatively small spread compared to other material classes, indicating that object material influences reconstruction quality.

Influence of Size and Geometry. While we find no direct relationship of object size to depth deviation, the most notable trend is seen with varying object geometry, where the deviation increases alongside the median surface incidence angle across all material classes. The backprojection algorithm assumes that similar energy amounts are received at a point across the majority of RX antennas. Received energy diminishes for surfaces oriented away from the antenna geometry, leading to variations based on antenna positioning. This, in turn, can lead to reduced confidence in the measurements, causing valid data to be filtered out along with noise.

Note that the MIMO radar we use has a large aperture and a high number of RX-TX antenna pairs, suggesting that stronger orientation-dependent effects may be observed with typical lower-resolution devices. To explore this further, we simulated various down-sampled antenna architectures and present their respective depth deviation in comparison to the fully occupied antenna array in the supplementary material.

7.3.3 Relation between RF Signal Response and Depth Deviation. Figure 5 visualizes potential correlations between the RF signal response and depth deviation in the *right* scatter plot. Focusing on the most prevalent material groups of polymer and metal objects, we generally find no direct relationship between signal magnitude and depth deviation. Polymer and metal objects have a large spread in the x- and y-axis, respectively, while the opposite axis has comparably low variation. The received signal magnitude may have a more significant influence on the reconstruction quality in less constrained scenarios, involving multiple objects and signal sources, where, for example, depth filtering becomes increasingly relevant. In our experiments, within the RF near field, we suggest that reconstruction quality is more closely related to the distribution of received signals across antennas, influenced primarily by the local antenna layout and characteristics.

To conclude, the analysis of MIMO imaging radar reconstruction quality is not as straightforward as that of optical sensors. We find no direct relationship between the RF signal response and the corresponding depth deviation after reconstruction and filtering on a global scale. Consistent patterns within object subsets suggest that the reconstruction quality of RF ToF sensors is primarily influenced by object geometry, while the impact of object material should not be overlooked, as it is a crucial factor for depth filtering.

To disentangle material and geometry-dependent effects at surface level, we suggest that material characterization is the first essential step in addressing this challenge; we will pick up on this topic in the subsequent application section.

8 Applications of MAROON

In this section, we highlight two applications utilizing the data from our proposed dataset. First, we briefly summarize the insights revealed from previous experiments:

- There is no single sensor modality that would consistently outperform the others. Each sensor has unique strengths and weaknesses related to the object's material, geometry, and distance from the sensor.
- NIR ToF displays systematic depth distortions due to effects of global radiance transport within partially transmissive media, whereas RF ToF reconstructions were mostly unaffected.
- RF ToF partially exhibits missing reconstructions compared to its optical counterpart, primarily due to the object geometry contributing to the reconstruction sparseness.

Given these insights, we anticipate that multimodal depth sensing amplifies the sensors' complementary strengths, hence providing notable benefits for *close-range* applications — similar to the multi-sensor design employed in *far-range* sensing for self-driving cars.

We provide two examples, where we build upon prior work and demonstrate the substantial role of the high-quality sensor co-localizations from MAROON for their successful implementation: first, we show how the dataset is utilized for achieving realistic RF simulations by adapting and extending the concurrent research of Hofmann et al. [2025]. Leveraging a pre-release of our dataset, Hofmann et al. propose a differentiable ray tracing pipeline to determine the material parameters of our captured objects.

Second, we present extended experiments for a recently proposed multimodal high-speed radar reconstruction method, MM-2FSK [Wirth et al. 2025], which utilizes only two frequencies as opposed to our employed 128 frequency-stepped backprojection. We note that we also examine the depth deviation in less computationally intensive versions of backprojection, by varying the frequency configuration, as detailed in the ablation study included in the supplementary material.

8.1 Material Characterization with Differentiable Ray Tracing

Accurate material modeling is vital not only for isolating the effects of mmWave signal interaction but also for high-fidelity radar simulation. Highly reflective materials, such as metals, produce vastly different radar returns compared to largely diffuse scatterers, such as objects made of wood or rubber. Therefore, Hofmann et al. [2025] propose a data-driven approach to determine reflective properties under mmWave radiation, such as permittivity and permeability.

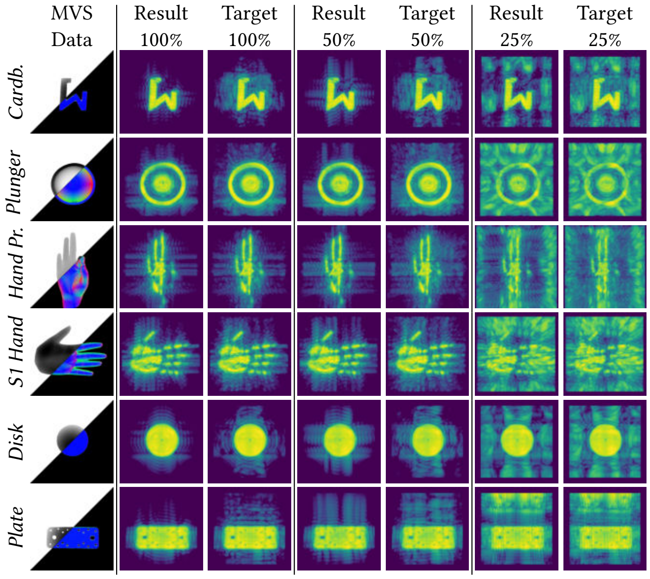


Fig. 7. Inverse radar rendering results for different antenna configuration. 100% considers all transmitting and receiving antennas in the MIMO array (94×94), 50% considers every second respective antenna in the array (47×47), and 25% only considers every fourth antenna in the array (24 × 24). Despite obvious artifacts and ambiguities due to a reduced antenna density, the optimization process still performs robustly, and thus generalizes to other MIMO configurations. Objects from top to bottom: *Cardboard*, *Plunger*, *Hand Printed F*, *S1 Hand Open Metal Disk (Thin)*, and *V1 Metal Plate*.

They initialize their differentiable optimization pipeline by simulating radar returns with randomized material properties and compare the result to the real RF ToF sensor data from MAROON, while utilizing the MVS data as ground-truth geometry. Analogous to neural networks, where parameters are iteratively optimized using gradient descent, they continuously update the material properties of the object until the difference between the simulated and measured radar returns is minimal. Notably, the loss is computed on the raw phasor data, instead of reconstructed images, which increased both robustness and fidelity of the optimization due to bypassing artifacts introduced by the reconstruction algorithm. To facilitate a close match between the simulated and real phasor data, the radar gain and a small registration offset of $\frac{\lambda}{2}$ along each principal axis, where λ is the longest wavelength emitted in a FSCW sequence, is optimized alongside the material properties [Hofmann et al. 2025]. This registration offset requires the error in the MVS data to be smaller than one wavelength to avoid getting stuck in local minima due to ambiguities from recurring wave patterns. Fortunately, we can safely assume this to be the case in MAROON with the calibration error ranging from 1–2 mm, as discussed in Section 5.2, which is half of the mean wavelength of ≈ 4 mm in the worst case.

To demonstrate the versatility of the dataset, in addition to the original experiments conducted by Hofmann et al. [2025], we utilized the high number of antenna signals available in MAROON to examine the impact of a varying antenna configurations and aperture sizes. To this end, we used 100%, 50%, and 25% of the antennas, which were simulated by selecting every, every second, or every fourth RX/TX antenna from the raw phasor data in the dataset, respectively. We showcase results for the three different antenna configurations and six different objects in Figure 7.

For a visualization of the respective antenna apertures, we refer to the supplementary material, where we also conducted ablation studies of more consumer-friendly aperture configurations.

8.2 Multimodal Depth Sensing

While the backprojection algorithm is employed in many near-field high-resolution RF ToF applications, its reconstruction time is typically orders of magnitude slower than the sensor's capture rate (see Table 1). As a robust and computationally efficient alternative, Wirth et al. [2025] introduce a multimodal image reconstruction method that builds upon the previous frequency shift keying (FSK) approach proposed by Bräunig et al. [2023]. By utilizing an optical depth camera as a secondary sensor, point-wise depth priors are integrated into the 2FSK signal processing pipeline, allowing just two frequencies to adjust these depth priors towards the target object's actual depth. The depth prior is essential to determine the correct period of the sinusoidal wave signal that is otherwise limited to a small unambiguous range. We refer the interested reader to [Wirth et al. 2025] for all technical details about the algorithm. The authors evaluated the proposed *MM-2FSK* method using the active stereo depth sensor and MIMO imaging radar from our pre-released dataset. In this section, we extend their work by comparing the method across all optical sensors in our dataset, simulating various capture scenarios influenced by the optical depth sensor.

Ablation with respect to Optical Depth Sensors. Drawing from insights about sensor-specific characteristics, we examine how different depth imagers affect depth deviations in the MM-2FSK method. We follow the evaluation procedure detailed in [Wirth et al. 2025], employing the most promising frequency configuration — specifically, two frequencies at 72 and 82 GHz, resulting in a frequency difference of $\Delta f = 10$ GHz.

In Figure 8, we display top-down views of the MM-2FSK reconstructions for three objects, overlaid with the ground-truth point cloud while varying the sensor that provides the depth prior. For the *Flowerpot (Transparent)* (*top row*), we notice numerous points reconstructed behind the object for all depth imagers except the ground truth. This transparency causes parts of the background or ground surface to be reconstructed, resulting in a depth prior positioned behind the ground truth. With limited unambiguous depth correction capabilities [Wirth et al. 2025], the MM-2FSK method can not correct outliers when the optical depth prior lies within a different signal period than the ground truth.

Similarly, for the *V2 Metal Plate*, large areas of its surface are reconstructed behind the object for both NIR ToF and active stereo sensors, due to multi-path effects and sensor oversaturation arising from the object's perfect specularity.

Lastly, there are a few depth outliers behind the *Bunny*, primarily associated with the NIR ToF and active stereo sensors, stemming from incorrect depth priors due to the triangulation of flying pixels, which the MM-2FSK method cannot correct.

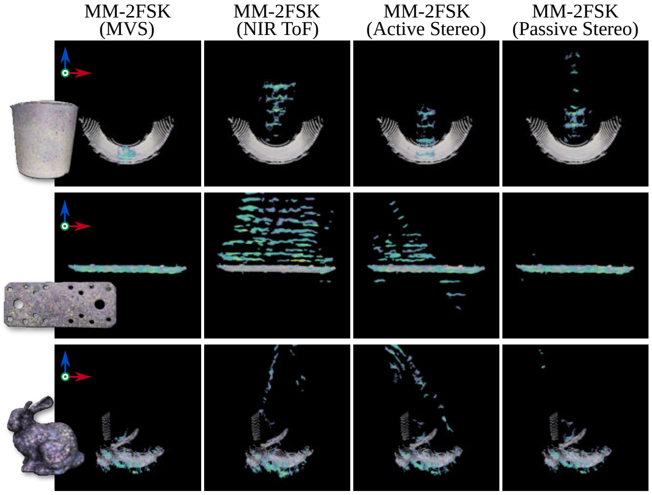


Fig. 8. Top-down views of the RF ToF reconstructions obtained with the MM-2FSK method, fused with the ground-truth MVS point cloud. From *left* to *right*, we vary the supporting sensor providing the depth prior. From *top* to *bottom*, we display the *Flowerpot (Transparent)*, *V2 Metal Plate*, and *Bunny* objects.

Depth Prior	C_g	C_s	P	P_e
MVS	**0.51**	**0.18**	**0.19**	**0.17**
NIR ToF	1.19	***1.67***	1.58	1.48
Active Stereo	***0.82***	1.74	1.36	***1.36***
Passive Stereo	0.92	1.91	***1.29***	1.41

Table 5. Ablation study of the MM-2FSK method with different depth priors, each from another optical depth imager. The mean depth deviation from the ground truth, given in centimeters, is averaged over all objects at 30 cm distance. The **best** and ***second best*** results per metric are highlighted.

Moreover, Table 5 lists the mean depth deviation of all objects in relation to the varying depth priors. By assessing the depth deviation always relative to the MVS setup, we expect that its respective depth prior yields the best performance.

Consistent with earlier evaluations (cf. Section 6.2.1), no single sensor outperforms others across all objects. For 3D errors ($C_g$and C_s), the active stereo and NIR sensors yield the best performance, while for projective errors (Pand P_e), both passive stereo and active stereo sensors provide the most accurate results.

9 Limitations

During sensor characterization, we observed that, even though the depth deviation of the MIMO imaging radar is on par with that of optical sensors, the reconstructions exhibit considerably more holes, where no valid depth is estimated. While it is intuitive to assume that reconstruction quality is influenced by object material (limiting the returned signal amount — akin to optically transmissive materials), we observe that in our experiments the object geometry is the primary factor of influence and has a greater impact than in optical sensors; however, disentangling the effects of geometry and material remains challenging, as precise impacts on fine-grained surface details cannot be easily assessed. Without a direct mapping between point targets and RX antennas, depth evaluation concerning these surface-level details is infeasible without backprojection, or any other depth processing algorithm. On the other hand, the reconstructed outcomes after signal processing may not align with reality, as these methods typically incorporate a systematic bias by relying on the Born approximation [Ahmed 2014]. To overcome these limitations, we highlighted one particular work of Hofmann et al. [2025], taking the first step towards automatic material characterization in the radio-frequency domain using inverse rendering. We anticipate that further analysis of these material parameters, combined with improvements of radio-frequency simulation frameworks [Schüßler et al. 2021], will considerably aid in disentangling the potential error sources behind missing reconstructions and enhancing current signal processing methods.

The proposed evaluation framework for sensor characterization is tailored to point cloud comparisons, and is, therefore, independent of the RF signal processing algorithm; however, it requires the spatial co-localization of sensors. To achieve the latter, it needs to be verified, whether the respective spatial calibration method may be applicable to other high-resolution radar systems; alternatively, it can be substituted with any other calibration method tailored to the radar system of interest.

Furthermore, the object reconstructions were evaluated solely for valid locations in the ground-truth data, excluding artifacts like ghost targets or other forms of noise that may arise from violations of the Born approximation. Lastly, we did not capture different orientations of flat objects, which would be an interesting future direction to investigate object orientation in isolation from geometry complexity.

10 Conclusion

We presented a novel multimodal dataset, MAROON, that allows us to characterize, for the first time, near-field MIMO imaging radars in direct relation with traditional depth imagers from the optical frequency domain for close-range applications. The dataset comprises depth images of a variety of objects, synchronously captured by four mutually calibrated depth imagers and a ground-truth multi-view stereo system. We subsequently analyzed the data within a comprehensive evaluation framework, offering quantitative and qualitative perspectives on each sensor's depth deviation across multiple metric types, objects, and object-to-sensor distances. The findings presented are based on aggregate trends and individual object analyses that contribute to the understanding of the addressed sensor characteristics; however, we believe that our dataset still invites further analysis, exploiting the high diversity of the 45 objects that could not be fully addressed in the scope of this paper.

Moreover, we presented two exemplary applications, utilizing the collected data. First, we built upon previous work to characterize the materials of our captured objects, which is an interesting future direction to disentangle material-specific effects from geometric influences. Second, we conducted extended experiments on a recently proposed multimodal depth estimation approach [Wirth et al. 2025], using our dataset as a baseline to evaluate its performance. In connection with this work, we examined the impact of different optical sensor modalities to identify suitable depth priors for radar signal processing.

We hope that by highlighting these promising research directions, along with the release of our MAROON dataset, our work will give rise to further study of multimodal sensor systems in a joint reference frame.

Acknowledgement

The authors would like to thank the Rohde & Schwarz GmbH & Co. KG (Munich, Germany) for providing the radar imaging devices.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1483 – Project-ID 442419336, EmpkinS.

The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center of the Friedrich-Alexander-Universität Erlangen-Nürnberg.

References

Sherif Sayed Ahmed. 2014. *Electronic Microwave Imaging with Planar Multistatic Arrays.* Logos Verlag, DEU.

Sherif Sayed Ahmed. 2021. Microwave Imaging in Security — Two Decades of Innovation. *IEEE Journal of Microwaves* 1, 1 (2021), 191–201. https://doi.org/10.1109/JMW.2020.3035790

Cyrus S. Bamji, Swati Mehta, Barry Thompson, Tamer Elkhatib, Stefan Wurster, Onur Akkaya, Andrew Payne, John Godbaz, Mike Fenton, Vijay Rajasekaran, Larry Prather, Satya Nagaraja, Vishali Mogallapu, Dane Snow, Rich McCauley, Mustansir Mukadam, Iskender Agi, Shaun McCarthy, Zhanping Xu, Travis Perry, William Qian, Vei-Han Chan, Prabhu Adepu, Gazi Ali, Muneeb Ahmed, Aditya Mukherjee, Sheethal Nayak, Dave Gampell, Sunil Acharya, Lou Kordus, and Pat O'Connor. 2018. IMpixel 65nm BSI 320MHz demodulated TOF Image sensor with 3μm global shutter pixels and analog binning. In *2018 IEEE International Solid-State Circuits Conference - (ISSCC).* 94–96. https://doi.org/10.1109/ISSCC.2018.8310200

Akanksha Bhutani, Sören Marahrens, Marius Kretschmann, Serdal Ayhan, Steffen Scherr, Benjamin Göttel, Mario Pauli, and Thomas Zwick. 2022. Applications of radar measurement technology using 24 GHz, 61 GHz, 80 GHz and 122 GHz FMCW radar sensors. *Technisches Messen* 89, 2 (2022), 107–121. https://doi.org/doi:10.1515/teme-2021-0034

DW Bliss and KW Forsythe. 2003. Multiple-input multiple-output (MIMO) radar and imaging: degrees of freedom and resolution. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 1. IEEE, 54–59.

Johanna Braeunig, Desar Mejdani, Daniel Krauss, Stefan Griesshammer, Robert Richer, Christian Schuessler, Julia Yip, Tobias Steigleder, Christoph Ostgathe, Bjoern M. Eskofier, and Martin Vossiek. 2023. Radar-based Recognition of Activities of Daily Living in the Palliative Care Context Using Deep Learning. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI).* 1–4. https://doi.org/10.1109/BHI58575.2023.10313506

Johanna Bräunig, Vanessa Wirth, Christoph Kammel, Christian Schüßler, Ingrid Ullmann, Marc Stamminger, and Martin Vossiek. 2023. An Ultra-Efficient Approach for High-Resolution MIMO Radar Imaging of Human Hand Poses. *IEEE Transactions on Radar Systems* 1 (2023), 468–480. https://doi.org/10.1109/TRS.2023.3309574

Anjun Chen, Xiangyu Wang, Kun Shi, Shaohao Zhu, Bin Fang, Yingfeng Chen, Jiming Chen, Yuchi Huo, and Qi Ye. 2023. ImmFusion: Robust mmWave-RGB Fusion for 3D Human Body Reconstruction in All Weather Conditions. In *2023 IEEE International Conference on Robotics and Automation (ICRA).* 2752–2758. https://doi.org/10.1109/ICRA48891.2023.10161428

Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. 2022. MmBody Benchmark: 3D Body Reconstruction Dataset and Analysis for Millimeter Wave Radar. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22).* Association for Computing Machinery, New York, NY, USA, 3501–3510. https://doi.org/10.1145/3503161.3548262

Chuang-Yuan Chiu, Michael Thelwell, Terry Senior, Simon Choppin, John Hart, and Jon Wheat. 2019. Comparison of depth cameras for three-dimensional reconstruction in medicine. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* 233, 9 (2019), 938–947. https://doi.org/10.1177/0954411919859922 arXiv:https://doi.org/10.1177/0954411919859922 PMID: 31250706.

Yoshana Deep, Patrick Held, Shobha Sundar Ram, Dagmar Steinhauser, Anshu Gupta, Frank Gruson, Andreas Koch, and Anirban Roy. 2020. Radar cross-sections of pedestrians at automotive radar frequencies using ray tracing and point scatterer modelling. *IET Radar, Sonar & Navigation* 14, 6 (2020), 833–844. https://doi.org/10.1049/iet-rsn.2019.0471 arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-rsn.2019.0471

Silvio Giancola, Matteo Valenti, and Remo Sala. 2018. *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies* (1st ed.). Springer Publishing Company, Incorporated.

Georg Halmetschlager-Funek, Markus Suchi, Martin Kampel, and Markus Vincze. 2019. An Empirical Evaluation of Ten Depth Cameras: Bias, Precision, Lateral Noise, Different Lighting Conditions and Materials, and Multiple Sensor Setups in Indoor Environments. *IEEE Robotics & Automation Magazine* 26, 1 (2019), 67–77. https://doi.org/10.1109/MRA.2018.2852795

Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Horaud. 2012. *Time-of-Flight Cameras: Principles, Methods and Applications.* Springer Publishing Company, Incorporated.

Jürgen Hasch, Eray Topak, Raik Schnabel, Thomas Zwick, Robert Weigel, and Christian Waldschmidt. 2012. Millimeter-Wave Technology for Automotive Radar Sensors in the 77 GHz Frequency Band. *IEEE Transactions on Microwave Theory and Techniques* 60, 3 (2012), 845–860. https://doi.org/10.1109/TMTT.2011.2178427

Nikolai Hofmann, Vanessa Wirth, Johanna Bräunig, Ingrid Ullmann, Martin Vossiek, Tim Weyrich, and Marc Stamminger. 2025. Inverse Rendering of Near-Field mmWave MIMO Radar for Material Reconstruction. *IEEE Journal of Microwaves* (2025), 1–17. https://doi.org/10.1109/JMW.2025.3535077

Intel 2023. *Intel® RealSense™ Product Family D400 Series.* Intel. https://www.intelrealsense.com/download/21345/?tmstv=1697035582

Henrik Wann Jensen, Stephen R. Marschner, Marc Levoy, and Pat Hanrahan. 2001. A practical model for subsurface light transport. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01).* Association for Computing Machinery, New York, NY, USA, 511–518. https://doi.org/10.1145/383259.383319

Uma S. Jha. 2018. The millimeter Wave (mmW) radar characterization, testing, verification challenges and opportunities. In *2018 IEEE AUTOTESTCON.* 1–5. https://doi.org/10.1109/AUTEST.2018.8532561

E.F. Knott, J.F. Schaeffer, and M.T. Tulley. 2004. *Radar Cross Section.* Institution of Engineering and Technology. https://books.google.de/books?id=0WuGjb8sqCUC

Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. 2023. HuPR: A Benchmark for Human Pose Estimation Using Millimeter Wave Radar. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).* 5704–5713. https://doi.org/10.1109/WACV56688.2023.00567

Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: ubiquitous gesture sensing with millimeter wave radar. *ACM Trans. Graph.* 35, 4, Article 142 (jul 2016), 19 pages. https://doi.org/10.1145/2897824.2925953

Teck-Yian Lim, Spencer A. Markowitz, and Minh N. Do. 2021. RaDICaL: A Synchronized FMCW Radar, Depth, IMU and RGB Camera Data Dataset With Low-Level FMCW Radar Signals. *IEEE Journal of Selected Topics in Signal Processing* 15, 4 (2021), 941–953. https://doi.org/10.1109/JSTSP.2021.3061270

Alvaro Lopez Paredes, Qiang Song, and Miguel Heredia Conde. 2023. Performance Evaluation of State-of-the-Art High-Resolution Time-of-Flight Cameras. *IEEE Sensors Journal* 23, 12 (2023), 13711–13727. https://doi.org/10.1109/JSEN.2023.3273165

Jonathan S. Lu, Patrick Cabrol, Daniel Steinbach, and Ravikumar V. Pragada. 2013. Measurement and Characterization of Various Outdoor 60 GHz Diffracted and Scattered Paths. In *MILCOM 2013 - 2013 IEEE Military Communications Conference.* 1238–1243. https://doi.org/10.1109/MILCOM.2013.212

Vanesa Lukinsone, Anna Maslobojeva, Uldis Rubins, Maris Kuzminskis, M. Osis, and Janis Spigulis. 2020. Remitted photon path lengths in human skin: in-vivo measurement data. *Biomedical Optics Express* 11 (05 2020). https://doi.org/10.1364/BOE.388349

Emidio Marchetti, Rui Du, Ben Willetts, Fatemeh Norouzian, Edward G. Hoare, Thuy Yung Tran, Nigel Clarke, Mikhail Cherniakov, and Marina Gashinova. 2018. Radar cross-section of pedestrians in the low-THz band. *IET Radar, Sonar & Navigation* 12, 10 (2018), 1104–1113. https://doi.org/10.1049/iet-rsn.2018.5016 arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-rsn.2018.5016

Microsoft 2022. *Azure Kinect DK hardware specifications.* Microsoft. https://learn.microsoft.com/en-us/azure/kinect-dk/hardware-specification

V. Mousavi, M. Khosravi, M. Ahmadi, N. Noori, S. Haghshenas, A. Hosseininaveh, and M. Varshosaz. 2018. The performance evaluation of multi-image 3D reconstruction software with different sensors. *Measurement* 120 (2018), 1–10. https://doi.org/10.1016/j.measurement.2018.01.058

Shree K. Nayar, Gurunandan Krishnan, Michael D. Grossberg, and Ramesh Raskar. 2006. Fast separation of direct and global components of a scene using high frequency illumination *(SIGGRAPH '06)*. Association for Computing Machinery, New York, NY, USA, 935–944. https://doi.org/10.1145/1179352.1141977

Fabio Remondino, Maria Grazia Spera, Erica Nocerino, Fabio Menna, and Francesco Nex. 2014. State of the art in high density image matching. *The Photogrammetric Record* 29, 146 (2014), 144–166. https://doi.org/10.1111/phor.12063 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/phor.12063

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:cs.CV/2401.14159

Rohde & Schwarz 2023. *R&S® QAR50 Quality automotive radome tester.* Rohde & Schwarz. https://www.rohde-schwarz.com/products/test-and-measurement/radome-tester/rs-qar50-quality-automotive-radome-tester_63493-1138625.html?change_c=true

Dominik Schwarz, Nico Riese, Ines Dorsch, and Christian Waldschmidt. 2022. System Performance of a 79 GHz High-Resolution 4D Imaging MIMO Radar With 1728 Virtual Channels. *IEEE Journal of Microwaves* 2, 4 (2022), 637–647. https://doi.org/10.1109/JMW.2022.3196454

Christian Schüßler, Marcel Hoffmann, Johanna Bräunig, Ingrid Ullmann, Randolf Ebelt, and Martin Vossiek. 2021. A Realistic Radar Ray Tracing Simulator for Large MIMO-Arrays in Automotive Environments. *IEEE Journal of Microwaves* 1, 4 (2021), 962–974. https://doi.org/10.1109/JMW.2021.3104722

Krishnasamy T. Selvan and Ramakrishna Janaswamy. 2017. Fraunhofer and Fresnel Distances: Unified derivation for aperture antennas. *IEEE Antennas and Propagation Magazine* 59, 4 (2017), 12–15. https://doi.org/10.1109/MAP.2017.2706648

Vasilii Semkin, Jaakko Haarla, Thomas Pairon, Christopher Slezak, Sundeep Rangan, Ville Viikari, and Claude Oestges. 2020. Analyzing Radar Cross Section Signatures of Diverse Drone Models at mmWave Frequencies. *IEEE Access* 8 (2020), 48958–48969. https://doi.org/10.1109/ACCESS.2020.2979339

Stereolabs 2023. *ZED X Datasheet.* Stereolabs. https://cdn2.stereolabs.com/assets/datasheets/zed-x-datasheet-march-2023.pdf

Shunqiao Sun, Athina P. Petropulu, and H. Vincent Poor. 2020. MIMO Radar for Advanced Driver-Assistance Systems and Autonomous Driving: Advantages and Challenges. *IEEE Signal Processing Magazine* 37, 4 (2020), 98–117. https://doi.org/10.1109/MSP.2020.2978507

Richard Szeliski. 2022. *Computer Vision - Algorithms and Applications, Second Edition.* Springer. 749–797 pages. https://doi.org/10.1007/978-3-030-34372-9

Klen Čopič Pucihar, Nuwan T. Attygalle, Matjaz Kljun, Christian Sandor, and Luis A. Leiva. 2022. Solids on Soli: Millimetre-Wave Radar Sensing through Materials. *Proc. ACM Hum.-Comput. Interact.* 6, EICS, Article 156 (jun 2022), 19 pages. https://doi.org/10.1145/3532212

Gustavo Velasco-Hernandez, De Jong Yeong, John Barry, and Joseph Walsh. 2020. Autonomous Driving Architectures, Perception and Data Fusion: A Review. In *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*. 315–321. https://doi.org/10.1109/ICCP51029.2020.9266268

Alexander Vilesov, Pradyumna Chari, Adnan Armouti, Anirudh Bindiganavale Harish, Kimaya Kulkarni, Ananya Deoghare, Laleh Jalilian, and Achuta Kadambi. 2022. Blending camera and 77 GHz radar sensing for equitable, robust plethysmography. *ACM Trans. Graph.* 41, 4, Article 36 (jul 2022), 14 pages. https://doi.org/10.1145/3528223.3530161

Te-Mei Wang and Zen-Chung Shih. 2021. Measurement and Analysis of Depth Resolution Using Active Stereo Cameras. *IEEE Sensors Journal* 21, 7 (2021), 9218–9230. https://doi.org/10.1109/JSEN.2021.3054820

Shunjun Wei, Zichen Zhou, Mou Wang, Jinshan Wei, Shan Liu, Jun Shi, Xiaoling Zhang, and Fan Fan. 2021. 3DRIED: A High-Resolution 3-D Millimeter-Wave Radar Dataset Dedicated to Imaging and Evaluation. *Remote Sensing* 13, 17 (2021). https://doi.org/10.3390/rs13173366

C.S. Williams and O.A. Becklund. 2002. *Introduction to the Optical Transfer Function.* SPIE Press. https://books.google.de/books?id=b5tVkUq3j4EC

N.J. Willis and H.D. Griffiths. 2007. *Advances in Bistatic Radar.* Institution of Engineering and Technology. https://books.google.de/books?id=HZYOVhgOmzwC

Vanessa Wirth, Johanna Bräunig, Danti Khouri, Florian Gutsche, Martin Vossiek, Tim Weyrich, and Marc Stamminger. 2024. Automatic Spatial Calibration of Near-Field MIMO Radar With Respect to Optical Sensors. *ArXiv* abs/2403.10981 (2024).

Vanessa Wirth, Johanna Bräunig, Martin Vossiek, Tim Weyrich, and Marc Stamminger. 2025. MM-2FSK: Multimodal Frequency Shift Keying for Ultra-Efficient and Robust High-Resolution MIMO Radar Imaging. arXiv:eess.SP/2511.01405 https://arxiv.org/abs/2511.01405

Emil Wolf. 1969. Three-dimensional structure determination of semi-transparent objects from holographic data. *Optics Communications* 1, 4 (1969), 153–156. https://doi.org/10.1016/0030-4018(69)90052-2

Di Wu, Matthew O'Toole, Andreas Velten, Amit Agrawal, and Ramesh Raskar. 2012. Decomposing global light transport using time of flight imaging. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 366–373. https://doi.org/10.1109/CVPR.2012.6247697

Zhiwei Xiong, Yueyi Zhang, Feng Wu, and Wenjun Zeng. 2017. Computational Depth Sensing : Toward high-performance commodity depth cameras. *IEEE Signal Processing Magazine* 34, 3 (2017), 55–68. https://doi.org/10.1109/MSP.2017.2669347

Pietro Zanuttigh, Ludovico Minto, Giulio Marin, Fabio Dominio, and Guido Cortelazzo. 2016. *Time-of-flight and structured light depth cameras: Technology and applications.* 1–355 pages. https://doi.org/10.1007/978-3-319-30973-6

Natnael S. Zewge, Youngmin Kim, Jintae Kim, and Jong-Hwan Kim. 2019. Millimeter-Wave Radar and RGB-D Camera Sensor Fusion for Real-Time People Detection and Tracking. In *2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA)*. 93–98. https://doi.org/10.1109/RITAPP.2019.8932892

Maxim Zhadobov, Nacer Chahat, Ronan Sauleau, Catherine Le Quement, and Yves Le Drean. 2011. Millimeter-wave interactions with the human body: state of knowledge and recent advances. *International Journal of Microwave and Wireless Technologies* 3, 2 (2011), 237–247. https://doi.org/10.1017/S1759078711000122

Yuri Álvarez López, María García Fernández, Raphael Grau, and Fernando Las-Heras. 2018. A Synthetic Aperture Radar (SAR)-Based Technique for Microwave Imaging and Material Characterization. *Electronics* 7, 12 (2018). https://doi.org/10.3390/electronics7120373

11 Spatially Resolved Depth Sensing

It is common practice for stereo sensors to contain two cameras, C_1 and C_2, of known relative spatial location. In the event of parallel optical axes, this location is defined as the baseline B. To compute depth, pixels in image of C_1 are matched to pixels of C_2, forming correspondence pairs. For every correspondence pair, the depth d is computed from the disparity D, which represents the difference between their pixel positions [Giancola et al. 2018]:

$$d = f\frac{B}{D} \ . \tag{11}$$

Spatial Resolution. The depth resolution δ_z of spatially resolved sensors is limited by the disparity resolution ΔD [Zanuttigh et al. 2016]:

$$\delta_z = \frac{z^2}{Bf}\Delta D \ . \tag{12}$$

We denote the ground-truth depth as z and the focal length as f. The disparity resolution is dependent on δ_x and δ_y. For camera-based systems, δ_x and δ_y are typically expressed through the optical transfer function (OTF) [Williams and Becklund 2002].

12 Time-resolved Sensors (Time-of-Flight)

Time-of-Flight sensors can be roughly categorized into direct Time-of-Flight (dToF) and indirect Time-of-Flight (iToF) depth sensing methods. DToF sensors transmit a signal pulse and directly measure the time it takes for the pulse to return. Due to their high cost, however, they are less commonly used in close-range applications. More cost-efficient than dTof are continuous wave (CW) signal modulations that measure time indirectly (iToF) based on the phase shift $\Delta\varphi$ between the transmitted and received signal [Zanuttigh et al. 2016]. The general form of a continuous sinusoidal carrier signal s_c can be described by two equal formulas of traveling time t and traveling distance ρ, respectively:

$$\begin{aligned} s_\mathrm{c}(t) &= A \cdot \cos(2\pi t f + \phi_\mathrm{c}) & (13)\\ &= A \cdot \cos(\underbrace{2\pi\frac{\rho}{\mathrm{c}}f + \phi_\mathrm{c}}_{\varphi}) = \widehat{s}_\mathrm{c}(\rho) \ . & (14)\end{aligned}$$

A and f are the known signal amplitude and frequency, respectively and c is the speed of light, φ is the phase and ϕ_c is a constant phase offset. As a transmitted signal $s_\mathrm{t} = s_\mathrm{c}(t_1) = \widehat{s}_\mathrm{c}(\rho_1)$ of known phase and amplitude reflects at a target, the received signal $s_\mathrm{r} = s_\mathrm{c}(t_2) = \widehat{s}_\mathrm{c}(\rho_2)$ has a relative traveling distance of $2 \cdot \Delta\rho = (\rho_2 - \rho_1)$ between the transmitter and receiver. The range, $\Delta\rho$, is related to the relative phase shift $\Delta\varphi$ [Zanuttigh et al. 2016] by:

$$\Delta\rho = c\frac{\Delta\varphi}{4\pi f} \ . \tag{15}$$

The general assumption of dToF sensors is that a signal directly reflects at the first target and therefore the range equals half of the traveling distance. The depth resolution of a ToF sensor is specific to the utilized wavelength and spatial arrangement of transmitters and receivers.

NIR AMCW Time-of-Flight. AMCW ToF algorithms usually operate on the SIMO principle, as they do not require as expensive sensor apertures as imaging radars, and often have more receivers and transmitters than can be effectively managed computationally in MIMO depth estimation algorithms [Zanuttigh et al. 2016]. The range resolution of a NIR AMCW ToF sensor can be expressed as [Lopez Paredes et al. 2023]:

$$\delta_z = \frac{c}{f_\mathrm{m}}\sqrt{\frac{P_l + P_a}{P_l} \cdot \frac{I}{k_o q_e \rho \Delta t}} \ . \tag{16}$$

Environment-specific parameters are the power of ambient light P_a, and the reflectivity of the target ρ. Hardware-specific parameters are the modulation frequency f_m, the power of the illumination unit P_l, the total illumination area I, the quantum efficiency q_e, the integration time Δt, and a constant parameter for the optical system, k_o. Due to unknown hardware-specific parameters, we were unable to determine the exact range resolution for NIR AMCW ToF (Azure Kinect) in Table 1 of the main paper. We refer to [Lopez Paredes et al. 2023] for an experimental approach of determining the effective range and lateral resolution.

MIMO FSCW Time-of-Flight. The spatial resolution of a square-shaped MIMO FSCW imaging sensor can be expressed as [Ahmed 2021]:

$$\delta_{x,y} = \frac{c}{4f_\mathrm{max}} \cdot \sqrt{4\left(\frac{z}{L}\right)^2 + 1} \tag{17}$$

$$\delta_z = \frac{0.5 \cdot c}{\Delta f + \left(1 - \frac{1}{\sqrt{1+0.5(L/z)^2}}\right) \cdot f_\mathrm{min}} \ . \tag{18}$$

We denote the size of the square aperture as L.

13 Sensor Parameters and Settings

The sensor settings in Table 1 of the main paper are chosen with respect to a trade-off between fair sensor comparability and practical applicability. We uniformly list the frame rate computed from the time takes to *capture* the relevant data of one depth frame. Note that this may not necessarily include the computation of depth. For instance, the QAR50 has a capture rate of ≈ 70 fps while the back-projection algorithm has an average computing time of 78 s such that the overall frame rate is below 1 fps. Furthermore, we manually adjusted each optical sensor's exposure time, if possible, to ensure similar lighting conditions. In summary, we selected the sensor settings that optimize quality while, when feasible, maintaining a comparable frame rate to that of the other sensors. Additionally, we adhered to the manufacturer's recommendations for optimal practical use in interactive applications.

13.1 Radar Field of View

Optical sensors typically model the field of view using a perspective camera model. In theory, MIMO radars can also be viewed as an array of small cameras such that the antenna aperture acts as a unified perspective camera, with its field of view defined by the union of all individual antenna frustums. In practice, modeling the complex antenna radiation pattern as a conventional camera

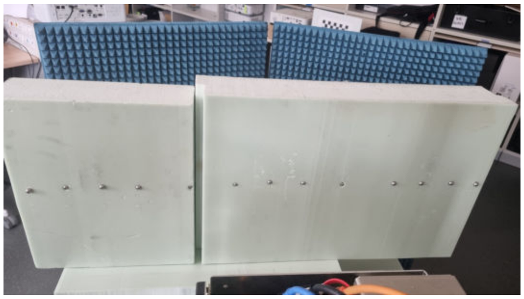


Fig. 9. Styrofoam board with mounted metal spheres of ⌀1 cm.

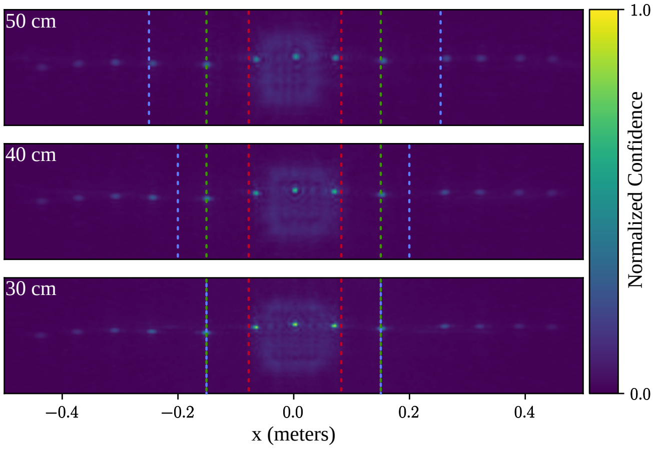


Fig. 10. Confidence map of multiple point scatterers, displaced along the x-axis of the antenna aperture for the three object-to-sensor distances of MAROON. The confidence values are normalized across the three reconstructions. The vertical, dashed lines mark the horizontal extent of the antenna aperture, reconstruction volume, and approximated 53 × 53° perspective camera frustum.

frustum is a crude approximation, as the extents of the visible area are not as straightforward to define as for optical sensors.

To demonstrate this, we conducted an experiment where we mounted several 1 cm diameter metal spheres on a styrofoam board at a fixed horizontal distance around the aperture origin, as shown in Figure 9. The aim of this experiment is to explore the maximum visible area by measuring the signal response of each pair of spheres placed on opposite sides of the origin. The signal response for each metal sphere is illustrated in Figure 10, as part of the confidence value after spatially resolving the raw signal using backprojection.

For point targets with uniform, view-independent scattering properties, the imaging radar's visible area encompasses all the mounted metal spheres, covering an horizontal area of approximately 90 cm (and potentially even further).

In contrast, targets with extended surfaces and non-uniform scattering properties are reconstructed within a more limited area that roughly corresponds to the size of the antenna aperture (marked in *red*). Here, this is observed for the styrofoam board, which reflects a minimal amount of the emitted signal and is typically considered nearly invisible. In this scenario, where only empty space is reconstructed alongside point targets, the signal response of the styrofoam board behaves similarly to other planar surface targets in MAROON.

In summary, the visible area of a MIMO radar has similarities with a continuous Gaussian function centered at the aperture origin. To determine the effective visible area, we compute the *full width at half maximum* (FWHM); here, it represents the horizontal extent of the reconstruction area where confidence values exceed 50 % of the maximum. For the three object-to-sensor distances of 30 cm, 40 cm, and 50 cm, this extent is approximately between [−0.15, 0.15] meters, aligning closely with the *green*-marked reconstruction volume used for evaluation. The corresponding fields of view of 53°, 41°, and 33° differ significantly across these distances, making it challenging to find a unified perspective camera frustum. A suitable perspective camera frustum would also need to fully encompass the 13.8 × 13.8 cm aperture at a distance of 0 cm, i.e., the aperture origin. Contrary to camera-based systems, where the spatial origin usually lies within the sensor extents, however, this frustum would yield an approximate field of view of 65°, with the camera origin located ≈ 10.8 cm behind the aperture. These observations suggest that an orthographic camera model, as utilized for backprojection, is a more suitable approximation for describing the visible volume of the RF ToF sensor.

However, to maintain consistency with the parameters given for camera-based systems, we assume an approximated 53° field of view in Table 1 of the main paper, which encompasses all of the extents measured with the FWHM and is highlighted in *blue* in Figure 10.

14 Dataset Post-processing and Evaluation

Example images of all 45 objects in MAROON can be found in Figure 20. We compare the reconstructions produced by the four presented depth imagers with a ground-truth reconstruction in a common metric space and describe the methods used in this process.

Projection into 3D. We acquire a point cloud of the object's surface utilizing the 2D depth and auxiliary data provided by MAROON. For a given pixel position (u, v) and its corresponding depth d from an optical depth sensor, we first verify its validity using the segmentation map of the same resolution — a step that has already been performed for radar during depth filtering. Subsequently we project each valid triple (u, v, d) back into 3D space using the given transformation matrix $\boldsymbol{T} \in \mathbb{R}^{4\times4}$:

$$\begin{pmatrix} x \\ y \\ d \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} & \boldsymbol{I} & & \boldsymbol{t} \\ 0 & 0 & 0 & 1 \end{pmatrix}^{-1}}_{T^{-1}} \begin{pmatrix} u \cdot a \\ v \cdot a \\ d \\ 1 \end{pmatrix} . \tag{19}$$

For all optical depth imagers, this equation is the inverse of a perspective transformation with intrinsic camera matrix $\boldsymbol{I} \in \mathbb{R}^{3\times3}$, pixel offset vector $\boldsymbol{t} \in \mathbb{R}^3 = \boldsymbol{0}$ and $a = d$. Analogously for radar data, the equation is the inverse of an orthographic transformation with a scale matrix $\boldsymbol{I}$, pixel offset $\boldsymbol{t}$, and $a = 1$.

Joint Alignment. To estimate the deviation of a sensor reconstruction $\boldsymbol{R}_s \in \mathbb{R}^{M\times 3}$ from the GT, $\boldsymbol{R}_g \in \mathbb{R}^{N\times 3}$, we use the previously determined spatial calibration parameters $\boldsymbol{K}_{g\to s} \in \mathbb{R}^{4\times 4}$ to transform $\boldsymbol{R}_g$ from the GT space g into the sensor space s:

$$\widetilde{\boldsymbol{R}}^{s}_{g} = \widetilde{\boldsymbol{R}}_{g}\boldsymbol{K}^{T}_{g\to s} \ . \tag{20}$$

$\widetilde{\boldsymbol{R}}$ denotes the homogeneous version of $\boldsymbol{R}$. We use the notation $\boldsymbol{R}^*$ to indicate a reconstruction that has been transformed to sensor space $*$.

14.1 Radar Depth Filtering

For our experiments and dataset post-processing, we chose an empirical threshold of −14 dB over all objects, which — to the best of our knowledge — has proven to yield the best balance of noise pruning while retaining relevant object measurements. In Figure 11, we show how the signal-to-noise ratio of the radar confidence map, i.e., the pixel-wise values of κ, behaves over different thresholds for the exemplary capture of the *S1 Hand Open*.

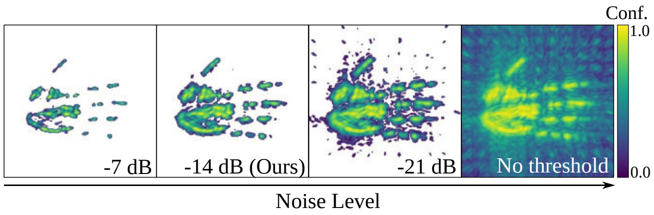


Fig. 11. Visualization of the 2D confidence map of the *S1 Hand Open* at various thresholds. The filtered confidence map is subsequently used to extract valid depth information.

Threshold	$\mathrm{C_g}$ (w = 2)	$\mathrm{C_s}$ (w = 1)	P (w = 1)	$\mathrm{P_e}$ (w = 1)	Weighted Mean
- 7 dB	1.37	**0.67**	**0.78**	**0.76**	0.99
-10 dB	1.07	***0.73***	***0.82***	***0.81***	0.90
-14 dB	0.82	0.9	0.95	0.85	**0.87**
-17 dB	0.68	1.05	1.03	0.95	***0.88***
-21 dB	***0.51***	1.54	1.16	1.07	0.96
—	**0.35**	5.75	1.45	1.45	1.87

Table 6. Ablation study with different signal thresholds used for depth filtering. The depth deviation is expressed in centimeters across the four metrics presented in the main paper, averaged for all objects at a 30 cm object-to-sensor distance. Additionally, we provide a weighted mean for each row, assigning double the weight, w, to $\mathrm{C_g}$, as it is most sensitive to point cloud completeness. The **best** and ***second best*** results per metric are highlighted.

Additionally, we performed an ablation study to evaluate how different thresholds impact the mean depth deviation across all MAROON objects at a 30 cm object-to-sensor distance. In Table 6, we present results for the four metrics discussed in the main paper. We include a weighted mean for each row, giving double weight to $\mathrm{C_g}$, as it is the only metric that is sensitive to the completeness of the point cloud. Notably, we find that the performance concerning $\mathrm{C_g}$ is inversely related to that of the other metrics, which are more sensitive to signal noise and depth quality. The best trade-off between completeness and noise is achieved with a threshold ranging from -14 dB to -17 dB.

14.2 Radar Material Classification

To investigate the radar signal response and depth deviation with respect to different materials, we divided the 45 objects of MAROON into six classes. These assignments are listed in Table 9. The goal of this classification is to highlight material differences on a coarse level, noting the large object variety that still persists within one material class. Furthermore, we list the objects that are larger than the antenna aperture. It is important to consider these objects when interpreting the depth deviation trends presented in the main paper, as their reconstructions may be incomplete due to the portions that fall outside the antenna aperture.

14.3 Additional Results

We provide additional quantitative results for all 45 objects with respect to the depth deviation from the ground truth in Table 10, Table 11, Table 12, and Table 13.

15 Extended Discussion of Depth Deviation

First, we present complementary perspectives on the data, where we put the depth deviation of all 45 objects into relation with the different metric types and, subsequently, the different types of depth imagers. We analyze each representation in turn, highlighting common trends, in combination with previously stated results of Section 6.2.

15.1 General Trends

Interpreting the extensive numerical data on depth deviations in Table 10, Table 11, Table 12, and Table 13 can be challenging, so we provide visual, complementary views in this section. In order to relate different quantities to each other, we use barycentric interpolation based on triples of metric types (Figure 12) and sensors (Figure 13), respectively. For each triple (μ_a, μ_b, μ_c), the mean values for depth deviation (μ) of each object are transformed to affine coordinates (w_a, w_b, w_c) by using the formula $w_{\{a,b,c\}} = \mu_{\{a,b,c\}}/(\mu_a + \mu_b + \mu_c)$. Circle locations closer to a triangle corner indicate higher relative depth deviation. Moreover, as the triples are drawn from a set of four, the triangles are arranged in the shape of an unfolded tetrahedron, highlighting that each triangle's contents can be seen as a projection of barycentric coordinates within a (3D) tetrahedron $w_{\{a,b,c,d\}} = w_{\{a,b,c,d\}}/(w_a + \cdots + w_d)$.

Interpretation of Metrics. In Figure 12, we provide a qualitative comparison of each sensor's depth deviation from GT with respect to the four presented metrics. In dense reconstructions, as is typical for optical depth sensors, metrics based on nearest neighbors (here, $\mathrm{C_g}$ and $\mathrm{C_s}$) are bound to be lower than those based on projection (P and $\mathrm{P_e}$); they also tend to be more resilient against noise. For RF reconstructions, however, that are prone to sparse depth maps, Chamfer distances often create false matches; accordingly,

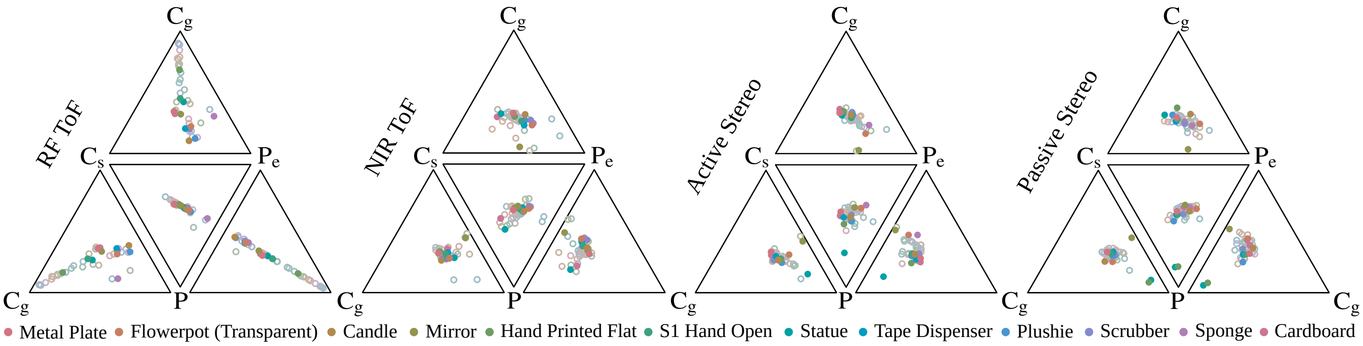


Fig. 12. A complementary view on the depth deviation across different *metric types*. For each triplet of metrics $M_i \in \{C_g, C_s, P, P_e\}$, we convert each mean depth deviation μ_i to affine coordinates, $\bar{\mu}_i = \mu_i / \sum_i \mu_i$, that map an object's errors in to a triangle whose corners correspond to metrics M_i. All 45 MAROON objects are shown as circles, with selected objects from Table 4 highlighted in solid colors. Samples closer to a triangle corner indicate a higher relative depth deviation in the corresponding metric.

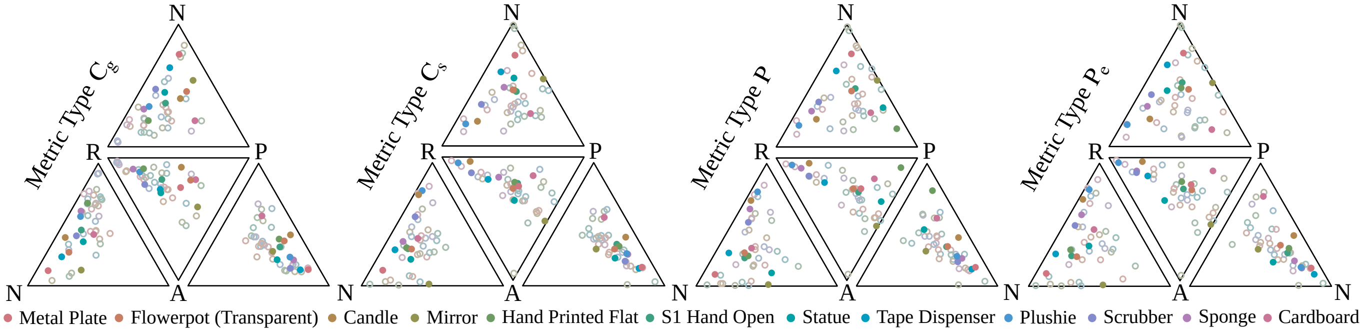


Fig. 13. A complementary view on the depth deviation across different *sensors*. The sensors are denoted as **R** (RF ToF), **N** (NIR ToF), **A** (Active Stereo), and **P** (Passive Stereo), respectively. Analogously to Figure 12, we convert the mean depth deviations μ_i to affine coordinates within triangles corresponding to all possible sensor triples. All 45 MAROON objects are shown as circles, with selected objects from Table 4 highlighted in solid colors. Samples closer to a triangle corner indicate a higher relative depth deviation for the corresponding sensor.

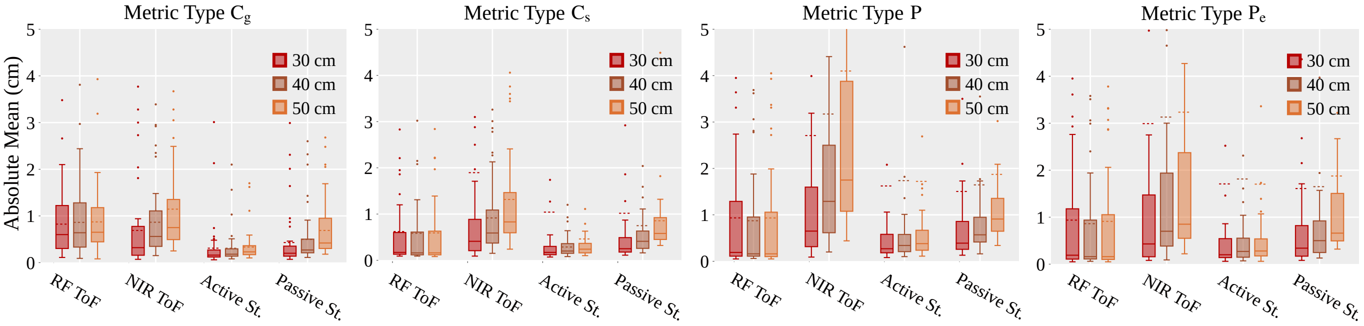


Fig. 14. Box plots, visualizing the distribution of the mean error across all objects with respect to different object-to-sensor distances. Solid (—) and dashed (- -) horizontal lines indicate the median and the mean of the distribution, respectively. The results are discussed in Section 15.

C_g dominates for RF ToF compared to the other metrics. For further discussion regarding the sparsity of RF reconstructions, see Section 7.3 in the main paper.

For optical sensors, a marginal trend towards the corners of C_s and P, away from the silhouette-resilient P_e and C_g, cf. Table 3, indicates the presence of noise at object silhouettes.

Relative Depth Deviation across Sensors. In Figure 13, we observe a considerable spread of depth deviations across different sensors. As noted in Section 6.2, the relative depth deviation between sensors ranges from 1.9 to 3.4 mm, that is, the variation in the depicted normalized error occurs within a comparatively small range of absolute errors. As a general trend, the two stereo sensors have the lowest depth deviation (see triangles N–R–A and N–R–P), with a moderate edge for active stereo, particularly for metrics P and P_e, see the A–P axis, where passive stereo‘s performance degrades. This is consistent with the results of Section 6.2.1 in the main paper, where we find that the active stereo has the highest number of best results. However, given that relative depth deviations between sensors differ by only a few millimeters, we conclude that the seemingly improved depth quality of active stereo sensors is of minimal significance. Moreover, an examination of the scatter plot reveals a marginal weight trend towards the corners of the two ToF sensors. Reconstruction errors between the two ToF sensors are highly object-dependent, and many objects, including many of the highlighted ones, demonstrate multi-path effects due to perfect signal reflections, retro-reflectivity, or partial signal transmission. Further details on this will be discussed in the next section.

Depth Deviation over Distance. In Figure 14, we observe that the depth deviation of the passive stereo and the NIR ToF sensor is considerably more distance-specific compared to active stereo and RF ToF. For passive stereo, this may be due to a decrease in effective spatial resolution, where δ_z directly depends on $\delta_{x,y}$, see Section 11. As the distance between the object and the sensor increases, $\delta_{x,y}$ decreases, resulting in a loss of high-frequency color details while the object appears smaller in the image. Compared to passive stereo, the active stereo sensor has a comparably higher effective resolution, assuming that the resolution of both sensors differs in accordance with δ_{xyz} in Table 1 of the main paper. The unique active NIR pattern may also be less sensitive to decreases in spatial resolution, maintaining the quality of correspondence matches. The trend for the NIR ToF sensor aligns with findings from Bamji et al. [2018] for 30–50 cm distances. However, we argue that absolute errors for a target with 20% reflectivity do not fully represent all objects in our experiments. We suggest that, in addition to the expected decrease in spatial resolution, a greater depth deviation with increasing distance arises from the signal-to-noise ratio with respect to environmental light, which typically decreases over distance due to the inverse-square law.

Notably, RF ToF does not exhibit a distance-dependent depth deviation, unlike its optical counterpart. This seems to contradict Table 1 of the main paper, where δ_z for RF ToF degrades more rapidly with depth, compared to optical sensors; however, the theoretical decay with depth stems from the worsening separability of neighboring point targets, which is not a pertinent scenario in our database, as the recorded targets primarily have smooth surfaces at locations where valid reconstructions are measured. Moreover, our setup minimizes mmWave interactions with external objects, limiting noise primarily to the object itself. As a result, the signal-to-noise ratio of RF ToF is considerably less sensitive to changes in object-to-sensor distance compared to NIR ToF, assuming no interference from external sources.

16 Discussion of the Influence of Geometry on Reconstruction Completeness and Depth Deviation

In Section 7.3 of the main paper, we discussed the influence of object geometry on the RF ToF sensor. Here, we extend our experiments to all four depth imagers.

Influence on Reconstruction Completeness. We visualize the extended results for all depth imagers in Figure 15, where we show the mean depth deviation with respect to C_g in conjunction with the median surface incidence angle. We generally observe lower errors for optical sensors compared to RF ToF, indicating that optical depth measurements tend to be more complete.

During dataset capture, we aligned the object surfaces with the RF ToF sensor aperture. Due to spatial constraints, all optical sensors were placed next to the antenna aperture, resulting in view directions that do not directly align with the majority of surface normals. This is further illustrated in Figure 16 on the *right*, which depicts the sensor placement. Consequently, the object measurements from optical sensors begin at a median surface incidence angle of approximately 20-30°. The absence of a significant proportion of smaller angles in the data complicates the identification of notable trends.

Influence on Depth Deviation. We conduct an additional experiment measuring per-angle depth deviation with respect to metric P, which is generally more sensitive to depth quality and noise than C_g. For each object, we compute the surface incidence angle and depth deviation for each point-wise measurement using the corresponding ground-truth normal. We then aggregate the point-wise measurements across all objects and cluster them into angle bins of 5°. After calculating the mean depth deviation for each angle bin, we normalize the results to $[0, 1]$ across all four sensors.

The findings are depicted in Figure 16, where the length of each bar represents the relative quantity of per-point measurements for each angle bin, offering insight into data distribution. We illustrate the sensor setup on the *right* to clarify each sensor's placement and viewing direction, from which we derive the surface incidence angle.

As previously noted for the median angle measurements, the main lobe of per-angle measurements is concentrated around 30° for optical sensors and 0° for the RF ToF sensor, due to the respective sensor placements. In general, we observe a more rapid decline in depth quality for active sensors (NIR ToF, active stereo, RF ToF) compared to passive stereo, which underscores their dependency on well-illuminated (or well-radiated) areas. Additionally, the depth quality of the RF ToF sensor significantly decreases at angles greater than 30°, rendering it more susceptible to object geometry than optical sensors, where we experience a decline in depth quality at angles greater than 60°.

17 Extended Discussion of MIMO Radar: Signal Response and Depth Deviation

In this section, we extend the experiments of Section 7.3 of the main paper, where we presented both, the signal magnitude and the depth deviation concerning object material, geometry, and size. To demonstrate that the signal magnitude is not only influenced by either object geometry or by object size — which would be possible due to the high diversity of captured objects that prevents us from isolating one variable while keeping the others constant — we visualize the signal magnitude in Figure 17 (*top*), in relation to object geometry (median surface incidence angle) and size (relative surface area).

A general trend on the x-axis shows an increase in the signal magnitude from left to right, particularly for objects with a surface

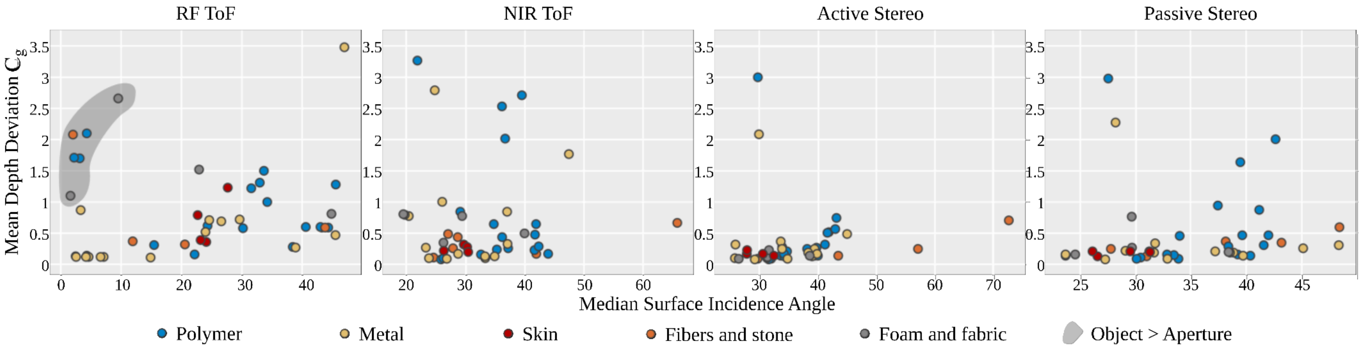


Fig. 15. Extended experiments for all four depth imagers, where object material, geometry (median surface incidence angle), and size (relative surface area) is put in relation to mean depth deviation. Measurements, where large objects appear outside the radar's antenna aperture, are highlighted in gray regions, as they exhibit higher depth deviations compared to the ground-truth reconstructions, which may extend beyond this aperture; this is attributed to the comparably small field of view and the surface reflection characteristics with respect to radio waves (see Section 13.1).

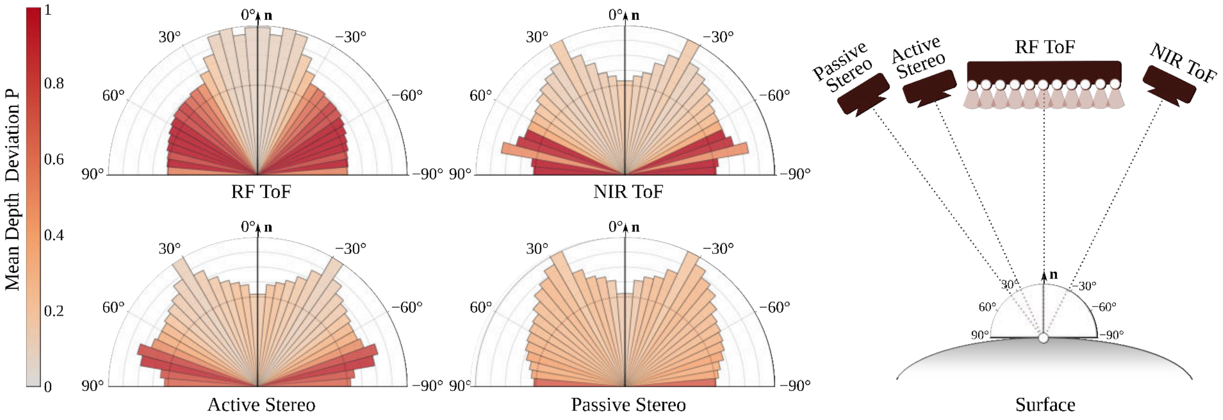


Fig. 16. Depth deviation with respect to metric P per 5°-binned surface incidence angle, normalized to $[0, 1]$ across the four presented depth imagers. The length of each bin indicates the relative data distribution, with the minimum bin length represented as a dotted hemisphere contour. The corresponding sensor setup is shown on the *right*, providing an intuitive understanding of the data distribution in relation to each sensor's viewing direction and the majority of ground-truth surface normals pointing into the direction of $\boldsymbol{n}$.

incidence angle greater than 10°. The majority of objects below this 10° angle on the y-axis exhibit significantly higher signal magnitudes, regardless of their relative surface area, hence indicating the influence of object geometry. On the *bottom* of Figure 17, we visualize the mean depth deviation in relation to object geometry and size. While a similar trend is observed for object geometry on the y-axis — where objects of lower surface incidence angle exhibit smaller errors — no overall trend appears on the x-axis, suggesting depth deviation is more influenced by object geometry instead of size.

17.1 Ablation Study with Reduced Antenna Architectures

To explore alternative consumer-friendly RF ToF devices, we experiment with different architectures, varying the aperture size and antenna density by selecting only a subset of antennas from the raw phasor measurements in MAROON. Two of the selected antenna architectures are shown in Figure 18, both with a comparable number of antennas.

We visually compare the reconstructions of four antenna configurations in Figure 19 for the *Hand Printed Flat* at 30 cm object-to-sensor distance. Reducing the aperture size (*right column*) results in a loss of continuous object geometry, causing the object to visually resemble a collection of point targets. Compared to architectures with larger apertures (*left column*), less of the surface details are preserved. Conversely, a decreasing antenna density introduces more localization ambiguities, leading to multiple reconstructions of the object, as seen with the replicated hands in the *lower left* reconstruction.

Further quantitative results are presented in Table 7, where we measure the mean depth deviation in centimeters across all objects at a distance of 30 cm while varying the antenna architecture. When comparing configurations with similar numbers of antennas, larger

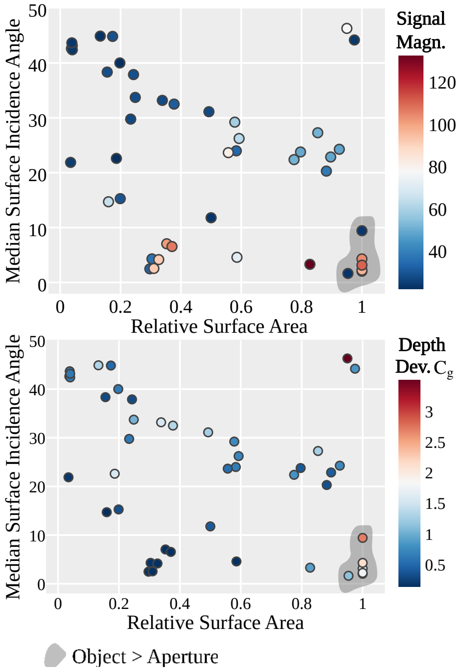


Fig. 17. The mean signal magnitude (*top*) and mean depth deviation (*bottom*), put in relation to object geometry (median surface incidence angle) and object size (relative surface area). Large objects outside the radar's antenna aperture exhibit higher depth deviations with respect to the ground-truth reconstructions that possibly extend beyond this aperture; this is attributed to the comparably small field of view and the surface reflection characteristics with respect to radio waves (see Section 13.1).

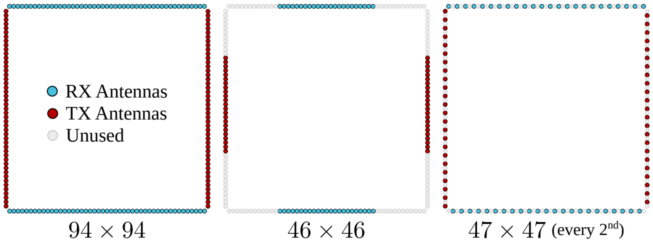


Fig. 18. We simulate more consumer-friendly antenna apertures and compare them to the full 94 × 94 antenna array. First, we reduce the aperture size by selecting a spatially centered antenna subset within the MIMO array. Second, we reduce the antenna density, using only every second antenna from the array, while preserving an aperture size comparable to that of the original array. Additionally, we maintain a similar number of antennas between both consumer-friendly variants to ensure a fair comparison.

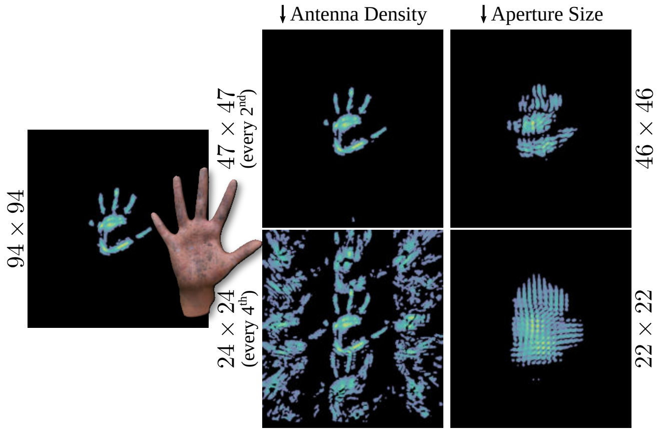


Fig. 19. Qualitative evaluation of different antenna configurations for the *Hand Printed Flat* object at 30 cm object-to-sensor distance, using only 50 % and 25 % of the original number of antennas.

aperture sizes exhibit lower depth deviation than higher-density configurations, as indicated by P and P_e; this supports prior qualitative observations that surface quality declines more rapidly with reduced aperture size. In contrast, decreasing the antenna density leads to a rapid rise in noise, likely due to the previously mentioned localization ambiguities, which tend to appear in areas where holes typically arise, as C_g behaves inversely proportional to C_s.

Antenna Config.	C_g	C_s	P	P_e
46 × 46	1.01	**1.09**	***1.09***	***1.08***
22 × 22	0.82	1.59	1.30	1.24
47 × 47 (every 2nd)	***0.73***	***1.2***	**1.02**	**0.95**
24 × 24 (every 4th)	**0.59**	5.97	1.41	1.35
94 × 94	0.82	0.90	0.94	0.85

Table 7. Ablation study on different antenna configurations. The depth deviation is expressed in centimeters across the four metrics presented in the main paper, averaged for all objects at a 30 cm object-to-sensor distance. The **best** and ***second best*** results per metric are highlighted.

17.2 Ablation Study with Reduced Frequencies

In this experiment, we adjust the frequency configuration of the RF ToF sensor to simulate various sensors. For this, we compute frequency subsets of the raw phasor data provided in the dataset. Following a similar approach to the antenna aperture ablations described in Section 17.1, we implement two key variations: first, by operating with a smaller bandwidth, and second, by varying the frequency differences through subsampling every second, fourth, eighth frequency, and so on.

The results are shown in Table 8, with mean depth deviations quantified in centimeters across all metrics and objects at a distance of 30 cm while varying the frequency configuration. Compared to the full frequency spectrum used in the main paper (*last row*), the 64-frequency-stepped configuration performs on par, indicating that a configuration with half the frequencies may serve as a viable alternative maintaining the same accuracy.

Additionally, we observe a general trend of increasing depth deviation with a decreasing number of frequencies. Among the variations, adjusting the frequency difference rather than the bandwidth yields better overall results and may present an interesting sensor configuration that could enhance the RF ToF sensor's capture rate — as fewer frequencies require less capture time.

Frequency Configuration (in GHz)	C_g	C_s	P	P_e
$f_m \in [72.00, 81.92]$, $\Delta f = 0.16$, $N_f = 64$	***0.81***	***0.91***	***0.95***	***0.86***
$f_m \in [72.00, 81.76]$, $\Delta f = 0.31$, $N_f = 32$	0.90	1.56	1.24	1.15
$f_m \in [72.00, 81.45]$, $\Delta f = 0.59$, $N_f = 16$	0.90	1.56	1.24	1.16
$f_m \in [72.00, 80.82]$, $\Delta f = 1.10$, $N_f = 8$	0.86	3.58	2.49	2.20
$f_m \in [72.00, 79.56]$, $\Delta f = 1.89$, $N_f = 4$	0.91	4.82	4.2	4.04
$f_m \in [76.96, 82.00]$, $\Delta f = 0.078$, $N_f = 64$	**0.76**	1.19	1.21	1.11
$f_m \in [79.48, 82.00]$, $\Delta f = 0.078$, $N_f = 32$	**0.76**	1.98	1.85	1.67
$f_m \in [80.74, 82.00]$, $\Delta f = 0.078$, $N_f = 16$	0.77	3.33	2.87	2.71
$f_m \in [81.37, 82.00]$, $\Delta f = 0.078$, $N_f = 8$	0.91	5.15	4.14	3.97
$f_m \in [81.69, 82.00]$, $\Delta f = 0.078$, $N_f = 4$	0.98	6.22	4.82	4.71
$f_m \in [72.00, 82.00]$, $\Delta f = 0.078$, $N_f = 128$	0.82	**0.90**	**0.94**	**0.85**

Table 8. Ablation study on various frequency configurations defined by the range of the modulation frequency f_m, frequency difference Δf, and the corresponding number of frequency steps N_f. Frequency units are given in GHz. The *last row* depicts the full-bandwidth frequency configuration from the main paper. The depth deviation is expressed in centimeters, averaged for all objects at a 30 cm object-to-sensor distance. The **best** and ***second best*** results per metric are highlighted.

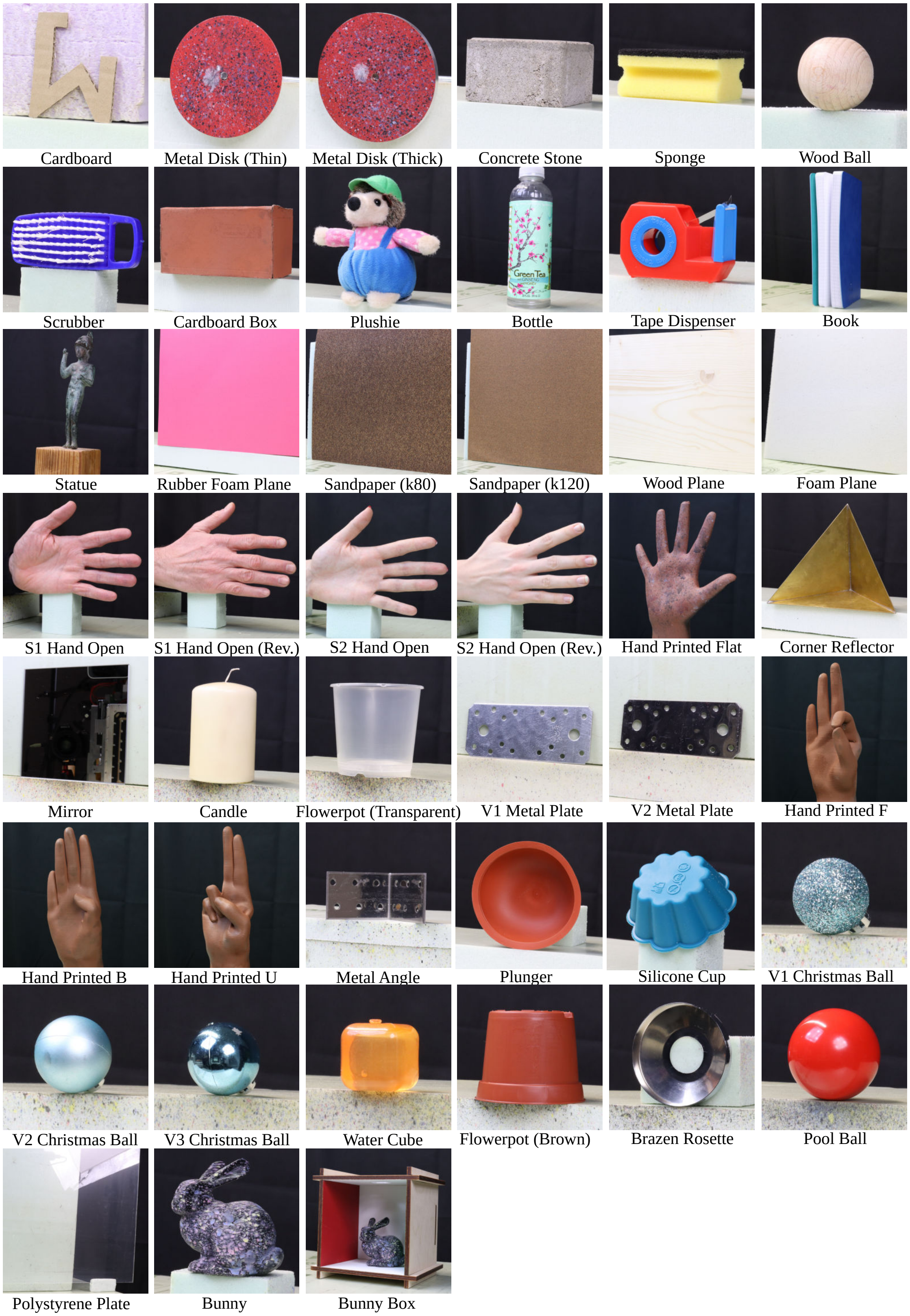


Fig. 20. Example images of all objects in MAROON.

Class	Objects	Additional Description
Metal	V1 Metal Plate, V2 Metal Plate	
	Metal Disk (Thin), Metal Disk (Thick)	
	Hand Printed: Flat, B, F, U	Coated with metal lacquer
	Brazen Rosette	
	Corner Reflector	
	Cardboard Box	Coated with metal lacquer
	Mirror	Metal surface beneath partially transmissive glass
	Metal Angle	
	Statue	
Fibers and stone	Wood Plane	
	Cardboard	
	Book	Primarily made out of paper
	Concrete Stone	
	Wood Ball	
	Bunny Box	Large wooden box, comparably small plastic bunny
Polymer	Plumber	
	Silicone Cup	
	Christmas Ball: V1, V2, V3	
	Candle	
	Bottle	
	Sandpaper (k120), Sandpaper (k80)	
	Flowerpot (Transparent), Flowerpot (Brown)	
	Polystyrene Plate	
	Water Cube	Water wrapped in a plastic cube
	Scrubber	
	Pool Ball	
	Bunny	
	Tape Dispenser	
Skin	S1 Hand Open, S1 Hand Open (Rev.)	
	S2 Hand Open, S2 Hand Open (Rev.)	
Foam and fabric (Primarily transmissive)	Rubber Foam Plane	
	Sponge	
	Plushie	
	Foam Plane	
Objects outside FOV	Polystyrene Plate	
	Sandpaper (k120), Sandpaper (k80)	
	Wood Plane	
	Foam Plane	
	Rubber Foam Plane	

Table 9. Assignment from objects to material classes with optional description about the assignment process.

	Metric Type	Cardboard	Metal Disk (Thin)	Metal Disk (Thick)	Concrete Stone	Sponge	Wood Ball
RF ToF	C_g	0.13 (± 0.06)	0.12 (± 0.04)	0.12 (± **0.05**)	0.14 (± 0.07)	1.52 (± 0.97)	0.59 (± 0.40)
NIR ToF		0.10 (± 0.06)	0.09 (± **0.03**)	0.08 (± **0.05**)	0.09 (± **0.04**)	0.79 (± 0.45)	0.17 (± 0.12)
Active Stereo		**0.08** (± **0.05**)	**0.06** (± 0.04)	**0.07** (± **0.05**)	**0.08** (± **0.04**)	**0.18** (± 0.17)	**0.13** (± **0.06**)
Passive Stereo		0.24 (± 0.11)	0.07 (± 0.04)	0.08 (± 0.07)	0.09 (± 0.06)	0.26 (± **0.15**)	0.34 (± 0.16)
RF ToF	C_s	0.15 (± 0.08)	0.13 (± **0.05**)	0.13 (± **0.06**)	0.15 (± 0.08)	0.54 (± 0.40)	**0.12** (± **0.05**)
NIR ToF		0.16 (± 0.18)	0.12 (± 0.13)	0.14 (± 0.20)	0.10 (± 0.06)	0.79 (± 0.42)	0.21 (± 0.19)
Active Stereo		**0.08** (± **0.05**)	**0.07** (± **0.05**)	**0.08** (± 0.07)	**0.08** (± **0.04**)	**0.17** (± **0.15**)	0.13 (± 0.07)
Passive Stereo		0.30 (± 0.13)	0.12 (± 0.17)	0.14 (± 0.33)	0.12 (± 0.08)	0.33 (± 0.18)	0.40 (± 0.20)
RF ToF	P	0.13 (± 0.14)	0.10 (± **0.11**)	0.11 (± **0.14**)	0.13 (± 0.15)	2.73 (± 2.47)	**0.10** (± **0.09**)
NIR ToF		0.19 (± 0.25)	0.13 (± 0.19)	0.16 (± 0.28)	**0.10** (± **0.10**)	1.32 (± 0.58)	0.32 (± 0.44)
Active Stereo		**0.10** (± **0.12**)	**0.08** (± 0.25)	**0.10** (± 0.17)	**0.10** (± 0.12)	**0.29** (± **0.42**)	0.20 (± 0.17)
Passive Stereo		0.36 (± 0.17)	0.14 (± 0.32)	0.18 (± 0.51)	0.16 (± 0.17)	0.47 (± 0.51)	0.65 (± 0.33)
RF ToF	P_e	0.12 (± 0.13)	0.10 (± 0.08)	0.10 (± 0.09)	0.13 (± 0.15)	2.93 (± 2.60)	**0.10** (± 0.09)
NIR ToF		0.08 (± 0.10)	0.12 (± **0.06**)	0.09 (± **0.06**)	**0.09** (± **0.07**)	1.69 (± **0.23**)	**0.10** (± 0.10)
Active Stereo		**0.06** (± **0.08**)	**0.06** (± **0.06**)	**0.06** (± **0.06**)	**0.09** (± 0.10)	**0.42** (± 0.50)	0.16 (± **0.07**)
Passive Stereo		0.38 (± 0.14)	0.08 (± 0.11)	0.08 (± 0.10)	0.10 (± 0.10)	0.55 (± 0.56)	0.69 (± 0.26)
RF ToF	P^*	+0.08 (± 0.14)	-0.06 (± **0.11**)	-0.05 (± **0.14**)	-0.06 (± 0.15)	+2.25 (± 2.47)	**-0.07** (± **0.09**)
NIR ToF		+0.15 (± 0.25)	-0.04 (± 0.19)	**-0.01** (± 0.28)	-0.08 (± **0.10**)	+1.30 (± 0.58)	+0.09 (± 0.44)
Active Stereo		**+0.04** (± **0.12**)	**-0.01** (± 0.25)	-0.02 (± 0.17)	**-0.04** (± 0.12)	**-0.15** (± **0.42**)	+0.17 (± 0.17)
Passive Stereo		+0.36 (± 0.17)	+0.07 (± 0.32)	+0.10 (± 0.51)	+0.15 (± 0.17)	+0.22 (± 0.51)	+0.64 (± 0.33)
RF ToF	$P_e{}^*$	+0.08 (± 0.13)	-0.09 (± 0.08)	-0.08 (± 0.09)	-0.07 (± 0.15)	+2.40 (± 2.60)	-0.07 (± 0.09)
NIR ToF		**-0.00** (± 0.10)	-0.12 (± **0.06**)	-0.08 (± **0.06**)	-0.08 (± **0.07**)	+1.69 (± **0.23**)	**-0.05** (± 0.10)
Active Stereo		**+0.00** (± **0.08**)	-0.04 (± **0.06**)	-0.05 (± **0.06**)	**-0.05** (± 0.10)	-0.37 (± 0.50)	+0.16 (± **0.07**)
Passive Stereo		+0.38 (± 0.14)	**+0.01** (± 0.11)	**+0.02** (± 0.10)	+0.08 (± 0.10)	**-0.34** (± 0.56)	+0.69 (± 0.26)

	Metric Type	Scrubber	Cardboard Box	Plushie	Bottle	Tape Dispenser	Book
RF ToF	C_g	0.58 (± 0.29)	0.12 (± 0.08)	0.81 (± 0.47)	1.22 (± 1.00)	0.31 (± 0.23)	0.37 (± 0.56)
NIR ToF		0.64 (± 0.34)	**0.11** (± **0.04**)	0.49 (± 0.21)	0.44 (± 0.51)	0.84 (± 0.30)	0.59 (± 0.65)
Active Stereo		0.20 (± 0.16)	0.12 (± 0.07)	**0.13** (± **0.14**)	**0.16** (± **0.21**)	0.16 (± 0.14)	**0.13** (± **0.11**)
Passive Stereo		**0.14** (± **0.11**)	0.29 (± 0.20)	0.19 (± 0.21)	0.28 (± 0.36)	**0.15** (± **0.11**)	0.15 (± 0.13)
RF ToF	C_s	0.97 (± 0.61)	**0.11** (± **0.05**)	2.21 (± 2.05)	0.49 (± 0.99)	0.55 (± 0.77)	**0.14** (± **0.08**)
NIR ToF		0.59 (± 0.30)	0.14 (± 0.16)	0.53 (± 0.28)	0.80 (± 1.32)	1.16 (± 0.60)	0.20 (± 0.17)
Active Stereo		**0.17** (± **0.11**)	0.15 (± 0.13)	**0.12** (± 0.38)	**0.20** (± **0.53**)	**0.17** (± **0.17**)	0.18 (± 0.15)
Passive Stereo		0.19 (± 0.13)	0.35 (± 0.24)	0.23 (± **0.27**)	0.41 (± 0.62)	0.22 (± 0.20)	0.24 (± 0.19)
RF ToF	P	1.28 (± 0.76)	0.38 (± 1.52)	3.64 (± 3.23)	0.64 (± 1.63)	0.67 (± 1.08)	**0.13** (± 0.31)
NIR ToF		0.91 (± 0.47)	**0.14** (± **0.19**)	0.85 (± **0.52**)	1.32 (± 2.00)	1.57 (± 0.70)	1.15 (± 1.42)
Active Stereo		**0.27** (± **0.34**)	0.19 (± 0.22)	**0.24** (± 1.25)	**0.49** (± 2.49)	**0.24** (± **0.35**)	0.19 (± **0.25**)
Passive Stereo		0.29 (± 0.38)	0.47 (± 0.55)	0.35 (± 0.54)	0.62 (± **0.97**)	0.29 (± **0.35**)	0.47 (± 1.21)
RF ToF	P_e	1.35 (± 0.66)	0.25 (± 1.19)	3.61 (± 3.23)	0.63 (± 1.63)	0.66 (± 1.07)	**0.13** (± 0.30)
NIR ToF		1.16 (± **0.27**)	**0.13** (± **0.07**)	0.82 (± 0.38)	1.16 (± 2.16)	1.70 (± 0.89)	1.34 (± 1.54)
Active Stereo		**0.22** (± **0.27**)	0.18 (± 0.12)	**0.16** (± 0.30)	**0.33** (± 1.13)	**0.15** (± **0.21**)	0.17 (± **0.23**)
Passive Stereo		0.24 (± 0.29)	0.42 (± 0.28)	0.18 (± **0.26**)	0.55 (± **0.86**)	0.20 (± 0.22)	0.33 (± 0.95)
RF ToF	P^*	+1.26 (± 0.76)	+0.28 (± 1.52)	+3.63 (± 3.23)	+0.52 (± 1.63)	+0.56 (± 1.08)	-0.10 (± 0.31)
NIR ToF		+0.89 (± 0.47)	**-0.09** (± **0.19**)	+0.80 (± **0.52**)	+1.30 (± 2.00)	+1.57 (± 0.70)	+1.13 (± 1.42)
Active Stereo		**-0.01** (± **0.34**)	-0.10 (± 0.22)	**-0.04** (± 1.25)	**+0.18** (± 2.49)	**+0.01** (± **0.35**)	**-0.07** (± **0.25**)
Passive Stereo		**-0.01** (± 0.38)	+0.42 (± 0.55)	+0.17 (± 0.54)	+0.51 (± **0.97**)	+0.19 (± **0.35**)	+0.39 (± 1.21)
RF ToF	$P_e{}^*$	+1.34 (± 0.66)	+0.14 (± 1.19)	+3.59 (± 3.23)	+0.52 (± 1.63)	+0.57 (± 1.07)	-0.10 (± 0.30)
NIR ToF		+1.16 (± **0.27**)	**-0.13** (± **0.07**)	+0.82 (± 0.38)	+1.15 (± 2.16)	+1.70 (± 0.89)	+1.33 (± 1.54)
Active Stereo		**-0.02** (± **0.27**)	-0.16 (± 0.12)	-0.06 (± 0.30)	**+0.21** (± 1.13)	**+0.03** (± **0.21**)	**-0.02** (± **0.23**)
Passive Stereo		-0.07 (± 0.29)	+0.42 (± 0.28)	**+0.02** (± **0.26**)	+0.52 (± **0.86**)	+0.08 (± 0.22)	+0.27 (± 0.95)

Table 10. We measure the depth deviation with respect to C_g, C_s, P, P_e and an additional signed version of P,P_e, which is denoted as P^*,$P_e{}^*$. All metrics are listed in the form ($\mu \pm \sigma$), consisting of the mean μ and standard deviation σ in centimeters, computed over the entire metric domain, respectively. The best results among all sensors of one metric type are highlighted in **bold** and the worst results are underlined.

.

	Metric Type	Statue	Rubber Foam Plane	Sandpaper (k80)	Sandpaper (k120)	Wood Plane	Foam Plane
RF ToF	C_g	0.27 (± 0.25)	1.10 (± 1.21)	1.70 (± 2.07)	1.71 (± 2.23)	2.08 (± 2.31)	2.66 (± 1.26)
NIR ToF		0.32 (± 0.28)	0.34 (± 0.07)	0.09 (± **0.04**)	**0.07** (± **0.03**)	0.48 (± 0.15)	0.80 (± 0.14)
Active Stereo		0.16 (± **0.13**)	**0.11** (± **0.05**)	**0.07** (± **0.04**)	0.08 (± 0.04)	0.13 (± **0.06**)	**0.08** (± **0.06**)
Passive Stereo		**0.13** (± **0.13**)	0.77 (± 0.67)	0.08 (± **0.04**)	0.10 (± 0.06)	**0.12** (± 0.10)	0.15 (± 0.09)
RF ToF	C_s	**0.17** (± **0.11**)	0.34 (± 0.72)	0.12 (± **0.05**)	0.12 (± **0.05**)	0.20 (± 0.12)	2.83 (± 0.89)
NIR ToF		0.43 (± 0.42)	0.38 (± 0.11)	0.11 (± 0.12)	**0.09** (± 0.12)	0.52 (± 0.16)	0.84 (± 0.15)
Active Stereo		0.19 (± 0.76)	**0.13** (± **0.08**)	**0.08** (± **0.05**)	**0.09** (± **0.05**)	0.15 (± **0.08**)	**0.09** (± 0.15)
Passive Stereo		0.18 (± 0.69)	0.83 (± 0.71)	0.12 (± 0.09)	0.11 (± 0.08)	**0.13** (± 0.11)	0.16 (± **0.12**)
RF ToF	P	**0.20** (± **0.26**)	1.62 (± 2.21)	**0.09** (± 0.11)	**0.08** (± 0.10)	0.18 (± 0.23)	3.95 (± 4.22)
NIR ToF		0.77 (± 3.13)	0.44 (± **0.19**)	0.10 (± 0.14)	0.09 (± 0.14)	0.65 (± 0.28)	1.13 (± 3.58)
Active Stereo		0.90 (± 5.17)	**0.16** (± 0.24)	**0.09** (± **0.09**)	0.10 (± **0.09**)	0.20 (± **0.12**)	0.21 (± 2.11)
Passive Stereo		1.43 (± 7.09)	0.91 (± 1.11)	0.13 (± 0.17)	0.13 (± 0.14)	**0.14** (± 0.18)	**0.19** (± **0.44**)
RF ToF	P_e	0.20 (± 0.27)	1.62 (± 2.21)	0.09 (± 0.11)	**0.08** (± 0.10)	0.18 (± 0.23)	3.95 (± 4.22)
NIR ToF		0.25 (± **0.13**)	0.43 (± 0.11)	**0.08** (± 0.10)	**0.08** (± 0.10)	0.64 (± 0.27)	0.91 (± 0.12)
Active Stereo		0.16 (± 0.21)	**0.16** (± **0.09**)	0.09 (± **0.08**)	0.10 (± **0.09**)	0.20 (± **0.10**)	**0.09** (± **0.10**)
Passive Stereo		**0.10** (± **0.13**)	0.94 (± 1.13)	0.12 (± 0.15)	0.11 (± 0.12)	**0.12** (± 0.13)	0.17 (± 0.20)
RF ToF	P^*	**-0.04** (± **0.26**)	-1.35 (± 2.21)	-0.04 (± 0.11)	-0.04 (± 0.10)	**+0.03** (± 0.23)	-0.66 (± 4.22)
NIR ToF		-0.23 (± 3.13)	+0.43 (± **0.19**)	-0.03 (± 0.14)	**+0.02** (± 0.14)	+0.65 (± 0.28)	+1.13 (± 3.58)
Active Stereo		+0.71 (± 5.17)	**-0.15** (± 0.24)	-0.06 (± **0.09**)	-0.08 (± **0.09**)	-0.19 (± **0.12**)	**+0.08** (± 2.11)
Passive Stereo		+1.34 (± 7.09)	+0.34 (± 1.11)	**+0.00** (± 0.17)	-0.08 (± 0.14)	+0.08 (± 0.18)	**+0.08** (± **0.44**)
RF ToF	P_e^*	-0.04 (± 0.27)	-1.35 (± 2.21)	-0.04 (± 0.11)	-0.04 (± 0.10)	**+0.03** (± 0.23)	-0.66 (± 4.22)
NIR ToF		-0.24 (± **0.13**)	+0.43 (± 0.11)	-0.04 (± 0.10)	**+0.00** (± 0.10)	+0.64 (± 0.27)	+0.91 (± 0.12)
Active Stereo		+0.07 (± 0.21)	**-0.16** (± **0.09**)	-0.06 (± **0.08**)	-0.09 (± **0.09**)	-0.20 (± **0.10**)	**-0.05** (± **0.10**)
Passive Stereo		**+0.03** (± **0.13**)	+0.35 (± 1.13)	**+0.01** (± 0.15)	-0.07 (± 0.12)	+0.07 (± 0.13)	+0.07 (± 0.20)

	Metric Type	S1 Hand Open	S1 Hand Open (Rev.)	S2 Hand Open	S2 Hand Open (Rev.)	Hand Printed Flat	Corner Reflector
RF ToF	C_g	0.36 (± 0.38)	1.23 (± 1.42)	0.39 (± 0.37)	0.79 (± 0.79)	0.71 (± 0.78)	3.48 (± 1.80)
NIR ToF		0.31 (± 0.14)	0.27 (± 0.12)	0.19 (± **0.10**)	0.21 (± **0.11**)	0.25 (± 0.12)	1.81 (± 1.01)
Active Stereo		**0.12** (± **0.09**)	0.16 (± 0.15)	**0.16** (± 0.13)	0.21 (± 0.16)	**0.09** (± **0.07**)	0.48 (± 0.62)
Passive Stereo		0.20 (± 0.16)	**0.12** (± **0.09**)	0.20 (± 0.16)	**0.20** (± 0.15)	0.21 (± 0.37)	**0.30** (± **0.29**)
RF ToF	C_s	0.22 (± 0.15)	**0.17** (± **0.11**)	0.20 (± **0.14**)	**0.15** (± **0.09**)	0.17 (± 0.13)	1.95 (± **0.94**)
NIR ToF		0.38 (± 0.26)	0.32 (± 0.24)	0.25 (± 0.23)	0.27 (± 0.24)	0.29 (± 0.20)	1.97 (± 1.09)
Active Stereo		**0.13** (± **0.10**)	0.18 (± 0.73)	**0.17** (± **0.14**)	0.22 (± 0.15)	**0.09** (± **0.06**)	**0.55** (± 1.54)
Passive Stereo		0.26 (± 0.22)	**0.17** (± 0.18)	0.25 (± 0.20)	0.25 (± 0.19)	0.18 (± 0.34)	0.61 (± 1.22)
RF ToF	P	**0.22** (± **0.25**)	**0.16** (± **0.21**)	**0.20** (± **0.24**)	**0.14** (± **0.17**)	**0.16** (± **0.20**)	3.31 (± **1.53**)
NIR ToF		0.52 (± 0.43)	0.47 (± 0.44)	0.39 (± 0.57)	0.35 (± 0.39)	0.33 (± 0.29)	2.87 (± 1.67)
Active Stereo		**0.22** (± 1.25)	0.30 (± 1.54)	0.29 (± 1.67)	0.33 (± 1.09)	**0.16** (± 1.30)	1.06 (± 3.92)
Passive Stereo		0.35 (± 0.41)	0.22 (± 0.38)	0.35 (± 0.85)	0.32 (± 0.48)	1.73 (± 8.95)	**0.99** (± 2.49)
RF ToF	P_e	0.22 (± 0.25)	**0.16** (± 0.20)	**0.20** (± 0.24)	**0.14** (± **0.16**)	0.16 (± 0.20)	3.14 (± 1.41)
NIR ToF		0.51 (± 0.27)	0.38 (± 0.21)	0.30 (± **0.22**)	0.25 (± 0.17)	0.30 (± **0.09**)	2.65 (± 1.29)
Active Stereo		**0.16** (± **0.24**)	0.20 (± 0.25)	**0.20** (± 0.23)	0.26 (± 0.30)	**0.08** (± 0.10)	**0.70** (± **0.92**)
Passive Stereo		0.25 (± 0.34)	**0.16** (± **0.17**)	0.21 (± 0.25)	0.25 (± 0.21)	0.17 (± 0.15)	1.26 (± 1.63)
RF ToF	P^*	-0.15 (± **0.25**)	**-0.06** (± **0.21**)	-0.10 (± **0.24**)	**-0.07** (± **0.17**)	-0.07 (± **0.20**)	+3.31 (± **1.53**)
NIR ToF		+0.49 (± 0.43)	+0.42 (± 0.44)	+0.32 (± 0.57)	+0.29 (± 0.39)	-0.30 (± 0.29)	+2.86 (± 1.67)
Active Stereo		**+0.06** (± 1.25)	-0.07 (± 1.54)	**-0.04** (± 1.67)	+0.12 (± 1.09)	**+0.04** (± 1.30)	**+0.82** (± 3.92)
Passive Stereo		+0.22 (± 0.41)	+0.17 (± 0.38)	+0.32 (± 0.85)	+0.29 (± 0.48)	+1.70 (± 8.95)	+0.88 (± 2.49)
RF ToF	P_e^*	-0.14 (± 0.25)	**-0.06** (± 0.20)	**-0.10** (± 0.24)	**-0.07** (± **0.16**)	-0.07 (± 0.20)	+3.14 (± 1.41)
NIR ToF		+0.51 (± 0.27)	+0.38 (± 0.21)	+0.28 (± **0.22**)	+0.24 (± 0.17)	-0.30 (± **0.09**)	+2.65 (± 1.29)
Active Stereo		**+0.00** (± **0.24**)	-0.13 (± 0.25)	-0.14 (± 0.23)	+0.11 (± 0.30)	**+0.00** (± 0.10)	**+0.45** (± **0.92**)
Passive Stereo		+0.05 (± 0.34)	+0.11 (± **0.17**)	+0.15 (± 0.25)	+0.23 (± 0.21)	+0.16 (± 0.15)	+1.23 (± 1.63)

Table 11. We measure the depth deviation with respect to C_g, C_s, P, P_e and an additional signed version of P,P_e, which is denoted as P^*,P_e^*. All metrics are listed in the form ($\mu \pm \sigma$), consisting of the mean μ and standard deviation σ in centimeters, computed over the entire metric domain, respectively. The best results among all sensors of one metric type are highlighted in **bold** and the worst results are underlined.

	Metric Type	Mirror	Candle	Flowerpot (Transparent)	V1 Metal Plate	V2 Metal Plate	Hand Printed F
RF ToF	C_g	**0.87** (± **0.26**)	1.50 (± 1.12)	1.31 (± 1.21)	0.12 (± **0.05**)	**0.12** (± **0.05**)	0.69 (± 0.86)
NIR ToF		3.77 (± 1.97)	2.04 (± 0.40)	2.73 (± 1.03)	0.77 (± 0.42)	0.74 (± 0.45)	**0.12** (± **0.09**)
Active Stereo		2.13 (± 1.52)	**0.26** (± **0.29**)	**0.74** (± **0.53**)	**0.08** (± 0.06)	0.30 (± 0.29)	0.17 (± 0.18)
Passive Stereo		2.31 (± 1.61)	1.64 (± 0.78)	2.01 (± 0.83)	0.13 (± 0.07)	0.15 (± 0.11)	0.20 (± 0.14)
RF ToF	C_s	**0.91** (± **0.14**)	5.57 (± 2.78)	1.86 (± 2.41)	0.13 (± **0.06**)	**0.13** (± **0.07**)	0.15 (± **0.10**)
NIR ToF		33.31 (± 9.07)	1.71 (± 0.49)	3.10 (± 1.22)	0.81 (± 0.43)	15.66 (± 17.32)	**0.12** (± **0.10**)
Active Stereo		30.21 (± 14.59)	**0.25** (± **0.26**)	**1.27** (± 1.78)	**0.09** (± 0.07)	5.54 (± 12.95)	0.21 (± 0.63)
Passive Stereo		27.02 (± 11.33)	1.28 (± 0.65)	1.86 (± **0.93**)	0.16 (± 0.11)	0.20 (± 0.14)	0.23 (± 0.16)
RF ToF	P	**0.93** (± **0.12**)	7.41 (± 3.79)	2.74 (± 3.66)	0.11 (± **0.12**)	**0.11** (± **0.12**)	**0.14** (± **0.20**)
NIR ToF		37.84 (± 14.84)	2.78 (± **0.35**)	5.24 (± 2.04)	0.95 (± 0.48)	16.23 (± 18.21)	0.17 (± 0.30)
Active Stereo		39.66 (± 24.75)	**0.42** (± 0.49)	**2.08** (± 2.30)	**0.10** (± 0.13)	6.16 (± 13.78)	0.56 (± 2.02)
Passive Stereo		30.82 (± 14.01)	2.10 (± 0.98)	3.50 (± **1.37**)	0.19 (± 0.15)	0.22 (± 0.20)	0.43 (± 0.73)
RF ToF	P_e	**0.93** (± **0.12**)	7.37 (± 3.85)	2.76 (± 3.66)	0.10 (± 0.11)	**0.11** (± **0.12**)	**0.13** (± 0.20)
NIR ToF		39.68 (± 6.57)	2.75 (± **0.15**)	6.18 (± 1.79)	0.79 (± 0.39)	24.92 (± 17.59)	**0.13** (± **0.16**)
Active Stereo		43.84 (± 20.28)	**0.31** (± 0.44)	**2.52** (± 1.77)	**0.07** (± **0.09**)	7.46 (± 14.94)	0.53 (± 1.07)
Passive Stereo		35.96 (± 7.83)	2.15 (± 0.65)	4.36 (± **0.71**)	0.15 (± **0.09**)	0.23 (± 0.19)	0.34 (± 0.65)
RF ToF	P^*	**+0.93** (± **0.12**)	+7.38 (± 3.79)	+2.68 (± 3.66)	-0.06 (± **0.12**)	**-0.06** (± **0.12**)	**-0.02** (± **0.20**)
NIR ToF		+37.84 (± 14.84)	+2.78 (± **0.35**)	+5.24 (± 2.04)	+0.95 (± 0.48)	+16.21 (± 18.21)	-0.11 (± 0.30)
Active Stereo		+39.58 (± 24.75)	**+0.37** (± 0.49)	**+2.00** (± 2.30)	**+0.00** (± 0.13)	+5.77 (± 13.78)	-0.03 (± 2.02)
Passive Stereo		+30.80 (± 14.01)	+2.10 (± 0.98)	+3.49 (± **1.37**)	+0.17 (± 0.15)	+0.19 (± 0.20)	**+0.02** (± 0.73)
RF ToF	$P_e{}^*$	**+0.93** (± **0.12**)	+7.34 (± 3.85)	+2.70 (± 3.66)	-0.06 (± 0.11)	**-0.06** (± **0.12**)	**-0.01** (± 0.20)
NIR ToF		+39.68 (± 6.57)	+2.75 (± **0.15**)	+6.18 (± 1.79)	+0.79 (± 0.39)	+24.90 (± 17.59)	-0.11 (± **0.16**)
Active Stereo		+43.84 (± 20.28)	**+0.27** (± 0.44)	**+2.52** (± 1.77)	**+0.01** (± **0.09**)	+7.08 (± 14.94)	-0.36 (± 1.07)
Passive Stereo		+35.96 (± 7.83)	+2.15 (± 0.65)	+4.36 (± **0.71**)	+0.15 (± **0.09**)	+0.21 (± 0.19)	-0.07 (± 0.65)

	Metric Type	Hand Printed B	Hand Printed U	Metal Angle	Plunger	Silicone Cup	V1 Christmas Ball
RF ToF	C_g	0.52 (± 0.70)	0.72 (± 0.80)	0.47 (± 0.28)	0.62 (± 0.60)	0.60 (± 0.47)	0.59 (± 0.40)
NIR ToF		**0.09** (± **0.06**)	**0.12** (± **0.09**)	0.90 (± 0.79)	**0.23** (± **0.13**)	0.16 (± **0.10**)	0.22 (± 0.14)
Active Stereo		0.12 (± 0.11)	0.15 (± 0.12)	**0.23** (± 0.18)	0.24 (± 0.17)	**0.13** (± **0.10**)	**0.13** (± 0.09)
Passive Stereo		0.19 (± 0.16)	0.18 (± 0.14)	0.24 (± **0.17**)	0.95 (± 0.64)	0.15 (± **0.10**)	**0.13** (± **0.05**)
RF ToF	C_s	0.17 (± **0.12**)	0.18 (± **0.13**)	1.20 (± 0.87)	1.18 (± 2.39)	1.43 (± 1.49)	0.75 (± 0.95)
NIR ToF		**0.10** (± **0.12**)	**0.14** (± 0.15)	1.02 (± 0.54)	**0.28** (± **0.19**)	0.21 (± 0.22)	0.41 (± 0.34)
Active Stereo		0.13 (± 0.25)	0.17 (± 0.51)	**0.29** (± **0.32**)	0.29 (± 0.32)	**0.13** (± **0.10**)	**0.14** (± **0.09**)
Passive Stereo		0.20 (± 0.21)	0.22 (± 0.22)	0.49 (± 0.57)	0.87 (± 0.63)	0.17 (± 0.14)	0.24 (± 0.15)
RF ToF	P	**0.17** (± **0.23**)	**0.19** (± **0.28**)	1.71 (± 1.47)	1.47 (± 2.94)	2.33 (± 2.79)	1.05 (± 1.41)
NIR ToF		**0.17** (± 0.33)	0.23 (± 0.45)	1.68 (± 1.29)	**0.42** (± **0.54**)	0.30 (± 0.38)	0.62 (± 0.64)
Active Stereo		0.26 (± 1.15)	0.46 (± 2.16)	**0.49** (± **0.69**)	0.55 (± 0.96)	**0.20** (± **0.28**)	**0.22** (± 0.25)
Passive Stereo		0.33 (± 0.54)	0.37 (± 0.62)	0.76 (± 0.88)	1.33 (± 1.35)	0.27 (± 0.37)	0.35 (± **0.22**)
RF ToF	P_e	0.17 (± 0.23)	0.19 (± 0.28)	1.79 (± 1.47)	1.62 (± 3.10)	2.55 (± 2.95)	1.05 (± 1.41)
NIR ToF		**0.08** (± **0.11**)	**0.16** (± **0.22**)	1.40 (± 0.76)	**0.27** (± **0.25**)	0.23 (± 0.27)	0.36 (± 0.37)
Active Stereo		0.18 (± 0.25)	0.28 (± 0.45)	**0.52** (± 0.74)	0.58 (± 0.78)	**0.16** (± **0.16**)	**0.15** (± **0.19**)
Passive Stereo		0.28 (± 0.40)	0.40 (± 0.58)	0.67 (± **0.50**)	1.71 (± 0.66)	0.25 (± 0.21)	0.44 (± 0.23)
RF ToF	P^*	**-0.02** (± **0.23**)	**+0.00** (± **0.28**)	+1.56 (± 1.47)	+0.89 (± 2.94)	+2.08 (± 2.79)	+0.96 (± 1.41)
NIR ToF		**+0.02** (± 0.33)	-0.01 (± 0.45)	+1.66 (± 1.29)	**+0.31** (± **0.54**)	+0.22 (± 0.38)	+0.50 (± 0.64)
Active Stereo		-0.03 (± 1.15)	+0.15 (± 2.16)	**+0.40** (± **0.69**)	-0.32 (± 0.96)	**+0.05** (± **0.28**)	**+0.18** (± 0.25)
Passive Stereo		+0.08 (± 0.54)	+0.15 (± 0.62)	+0.72 (± 0.88)	-0.89 (± 1.35)	+0.17 (± 0.37)	+0.34 (± **0.22**)
RF ToF	$P_e{}^*$	-0.01 (± 0.23)	**+0.00** (± 0.28)	+1.63 (± 1.47)	+0.96 (± 3.10)	+2.27 (± 2.95)	+0.96 (± 1.41)
NIR ToF		-0.02 (± **0.11**)	-0.04 (± **0.22**)	+1.37 (± 0.76)	**+0.26** (± **0.25**)	+0.19 (± 0.27)	+0.30 (± 0.37)
Active Stereo		-0.08 (± 0.25)	-0.02 (± 0.45)	**+0.39** (± 0.74)	-0.55 (± 0.78)	**+0.12** (± **0.16**)	**+0.14** (± **0.19**)
Passive Stereo		**-0.00** (± 0.40)	+0.05 (± 0.58)	+0.65 (± **0.50**)	-1.71 (± 0.66)	+0.22 (± 0.21)	+0.44 (± 0.23)

Table 12. We measure the depth deviation with respect to C_g, C_s, P, P_e and an additional signed version of P,P_e, which is denoted as P^*,$P_e{}^*$. All metrics are listed in the form ($\mu \pm \sigma$), consisting of the mean μ and standard deviation σ in centimeters, computed over the entire metric domain, respectively. The best results among all sensors of one metric type are highlighted in **bold** and the worst results are underlined.

	Metric Type	V2 Christmas Ball	V3 Christmas Ball	Water Cube	Flowerpot (Brown)	Brazen Rosette	Pool Ball
RF ToF	C_g	0.59 (± 0.40)	0.60 (± 0.41)	**0.16** (± **0.11**)	1.00 (± 0.90)	**0.11** (± **0.08**)	1.28 (± 0.78)
NIR ToF		**0.28** (± 0.18)	0.47 (± 0.24)	3.00 (± 0.44)	0.15 (± **0.08**)	0.94 (± 0.35)	0.64 (± **0.17**)
Active Stereo		0.51 (± 0.22)	0.50 (± 0.20)	0.56 (± 0.17)	**0.12** (± 0.21)	0.36 (± 0.24)	**0.31** (± 0.25)
Passive Stereo		0.46 (± **0.17**)	**0.30** (± **0.13**)	0.46 (± 0.24)	0.45 (± 0.25)	0.18 (± 0.13)	0.87 (± 0.30)
RF ToF	C_s	**0.10** (± **0.03**)	**0.10** (± **0.03**)	**0.12** (± **0.06**)	0.53 (± 1.23)	**0.11** (± **0.05**)	**0.09** (± **0.03**)
NIR ToF		0.80 (± 0.80)	1.66 (± 1.65)	2.88 (± 0.54)	0.22 (± 0.27)	8.11 (± 11.87)	0.69 (± 0.35)
Active Stereo		0.43 (± 0.23)	0.39 (± 0.22)	0.54 (± 0.27)	**0.10** (± **0.09**)	0.46 (± 0.45)	0.33 (± 0.27)
Passive Stereo		0.46 (± 0.21)	0.39 (± 0.21)	0.52 (± 0.28)	0.49 (± 0.27)	0.29 (± 0.29)	0.71 (± 0.35)
RF ToF	P	**0.07** (± **0.06**)	**0.08** (± **0.08**)	**0.10** (± **0.10**)	0.77 (± 1.96)	**0.09** (± **0.13**)	**0.05** (± **0.07**)
NIR ToF		1.05 (± 1.29)	3.19 (± 6.90)	3.99 (± 0.84)	0.37 (± 0.46)	9.03 (± 12.38)	1.06 (± 0.48)
Active Stereo		0.66 (± 0.24)	0.64 (± 0.25)	0.89 (± 0.35)	**0.16** (± **0.26**)	0.72 (± 0.82)	0.50 (± 0.38)
Passive Stereo		0.72 (± 0.22)	0.59 (± 0.23)	0.84 (± 0.35)	0.67 (± 0.37)	0.39 (± 0.48)	1.15 (± 0.30)
RF ToF	P_e	**0.07** (± **0.06**)	**0.08** (± **0.08**)	**0.10** (± 0.10)	0.78 (± 1.97)	**0.09** (± **0.13**)	**0.05** (± **0.07**)
NIR ToF		0.36 (± 0.57)	1.77 (± 2.11)	4.97 (± 0.27)	0.15 (± **0.11**)	15.00 (± 13.89)	0.96 (± 0.18)
Active Stereo		0.75 (± 0.21)	0.73 (± 0.15)	0.86 (± 0.19)	**0.10** (± 0.12)	0.76 (± 0.78)	0.52 (± 0.37)
Passive Stereo		0.85 (± 0.16)	0.70 (± 0.18)	1.17 (± **0.05**)	0.81 (± 0.25)	0.40 (± 0.44)	1.29 (± 0.18)
RF ToF	P^*	**-0.05** (± **0.06**)	**-0.05** (± **0.08**)	**+0.08** (± **0.10**)	+0.67 (± 1.96)	**-0.02** (± **0.13**)	**+0.01** (± **0.07**)
NIR ToF		+0.68 (± 1.29)	+3.08 (± 6.90)	+3.99 (± 0.84)	+0.36 (± 0.46)	+9.01 (± 12.38)	+1.05 (± 0.48)
Active Stereo		+0.66 (± 0.24)	+0.63 (± 0.25)	+0.88 (± 0.35)	**-0.07** (± **0.26**)	-0.62 (± 0.82)	+0.48 (± 0.38)
Passive Stereo		+0.72 (± 0.22)	+0.59 (± 0.23)	+0.83 (± 0.35)	+0.66 (± 0.37)	-0.23 (± 0.48)	+1.15 (± 0.30)
RF ToF	${P_e}^*$	**-0.05** (± **0.06**)	**-0.05** (± **0.08**)	**+0.08** (± 0.10)	+0.68 (± 1.97)	**-0.02** (± **0.13**)	**+0.01** (± **0.07**)
NIR ToF		+0.17 (± 0.57)	+1.72 (± 2.11)	+4.97 (± 0.27)	+0.15 (± **0.11**)	+14.97 (± 13.89)	+0.96 (± 0.18)
Active Stereo		+0.75 (± 0.21)	+0.73 (± 0.15)	+0.86 (± 0.19)	**-0.06** (± 0.12)	-0.70 (± 0.78)	+0.52 (± 0.37)
Passive Stereo		+0.85 (± 0.16)	+0.70 (± 0.18)	+1.17 (± **0.05**)	+0.81 (± 0.25)	-0.30 (± 0.44)	+1.29 (± 0.18)

	Metric Type	Polystyrene Plate	Bunny Box	Bunny
RF ToF	C_g	**2.10** (± 2.38)	**0.32** (± **0.28**)	0.28 (± 0.22)
NIR ToF		3.28 (± **2.15**)	0.50 (± 0.29)	0.26 (± 0.14)
Active Stereo		3.01 (± 3.05)	0.39 (± 0.40)	0.12 (± 0.10)
Passive Stereo		2.99 (± 2.76)	0.39 (± 0.62)	**0.08** (± **0.05**)
RF ToF	C_s	**0.14** (± **0.07**)	**0.30** (± **0.26**)	0.50 (± 0.41)
NIR ToF		2.50 (± 0.81)	0.48 (± 0.37)	0.31 (± 0.20)
Active Stereo		1.74 (± 1.97)	0.37 (± 0.35)	**0.12** (± **0.09**)
Passive Stereo		2.92 (± 4.87)	0.61 (± 1.08)	0.13 (± 0.13)
RF ToF	P	**0.12** (± **0.10**)	1.32 (± 2.45)	0.74 (± 0.69)
NIR ToF		17.95 (± 24.15)	1.91 (± 2.83)	0.45 (± 1.77)
Active Stereo		10.63 (± 15.21)	**0.99** (± **1.32**)	**0.16** (± **0.20**)
Passive Stereo		10.08 (± 11.66)	1.05 (± 1.88)	0.17 (± 0.27)
RF ToF	P_e	**0.12** (± **0.10**)	1.12 (± 2.15)	0.74 (± 0.69)
NIR ToF		17.86 (± 24.12)	**0.93** (± **0.87**)	0.38 (± **0.12**)
Active Stereo		10.48 (± 15.02)	1.46 (± 1.54)	0.11 (± **0.12**)
Passive Stereo		9.81 (± 11.41)	2.68 (± 2.54)	**0.10** (± **0.12**)
RF ToF	P^*	**+0.11** (± **0.10**)	+1.15 (± 2.45)	+0.69 (± 0.69)
NIR ToF		+17.95 (± 24.15)	+1.73 (± 2.83)	-0.24 (± 1.77)
Active Stereo		+10.63 (± 15.21)	-0.83 (± **1.32**)	-0.10 (± **0.20**)
Passive Stereo		+10.08 (± 11.66)	**-0.78** (± 1.88)	**+0.08** (± 0.27)
RF ToF	${P_e}^*$	**+0.11** (± **0.10**)	+1.03 (± 2.15)	+0.69 (± 0.69)
NIR ToF		+17.86 (± 24.12)	**+0.92** (± **0.87**)	-0.38 (± **0.12**)
Active Stereo		+10.48 (± 15.02)	-1.46 (± 1.54)	-0.08 (± **0.12**)
Passive Stereo		+9.81 (± 11.41)	-2.48 (± 2.54)	**+0.07** (± **0.12**)

Table 13. We measure the depth deviation with respect to C_g, C_s, P, P_e and an additional signed version of P,P_e, which is denoted as P^*,${P_e}^*$. All metrics are listed in the form $(\mu \pm \sigma)$, consisting of the mean μ and standard deviation σ in centimeters, computed over the entire metric domain, respectively. The best results among all sensors of one metric type are highlighted in **bold** and the worst results are underlined.

.