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

Vanessa Wirth[1], Johanna Bräunig[3], Martin Vossiek[2], Tim Weyrich*[1,4], and Marc Stamminger*[1]

[1]Visual Computing Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2]Institute of Microwaves and Photonics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[3]fiveD, Germany
[4]University College London (UCL), United Kingdom

***Abstract*—Accurate reconstruction of static and rapidly moving targets demands three-dimensional imaging solutions with high temporal and spatial resolution. Radar sensors are a promising sensing modality because of their fast capture rates and their independence from lighting conditions. To achieve high spatial resolution, MIMO radars with large apertures are required. Yet, they are infrequently used for dynamic scenarios due to significant limitations in signal processing algorithms. These limitations impose substantial hardware constraints due to their computational intensity and reliance on large signal bandwidths, ultimately restricting the sensor's capture rate. One solution of previous work is to use few frequencies only, which enables faster capture and requires less computation; however, this requires coarse knowledge of the target's position and works in a limited depth range only. To address these challenges, we extend previous work into the multimodal domain with MM-2FSK, which leverages an assistive optical depth sensing modality to obtain a depth prior, enabling high framerate capture with only few frequencies. We evaluate our method using various target objects with known ground truth geometry that is spatially registered to real millimeter-wave MIMO radar measurements. Our method demonstrates superior performance in terms of depth quality, being able to compete with the time- and resource-intensive measurements with many frequencies.**

***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]–[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, for instance in security screening [14], MIMO radars typically employ a large number of transmitting (TX) and receiving (RX) antennas. Traditional radar imaging techniques, such as backprojection [14], [15], rely on these dense antenna arrays to leverage the numerous TX-RX combinations for precise reconstruction; however, this comes at the expense of significant computational resources. Such methods also often require many distinct transmission frequencies, which limits the sensor's capture rate and renders the algorithm unsuitable for rapidly moving targets.

An alternative research direction is to improve resolution capabilities with sparse, low-cost antenna arrays that can operate at high capture rates. Recent advancements in deep learning focus on smooth neural target representations that surpass the spatial resolution limitations of conventional signal processing techniques. One approach is to employ generative methods, such as implicit neural representations [16] or conditional generative adversarial networks [17], 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 [18] or range-Doppler maps [19], while also incorporating additional modalities such as LiDAR and cameras [20]. 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

*The authors contributed equally to this work

scenarios. Moreover, the training process is computationally intensive.

Among traditional approaches, Bräunig et al. [21] 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, making it up to 1000 times faster than backprojection [21], while ensuring rapid capture times due to the limited set of transmitted frequencies.

A key limitation of the 2FSK method, however, is its assumption that the depth of the captured object is roughly known. Follow-up work [22] (*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, resembling a plane or depth slice in 3D space, makes both methods suitable primarily for flat targets with limited depth extent, as shown for hands in [21], [22]. Consequently, their applicability is restricted to a narrow range of target geometries and can lead to inaccuracies in uncertain environments.

Our work extends the 2FSK technique into the multimodal domain, aiming towards a broadly applicable method that offers reliable fast-capture and fast-reconstruction performance with respect to unknown static and dynamic environments. We integrate a secondary depth sensing modality, such as an optical RGB-D camera, to obtain a depth prior. Optical depth sensors offer higher spatial resolution compared to existing imaging radars. However, their temporal resolution is significantly lower. The such obtained depth prior allows our method to handle objects with varying geometries in the depth direction. We refer to this approach as *multimodal 2FSK* (MM-2FSK).

We provide a comprehensive evaluation based on a dataset [23] that provides ground-truth geometry of static objects, spatially aligned with real-world measurements from a mmWave high-resolution MIMO imaging radar with frequency-stepped continuous-wave (FSCW) signal modulation. In this evaluation, we compare our method with 2FSK, 3FSK, and traditional backprojection. In addition, we investigate the influence of signal bandwidth beyond theoretical analysis and provide ablation studies focused on different frequency configurations.

In summary, our contributions are the following:

1) A novel multimodal signal processing method that incorporates a mmWave FSCW MIMO radar along with an optical depth camera as an assistive modality; for evaluation, we use an active stereo RGB-D camera.
2) A method for robust, high-speed radar imaging of arbitrary objects without requiring additional knowledge about the capture environment, i.e. the object position and depth variation over surface.
3) A comprehensive evaluation of various static objects: 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. [21]. Subsequently, we derive our MM-2FSK method from this principle. The key differences between the two algorithms are highlighted in Figure 1.

Both methods are designed for high-resolution MIMO imaging radars that utilize multiple transmit-receive (TX-RX) antenna pairs for imaging. For simplicity, we omit the repetitive calculations over multiple antenna pairs – typically performed for 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 [21].

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. In a MIMO antenna configuration, the corresponding baseband signals are transmitted from a TX antenna, $\boldsymbol{r}_{\mathrm{TX}} \in \mathbb{R}^3$, reflect off the first point target located at $\boldsymbol{p} \in \mathbb{R}^3$, and are subsequently received by each RX antenna, $\boldsymbol{r}_{\mathrm{RX}} \in \mathbb{R}^3$. 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 ϕ_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 $\widetilde{\boldsymbol{p}} \in \boldsymbol{\mathcal{P}} = \{(x, y, \widetilde{d})\}$ with a corresponding scalar depth prior $\widetilde{d}$, two signal hypotheses, w_1, w_2, are computed from the round-trip distance $\widetilde{\rho}$ between an TX-RX antenna pair and $\widetilde{\boldsymbol{p}}$:

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

The hypotheses are correlated with the baseband signals as follows:

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

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, $\Delta\rho = (\rho - \widetilde{\rho}) = 2\Delta d$, and correspondingly depth Δd:

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

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

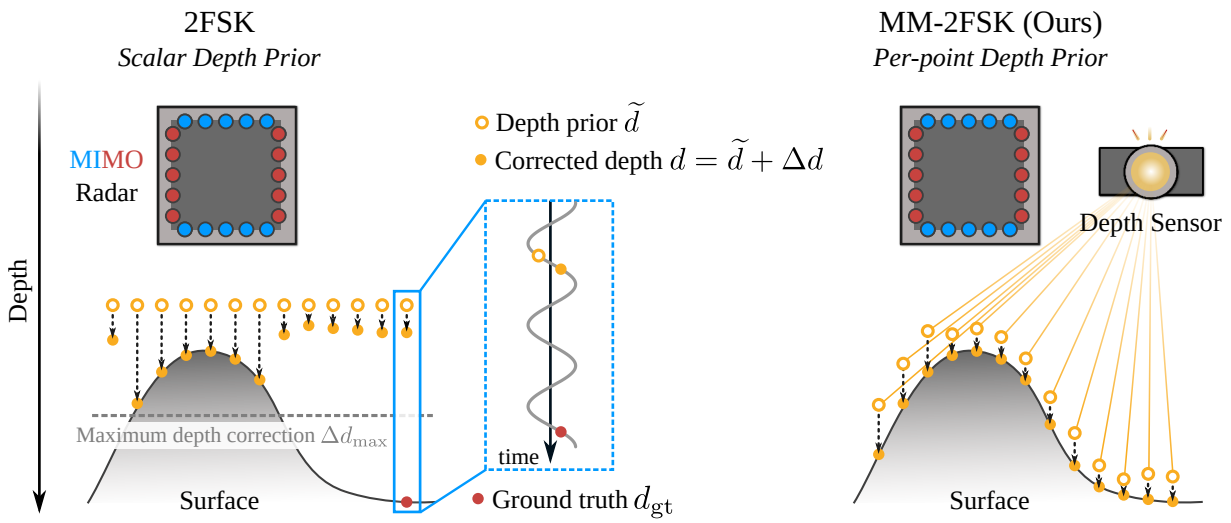

Fig. 1. In our work, we extend the 2FSK imaging principle to the multimodal domain (MM-2FSK). Given a unified scalar depth prior, $\widetilde{d}$, for each point, the 2FSK method iteratively refines the current estimate with a per-point depth correction factor, Δd, up to a limited extent, given by the 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.

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

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

Computing the phase $\Delta\varphi_1$ or $\Delta\varphi_2$ involves inverse trigonometric functions to determine the angle of the residual complex phasor $c_i(\widetilde{\rho})$. Due to the 2π 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 Equation 5 by assuming $\Delta\varphi_i$ approaches 2π, which yields $\mathrm{c}/(2 \cdot f_i)$.

High-resolution MIMO imaging radars typically operate in the GHz to THz range [14], meaning this maximum correction factor can be quite small. Thus, Bräunig et al. [21] introduced the concept of calculating a differential complex phasor from the two single-frequency residual phasors based on Equation 3 and Equation 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. Figure 2), allowing a depth correction Δd to be within the so-called *maximum unambiguous depth correction* $\Delta d_{\max}$, which now solely dependends 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 Equation 8 from [21], 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 Figure 2, where we simplify the illustration of an analytic

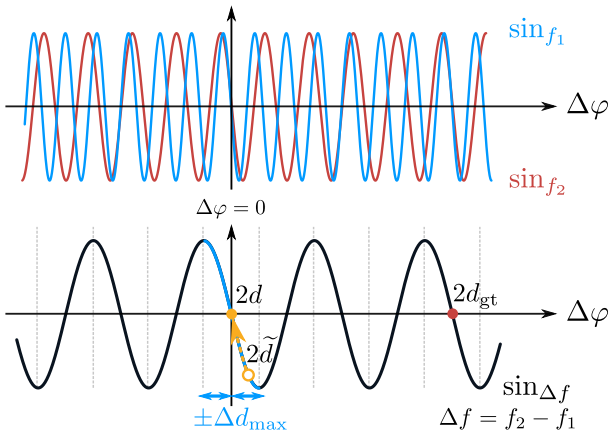

Fig. 2. Simplified visualization of the 2FSK depth correction process, where complex, analytic signals are schemed as periodic, real-valued sine waves. The 2FSK principle computes the depth correction Δd based on the residual of the phase, $\Delta\varphi$, that remains after correlating the two single-frequency 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 Δf 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 2π-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.

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. To address this challenge, we introduce the MM-2FSK method next.

B. Multimodal Frequency Shift Keying (MM-2FSK)

As our method is tailored to high-speed radar imaging, we first describe theoretical details of our algorithm, followed by its efficient implementation on the graphics card.

1) Algorithm: In contrast to the 2FSK approach, which relies on a single scalar depth prior, we propose utilizing per-pixel depth measurements obtained from a depth camera that is spatially calibrated with a MIMO imaging radar.

This spatial calibration can be achieved, for example, using target-based methods, such as in [24], where spherical calibration targets composed of metallic and styrofoam-based materials are combined and symmetrically mounted on a board, to calibrate near-field MIMO imaging radars in conjunction with optical depth sensors.

By employing this secondary sensor, we first acquire an optical depth map $\boldsymbol{D}_o \in \mathbb{R}^{H\times W}$, with pixels (u, v) and corresponding depth d. We then compute the associated point cloud $\boldsymbol{p}_o \in \boldsymbol{P}_o \in \mathbb{R}^{N\times 3}$ by back-projecting each triplet (u, v, d) utilizing the intrinsic calibration parameters of the depth camera:

$$\boldsymbol{p}_o = \begin{pmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} u \cdot d \\ v \cdot d \\ d \end{pmatrix}, \tag{9}$$

where f_u, f_v are the focal lengths and c_u, c_v are the principal point offsets of the depth camera model. Note that the depth image may contain invalid pixels due to the sensitivity of optical depth sensors to environmental lighting and reflective materials; such pixels are simply skipped.

To generate a depth prior for radar imaging, the resulting point cloud is converted to a closed triangle mesh. To this end, we triangulate the point cloud (in 2D) using Delaunay triangulation [25]. This triangulation computes the 2D convex hull of $\boldsymbol{P}_o$, effectively filling in depth gaps and preventing surface holes. A visualization of such triangulation is given in Figure 3. Next, we use the extrinsic parameters obtained from spatial calibration [24] to transform the triangulated point cloud into the radar's coordinate space:

$$\boldsymbol{P}_o^r = [\boldsymbol{R} \mid \boldsymbol{t}]\boldsymbol{P}_o\,. \tag{10}$$

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 mapping of spatial coordinates to radar image pixels – typically represented as an orthographic camera model – we rasterize the triangulated point cloud $\boldsymbol{P}_o^r$.

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 $\widetilde{d}$, as illustrated in Figure 3. Finally, we proceed with the depth correction in analogy to Equation 6 of the 2FSK method.

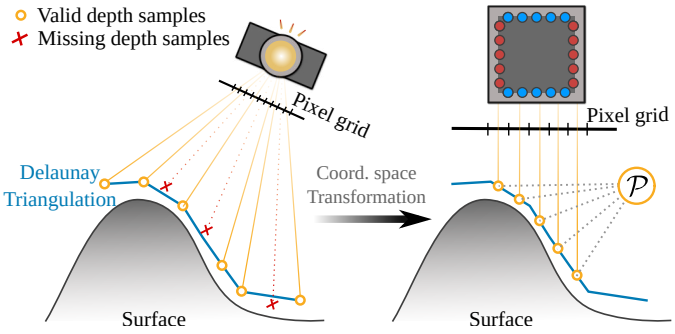

Fig. 3. Visualization of the depth prior generation: We first create a closed triangle mesh using Delaunay triangulation on 2D pixels corresponding to valid 3D point samples from the optical depth sensor; illustrated here in 2D as blue line sets. The mesh is then transformed into the radar's coordinate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

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 optical depth camera 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.

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 $\mathcal{P} \in \mathbb{R}^{H'W'\times 3}$, 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 $\boldsymbol{C}_1 \in \mathbb{C}^{H'W'}$, $\boldsymbol{C}_2 \in \mathbb{C}^{H'W'}$

$\boldsymbol{p} \leftarrow \mathcal{P}[j]$;
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
$\boldsymbol{C}_1[j] \leftarrow$ warp_reduce_sum$(\bar{c}_1)/(T \cdot R)$;
$\boldsymbol{C}_2[j] \leftarrow$ warp_reduce_sum$(\bar{c}_2)/(T \cdot R)$;

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.

In MIMO radar imaging algorithms [14], [21], 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 Equation 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 Equation 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 [23], along with the implementation details of the algorithms that we use in our evaluation: backprojection [14], [15], 2FSK [21], 3FSK [22], and MM-2FSK. 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.

A. Dataset

We validate the proposed method using the MAROON dataset [23], which comprises real sensor measurements of 45 distinct static household and construction objects of varying surface geometry. These measurements were collected from a high-resolution MIMO radar, which was synchronized with three different spatially calibrated depth cameras and a ground-truth measurement system. While the target objects were captured at various distances [23], we focus on the object measurements taken at approximately 30 cm from the MIMO radar.

The QAR50 MIMO radar submodule utilized in MAROON features an aperture consisting of 94×94 TX-RX antenna pairs and employs a frequency-stepped continuous-wave (FSCW) signal modulation, operating across 128 discrete frequencies from 72 to 82 GHz. To assess the imaging accuracy of the proposed method, we utilize ground-truth measurements of the target objects, obtained from a multi-view stereo system composed of five digital single-lens reflex (DSLR) cameras. Depending on the experiment, we chose either the ground-truth system or the Realsense D435i active stereo depth camera as secondary modalities for our MM-2FSK approach.

B. Implementation Details

For the MIMO imaging radar, the dataset provides 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, we sample points within a $30 \times 30 \times 20$ cm volume centered around the object, yielding a voxel grid of dimensions $301 \times 301 \times 201$. This grid is then projected to a $301 \times 301 = H' \times W'$ depth map using maximum intensity projection. For the 2FSK, 3FSK, and MM-2FSK methods, we correspondingly reconstruct the 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, 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 of the resulting mean phasor. Finally, we keep the depth values with a magnitude higher than -14 dB relative to the maximum.

In terms of runtime, our implementation for depth estimation with two-frequency backprojection takes approximately 1430 ms, when using an NVIDIA GeForce RTX 3080 graphics card (10GB VRAM) and an Intel Xeon W-1390P (3.50 GHz) processor. 3FSK achieves a runtime of about 7 ms and the (MM-)2FSK methods achieve a runtime of about 4 ms.

IV. Evaluation

In the following sections, we will describe the evaluation metrics and experimental results. We present two key experiments: first, we conduct an ablation study to explore the accuracy of MM-2FSK while varying the frequency differences. Second, we compare our method against the 2FSK [21] and 3FSK [22] approaches, and traditional backprojection (BP) [14], [15].

Our evaluation consists of six representative frequency configurations, which are listed in Table I. For (MM-)2FSK and BP, we will use the terminology $2FSK\Delta f$, e.g., FSKΔ0.5 to denote a configuration with 0.5 GHz frequency difference. For

TABLE I
FREQUENCY CONFIGURATIONS, UTILIZED FOR ALL SUBSEQUENT EXPERIMENTS.

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

3FSK, we denote the lowest and highest frequency differences as 3FSK($\Delta f_{\min}$, $\Delta f_{\max}$).

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 [23] 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 $\boldsymbol{P}_{\text{gt}} \in \mathbb{R}^{M\times 3}$:

$$\mathrm{C} = \frac{1}{N} \sum_{\boldsymbol{p}_r \in \boldsymbol{P}_r} \min_{\boldsymbol{p}_{\text{gt}} \in \boldsymbol{P}_{\text{gt}}} \|\boldsymbol{p}_r - \boldsymbol{p}_{\text{gt}}\|_2 \,. \tag{11}$$

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. [23]). The Chamfer distance is measured in both directions: from the ground-truth point cloud to the radar reconstruction, denoted as $\mathrm{C_g}$, and vice versa, denoted as $\mathrm{C_s}$.

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:

$$\mathrm{P} = \frac{1}{H' \cdot W'} \sum_{u=0}^{H'-1} \sum_{v=0}^{W'-1} |\boldsymbol{D}_r(u,v) - \boldsymbol{D}_{\text{gt}}(u,v)| \,. \tag{12}$$

Following the methodology of [23], 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 $\mathrm{P_e}$, respectively.

B. Ablation with respect to Frequency Differences

We assess the MM-2FSK algorithm using the frequency configurations outlined in Table I 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 measurement system. The results are summarized in Table II, showcasing performance across all four metrics, averaged over the 45 objects of the 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 I), 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 [22], 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 Figure 4. Reconstructions with higher frequency differences exhibit fewer noise artifacts.

TABLE II
ABLATION STUDY OF THE MM-2FSK METHOD WITH DIFFERENT FREQUENCY CONFIGURATIONS. ALL METRICS ARE GIVEN IN CENTIMETERS AND ARE AVERAGED OVER ALL OBJECTS AT 30 CM DISTANCE. THE **BEST** RESULTS PER METRIC ARE HIGHLIGHTED.

	$\mathrm{C_g}$	$\mathrm{C_s}$	P	$\mathrm{P_e}$
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**

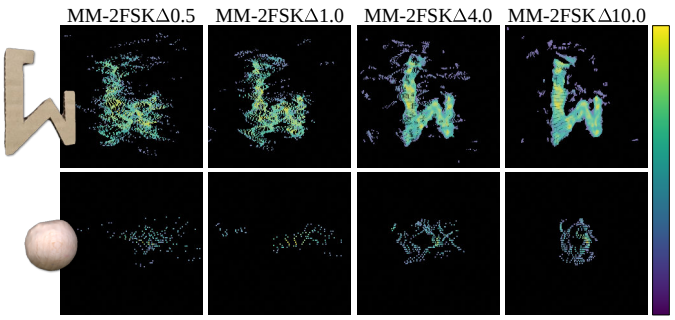

Fig. 4. MM-2FSK reconstructions for the *Cardboard* and *Wood Ball* objects, compared across different frequency configurations. The point clouds are color-coded based on the residual phasor magnitude, which approximately corresponds to the intensity of the signal. Higher bandwidths exhibit fewer artifacts as they are less sensible to noisy phase variations.

C. Comparison with 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 active stereo depth camera. Our evaluation focuses on three frequency configurations from Table I: the first, where $\widetilde{d}$ 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 $\mathrm{BP_{max}}$, utilizing the maximum of 128 frequency steps.

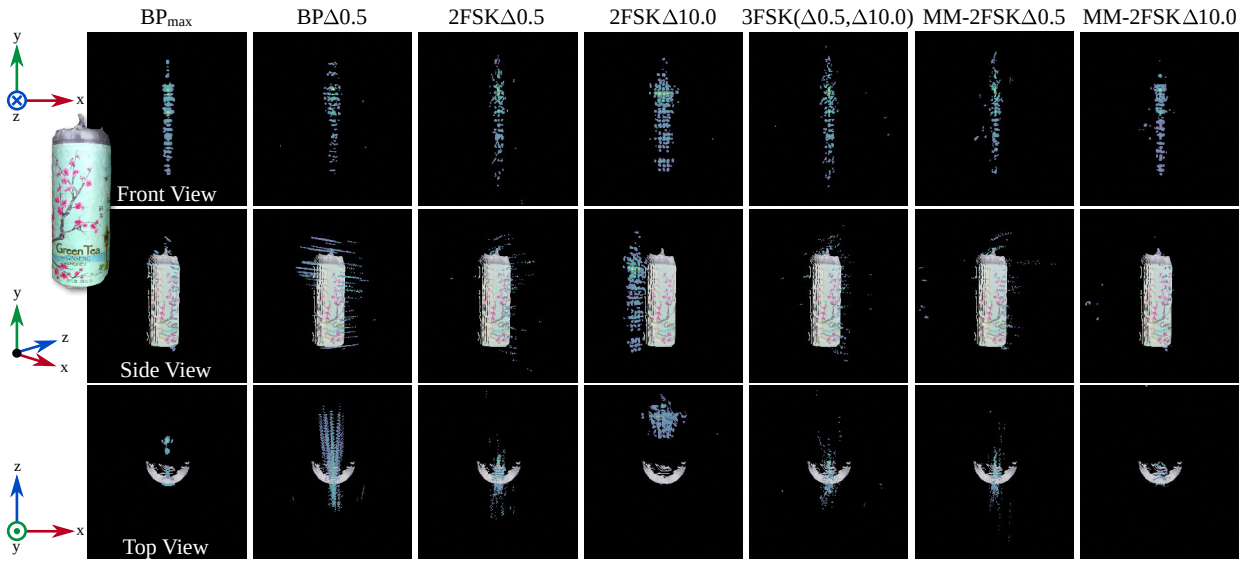

Fig. 5. Comparison of the reconstructed point clouds for backprojection, 2FSK, 3FSK, and MM-2FSK across different frequency configurations and views, with standard BP using the full 10 GHz frequency spectrum at 128 frequency steps. The radar point clouds are color-coded based on the residual phasor magnitude, approximately corresponding to the intensity of the signal. 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.

In Figure 5, 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 BP$_{max}$.

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

TABLE III

THE MEAN RECONSTRUCTION ERROR IN CENTIMETERS, EVALUATED FOR BACKPROJECTION, 2FSK AND, MM-2FSK AGAINST THE GROUND-TRUTH SETUP. EACH METRIC IS AVERAGED ACROSS ALL MAROON OBJECTS AT A SENSOR DISTANCE OF 30 CM. THE BEST RESULTS PER METRIC AMONG ALL HIGH-SPEED IMAGING METHODS ARE HIGHLIGHTED.

	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**

V. DISCUSSION

Our experiments demonstrate that the proposed MM-2FSK algorithm outperforms comparable radar-only algorithms. While integrating a complementary depth sensor yields superior results, it also renders the algorithm prone to its limitations, for example in scenarios with unsuitable lighting conditions or highly reflective materials. Although our triangulation method may still provide reasonable depth values to fill in the 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

color data for example – as most depth cameras provide color information alongside depth.

Moreover, we recognize that sensor fusion with an optical depth camera limits radar-specific characteristics, such as signal transmission, which is desirable in applications like security scanning or medical imaging. An intriguing research direction would be to investigate sensor solutions with similar transmission properties, like time-of-flight cameras operating in the infrared frequency spectrum. Ultimately, it is essential to carefully evaluate the trade-off between the desirable characteristics of the radar sensor and the constraints imposed by the supporting depth sensor 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 MIMO radar imaging [21].

Leveraging the capabilities of an assistive optical depth camera, our proposed MM-2FSK algorithm overcomes current limitations of the 2FSK [21] approach with respect to the maximum unambiguous depth correction factor. By employing geometric processing methods such as triangulation, we utilize the captured optical depth maps to create a per-point depth prior from the perspective of the radar sensor, thereby addressing potential shortcomings of the depth sensor through a hole-filling method. 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.

Evaluating our method with the diverse set of objects in the MAROON dataset [23], we conducted experiments using a high-resolution MIMO imaging radar in conjunction with an active stereo depth camera. 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.

Acknowledgement

The authors would like to express their gratitude to Paul Himmler for the insightful discussions.

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

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, J. E. Lenssen, M. Niemeyer, Y. Liao, C. Rupprecht, C. Theobalt, G. Pons-Moll, J.-B. Huang, V. Golyanik, and E. Ilg, "Recent trends in 3d reconstruction of general non-rigid scenes," *Computer Graphics Forum*, vol. 43, no. 2, p. e15062, 2024.

[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 Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–21, 2024.

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

[4] M. Gambietz, A. Dröge, C. Schüßler, M. Stahlke, V. Wirth, J. Miehling, M. Vossiek, and A. D. Koelewijn, "Unobtrusive gait reconstructions using radar-based optimal control simulations," in *2024 58th Asilomar Conference on Signals, Systems, and Computers*, pp. 1505–1509, 2024.

[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," *Biomedical Signal Processing and Control*, vol. 96, p. 106508, 2024.

[6] V. Wirth, A.-M. Liphardt, B. Coppers, J. Bräunig, S. Heinrich, S. Leyendecker, A. Kleyer, G. Schett, M. Vossiek, B. Egger, and M. Stamminger, "Sharpy: Shape reconstruction and hand pose estimation from rgb-d with uncertainty," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2625–2633, October 2023.

[7] U. Phutane, A.-M. Liphardt, J. Bräunig, J. Penner, M. Klebl, K. Tascilar, M. Vossiek, A. Kleyer, G. Schett, and S. Leyendecker, "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.

[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 *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, (New York, NY, USA), p. 3501–3510, Association for Computing Machinery, 2022.

[9] A. Chen, X. Wang, K. Shi, S. Zhu, B. Fang, Y. Chen, J. Chen, Y. Huo, and Q. Ye, "Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2752–2758, 2023.

[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 *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5704–5713, 2023.

[11] L. Engel, J. Mueller, E. J. F. Rendon, E. Dorschky, D. Krauss, I. Ullmann, B. M. Eskofier, and M. Vossiek, "Advanced millimeter wave radar-based human pose estimation enabled by a deep learning neural network trained with optical motion capture ground truth data," *IEEE Journal of Microwaves*, vol. 5, no. 2, pp. 373–387, 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 *2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA)*, pp. 93–98, 2019.

[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 *Proceedings of the 3rd ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, mmNets '19, (New York, NY, USA), p. 51–56, Association for Computing Machinery, 2019.

[14] S. S. Ahmed, "Microwave imaging in security — two decades of innovation," *IEEE Journal of Microwaves*, vol. 1, no. 1, pp. 191–201, 2021.

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

[16] S. M. Farrell, V. Boominathan, N. Raymondi, A. Sabharwal, and A. Veeraraghavan, "Coir: Compressive implicit radar," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 9, pp. 7316–7327, 2025.

[17] J. Guan, S. Madani, S. Jog, S. Gupta, and H. Hassanieh, "Through fog high-resolution imaging using millimeter wave radar," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11470, 2020.

[18] D. Borts, E. Liang, T. Broedermann, A. Ramazzina, S. Walz, E. Palladin, J. Sun, D. Brueggemann, C. Sakaridis, L. Van Gool, M. Bijelic, and F. Heide, "Radar fields: Frequency-space neural scene representations for fmcw radar," in *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, (New York, NY, USA), Association for Computing Machinery, 2024.

[19] T. Huang, J. Miller, A. Prabhakara, T. Jin, T. Laroia, Z. Kolter, and A. Rowe, "Dart: Implicit doppler tomography for radar novel view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24118–24129, June 2024.

[20] M. Rafidashti, J. Lan, M. Fatemi, J. Fu, L. Hammarstrand, and L. Svensson, "Neuradar: Neural radiance fields for automotive radar point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2488–2498, June 2025.

[21] J. Bräunig, V. Wirth, C. Kammel, C. Schüßler, I. Ullmann, M. Stamminger, and M. Vossiek, "An ultra-efficient approach for high-resolution mimo radar imaging of human hand poses," *IEEE Transactions on Radar Systems*, vol. 1, pp. 468–480, 2023.

[22] 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 *2024 21st European Radar Conference (EuRAD)*, pp. 332–335, 2024.

[23] V. Wirth, J. Bräunig, M. Vossiek, T. Weyrich, and M. Stamminger, "Maroon: A framework for the joint characterization of near-field high-resolution radar and optical depth imaging techniques," 2024.

[24] V. Wirth, J. Bräunig, D. Khouri, F. Gutsche, M. Vossiek, T. Weyrich, and M. Stamminger, "Automatic spatial calibration of near-field mimo radar with respect to optical sensors," *ArXiv*, vol. abs/2403.10981, 2024.

[25] B. Delaunay, "Sur la sphère vide," *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na*, vol. 1934, no. 6, pp. 793–800, 1934.