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

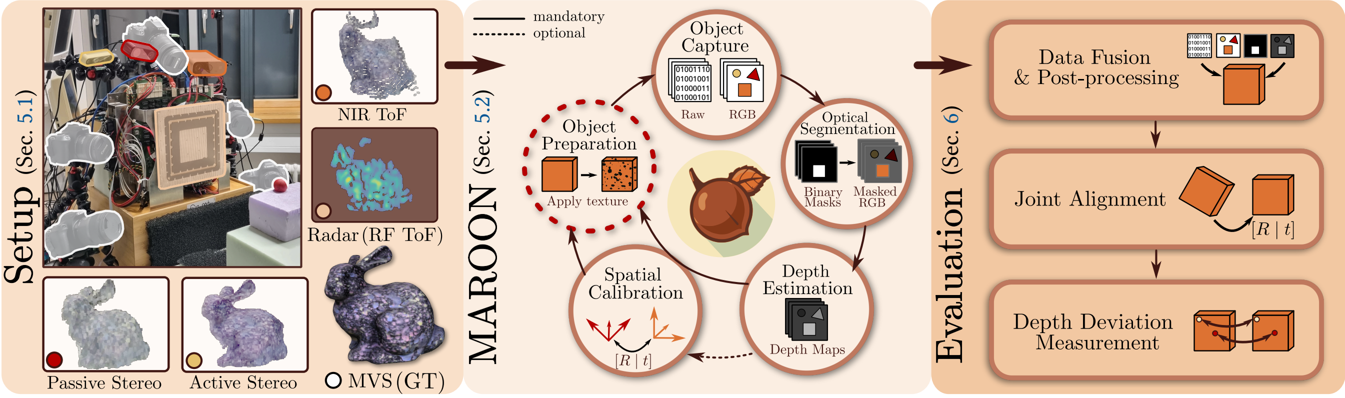


Fig. 1. Recent developments for near-field imaging radars enabled the acquisition of high-resolution depth images, and the sensors are now increasingly gaining attention as complementary modalities to optical depth sensing. Direct comparisons from our MAROON dataset, however, highlight significant differences 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 are made public in form of a multimodal dataset, called MAROON, which can be accessed at: https://vwirth.github.io/maroon.

CCS Concepts: • **Computing methodologies** → **3D imaging**; **Perception**; **Active vision**; Visual inspection;

Additional Key Words and Phrases: Time of flight, radar, radio frequency, mmWave, MIMO, depth camera, RGB-D, spatial calibration, multimodal, sensor fusion, near field, high resolution

ACM Reference Format:
Vanessa Wirth, Johanna Bräunig, Nikolai Hofmann, Martin Vossiek, Tim Weyrich, and Marc Stamminger. 2026. MAROON: A Dataset for the Joint Characterization of Near-Field High-Resolution Radio-Frequency and Optical Depth Imaging Techniques. *ACM Trans. Graph.* 45, 2, Article 23 (March 2026), 18 pages. https://doi.org/10.1145/3796224

Tim Weyrich and Marc Stamminger both contributed equally to this research.

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

Authors' Contact Information: Vanessa Wirth (corresponding author), Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany; e-mail: vanessa.wirth@fau.de; Johanna Bräunig, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany; e-mail: johanna.braeunig@fau.de; Nikolai Hofmann, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany; e-mail: nikolai.hofmann@fau.de; Martin Vossiek, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany; e-mail: martin.vossiek@fau.de; Tim Weyrich, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany and University College London, UK; e-mail: tim.weyrich@fau.de; Marc Stamminger, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany; e-mail: marc.stamminger@fau.de.

ACM 0730-0301/2026/03-ART23
https://doi.org/10.1145/3796224

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** (**RF**) 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 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.

Drawing on prior research about wavelength-specific strengths and weaknesses, this article 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 RF **Time-of-Flight** (**ToF**) sensor. 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:

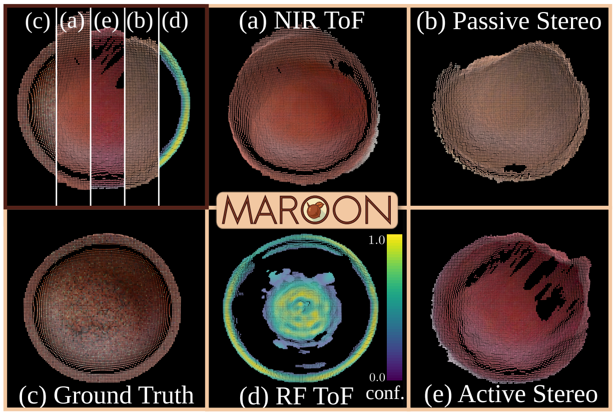


Fig. 2. Example data of the *Plunger* object from the MAROON dataset. In the *upper left*, all reconstructions are spatially aligned with respect to the RF ToF coordinate system. The RF ToF colorscale encodes the normalized reconstruction confidence (cf. Section 4.2.2).

(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 RF and optical domains.
(3) To characterize the under-researched effects of millimeter waves in near-field imaging radars, e.g., object materials in the RF domain, akin to studies on **bi-directional reflectance distribution functions** (**BRDFs**) in optics.
(4) To improve RF signal simulations by supplying raw data for modeling high- and 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 RF 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 release this dataset alongside the raw radar measurements to facilitate exploration of various signal reconstruction and filtering techniques: https://vwirth.github.io/maroon.

- 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, RF 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 article pursues a unified terminology, summarized in the table below and used in the remainder of this article.

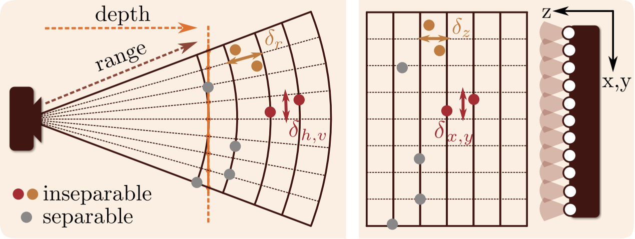


Fig. 3. Visualization of the effects caused by limited spatial resolution for multiple point targets. Optical sensors (*left*) divide spatial resolution into range resolution, δ_r, and perspective pixel resolution, $\delta_{h,v}$. Contrary to that, near-field imaging radars (*right*) refer to range δ_z 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.

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 $\boldsymbol{\delta}$. We 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. The resolution of near-field MIMO imaging radars is defined with respect to the three orthogonal axes x, y, z, where z refers to the depth and x, y are parallel to the antenna aperture. Sharing the same terminology as far-field RF ToF sensors that measure range, resolution is divided into range δ_z and cross-range [Ahmed 2014] $\delta_{x,y}$ resolution, respectively. Optical sensors define either depth or range resolution, where the latter is denoted as δ_r. Due to perspective, depth differs from range; within this work, we consider resolution at the sensor center, where $\delta_z \approx \delta_r$ (cf. Figure 3). Furthermore, optical sensors refer to pixel resolution along the perspective horizontal (δ_h) and vertical (δ_v) axes, respectively, of which we assume $\delta_{h,v} \approx \delta_{x,y}$ holds in the center.

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.

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 NIR light spectrum and the millimeter-wave (mmWave) RF 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 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

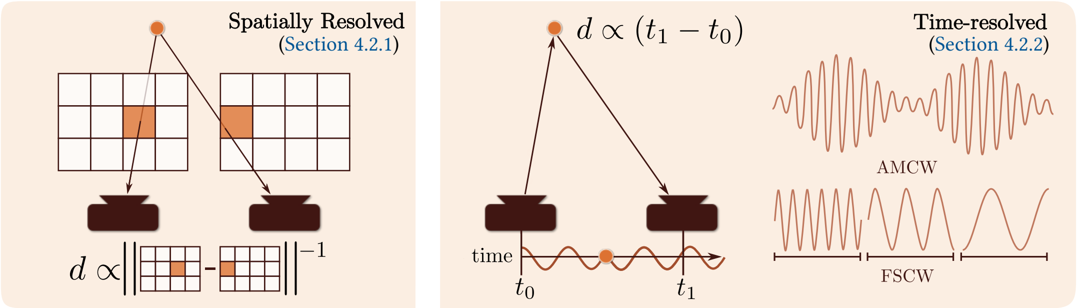


Fig. 4. Overview of the two depth sensing categories considered in this work. Spatially resolved methods compute the depth from disparity in the pixel positions. Time-resolved methods measure the depth through the round-trip propagation time of the received continuous wave (CW) signal. The types of wave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

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 Section S11 of the supplementary material.

4.2.2 Time-Resolved Sensors (Time-of-Flight). ToF 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 Section S2. 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 at receiving time to yield m_r. It is then cross-correlated with a so-called signal hypothesis s_h, which is commonly chosen as the currently transmitted signal s_t, such that the output of this correlation, c_r, describes the signal similarity from which the relative phase shift x is inferred [Horaud et al. 2016]:

$$c_\mathrm{r}(x) = \int_{-T_\mathrm{m}/2}^{+T_\mathrm{m}/2} m_\mathrm{r}(t) s_\mathrm{h}(t-x) dt \,, \tag{4}$$

where 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

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

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

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:

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

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.

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, e.g., through integration of multimodal sensor data.

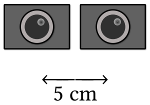
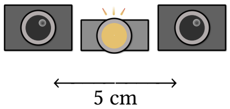
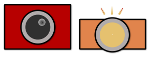
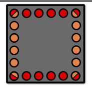
In the course of this section, we outline the capture setup and data acquisition pipeline, depicted in Figure 1, to aid future research on the publicly accessible data. Four single-view depth sensors are used in our experiments: Stereolabs ZED X [Stereolabs 2023] (Passive Stereo), Intel Realsense D435i [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 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

Table 1. Overview of the Sensors and Their Parameters Used in Our Experiments

		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		 5 cm	 5 cm	 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* $\boldsymbol{\delta_x \times \delta_y \times \delta_z}$	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

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 Section S3.1 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 Sections S1 and S2 of 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].

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 tradeoff between fair sensor comparability and practical applicability (see supplementary Section S3). An overview of the chosen settings and 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 in 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 by 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 $30 \times 30 \times 20$ 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 Section S4.1 of 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 GT 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.

Table 2. Statistics of the MAROON Dataset

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

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 GT 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 supplementary Section S4), 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

Table 3. Categorization of the Presented Metrics with Respect to Their Sensitivities

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

In addition to depth, a 3D error evaluates errors along the cross-depth axes.

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 supplementary Section S4). 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:

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

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

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.

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 RF 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 Figure S12 of 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 Figure S6 of the supplementary material.

6.2.2 RF Signal Response. The point-wise confidence of the BP 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 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 Table S4 of the supplementary material. The goal 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 Section S6 of the 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

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

	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**)

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 7.1.

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 Section S5.1 of 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

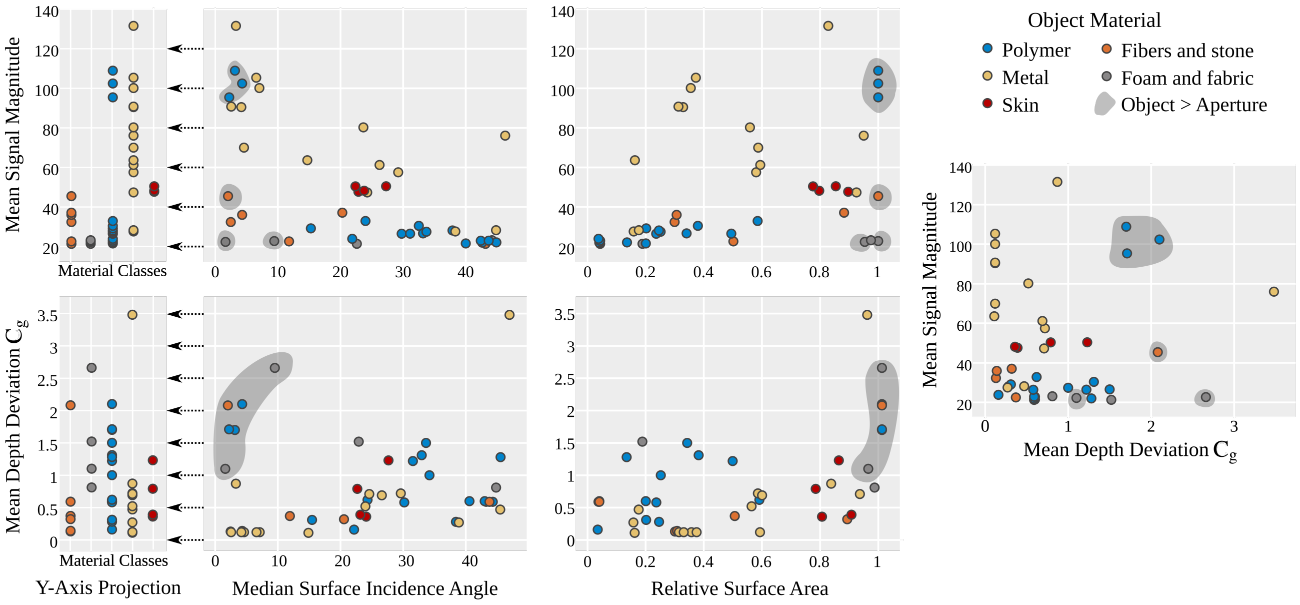


Fig. 5. On the *left*, object material, geometry (median surface incidence angle), and size (relative surface area) are put in relation to received signal response (mean signal magnitude, *top row*) and mean depth deviation (*bottom row*). On the *right*, both quantities are directly compared to each 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 GT 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.

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 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 Section S4.3 includes signed versions of P and P_e for all MAROON objects.

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

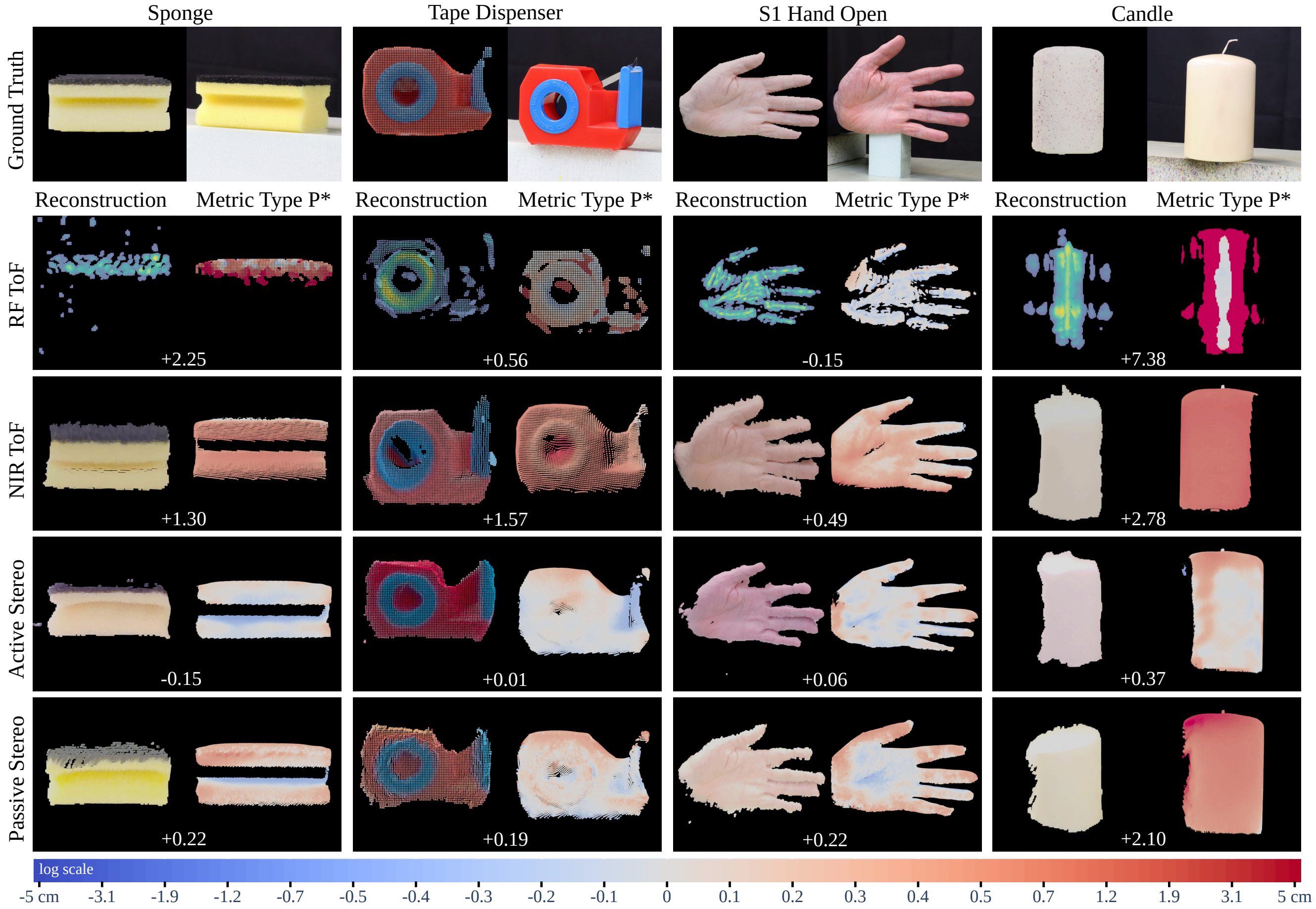


Fig. 6. For selected objects, we show the reconstructed point clouds (*left*) next to their deviation from to the MVS reconstruction (*right*). The signed depth deviation P^* is given for each pixel (u, v) in centimeters. All measurements in the domain $\boldsymbol{M}^+(u, v)$ are projected onto the 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.

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

The results in Table 4 suggest that RF ToF reconstructions are generally less complete than those of optical sensors, as further 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 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 RCS [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 Section S7 of 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 BP 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.

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 Section S7.1 of the supplementary material. Lastly, we note that in the results, instances of object self-shadowing, that would contribute to increased depth deviation, are infrequent. While the response of backprojection to partial occlusions is an important consideration, the overall impact on the results is minimized given the characteristics of the primarily smooth objects in our dataset.

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 multisensor 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 colocalizations 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 Section S7.2 of 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.

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.

8.2 Multimodal Depth Sensing

While the BP 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

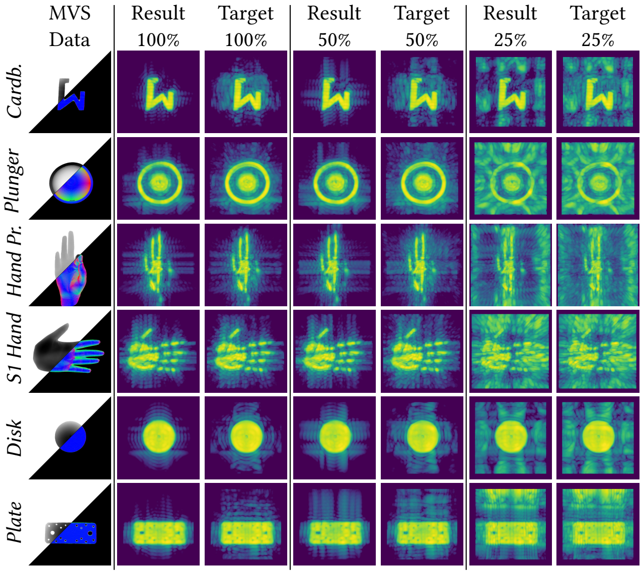


Fig. 7. Inverse radar rendering results for different antenna configuration. 100% considers all transmitting and receiving antennas in the MIMO array (94 × 94), 50% considers every second respective antenna in the array (47 × 47), and 25% only considers every fourth antenna in the array (24 × 24). Despite 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*.

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 prereleased 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. The transparency causes parts of the background or ground surface to be reconstructed, resulting in a depth prior positioned behind the ground truth. With limited 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 surface areas 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.

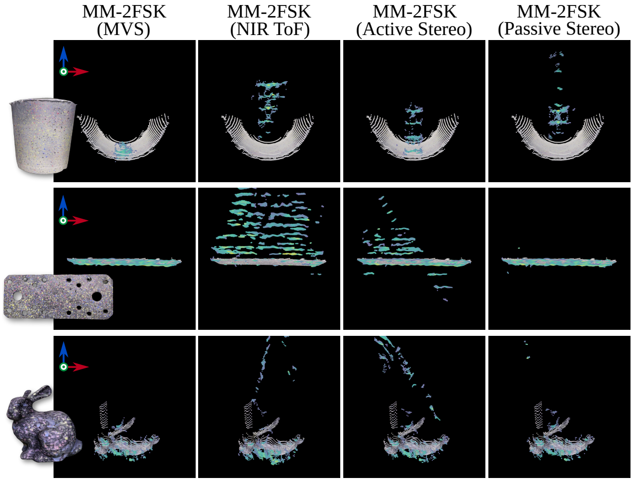


Fig. 8. Top-down views of the RF ToF reconstructions obtained with the MM-2FSK method, fused with the ground-truth MVS point cloud. *Left* to *right*: we vary the supporting sensor providing the depth prior. *Top* to *bottom*: we display the *Flowerpot (Transparent)*, *V2 Metal Plate*, and *Bunny* objects.

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.

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 (P and 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 BP, 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,

Fig. 5. Ablation Study of the MM-2FSK Method with Different Depth Priors, Each from Another Optical Depth Imager

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

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.

we highlighted one particular work of Hofmann et al. [2025], taking the first step towards automatic material characterization in the RF domain using inverse rendering. We anticipate that further analysis of these material parameters, combined with improvements of RF 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, 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 avenue to explore 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 GT MVS 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 article.

Moreover, we presented two representative 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 studies 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. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b175dc. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683.

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. DOI: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, et al. 2018. IMpixel 65nm BSI 320MHz demodulated TOF image sensor with 3μm global shutter pixels and analog binning. In *Proceedings of the 2018 IEEE International Solid-State Circuits Conference - (ISSCC)*. 94–96. DOI: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. DOI:https://doi.org/doi:10.1515/teme-2021-0034

D. W. Bliss and K. W. Forsythe. 2003. Multiple-input multiple-output (MIMO) radar and imaging: Degrees of freedom and resolution. In *Proceedings of the 37th Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 1. IEEE, 54–59.

Johanna Bräunig, Desar Mejdani, Daniel Krauss, Stefan Griesshammer, Robert Richer, Christian Schuessler, Julia Yip, Tobias Steigleder, Christoph Ostgathe, Björn M. Eskofier, et al. 2023. Radar-based recognition of activities of daily living in the palliative care context using deep learning. In *Proceedings of the 2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. 1–4. DOI: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. DOI: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 *Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA)*. 2752–2758. DOI: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 (MM '22)*. ACM, New York, NY, USA, 3501–3510. DOI: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. DOI: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. arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-rsn.2019.0471 DOI:https://doi.org/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. DOI: 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. DOI: 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* 5, 2 (2025), 1–17. DOI: https://doi.org/10.1109/JMW.2025.3535077

Radu Horaud, Miles Hansard, Georgios Evangelidis, and Menier Clément. 2016. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine Vision and Applications* 27, 7 (10 2016), 1005–1020. DOI: https://doi.org/10.1007/s00138-016-0784-4

Intel 2023. *Intel® RealSense™ Product Family D400 Series.* Intel. Retrieved February 19, 2026 from https://www.intel.com/content/www/us/en/products/sku/190004/intel-realsense-depth-camera-d435i/specifications.html

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).* ACM, New York, NY, USA, 511–518. DOI: 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. DOI: 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. Retrieved from 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 *Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).* 5704–5713. DOI: 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 Transactions on Graphics* 35, 4, Article 142 (Jul 2016), 19 pages. DOI: 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. DOI: https://doi.org/10.1109/JSTSP.2021.3061270

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 - Proceedings of the 2013 IEEE Military Communications Conference.* 1238–1243. DOI: 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, 5 (05 2020), 2866–2873. DOI: 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. arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-rsn.2018.5016 DOI: https://doi.org/10.1049/iet-rsn.2018.5016

Microsoft 2022. *Azure Kinect DK Hardware Specifications.* Microsoft. Retrieved February 19, 2026 from 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, 1 (2018), 1–10. DOI: 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 . In *Proceedings of the 33rd ACM Special Interest Group on Graphics and Interactive Techniques (SIGGRAPH '06).* ACM, New York, NY, USA, 935–944. DOI: 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. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/phor.12063 DOI: https://doi.org/10.1111/phor.12063

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded SAM: Assembling open-world models for diverse visual tasks. arXiv:abs/2401.14159. Retrieved from https://arxiv.org/abs/2401.14159

Rohde & Schwarz 2023. *R&S® QAR50 Quality automotive radome tester.* Rohde & Schwarz. Retrieved February 19, 2026 from 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. DOI: 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. DOI: 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. DOI: 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, 1 (2020), 48958–48969. DOI: https://doi.org/10.1109/ACCESS.2020.2979339

Stereolabs 2023. *ZED X Datasheet.* Stereolabs. Retrieved February 19, 2026 from https://www.stereolabs.com/en-de/store/products/zed-x-mini-stereo-camera

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. DOI: https://doi.org/10.1109/MSP.2020.2978507

Richard Szeliski. 2022. *Computer Vision - Algorithms and Applications (2nd ed.).* Springer. 749–797 pages. DOI: 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. *Proceedings of the ACM on Human-Computer Interaction* 6, EICS, Article 156 (Jun 2022), 19 pages. DOI: 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 *Proceedings of the 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP).* 315–321. DOI: 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 Transactions on Graphics* 41, 4, Article 36 (Jul 2022), 14 pages. DOI: 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. DOI: 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://www.mdpi.com/about/announcements/784

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. In *Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '24).* 8322–8329. Retrieved from https://ieeexplore.ieee.org/abstract/document/10801705

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:abs/2511.01405. Retrieved from 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. DOI: 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 *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition.* 366–373. DOI: 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. DOI: 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*. Springer. 1–355 pages. DOI: 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 *Proceedings of the 2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA)*. 93–98. DOI: 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. DOI: 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://www.mdpi.com/about/announcements/784

Received 31 October 2024; revised 19 January 2026; accepted 24 January 2026