This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

IEEE Journal of Microwaves

MTT-S MHz to THz Community

Received 3 November 2025; revised 16 May 2026; accepted 8 July 2026.

Digital Object Identifier 10.1109/JMW.2026.3713325

# MM-2FSK: Multimodal Frequency Shift Keying for Ultra-Efficient and Robust High-Resolution MIMO Radar Imaging

**VANESSA WIRTH [1], JOHANNA BRÄUNIG [3], MARTIN VOSSIEK [3] (Fellow, IEEE), TIM WEYRICH [1,4], AND MARC STAMMINGER [1]**

*(Regular Paper)*

[1]Visual Computing Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany
[2]fiveD GmbH, 91052 Erlangen, Germany
[3]Institute of Microwaves and Photonics, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany
[4]University College London (UCL), WC1E 6BT London, U.K.

CORRESPONDING AUTHOR: Vanessa Wirth (e-mail: vanessa.wirth@fau.de).

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

*(Tim Weyrich and Marc Stamminger contributed equally to this work.)*

**ABSTRACT** Accurate reconstruction of dynamic targets demands three-dimensional imaging at high temporal resolution. Radar is a particularly attractive sensing modality for this purpose due to its fast capture capabilities and robustness under varying lighting and weather conditions. Consequently, it has been successfully employed in a variety of dynamic applications, such as automotive and robotic perception. Nevertheless, dynamic sensing remains comparatively unexplored in operating regimes requiring near-field, high-fidelity surface reconstruction, as commonly targeted by large-aperture mmWave MIMO radars. Here, achieving high frame rates is challenged by the substantial computational cost of reconstruction, imposing stringent constraints on bandwidth, compute resources, and ultimately sensor capture rate. To enable fast captures at reduced computation cost, prior work explored few-tone imaging, reducing the number of transmitted frequencies at the expense of increased range-phase ambiguity; however, this requires coarse knowledge of the target position and is primarily designed to reconstruct flat surface geometries. We present MM-2FSK, which extends two-tone stepped frequency imaging with an external modality-agnostic per-point depth prior for precise near-field 3D imaging of geometrically complex surfaces. To address range-phase ambiguities, the prior is processed to mitigate inter-modality data mismatch, transformed into a radar-compatible representation, and integrated into signal processing as a phase constraint. In experiments with a mmWave FSCW MIMO radar and with representative priors obtained from a co-located optical sensor, we demonstrate superior depth accuracy of our method over comparable near-field few-tone radar imaging approaches in static and dynamic scenes, competitive with time- and resource-intensive measurements of many-tone backprojection.

**INDEX TERMS** 3D reconstruction, depth cameras, frequency shift keying, mimo radar, multimodal, radar imaging, sensor fusion.

## I. INTRODUCTION

In recent years, the reconstruction of dynamic targets using contactless sensors has gained significant attention, influencing research in many areas, including entertainment (e.g., computer games, AR/VR), autonomous agents, human-computer interaction, and medical diagnosis [1], [2].

Among these, applications involving critical decisions based on complex movements, such as human gait analysis [3], [4] and clinical hand function assessments [5], [6], [7], place a particularly high demand on fast and precise sensing techniques to ensure the reliability of such decisions.

Millimeter-wave (mmWave) multiple-input multiple-output (MIMO) radars offer a viable solution, providing spatial resolution beyond the capabilities of traditional monostatic antenna systems, and enabling the distinction of static and dynamic targets. Moreover, radar systems can analyze motion via the Doppler effect, making them well-suited for dynamic environments, compared to other modalities such as conventional LiDAR or RGB-D cameras. Thus, MmWave MIMO radar systems have been utilized for 3D human body reconstruction [8], [9], pose estimation [10], [11], people tracking [12], and activity recognition [13].

To achieve high-resolution three-dimensional imaging, MIMO radars typically employ a large number of transmitting (TX) and receiving (RX) antennas. With the ongoing move towards larger apertures [14], these systems increasingly operate in the sensor's near field. Traditional near-field radar imaging techniques, such as backprojection [15], [16], rely on these large and dense antenna arrays to leverage the numerous TX-RX combinations for precise reconstruction; however, the algorithmic complexity of the volumetric space, that is sampled in near-field backprojection, scales with the number of RX-TX antenna pairs, implying that precise 3D imaging capabilities of high-resolution antenna arrays come along with the expense of significant computational resources for reconstruction. Moreover, in representative applications such as security screening [16], where frequency-stepped continuous-wave (FSCW) radars rely on time-division multiplexing (TDM), high-fidelity 3D imaging via backprojection requires a high bandwidth and a substantial number of sequentially transmitted frequency tones. Consequently, this setting imposes high constraints on the attainable capture rate and, in turn, renders these sensors currently impractical for dynamic targets. In this work, we address these limitations, aiming to meet the stringent demands for fast capture rates and computational efficiency alongside high-fidelity reconstruction for dynamic scenes.

A line of prior work aims to reduce computation and capture time by utilizing sparse, low-cost antenna arrays. Recent advancements in deep learning focus on smooth reconstruction representations that surpass the spatial resolution limitations of conventional signal processing techniques. One approach is to employ generative methods, such as implicit neural representations [17] or conditional generative adversarial networks [18], to recover the spatially resolved reflective properties of targets at super-resolution. Another line of work uses Neural Radiance Fields (NeRFs) as compact geometric representations to simulate novel views and synthesize raw frequency-space measurements [19] or range-Doppler maps [20], while also incorporating additional modalities such as LiDAR and cameras [21]. Due to their data-driven learning processes, these approaches typically require a substantially high number of radar measurements, which currently limits their application primarily to autonomous driving scenarios. Moreover, the training process is computationally intensive.

Closer to our work, another line of research utilizes dense antenna arrays in combination with a reduced set of frequency tones. Among these approaches, Bräunig et al. [22] introduced an imaging technique for large, densely-packed antenna arrays of high-resolution MIMO radars, which focuses on high-speed pose tracking of both static and dynamic human hands. The proposed *2FSK* method operates using just two neighboring frequencies, based on the principle of continuous-wave (CW) Frequency Shift Keying (FSK). This approach significantly speeds up the reconstruction process by a factor of 1000 compared to backprojection [22], while ensuring rapid capture times due to the limited set of sequentially transmitted frequency tones

A key limitation of the 2FSK method, however, is its assumption that the depth of the captured object is approximately known. Follow-up work [23] (*3FSK*) aims to improve robustness by requiring more complex hardware configurations, that is, a larger signal bandwidth and three representative frequencies with specific frequency displacements, allowing for a less accurate depth prior.

The initial scalar depth prior, used by 2FSK and 3FSK, resembles a plane or depth slice in in a discretized, volumetric 3D space; hence, it renders both methods suitable primarily for flat targets with limited depth extent, as shown for hands in [22], [23]. Consequently, their applicability is restricted to a narrow range of target geometries and can lead to inaccuracies in environments, where depth is uncertain.

We extend 2FSK to a multimodal setting (MM-2FSK) that preserves the two-tone capture and reconstruction speed while supporting complex, spatially varying surface geometry. The key idea is to incorporate an absolute per-point surface-depth prior from a secondary modality into the 2FSK phase model, thereby enabling fine-grained relative depth estimation of that surface via the FSK principle. We revisit the conditions for unambiguous depth recovery with two tones and detail how to embed this external prior into the spatial sensor coordinate system and the signal model of the primary radio-frequency modality, while explicitly accounting for phase-range ambiguities and data mismatch between modalities. Specifically, our method targets high-resolution, near-field mmWave FSCW MIMO imaging radars as primary modality, while we rely on a secondary sensor that supplies only the absolute surface prior. In principle, this prior is modality-agnostic and can be sourced from any secondary sensor, including low-resolution measurements, e.g., from another mmWave sensor, when sensing characteristics within similar frequency bands are favored. In practice, we demonstrate multimodal efficiency by using a secondary sensor from the optical spectrum, i.e. an RGB-D camera, due to its high spatial resolution.

We provide a comprehensive evaluation based on a dataset [14] that includes ground-truth geometry of static objects, spatially aligned with real-world measurements from a mmWave high-resolution MIMO imaging radar operating in its near field. In this evaluation, we compare MM-2FSK against 2FSK, 3FSK, and traditional near-field backprojection. In addition, we investigate the influence of signal bandwidth beyond theoretical analysis and provide ablation studies focused on different frequency configurations. Lastly,

IEEE Journal of Microwaves

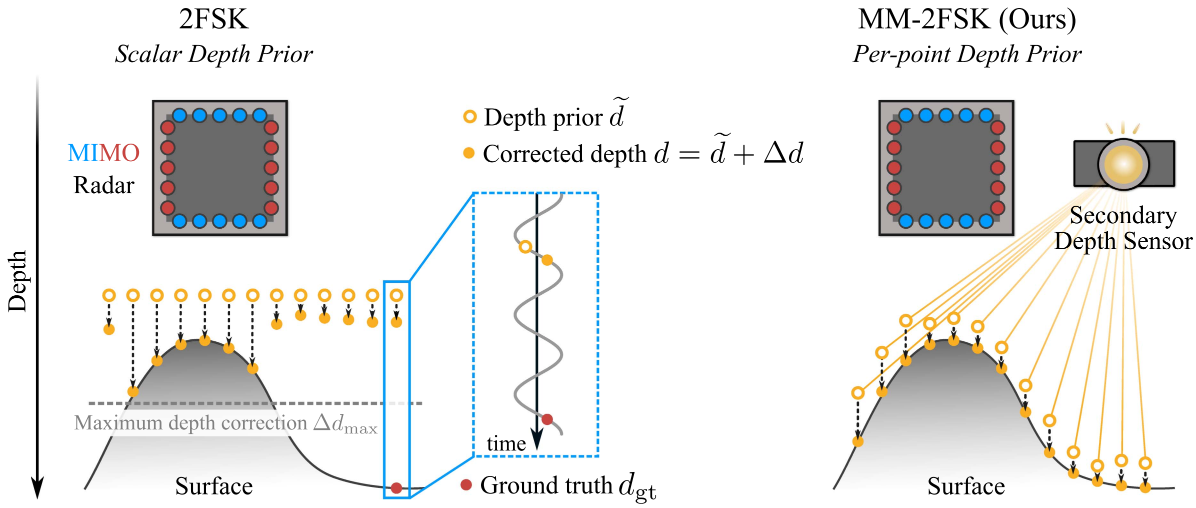


**FIGURE 1.** In our work, we extend the 2FSK imaging principle to the multimodal domain (MM-2FSK). Given a unified scalar depth prior, $\tilde{d}$, for each point, the 2FSK method iteratively refines the current estimate with a per-point depth correction factor, $\Delta d$, up to a limited extent, given by the maximum unambiguous depth correction, $d_{\max}$. In contrast, our method receives per-point depth priors from a secondary depth sensor, without requiring knowledge about the target position, and is more robust towards targets of varying surface depth.

we demonstrate effective, robust reconstruction of moving targets in dynamic settings.

In summary, our contributions are the following:

1) A novel multimodal signal processing method that incorporates a mmWave FSCW MIMO radar along with a secondary, assistive modality. In experiments, we selected secondary sensors from the optical frequency spectrum: an active stereo camera, a time-of-flight camera, and a multi-view stereo system.
2) A method for robust, high-speed radar imaging of arbitrary object geometries without requiring additional knowledge about the capture environment, i.e. the object position and surface depth variation.
3) A comprehensive evaluation in static and dynamic settings, which includes an ablation study over different frequency configurations and comparison to state-of-the-art radar imaging methods, i.e. 2FSK, 3FSK, and backprojection.

## II. FREQUENCY SHIFT KEYING FOR MIMO RADAR IMAGING

In the following section, we first address the theoretical foundations of the 2FSK imaging principle of Bräunig et al. [22]. Subsequently, we derive our MM-2FSK method from this principle. The key differences between the two algorithms are highlighted in Fig. 1.

Both methods are designed for high-resolution MIMO imaging radars that utilize multiple TX-RX antenna pairs for imaging. For simplicity, we omit the repetitive calculations over multiple antenna pairs– typically performed for near-field backprojection and related imaging methods–and present exemplary equations using just one TX-RX antenna pair. For a more detailed derivation of the equations, we refer to [22].

### A. FREQUENCY SHIFT KEYING WITH TWO NEIGHBORING FREQUENCIES (2FSK)

The 2FSK approach uses two transmitted signals of discrete neighboring frequencies, $f_1$ and $f_2$. As is typical in MIMO antenna configurations, the corresponding baseband signals of these two frequencies are first transmitted from a TX antenna, positioned at $\boldsymbol{r}_{\mathrm{TX}} \in \mathbb{R}^3$. These signals then propagate through space until they reflect off the first point target located at $\boldsymbol{p} \in \mathbb{R}^3$. Finally, the signal reflections are received by all RX antennas, with each positioned at a different location $\boldsymbol{r}_{\mathrm{RX}} \in \mathbb{R}^3$. After signal demodulation, the baseband signals, $s_i$ with $i \in \{1, 2\}$, can be expressed in analytic notation as follows:

$$s_i = A_i \exp\left(-j2\pi f_i \frac{\rho}{\mathrm{c}} + \phi_c\right) , \tag{1}$$

where $A_i$ is the amplitude, c is the speed of light, and $\phi_c$ is a constant phase offset. The traveled round-trip distance to $\boldsymbol{p}$, defined as $\rho = \|\boldsymbol{r}_{\mathrm{TX}} - \boldsymbol{p}\|_2 + \|\boldsymbol{r}_{\mathrm{RX}} - \boldsymbol{p}\|_2$, relates to the target depth $d$ by $\rho = 2d$, assuming far-field conditions where depth approximates range.

Given a set of candidate point target locations $\tilde{\boldsymbol{p}} \in \mathcal{P} = \{(x, y, \tilde{d})\}$ with a corresponding scalar depth prior $\tilde{d}$, two signal hypotheses, $w_1, w_2$, are computed from the round-trip distance $\tilde{\rho}$ between a TX-RX antenna pair and $\tilde{\boldsymbol{p}}$:

$$w_i(\tilde{\rho}) = \exp\left(-j2\pi f_i \frac{\tilde{\rho}}{\mathrm{c}}\right) . \tag{2}$$

The hypotheses are correlated with the baseband signals as follows:

$$c_i(\tilde{\rho}) = s_i w_i(\tilde{\rho})* = \exp\left(-j2\pi f_i \frac{(\rho - \tilde{\rho})}{\mathrm{c}}\right) , \tag{3}$$

where $*$ denotes the complex conjugate. The resulting complex signal contains a residual phase $\Delta\varphi_i$ that is proportional to a correction factor for distance, $\Delta\rho = (\rho - \widehat{\rho}) = 2\Delta d$, and, correspondingly, a factor for depth $\Delta d$:

$$2\pi f_i \frac{\Delta\rho}{\mathrm{c}} = 2\pi f_i \frac{2\Delta d}{\mathrm{c}} = \Delta\varphi_i \tag{4}$$

$$\Leftrightarrow \Delta d = \frac{\mathrm{c}\Delta\varphi_i}{4\pi f_i}, \tag{5}$$

With this information, each per-point depth estimate $d$ can be refined as follows:

$$d = \widetilde{d} + \Delta d\,. \tag{6}$$

Computing the phase $\Delta\varphi_i$ involves inverse trigonometric functions to determine the angle of the residual complex phasor $c_i(\widetilde{\rho})$. Due to the $2\pi$ phase ambiguities arising from its repetitive nature, these angles are typically restricted to the first period of the residual phasor. Consequently, the maximum correction factor for depth in one direction can be derived from (5) by assuming $\Delta\varphi_i$ approaches $2\pi$, which yields $\mathrm{c}/(2 \cdot f_i)$.

High-resolution MIMO imaging radars typically operate in the GHz to THz range [16], implying that this maximum correction factor can be small, i.e., within millimeters. Thus, Bräunig et al. [22] introduced the concept of calculating a differential complex phasor from the two single-frequency residual phasors based on (3) and (4) as follows:

$$c_{\Delta f}(\widetilde{\rho}) = c_2(\widetilde{\rho})c_1(\widetilde{\rho})* = \exp\left(-j2\pi\Delta f \frac{2\Delta d}{\mathrm{c}}\right). \tag{7}$$

The complex phasor of frequency difference $\Delta f = f_2 - f_1$ resembles a signal of considerably lower frequency (cf. Fig. 2), allowing a depth correction $\Delta d$ to be within the so-called *maximum unambiguous depth correction* $\Delta d_{\max}$, which now solely depends on the signal bandwidth:

$$\Delta d = \frac{\mathrm{c}\Delta\varphi_{\Delta f}}{4\pi\Delta f} \Rightarrow \Delta d_{\max} = \frac{\mathrm{c}}{2\cdot 2\Delta f}. \tag{8}$$

It is noteworthy that the right side corresponds to (8) from [22], additionally divided by a factor of 2 since we consider depth correction in both directions, yielding $\Delta d \in [-\Delta d_{\max}, +\Delta d_{\max}]$.

We illustrate the intuition behind the depth correction in Fig. 2, where we simplify the illustration of an analytic complex signal to a simple sine wave. As the depth correction is limited to the first period of the residual phasor $c_{\Delta f}$, the 2FSK algorithm does not necessarily converge to the ground-truth depth $d_{\max}$ of each point target, which may lie outside this period. Consequently, depth correction can fail and may even inadvertently adjust prior depth estimates in the wrong direction.

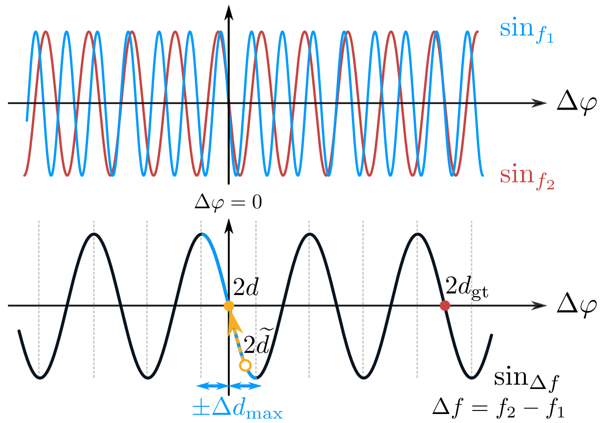


**FIGURE 2. Simplified visualization of the 2FSK depth correction process, where complex, analytic signals are schemed as periodic, real-valued sine waves. The 2FSK principle computes the depth correction $\Delta d$ based on the residual of the phase, $\Delta\varphi$, that remains after correlating the two single-frequency signals with a signal hypothesis, constructed with the depth prior $\widetilde{d}$. The *top row* depicts the two residual signals at frequencies $f_1$ and $f_2$, respectively. Using these, a complex differential signal with frequency $\Delta f$ is calculated, as depicted in the *bottom row*. This differential signal is used to adjust the current depth guess and is constrained by the maximum unambiguous depth correction, $\pm d_{\max}$. The correction factor is centered around the zero-crossing of the signal within the first period–or here, half of the period, due to signal simplification– pointing into the direction where the residual phase yields zero. Due to the $2\pi$-periodicity of the continuous signal, the residual phase corresponding to the ground truth depth, $d_{\max}$ may lie within a different signal period, resulting in the depth correction not producing the intended outcome.**

## B. MULTIMODAL FREQUENCY SHIFT KEYING (MM-2FSK)

Our method extends high-speed mmWave radar imaging to variable surface geometry by utilizing per-point depth priors from a secondary modality, ensuring the precision meets the unambiguous range-phase requirement for depth correction using a FSCW signal modulation. We detail how to embed this prior into the near-field MIMO coordinate and signal processing model while mitigating errors induced by sensor data mismatches in sampling and resolution.

Furthermore, we target the method's practical operation in dynamic scenes, where concurrent reconstruction and capture is desirable, by providing an efficient GPU implementation that integrates readily with matched-filter–based imaging, including backprojection, 2FSK, and MM-2FSK.

### 1) ALGORITHM

Unlike 2FSK, which relies on a single scalar depth prior, MM-2FSK leverages per-pixel depth from a secondary modality; in the following, we describe the algorithm using a generalized, modality-agnostic formulation. We assume that the secondary sensor is spatially calibrated with the MIMO imaging radar. Such spatial calibration can be obtained, for instance, with target-based procedures suitable for near-field operation, e.g., by employing symmetrically arranged spherical calibration targets composed of metallic and styrofoam materials [24]. The resulting extrinsic parameters allow transforming the 3D points from the secondary modality into the coordinate system of the primary modality, i.e. the imaging radar.

We assume the secondary modality provides a depth map $\boldsymbol{D}_s \in \mathbb{R}^{H\times W}$, where each pixel coordinate $(u, v)$ with $u \in [0, W-1]$ and $v \in [0, H-1]$ is associated with a metric depth scalar $d$. We adopt a generalized ray model to transform each per-pixel depth measurement $(u, v, d)$ into its

IEEE Journal of Microwaves

respective 3D point, while considering different image projection models. The 3D point is described by a ray in the sensor's coordinate system that is defined by an origin, $\boldsymbol{o}(u, v) \in \mathbb{R}^3$ and a direction $\boldsymbol{\omega}(u, v) \in \mathbb{R}^3$, that is scaled by the depth $d$, such that:

$$\boldsymbol{r}(u, v, d) = \boldsymbol{o}(u, v) + d \cdot \boldsymbol{\omega}(u, v) \; . \tag{9}$$

For range sensing modalities, where $\boldsymbol{o}(u, v)$ and $\boldsymbol{\omega}(u, v)$ are directly provided by the sensor, this formulation applies in a straightforward manner. We further specialize the model to two common projection types used in high-resolution depth imaging: perspective and orthographic projection.

Perspective projection is typically associated with sensors based on pixel-array imaging hardware, such as standard cameras. Here, the ray that points to the corresponding 3D point of a pixel is typically derived from the intrinsic calibration parameters of a pinhole camera model:

$$\boldsymbol{o}(u, v) = (0, 0, 0)^T \tag{10}$$

$$\boldsymbol{\omega}(u, v) = \begin{pmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} , \tag{11}$$

where $f_u, f_v$ are the focal lengths and $c_u, c_v$ are the principal point offsets.

Another class of sensors is based on orthographic projection models. Typical representatives are imaging radar systems that reconstruct the scene via near-field backprojection, followed by Maximum Intensity Projection (MIP) [14] along the depth axis of a Cartesian coordinate system defining the spatial extent and location of the reconstruction volume around the target of interest. For orthographic projection, the generalized ray model maps the origin to metric coordinates along the lateral dimensions, and the ray direction reduces to an identity vector:

$$\boldsymbol{o}(u, v) = \begin{pmatrix} \frac{(N_x-1)}{r-l} & 0 & 0 & \frac{(N_x-1)\cdot l}{l-r} \\ 0 & \frac{(N_y-1)}{b-t} & 0 & \frac{(N_y-1)\cdot b}{t-b} \\ 0 & 0 & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} u \\ v \\ 0 \\ 1 \end{pmatrix} , \tag{12}$$

$$\boldsymbol{\omega}(u, v) = (0, 0, 1)^T \tag{13}$$

where $(l, r)$ and $(t, b)$ are the spatial min-max extents of the reconstruction volume along the Cartesian x- and y-axis, respectively, and $N_x, N_y$ denote the respective number of discretized voxels within these extents.

Moreover, a depth image from a secondary modality may contain invalid pixels, e.g., due to the sensitivity to environmental factors; such pixels are simply skipped.

To generate a spatially coherent depth prior for radar imaging, we convert the resulting 3D point cloud $\boldsymbol{P}_s \in \mathbb{R}^{N\times 3}$ to a closed triangle mesh. To this end, we triangulate the point cloud in 2D using Delaunay triangulation [25] based on the corresponding pixel coordinates $(u, v)$. This triangulation computes the 2D convex hull of $\boldsymbol{P}_s$, effectively filling in depth

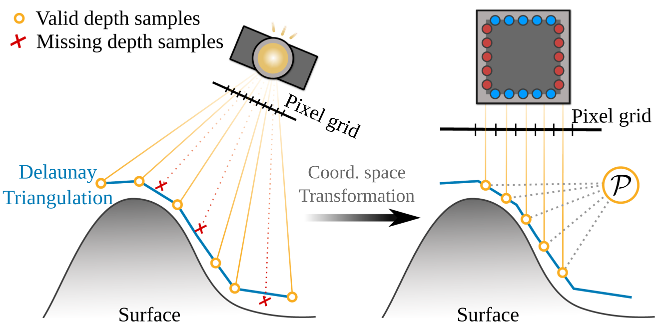


**FIGURE 3. Visualization of the depth prior generation: We first create a closed triangle mesh using Delaunay triangulation on 2D pixels corresponding to valid 3D point samples from the optical depth sensor; illustrated here in 2D as blue line sets. The mesh is then transformed into the radar's coordinate space and re-sampled via rasterization on the radar pixel grid to generate the candidate point set $\mathcal{P}$.**

gaps and preventing surface holes. A visualization of such triangulation is given in Fig. 3. Next, we use the extrinsic parameters obtained from spatial calibration [24] to transform the triangulated point cloud into the radar's coordinate space:

$$\boldsymbol{P}_s^r = [\boldsymbol{R} \mid \boldsymbol{t}]\boldsymbol{P}_s \, . \tag{14}$$

The extrinsic parameters consist of a rotation $\boldsymbol{R} \in \mathbb{R}^{3\times 3}$ and translation $\boldsymbol{t} \in \mathbb{R}^3$.

We then construct the set of candidate point target locations $\mathcal{P}$. Since the final output of high-resolution MIMO radars is typically an image, we compute $\mathcal{P}$ by sampling a user-defined $H' \times W'$ pixel grid of Cartesian 2D coordinates $(x, y)$, centered around the antenna aperture. Given the orthographic mapping of spatial coordinates to radar image pixels, we rasterize the triangulated point cloud $\boldsymbol{P}_s^r$.

Specifically, we render a depth map with barycentrically interpolated depth values, based on the triangle topology, to yield $\mathcal{P} \in \mathbb{R}^{H'W'\times 3}$. This essentially becomes our set of point candidates with *per-point* depth prior $\tilde{d}$, as illustrated in Fig. 3. Finally, we proceed with the depth correction in analogy to (6) of the 2FSK method.

Note that in contrast to 2FSK and 3FSK our depth prior is not constant and can thus represent non-flat shapes with larger depth range. As long as the noise parameters of the secondary modality and any spatial calibration errors remain within the maximum unambiguous depth correction factor, the final depth estimates of the MM-2FSK method are expected to be close to the ground truth.

### 2) EFFICIENT IMPLEMENTATION

Our method utilizes the single-instruction multiple-threads (SIMT) instructions of graphics processing units (GPUs), implemented by using NVIDIA CUDA as domain specific language. In this section, we put emphasis on the CUDA implementation of the baseband signal correlation kernel, which is commonly the runtime bottleneck in reconstruction using high-resolution imaging radars.

**Algorithm 1:** (MM-)2FSK for One CUDA Thread.

**Input:** Thread ID $i \in [0, 31]$, warp ID $j \in \mathbb{N}_0$, baseband signals $\boldsymbol{S}_1, \boldsymbol{S}_2 \in \mathbb{C}^{T \times R}$, point candidates $\boldsymbol{\mathcal{P}} \in \mathbb{R}^{H'W' \times 3}$, antenna positions $\boldsymbol{R}_{\mathrm{TX}} \in \mathbb{R}^{T \times 3}$ and $\boldsymbol{R}_{\mathrm{RX}} \in \mathbb{R}^{R \times 3}$

**Data:** Shared memory buffer $\boldsymbol{D}_{\boldsymbol{r}_{\mathrm{TX}}} \in \mathbb{R}^{T}$, $\boldsymbol{D}_{\boldsymbol{r}_{\mathrm{RX}}} \in \mathbb{R}^{R}$

**Output:** Correlations $\boldsymbol{C}_1 \in \mathbb{C}^{H'W'}$, $\boldsymbol{C}_2 \in \mathbb{C}^{H'W'}$

$\boldsymbol{p} \leftarrow \boldsymbol{\mathcal{P}}[j]$;

**for** $k \leftarrow 0, T/32$ **do**

  $\boldsymbol{D}_{\boldsymbol{r}_{\mathrm{TX}}}[i + 32 \cdot k] \leftarrow \|\boldsymbol{R}_{\mathrm{TX}}[i + 32 \cdot k] - \boldsymbol{p}\|_2$;

**end**

**for** $k \leftarrow 0, R/32$ **do**

  $\boldsymbol{D}_{\boldsymbol{r}_{\mathrm{RX}}}[i + 32 \cdot k] \leftarrow \|\boldsymbol{p} - \boldsymbol{R}_{\mathrm{RX}}[i + 32 \cdot k]\|_2$;

**end**

$\bar{c}_1, \bar{c}_2 \leftarrow 0$;

**for** $k \leftarrow 0, (T \cdot R)/32$ **do**

  $\widetilde{\rho} \leftarrow \boldsymbol{D}_{\boldsymbol{r}_{\mathrm{TX}}}[i + 32 \cdot k] + \boldsymbol{D}_{\boldsymbol{r}_{\mathrm{RX}}}[i + 32 \cdot k]$;

  $\bar{c}_1 \leftarrow \bar{c}_1 + c_1(\widetilde{\rho})$; `// Equation 3 for` $f_1$

  $\bar{c}_2 \leftarrow \bar{c}_2 + c_2(\widetilde{\rho})$; `// Equation 3 for` $f_2$

**end**

$\boldsymbol{C}_1[j] \leftarrow \mathrm{warp\_reduce\_sum}(\bar{c}_1)/(T \cdot R)$;

$\boldsymbol{C}_2[j] \leftarrow \mathrm{warp\_reduce\_sum}(\bar{c}_2)/(T \cdot R)$;

In MIMO radar imaging algorithms [16], [22], determining the spatial position of a point target typically requires performing correlation across the entire antenna aperture configuration. Specifically, for each candidate point target $\boldsymbol{p} \in \mathcal{P}$, the residual phasors from (3) are averaged over multiple TX-RX antenna positions. To optimize performance on the GPU, we parallelize the iterations over point targets and antenna pairs, utilizing the shared memory features of the GPU architecture.

NVIDIA graphics cards consist of SIMT units called *warps*, which include 32 parallel threads grouped into blocks that share fast-access memory. In our GPU kernel, as shown in Algorithm 1, each point $\boldsymbol{p}$ is processed by an entire warp, distributing the computations over multiple TX-RX antenna pairs. Given the known $T \times R$ MIMO antenna architecture, each thread pre-computes a subset of one-directional antenna paths from any TX antenna to $\boldsymbol{p}$ and from $\boldsymbol{p}$ to any RX antenna, stored in $\boldsymbol{D}_{\boldsymbol{r}_{\mathrm{TX}}}$ and $\boldsymbol{D}_{\boldsymbol{r}_{\mathrm{RX}}}$.

By sharing memory within the warp, each thread computes a partial summation of the residual phasors over all possible $(T \cdot R)/32$ TX-RX combinations. Finally, we utilize CUDA warp functions for summing the residual phasors in `warp_reduce_sum` and then average the result. The depth correction step, calculated from (7), is then performed after the kernel, as we use the two intermediate residual phasor sets, $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, for additional depth filtering, as discussed in the next section.

## III. EXPERIMENTAL SETUP

In the following sections, we describe the measurement setup derived from the MAROON dataset [14], along with the implementation details of the algorithms that we use in our evaluation. We compare our method against:

1) Near-field backprojection ((1) in [16]), followed by MIP ((8) in [14]) along the depth axis.
2) 2FSK ((6)–(8) in this work).
3) 3FSK (Fig. 2 in [23]).

Similar to ours, the 3FSK method extends the 2FSK approach by utilizing three representative frequencies with specific frequency displacements, resulting in three frequency differences. This allows for depth correction to be performed twice: first, by using the initial scalar depth value with a low frequency difference and, second, by utilizing the per-point corrected depth prior with two high frequency differences.

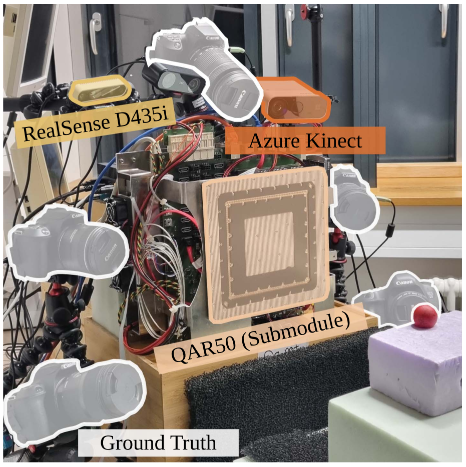


**FIGURE 4.** Sensor configuration from the MAROON dataset used in our experiments. For static-scene evaluation with MM-2FSK, we utilize per-point depth priors from an Intel RealSense D435i (active stereo) camera, selected for its low-noise characteristics [14]. For evaluating dynamic scenes, we pair the QAR50 radar submodule with a Microsoft Azure Kinect (optical time-of-flight) camera due to its comparatively simple time-synchronization support.

### A. CAPTURE SETUP

We validate the proposed method using the sensor configuration from the MAROON dataset [14]. The measurements were collected from a high-resolution MIMO radar, which is depicted in Fig. 4, where it is arranged alongside three spatially calibrated depth cameras and a ground-truth multi-view stereo (MVS) measurement system composed of five digital single-lens reflex (DSLR) cameras.

The QAR50 MIMO radar submodule features a $13.8 \times 13.8$ cm square-shaped aperture consisting of $94 \times 94$ TX-RX antenna pairs. The sensor operates in its near field. It employs a FSCW signal modulation with a maximum of 128 frequency tones ranging from 72 to 82 GHz. Further sensor specifications are listed in Table 1 of Wirth et al. [14].

To evaluate static scenes of varying geometry, we rely on the sensor captures from MAROON that comprise real sensor measurements of 45 distinct household and construction objects. Each target is positioned such that its geometric center coincides with the aperture center of the antenna array. To

IEEE Journal of Microwaves

**TABLE 1.** **Frequency Configurations, Utilized for All Subsequent Experiments**

| $\Delta f$ (GHz) | $f_1$ (GHz) | $f_2$ (GHz) | $\Delta d_{\text{max}}$ (cm) |
|---|---|---|---|
| $\approx$ 0.5 | 81.45 | 82.00 | 13.60 |
| $\approx$ 1.0 | 80.98 | 82.00 | 7.32 |
| $\approx$ 2.0 | 79.95 | 82.00 | 3.66 |
| $\approx$ 4.0 | 77.91 | 82.00 | 1.83 |
| $\approx$ 8.0 | 73.97 | 82.00 | 0.93 |
| $\approx$ 10.0 | 72.00 | 82.00 | 0.75 |

align the centers along the height dimension, the targets are supported by styrofoam boards (cf. Fig. 4).

While the target objects were captured at various distances [14], we focus on the object measurements taken at approximately 30 cm radial distance from the MIMO radar. To assess the imaging accuracy of the proposed method, we utilize ground-truth measurements of the target objects, obtained from the MVS system.

For evaluating dynamic scenes, we manually captured two additional sequences of radial hand motion, using a subset of the sensor arrangement given in Fig. 4, where we only utilize the QAR50 and the Azure Kinect time-of-flight sensor due to the availability of time-synchronization support of both sensors. In these measurements, the hand pose is held constant and the arm position is varied back and forth between approximately 20 cm and 60 cm radial distance to the imaging radar. In the absence of reliable surface ground truth–since dynamic scenes are challenging to reconstruct with traditional MVS–we perform a relative comparison against high-resolution backprojection, utilizing the full signal bandwidth and 128 frequency steps.

Depending on the experiment, we utilize different secondary sensors for MM-2FSK: for static scenes, we use either the ground-truth system or the noise-resilient RealSense D435i active stereo depth camera. For dynamic scenes requiring inter-sensor time synchronization, we leverage the integrated timestamp mechanisms of the Azure Kinect camera. Prior to data capture, we determine the time offset between relative Azure Kinect timestamps and the QAR50 timestamps by using the motion trajectory of a metal sphere pendulum as temporal calibration target. For evaluation, the modality-wise captures are then paired by nearest-neighbor timestamp matching. We note that typical temporal data mismatch is expected even after calibration, caused by varying capture latency and the different frame rates of the optical sensor (30 fps) and the radar ($\approx$ 70 fps).

### B. IMPLEMENTATION DETAILS

The MIMO imaging radar measurements provide raw phasor data in the form of a $94 \times 94 \times 128$ complex tensor, which we reduce to include only the two or three relevant frequencies, resulting in a $94 \times 94 \times 2$ or $94 \times 94 \times 3$ tensor, respectively, depending on the radar imaging method.

For backprojection of static scenes, we sample points within a $30 \times 30 \times 20$ cm volume centered around the object, yielding a voxel grid of dimensions $N_x = N_y = 301$ and $N_z = 201$. This grid is then projected to a $H' \times W' = N_x \times N_y$ depth map using Maximum Intensity Projection along the depth axis (z-axis).

For backprojection of dynamic scenes, we sample points within a reconstruction volume of $l = b = -15$ cm, $r = t = 15$ cm, $n = 20$ cm, and $f = 60$ cm, where the pairs $(l, b)$, $(t, b)$, $(n, f)$ are the min-max extents along the Cartesian xyz-axes, respectively. The voxel density is kept at $N_x = N_y = 301$, consistent with the static case, while $N_z = 401$ is increased, effectively yielding twice as many voxels to accommodate the increased z-extent that is necessary to construct the radial hand motion.

For the 2FSK, 3FSK, and MM-2FSK methods in both static and dynamic scenarios, we correspondingly reconstruct the $H' \times W'$ depth map directly. To investigate realistic scenarios, where the object placement is not trivial to assess, we use a 2FSK/3FSK depth prior of $\widetilde{d} = 40$ cm, which means, for the static scenario we expect a depth correction factor of $\Delta d \approx 10$ cm based on the ground truth at approximately 30 cm depth.

To filter out clutter and noise, we employ a depth filtering threshold across all four imaging methods, applied to the magnitude of the residual complex phasor after spatially resolving the signal. Specifically, we use the CUDA kernel listed in Algorithm 1 to average the residual phasors across all TX-RX antenna and frequency combinations, then compute the magnitude $V$ of the resulting mean phasor. Finally, we keep the depth values with with respect to a decibel threshold, $20 \cdot \log_{10}(\frac{V}{V_{\max}}) \geq -14$ dB, relative to the maximum $V_{\max}$ per object reconstruction.

In terms of average runtime, our implementation for depth estimation with two-frequency backprojection takes approximately 1430 ms, when using an NVIDIA GeForce RTX 3080 graphics card (10 GB VRAM) and an Intel Xeon W-1390P (3.50 GHz) processor. 3FSK achieves a runtime of about 70 ms and the (MM-)2FSK methods achieve a runtime of about 40 ms.

## IV. EVALUATION

This section presents three key experiments: first, an ablation study to explore the accuracy of MM-2FSK while varying the frequency differences. Second, a static-scene comparison of our method against the state of the art, i.e. 2FSK [22] and 3FSK [23] approaches, and traditional backprojection (BP) [15], [16]. Third, an extension of this state-of-the-art comparison to a dynamic radial hand motion setting. We perform evaluation using six representative frequency configurations, which are listed in Table 1.

For (MM-)2FSK and BP, we use the terminology *2FSK*$\Delta f$, e.g., FSK$\Delta$0.5 to denote a configuration with 0.5 GHz frequency difference. Furthermore, we denote the configuration of the lowest and highest frequency differences of 3FSK as 3FSK$(\Delta f_{\min}, \Delta f_{\max})$.

**TABLE 2. Ablation Study of the MM-2FSK Method, Using Depth Priors From MVS in Combination With Different Frequency Configurations. All Metrics Are Given in Centimeters and Are Averaged Over All Static Objects at 30 cm Distance**

| | $C_g$ | $C_s$ | P | $P_e$ |
|---|---|---|---|---|
| MM-2FSKΔ0.5 | 0.72 | 2.14 | 2.15 | 1.69 |
| MM-2FSKΔ1.0 | 0.74 | 1.26 | 1.74 | 1.44 |
| MM-2FSKΔ2.0 | 0.57 | 0.60 | 0.77 | 0.70 |
| MM-2FSKΔ4.0 | 0.53 | 0.35 | 0.41 | 0.37 |
| MM-2FSKΔ8.0 | 0.54 | 0.22 | 0.24 | 0.21 |
| MM-2FSKΔ10.0 | **0.51** | **0.18** | **0.19** | **0.17** |

The best results per metric are highlighted.

### A. METRICS

We follow a similar evaluation procedure as outlined in the MAROON dataset, utilizing the corresponding metrics: the one-directional Chamfer distance and the projective error. Interested readers are referred to [14] for a detailed discussion on the interpretation of these metrics.

The one-directional Chamfer distance quantifies the mean point-wise euclidean norm between point cloud $\boldsymbol{P}_r \in \mathbb{R}^{N\times 3}$, and point cloud $\boldsymbol{P}_{\text{gt}} \in \mathbb{R}^{M\times 3}$:

$$C = \frac{1}{N} \sum_{\boldsymbol{p}_r \in \boldsymbol{P}_r} \min_{\boldsymbol{p}_{\text{gt}} \in \boldsymbol{P}_{\text{gt}}} \|\boldsymbol{p}_r - \boldsymbol{p}_{\text{gt}}\|_2 \ . \tag{15}$$

To compute this, we transform the radar depth maps back into point cloud representation, where we compare them against the re-sampled ground-truth object reconstruction of similar point density (cf. [14]). The Chamfer distance is measured in both directions: from the ground-truth point cloud to the radar reconstruction, denoted as $C_g$, and vice versa, denoted as $C_s$.

The projective error is calculated on the respective depth maps, $\boldsymbol{D}_r$ and $\boldsymbol{D}_{\text{gt}}$, with the ground-truth depth map obtained by rasterizing the point cloud with respect to the radar pixel grid:

$$P = \frac{1}{H' \cdot W'} \sum_{u=0}^{H'-1} \sum_{v=0}^{W'-1} |\boldsymbol{D}_r(u, v) - \boldsymbol{D}_{\text{gt}}(u, v)| \ . \tag{16}$$

Following the methodology of [14], we measure the projective error on the masked object depth maps, once with and without performing additional mask erosion to mitigate silhouette artifacts; we denote the resulting metrics as P and $P_e$, respectively.

### B. ABLATION WITH RESPECT TO FREQUENCY DIFFERENCES

We assess the MM-2FSK algorithm in the static scenario using the frequency configurations outlined in Table 1 to simulate different radar systems. To isolate the impact of the frequency configuration from sensor characteristics, we utilize per-point depth priors obtained from the ground-truth MVS measurement system. The results are summarized in Table 2, showcasing performance across all four metrics, averaged over the 45 objects of the MAROON dataset.

We observe a trend towards better performance at higher frequency differences, with the MM-2FSKΔ10.0 method achieving the best results, yielding reconstruction errors in millimeter range, with a maximum pixel-wise depth error of only 1.9 mm with respect to P. We suggest this trend is related to the maximum unambiguous depth correction (cf. Table 1), which decreases as frequency difference increases, thereby constraining the radar depth variance. Specifically, the maximum unambiguous depth correction is inversely proportional to the phase sensitivity [23], which means that larger frequency differences are less sensitive to phase variations due to clutter and noise. This phenomenon becomes more evident when visualizing the corresponding point clouds of the reconstructions, as shown in Fig. 5. Reconstructions with higher frequency differences exhibit fewer noise artifacts.

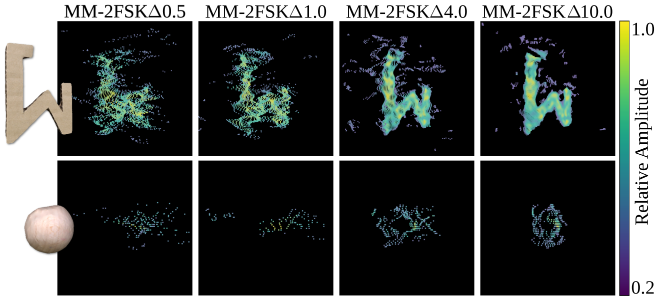


**FIGURE 5. Static MM-2FSK reconstructions with depth priors obtained from MVS for the Cardboard and Wood Ball objects, compared across different frequency configurations. The displayed point clouds are colored by the relative amplitude $V/V_{\max}$, where $V_{\max}$ is the maximum amplitude per reconstruction. We visualize a high dynamic range of [−14 dB, 0 dB], mapped to linear values in [0.2, 1.0]. We conclude that higher bandwidths exhibit fewer artifacts as they are less sensible to noisy phase variations.**

### C. STATIC COMPARISON AGAINST THE STATE OF THE ART

We compare the performance of MM-2FSK against 2FSK, 3FSK and BP, using per-point depth priors obtained from the RealSense active stereo depth camera. Our evaluation focuses on three frequency configurations from Table 1: the first, where $\tilde{d}$ lies within the maximum unambiguous depth correction ($\Delta f = 0.5$), the second, where it narrowly exceeds this factor ($\Delta f = 1.0$), and finally, the configuration where MM-2FSK performed best in previous ablation ($\Delta f = 10.0$). Additionally, we present reference radar reconstructions derived from the significantly more resource-intensive backprojection $\text{BP}_{\max}$, utilizing the maximum of 128 frequency steps.

In Fig. 6, we present exemplary reconstructions of the Bottle object generated by backprojection, 2FSK, 3FSK, and MM-2FSK. Among these methods with the same frequency configuration, MM-2FSK is the closest to the ground truth; at $\Delta f = 10.0$ GHz (*rightmost column*), it avoids reconstructing the partially transmissive plastic, resulting in a closer match to the ground-truth reconstruction than $\text{BP}_{\max}$.

The qualitative observations align with the quantitative evaluations in Table 3, which detail the reconstruction errors in centimeters across all objects; here, MM-2FSKΔ10.0

IEEE Journal of Microwaves
MTT-S MHz to THz Community

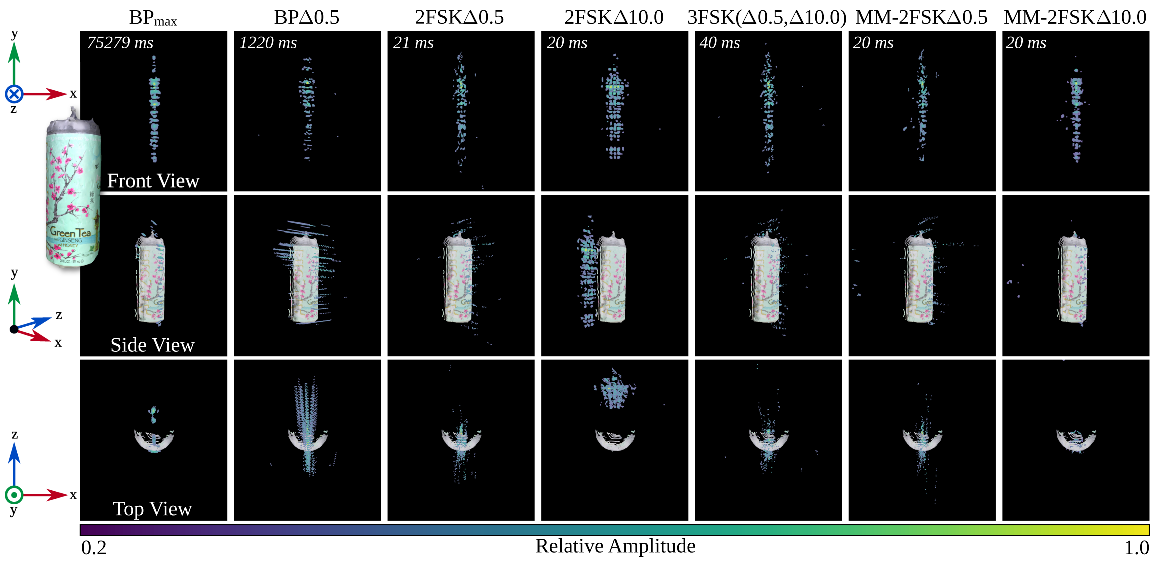


**FIGURE 6.** Static object comparison of the reconstructed point clouds for backprojection, 2FSK, 3FSK, and MM-2FSK (priors from RealSense) across different frequency configurations and views, with $BP_{max}$ using the full 10 GHz frequency spectrum at 128 frequency steps. The displayed point clouds are colored by the relative amplitude $V/V_{max}$, where $V_{max}$ is the maximum amplitude per reconstruction. We visualize a high dynamic range of [−14 dB, 0 dB], mapped to linear values in [0.2, 1.0]. Furthermore, we denote the reconstruction times of every method in the top left corner of the first row. Side and top views overlay the reconstructed point cloud with the ground-truth point cloud for the Bottle object at 30 cm object-to-sensor distance. The MM-2FSK method exhibits fewer artifacts and is closer to the ground truth than backprojection and 2FSK at the same frequency configuration. Additionally, depending on the frequency difference, MM-2FSK performs as well as or better than 3FSK, which employs a greater number of frequencies.

**TABLE 3.** Quantitative Evaluation Against the State of the Art. The Mean Reconstruction Error Is Given in Centimeters, Evaluated for Backprojection, 2FSK and, MM-2FSK (priors From RealSense) Against the Ground-Truth Setup. Each Metric Is Averaged Across All Static MAROON Objects at a Sensor Distance of 30 cm

| | $C_g$ | $C_s$ | P | $P_e$ |
|---|---|---|---|---|
| $BP_{max}$ | 0.82 | 0.90 | 0.94 | 0.85 |
| BPΔ0.5 | **0.69** | 4.27 | 3.34 | 3.18 |
| BPΔ1.0 | 0.79 | 5.29 | 3.70 | 3.40 |
| BPΔ10.0 | 1.02 | 6.35 | 4.98 | 4.87 |
| 2FSKΔ0.5 | 0.90 | 2.36 | 2.55 | 2.10 |
| 2FSKΔ1.0 | 8.68 | 12.27 | 12.87 | 12.82 |
| 2FSKΔ10.0 | 9.80 | 9.69 | 10.38 | 10.38 |
| 3FSK(Δ0.5,Δ10.0) | 0.83 | 2.54 | 2.63 | 2.30 |
| 3FSK(Δ1.0,Δ10.0) | 8.37 | 12.47 | 12.88 | 12.87 |
| 3FSK(Δ2.0,Δ10.0) | 6.43 | 7.90 | 8.51 | 8.52 |
| MM-2FSKΔ0.5 | 0.95 | 2.95 | 3.13 | 2.39 |
| MM-2FSKΔ1.0 | 0.98 | 2.72 | 2.84 | 2.28 |
| MM-2FSKΔ10.0 | 0.82 | **1.74** | **1.36** | **1.01** |

The best results per metric among all high-speed imaging methods are highlighted.

**TABLE 4.** Quantitative Evaluation Against the State of the Art in a Dynamic Setting. The Mean Reconstruction Error Is Given in Centimeters, Evaluated for Backprojection, 2FSK and, MM-2FSK (Priors From Azure Kinect) Against $BP_{max}$. Each Metric Is Averaged Across Five Frames of Each Radial Hand Motion Sequence

| | $C_g$ | $C_s$ | P |
|---|---|---|---|
| BPΔ0.5 | 0.37 | 15.12 | 3.89 |
| BPΔ1.0 | 0.51 | 15.71 | 5.67 |
| BPΔ10.0 | 0.34 | 15.16 | 5.23 |
| 2FSKΔ0.5 | 2.48 | 3.80 | 8.01 |
| 2FSKΔ1.0 | 3.97 | 4.92 | 9.52 |
| 2FSKΔ10.0 | 5.43 | 5.61 | 11.61 |
| 3FSK(Δ0.5,Δ10.0) | 2.24 | 3.98 | 3.75 |
| 3FSK(Δ1.0,Δ10.0) | 3.88 | 5.21 | 8.46 |
| 3FSK(Δ2.0,Δ10.0) | 4.43 | 5.60 | 10.21 |
| MM-2FSKΔ0.5 | 0.28 | 2.36 | **1.01** |
| MM-2FSKΔ1.0 | **0.21** | 1.77 | 4.47 |
| MM-2FSKΔ10.0 | 0.59 | **1.59** | 3.84 |

The best results per metric among all high-speed imaging methods are highlighted.

outperforms all approaches of similar number of frequencies across three metrics. Additionally, its performance in pixel-wise depth estimation, measured by metrics P and $P_e$, approaches that of $BP_{max}$, with only +4.2 mm and +1.6 mm additional error relative to the ground truth, respectively, despite utilizing significantly fewer frequencies.

In terms of different frequency configurations, our method proves to be more robust than other approaches, either matching or exceeding their performance. In contrast, both 2FSK and 3FSK show significantly poorer results when the depth prior falls outside the maximum unambiguous depth correction range, particularly when $\Delta f > 0.5$ GHz.

### D. DYNAMIC COMPARISON AGAINST THE STATE OF THE ART

We evaluate two radial hand-motion captures, comparing 2FSK, 3FSK, and BP against MM-2FSK, using time-synchronized depth priors from the Azure Kinect camera. All metrics are reported relative to $BP_{max}$, established as our best proxy for ground truth (see Table 3).

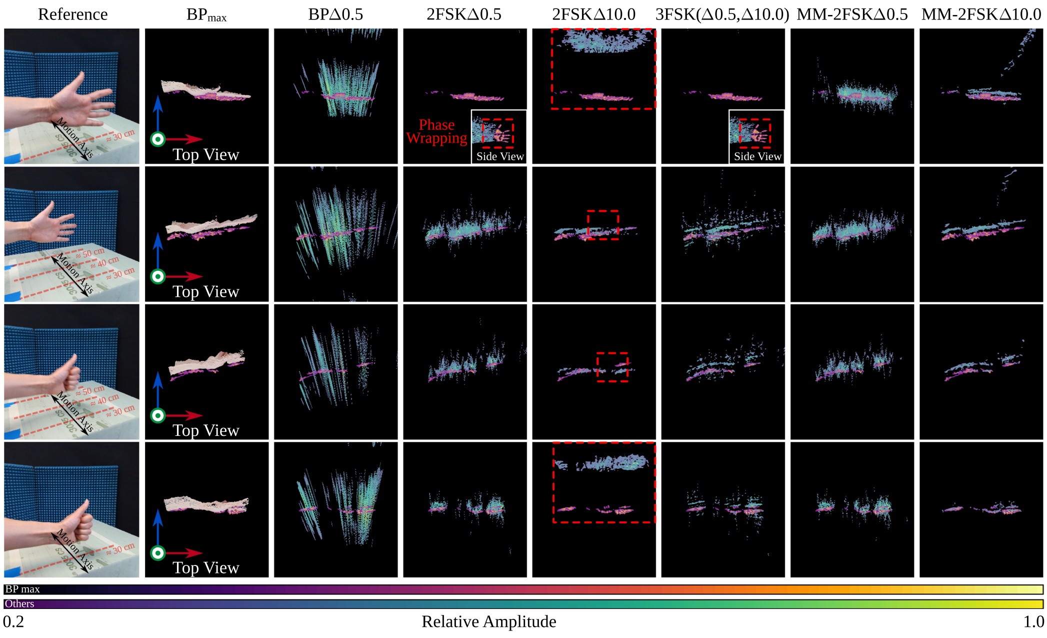


**FIGURE 7.** **Qualitative results of two dynamic sequences with radial hand motion, where reference views originate from Azure Kinect RGB frames, with dashed red overlays marking approximate target-to-radar ranges; the 40 cm range aligns most closely with the scalar depth priors used in 2FSK and 3FSK. We present top-down views of the reconstructed point clouds using backprojection, 2FSK, 3FSK, and MM-2FSK (priors from Azure Kinect) across different frequency configurations. We employ a distinct colormap for $BP_{max}$ to differentiate its measurements from few-tone reconstructions. We overlay $BP_{max}$ with per-point Azure Kinect measurements (*Second column from left*), highlighting data mismatch. Furthermore, we overlay few-tone reconstructions with $BP_{max}$ to assess relative depth deviation. Across all few-tone settings, MM-2FSK shows the closest agreement with $BP_{max}$, exhibiting reduced noise artifacts and preserving the reconstructed hand geometry without (partial) phase-wrapping-induced distortions (*red dashed boxes*). Within MM-2FSK variants (*two rightmost columns*), a trade-off is observed between reconstructions with lower noise levels (MM-2FSKΔ10.0) and increased robustness to inter-modality depth mismatch due to higher maximum unambiguous depth range (MM-2FSKΔ0.5).**

In Table 4, we list the mean reconstruction error averaged over five frames of each hand-motion sequence, sampled at timestamps corresponding to 0 %, 25 %, 50 %, 75 %, and 95 % of the overall sequence time. In the absence of a labeled foreground mask for the human hand, we omit computation of the $P_e$ metric. Delivering the best results across all metrics, MM-2FSK reconstruction quality is the closest to $BP_{max}$ among all few-tone imaging methods. On the other hand, 2FSK and 3FSK become increasingly difficult to adapt to dynamic scenarios, leading to larger depth deviations from $BP_{max}$ than in static experiments.

In Fig. 7, we present additional reconstructions of the two radial hand motions. Consistent with the experiments in a static setting, MM-2FSK is more robust than other few-tone methods, exhibiting lower noise and fewer phase-wrapping errors. We also observe the expected frequency configuration trade-offs from prior ablations: a smaller frequency difference $\Delta f$ increases noise but yields a higher maximum unambiguous depth correction, enabling greater tolerance to uncertainty in the depth prior. This is evident in the two rightmost columns of Fig. 7: MM-2FSKΔ10.0 shows misalignment due to data mismatch between the primary and secondary modality, whereas MM-2FSKΔ0.5 includes more points overlapping with $BP_{max}$ at the cost of higher noise. These results suggest that the frequency configuration should be selected not only according to spatial but also temporal mismatch when fusing modalities into a common reference frame.

## V. DISCUSSION

Our experiments demonstrate that the proposed MM-2FSK algorithm outperforms comparable radar-only algorithms. While integrating a secondary modality yields superior results, it also renders the algorithm prone to sensing limitations of this modality; in the case of an optical sensor, representative limitations may arise from unsuitable lighting conditions or highly reflective materials. Although our triangulation method may still provide reasonable depth values to fill in missing depth priors, an alternative approach could involve fusing the 3FSK method with our work, provided that the radar sensor supports high bandwidth and signal modulation at non-equidistant frequency steps.

Furthermore, the triangulation method does not respect object boundaries, such that in complex scenarios with multiple surface targets, depth interpolation using triangle topology may yield insufficient depth priors. An interesting future task could be the incorporation of semantic knowledge about the environment, as achieved by object segmentation.

Based on our experiments with optical depth sensors, we further recognize that sensor fusion with complementary modalities limits radar-specific characteristics, such as signal transmission, which is desirable in applications like security scanning or medical imaging. An intriguing extension of our work would be to investigate assistive modalities with transmission properties closer to the primary modality, for example from the Terahertz frequency bands. Ultimately, it

IEEE Journal of Microwaves

is essential to carefully evaluate the trade-off between the desirable characteristics of the radar sensor and the constraints imposed by the supporting modality for each application individually.

## VI. CONCLUSION

In this work, we address the increasing demand for radar sensors capable of high-speed capture and reconstruction to enable fast and accurate depth sensing of both static and dynamic targets by presenting a novel multimodal signal processing method based on frequency shift keying principles for near-field mmWave MIMO radar imaging [22].

Leveraging the capabilities of an assistive secondary modality, our proposed MM-2FSK algorithm overcomes current limitations of the 2FSK [22] approach with respect to the maximum unambiguous depth correction factor. By incorporating geometry processing into the reconstruction pipeline, we exploit additionally captured depth maps to create per-point depth priors from the perspective of a near-field, high-resolution radar. Collectively, these geometry-inspired approaches reduce the risk of erroneous measurements and range-phase ambiguities during MM-2FSK processing. This simple yet effective approach allows us to generalize our MM-2FSK extension to capture environments where neither the object's position nor its geometry is known in advance.

We evaluate the proposed method on a diverse set of static objects in the MAROON dataset and on dynamic radial hand-motion captures, using a high-resolution MIMO radar in conjunction with multiple secondary sensors from the optical frequency spectrum: multi-view stereo, active stereo, and optical time-of-flight. Our results demonstrate that our multimodal imaging approach outperforms comparable related work in terms of depth quality and performs only marginally worse than backprojection with maximum frequency steps, despite using fewer frequencies, therefore significantly lowering the capture and computation time. In summary, we believe our method holds great potential for future applications in multi-sensor target tracking.

## ACKNOWLEDGMENT

The authors would like to express their gratitude to Paul Himmler for the insightful discussions. 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 of the Friedrich-Alexander-Universität Erlangen-Nürnberg.

## REFERENCES

[1] R. Yunus et al., "Recent trends in 3D reconstruction of general non-rigid scenes," *Comput. Graph. Forum*, vol. 43, no. 2, 2024, Art. no. e15062.

[2] Y. Wang, Y. Tian, J. Chen, K. Xu, and X. Ding, "A survey of visual SLAM in dynamic environment: The evolution from geometric to semantic approaches," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 2523221.

[3] A.-K. Seifert, M. G. Amin, and A. M. Zoubir, "Toward unobtrusive in-home gait analysis based on radar micro-doppler signatures," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 9, pp. 2629–2640, Sep. 2019.

[4] M. Gambietz et al., "Unobtrusive gait reconstructions using radar-based optimal control simulations," in *Proc. 58th Asilomar Conf. Signals Syst. Comput.*, 2024, pp. 1505–1509.

[5] G. Amprimo, G. Masi, G. Pettiti, G. Olmo, L. Priano, and C. Ferraris, "Hand tracking for clinical applications: Validation of the Google mediapipe hand (GMH) and the depth-enhanced GMH-D frameworks," *Biomed. Signal Process. Control*, vol. 96, 2024, Art. no. 106508.

[6] V. Wirth et al., "ShaRPy: Shape reconstruction and hand pose estimation from RGB-D with uncertainty," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2023, pp. 2625–2633.

[7] U. Phutane et al., "Evaluation of optical and radar based motion capturing technologies for characterizing hand movement in rheumatoid arthritis–A pilot study," *Sensors*, vol. 21, no. 4, 2021, Art. no. 1208.

[8] A. Chen, X. Wang, S. Zhu, Y. Li, J. Chen, and Q. Ye, "mmBody benchmark: 3D body reconstruction dataset and analysis for millimeter wave radar," in *Proc. 30th ACM Int. Conf. Multimedia*, New York, NY, USA, 2022, pp. 3501–3510.

[9] A. Chen et al., "ImmFusion: Robust mmWave-RGB fusion for 3D human body reconstruction in all weather conditions," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 2752–2758.

[10] S.-P. Lee, N. P. Kini, W.-H. Peng, C.-W. Ma, and J.-N. Hwang, "HuPR: A benchmark for human pose estimation using millimeter wave radar," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 5704–5713.

[11] L. Engel et al., "Advanced millimeter wave radar-based human pose estimation enabled by a deep learning neural network trained with optical motion capture ground truth data," *IEEE J. Microwaves*, vol. 5, no. 2, pp. 373–387, Mar. 2025.

[12] N. S. Zewge, Y. Kim, J. Kim, and J.-H. Kim, "Millimeter-wave radar and RGB-D camera sensor fusion for real-time people detection and tracking," in *Proc. 7th Int. Conf. Robot Intell. Technol. Appl.*, 2019, pp. 93–98.

[13] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "RadHAR: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proc. 3rd ACM Workshop Millimeter-Wave Netw. Sens. Syst.*, New York, NY, USA, 2019, pp. 51–56.

[14] V. Wirth, J. Bräunig, N. Hofmann, M. Vossiek, T. Weyrich, and M. Stamminger, "MAROON: A dataset for the joint characterization of near-field high-resolution radio-frequency and optical depth imaging techniques," *ACM Trans. Graph.*, vol. 45, pp. 1–18, Mar. 2026.

[15] E. Wolf, "Three-dimensional structure determination of semi-transparent objects from holographic data," *Opt. Commun.*, vol. 1, no. 4, pp. 153–156, 1969.

[16] S. S. Ahmed, "Microwave imaging in security–Two decades of innovation," *IEEE J. Microwaves*, vol. 1, no. 1, pp. 191–201, Jan. 2021.

[17] S. M. Farrell, V. Boominathan, N. Raymondi, A. Sabharwal, and A. Veeraraghavan, "CoIR: Compressive implicit radar," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 9, pp. 7316–7327, Sep. 2025.

[18] J. Guan, S. Madani, S. Jog, S. Gupta, and H. Hassanieh, "Through fog high-resolution imaging using millimeter wave radar," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11461–11470.

[19] D. Borts et al., "Radar fields: Frequency-space neural scene representations for FMCW radar," in *Proc. ACM SIGGRAPH Conf. Papers*, New York, NY, USA, 2024, pp 1–10.

[20] T. Huang et al., "DART: Implicit doppler tomography for radar novel view synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 24118–24129.

[21] M. Rafidashti, J. Lan, M. Fatemi, J. Fu, L. Hammarstrand, and L. Svensson, "NeuRadar: Neural radiance fields for automotive radar point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 2488–2498.

[22] J. Bräunig et al., "An ultra-efficient approach for high-resolution MIMO radar imaging of human hand poses," *IEEE Trans. Radar Syst.*, vol. 1, pp. 468–480, 2023.

[23] J. Bräunig, V. Wirth, M. Stamminger, I. Ullmann, and M. Vossiek, "An efficient yet high-performance method for precise radar-based imaging of human hand poses," in *Proc. 21st Eur. Radar Conf.*, 2024, pp. 332–335.

[24] V. Wirth et al., “Automatic spatial calibration of near-field MIMO radar with respect to optical sensors,” *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024, pp. 8322–8329.

[25] B. Delaunay, “Sur la sphère vide,” *Bull. de l’Académie des Sci. de l’URSS. Classe des sci. mathématiques et na*, vol. 1934, no. 6, pp. 793–800, 1934.

**VANESSA WIRTH** was born in Schwabach, Germany, in 1996. She received the M.Sc. degree in computer science from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany, in collaboration with the Technical University of Munich, Munich, Germany. She is currently working toward the Ph.D. degree with the Chair of Visual Computing, FAU. Her research interests include the intersection of computer graphics, computer vision, and machine learning, and focuses on surface reconstruction and motion tracking of nonrigid objects using radar-based and optical sensor systems.

**JOHANNA BRÄUNIG** was born in Bamberg, Germany, in 1993. She received the M.Sc. degree in electrical engineering from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany, in 2020, and the Ph.D. degree from the Institute of Microwaves and Photonics, FAU, in Jan. 2025. Her research interests include radar imaging and signal processing, focusing radar-based motion tracking and the application of radar technology in medical engineering and machine learning. She was the recipient of the Young Engineer Prize at the European Radar Conference in 2023.

**MARTIN VOSSIEK** (Fellow, IEEE) received the Ph.D. degree from Ruhr-Universität Bochum, Germany, in 1996. In 1996, he joined Siemens Corporate Technology, Munich, Germany, where he led the Microwave Systems Group, from 2000 to 2003. Since 2003, he has been a Full Professor with the Clausthal University of Technology, Clausthal-Zellerfeld, Germany. Since 2011, he has been the Chair with the Institute of Microwaves and Photonics (LHFT), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany. He has authored or coauthored more than 450 publications, and his research has resulted in more than 100 granted patents. His research interests include radar, microwave systems, wave-based imaging, transponders, RF identification, communication, and wireless sensor and localization systems. He is also the Spokesperson for the Collaborative Research Centre (CRC 1483) EmpkinS, where more than 80 researchers are developing innovative wireless and wave-based sensor technologies for applications in medicine and psychology. He is also a member of the German National Academy of Science and Engineering (ACATECH) and the German Research Foundation (DFG) Review Board 4.42-02, Communication Technology and Networks, Microwave Technology and Photonic Systems, Signal Processing and Machine Learning for Information Technology. Since Apr. 2024, he has been a Spokesperson for the DFG Review Board 4.42 for electrical engineering and information technology. He is actively involved with the IEEE Microwave Theory and Technology (MTT) Technical Committees, including MTT-24 (Microwave/mm-wave Radar, Sensing, and Array Systems), MTT-27 (Connected and Autonomous Systems, where he was the Founding Chair), and MTT-29 (Microwave Aerospace Systems). He is also on the advisory board of the IEEE CRFID Technical Committee on Motion Capture & Localization. He was the recipient of numerous Best Paper awards and other Distinctions. In 2019, the IEEE MTT Society (MTT-S) honored him with the Microwave Application Award for his pioneering research in wireless local positioning systems. He was on Organizing Committees and Technical Program Committees for many international conferences and has contributed as a Reviewer for various technical journals. From 2013 to 2019, he was an Associate Editor for IEEE Transactions on Microwave Theory and Techniques, and has been the Associate Editor-in-Chief of IEEE Transactions on Radar Systems since Oct. 2022.

**TIM WEYRICH** received the PhD degree from ETH Zurich, Switzerland, in 2006. He was a Postdoctoral Teaching Fellow with Princeton University, working with Princeton Computer Graphics Group. He currently holds the Chair of Digital Reality with the Department of Computer Science of Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and is also an Associate Member with the FAU’s Department for Digital Humanities and Social Studies. He also retains a part-time assignment as Professor of visual computing with Virtual Environments and Computer Graphics Group, Department of Computer Science, University College London (UCL), a position he originally held till 2021, after joining UCL, in 2008. His research interests include physically-based rendering, computational photography, appearance modelling and fabrication, point-based graphics, cultural heritage analysis and digital humanities. He was an Associate Editor for IEEE Transactions of Visualization and Computer Graphics (TVCG), *Computer and Graphics* (CAG), Elsevier, and *Computer Graphics Forum* (CGF), Wiley/Eurographics. He is also a member of the Executive Committee of Eurographics and member of the Steering Board of the Eurographics Workshop on Graphics and Cultural Heritage.

**MARC STAMMINGER** is currently the Chair of visual computing with Friedrich-Alexander-Universität Erlangen-Nürnberg. He has authored or coauthored more than 200 papers in leading journals and conference proceedings. His research interests include the various areas of visual computing, such as real-time computer graphics, neural and inverse rendering, physical light and microwave simulation, and virtual and augmented reality. He is also an Associate Editor of the *Journal of Computer Graphics Techniques*. He was a Program Chair for Eurographics, High Performance Graphics, and the Eurographics Symposium of Rendering, and has participated in the Program Committees of all major computer graphics conferences. He has been elected to the Eurographics Executive Committee and is the Eurographics Professional Board chair.